UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXV CICLO DEL DOTTORATO DI RICERCA IN
SCIENZE DELL'INTERPRETAZIONE E DELLA TRADUZIONE

CONCORDANCING SOFTWARE IN PRACTICE:
AN INVESTIGATION OF SEARCHES AND TRANSLATION PROBLEMS ACROSS EU OFFICIAL LANGUAGES

Settore scientifico-disciplinare: L-LIN/12

DOTTORANDA
PAOLA VALLI

COORDINATORE e SUPERVISORE DI TESI
PROF. FEDERICA SCARPA

CO-SUPERVISORE
DR. GIUSEPPE PALUMBO

ANNO ACCADEMICO 2011/2012

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXV CICLO DEL DOTTORATO DI RICERCA IN
SCIENZE DELL'INTERPRETAZIONE E DELLA TRADUZIONE

CONCORDANCING SOFTWARE IN PRACTICE:
AN INVESTIGATION OF SEARCHES AND TRANSLATION PROBLEMS ACROSS EU OFFICIAL LANGUAGES

Settore scientifico-disciplinare: L-LIN/12

DOTTORANDA
PAOLA VALLI

COORDINATORE e SUPERVISORE DI TESI
PROF. FEDERICA SCARPA

CO-SUPERVISORE
DR. GIUSEPPE PALUMBO

ANNO ACCADEMICO 2011/2012

To YOU,

who have changed my life

for the better

Think and wonder, wonder and think.

Dr. Seuss

ABSTRACT

The present work reports on an empirical study aimed at investigating translation problems across multiple language pairs. In particular, the analysis is aimed at developing a methodological approach to study concordance search logs taken as manifestations of translation problems and, in a wider perspective, information needs. As search logs are a relatively unexplored data type within translation process research, a controlled environment was needed in order to carry out this exploratory analysis without incurring in additional problems caused by an excessive amount of variables. The logs were collected at the European Commission and contain a large volume of searches from English into 20 EU languages that staff translators working for the EU translation services submitted to an internally available multilingual concordancer. The study attempts to (i) identify differences in the searches (i.e. problems) based on the language pairs; and (ii) group problems into types. Furthermore, the interactions between concordance users and the tool itself have been examined to provide a translation-oriented perspective on the domain of Human-Computer Interaction.

The study draws on the literature on translation problems, Information Retrieval and Web search log analysis, moving from the assumption that in the perspective of concordance searching, translation problems are best interpreted as information needs for which the concordancer is chosen as a form of external support. The structure of a concordance search is examined in all its parts and is eventually broken down into two main components: the 'Search Strategy' component and the 'Problem Unit' component. The former was further analyzed using a mainly quantitative approach, whereas the latter was addressed from a more qualitative perspective. The analysis of the Problem Unit takes into account the length of the search strings as well as their content and linguistic form, each addressed with a different methodological approach. Based on the understanding of concordance searches as manifestations of translation problems, a user-centered classification of translation-oriented information needs is developed to account for as many "problem" scenarios as possible.

According to the initial expectations, different languages should experience different problems. This assumption could not be verified: the 20 different language pairs considered in this study behaved consistently on many levels and, due to the specific research environment, no definite conclusions could be reached as regards the role of the language family criterion for problem identification. The analysis of the 'Problem Unit' component has highlighted automatized support for translating Named Entities as a possible area for further research in translation technology and the development of computer-based translation support tools. Finally, the study indicates (concordance) search logs as an additional data type to be used in experiments on the translation process and for triangulation purposes, while drawing attention on the concordancer as a type of translation aid to be further fine-tuned for the needs of professional translators.

KEYWORDS

Translation Studies, Translation Process, Translation Problems, Concordancing Tool, European Union, Search Logs, Information Need

RIASSUNTO

Il presente lavoro consiste in uno studio empirico sui problemi di traduzione che emergono quando si considerano diverse coppie di lingue e in particolare sviluppa una metodologia per analizzare i *log* di ricerche effettuate dai traduttori in un software di concordanza (*concordancer*) quali manifestazioni di problemi di traduzione che, visti in una prospettiva più ampia, si possono anche considerare dei "bisogni d'informazione" (*information needs*). I log di ricerca costituiscono una tipologia di dato ancora relativamente nuova e inesplorata nell'ambito delle ricerche sul processo di traduzione e pertanto è emersa la necessità di svolgere un'analisi di tipo esplorativo in un contesto controllato onde evitare le problematiche aggiuntive derivanti da un numero eccessivo di variabili. I log di ricerca sono stati raccolti presso la Commissione europea e contengono quantitativi ingenti di ricerche effettuate dai traduttori impiegati presso i servizi di traduzione dell'Unione europea in un concordancer multilingue disponibile come risorsa interna. L'analisi si propone di individuare le differenze nelle ricerche (e quindi nei problemi) a seconda della coppia di lingue selezionata e di raggruppare tali problemi in tipologie. Lo studio fornisce inoltre informazioni sulle modalità di interazione tra gli utenti e il software nell'ambito di un contesto traduttivo, contribuendo alla ricerca nel campo dell'interazione uomo-macchina (*Human-Computer Interaction*).

Il presente studio trae spunto dalla letteratura sui problemi di traduzione, sull'estrazione d'informazioni (*Information Retrieval*) e sulle ricerche nel Web e si propone di considerare i problemi di traduzione associati all'impiego di uno strumento per le concordanze quali bisogni di informazione per i quali lo strumento di concordanze è stato scelto come forma di supporto esterna. Ogni singola ricerca è stata esaminata e scomposta in due elementi principali: la "strategia di ricerca" (*Search Strategy*) e l'"unità problematica" (*Problem Unit*) che vengono studiati rispettivamente usando approcci prevalentemente di tipo quantitativo e qualitativo. L'analisi dell'unità problematica prende in considerazione la lunghezza, il contenuto e la forma linguistica delle stringhe, analizzando ciascuna con una metodologia di lavoro appositamente studiata. Avendo interpretato le ricerche di concordanze quali manifestazioni di bisogni d'informazione, l'analisi prosegue con la definizione di una serie di categorie di bisogni d'informazione (o problemi) legati alla traduzione e incentrati sul singolo utente al fine di includere quanti più scenari di ricerca possibile.

L'assunto iniziale in base al quale lingue diverse manifesterebbero problemi diversi non è stato verificato empiricamente in quanto le 20 coppie di lingue esaminate hanno mostrato comportamenti alquanto similari nei diversi livelli di analisi. Vista la peculiarità dei dati utilizzati e la specificità dell'Unione europea come contesto di ricerca, non è stato possibile ottenere conclusioni definitive in merito al ruolo delle famiglie linguistiche quali indicatori di problemi, rispetto ad altri criteri di classificazione. L'analisi dell'unità problematica ha evidenziato le entità denominate (*Named Entities*) quale possibile oggetto di futuri progetti di ricerca nell'ambito delle tecnologie della traduzione. Oltre a offrire un contributo per i futuri sviluppi nell'ambito dei supporti informatici alla traduzione, con il presente studio si è voluto altresì presentare i log delle ricerche (di concordanze) quale tipologia aggiuntiva di dati per lo studio del processo di traduzione e per la triangolazione dei risultati empirico-sperimentali, cercando anche di suggerire possibili tratti migliorativi dei software di concordanza sulla base dei bisogni di informazione riscontrati nei traduttori.

ACKNOWLEDGMENTS

On second thoughts, I think this section would be better entitled "Opening Credits".

In my early days as a PhD student, I remember reading (or hearing) somewhere that a PhD is "a solitary endeavor". Needless to say, this statement added a grim note to the seemingly long years already looming ahead. With hindsight, I would rephrase it as "solitary *mental* endeavor" but definitely not as "lonely" as that statement made it sound in the very beginning. In fact, this adventure has been more like directing a movie than sitting in a library cubicle chewing on a pencil. And if you are reading this, it means that the movie has become a reality and many people deserve to be thanked.

The name of this work, "Ph.D.", sounds like the movie title following the award in the category of "Dr." — Director, of course. In a similar fashion, this introductory section will most likely resemble an acceptance speech. At this point, the person standing before the audience is in a state of frenzy, trying to thank all the people who contributed to the movie, desperately trying not to forget anyone and squeezing in as many names as possible before the conductor brandishes the infamous stick, the music starts to play and your time is up. However witty, it would not be an acceptance speech if it didn't start with the proverbial opening line *I would like to thank*...

Undoubtedly, the persons I have to thank the most are Producer Prof. Federica Scarpa and Co-producer Dr. Giuseppe Palumbo who initially suggested that I consider making a movie after graduating and, when I said 'ok', let me spread my wings to pursue some kind of idea that had taken shape in my head. Without them – and the University of Trieste as a Production Company – this movie would not exist.

The original idea came when I was preparing to audition for a role in this movie, so the first person I should be thanking is probably Prof. Lynne Bowker for the papers she has written. Before I could even go to my audition, I visited another place: Luxembourg. This is where I met the soon-to-be Production Management crew at the European Parliament and Commission. It has been a key experience for the future development of the plot because after becoming the director of the movie I decided that Luxembourg would be where most of the action would take place. Denis Navarre, my Production Designer, Fons De Vuyst and Josep Bonet, the Executive Producers, are the people I owe most during my research stays but I am also grateful to all the nice people who put up with me while I was there, gave me some of their time and made sure I got regularly fed at lunch: Paula Álvarez, Andreas Eisele, Hilário Fontes, Mark Röder, Micha, Luis and Bart. Not less helpful was the Second Unit at the European Parliament: Lucia Magris, Franco Urzì, the whole Italian unit at DG TRAD, Ivo Tampieri and Felicity Hands. I am also thankful to Christine Laaboudi from the Publication Office, Aleksandra Kowalska from the EC and Eric Davies for their additional help. A special thank goes to Alexandros, the Production Assistant, for listening to my initial rants about the unlikely storyboard I was working on. Believe it or not, you gave me the moral support I very much needed back then, you inspired me to push myself and try out new and unbeaten paths and were always available whenever I needed some help. Parts of this work virtually bear your name, so, yes, you can now add a checkmark next to 'PhD' on your own to-do list.

However, Luxembourg turned out to be just one of the shooting locations for this movie. Every other shooting location — Copenhagen, Saarbücken and Rome — has been crucial for the development of the plot not so much for the work carried out on site but for the film crew that I had the pleasure to work with and who played a decisive role in the outcome of the movie.

Not only did I get to spend time in one of the cradles of process research, but I also had the privilege to work with an outstanding guest Production Supervisor: Prof. Arnt Lykke Jakobsen. I can see the conductor wobbling the stick so I shall just say that I don't have enough words to express my gratitude to him for everything that he has taught me academically, professionally and personally. I am also thankful to the rest of the CRITT group & Co. — in particular Michael, Kristian, Barbara and Esben. My gratitude then goes to Prof. Erich Steiner for accepting me as part of his thriving research team, who quickly adopted me: Katja, José, Kerstin, Peggy, Katrin, Jörg, Marilisa, Pauline, Tabea and all my fellow students and researchers at DFKI, FR 4.6 and COLI departments. In Rome, I found the best Special Effects crew at Translated. I truly thank them for the time they shared with me, the opportunities and help they provided and for the good times we had.

A very special thank goes to the Animation Unit, a group of selected few who volunteered to make this study possible by helping me with the very much needed scripts and for their patience while I learned to use the command line and read a little of at least four new languages — only this time, scripting languages: Denis (EC), Dan and Michael (CBS), Sabine, Philip, Hannah, Katja, Richard and Nora (UniSaarland), Antonio, Gianluca, Alberto and Evan (Translated). At different stages, a vital role was played by the Number Chrunching Unit, charged with the burdensome task of making some sense out of my attempts at statistical analysis. For this, I am grateful first and foremost to Vahram (UniSaarland) but also to Prof. Gabriella Schoier (UniTs), Laura (CBS), Hannah and Marilisa (UniSaarland), Marco Trombetti (Translated) and Marco Turchi (FBK). At this point, I should probably mention the person who made it possible for me to learn any programming in the first place by making sure I would get my current laptop (and by pimping it up): thank you uncle Sandro!

A mention for outstanding achievement goes to my Stuntman, Max, who not only took care of some Visual Effects and Animations but also literally saved this movie on several occasions when my equipment stopped working, preferably when I was abroad or about to leave. Thank you, Max, for your relentless help, incredible patience, brilliant teaching and precious tips but, most importantly, for always being there.

I am also indebted to the Logistics Unit that was always ready to take care of transfers of any kind and magnitude. Thank you, Dad. Thank you, Mom, for taking care of me and supporting me during this whole adventure with warmth and kindness. Thank you to Gabri for providing me with a special shelter for the final rush. Thank you to the Sound Unit, Riccardo, Solo, Nicolas and Dan who added a new soundtrack to my life and made sure I would make up for the interminable hours of sitting at my desk with some salsa.

Last but not least, I would like to express my gratitude to my official and unofficial families, Cecilia for her unconditional love, Nathaniel for his unrelenting support, Francesco for the many car rides, all the people whose paths have crossed mine in the past three years and who made a difference in my life, my old and new friends, PhD colleagues and anyone whose name I have failed to mention explicitly. I have had the privilege to exchange ideas with remarkable people and professors, among whom the late Prof. Miriam Schlesinger who pointed out, when I was only at the end of my first year, that the focus of my work would be on the methodology and, unsurprisingly, she was right. The last thought goes to 'PhD Comics' for their tongue-in-cheek truths and all PhD students whom I met along the way and who have already delivered their speech, are awaiting the award ceremony or are still working on their own movie.

The movie script a few pages ahead is the result of a combination of solitary work, a great deal of thinking and good spells of teamwork but, ultimately, any fault in the present work is entirely my own.

February 2013

TABLE OF CONTENTS

ABSTRACT IV
RIASSUNTO V
ACKNOWLEDGMENTS VI
LIST OF TABLES XI
LIST OF FIGURES XIII
LIST OF ABBREVIATIONS XV

CHAPTER 1: INTRODUCTION 1
1.1 AIMS, HYPOTHESES and RESEARCH QUESTIONS 2
1.2 SCOPE of the STUDY 3
1.2.1 *WHAT*: the DATA 4
1.2.2 *WHERE*: the DATA SOURCE 4
1.2.3 *WHO*: the TRANSLATORS 4
1.2.4 *WHEN*: the TIME FRAME 4
1.2.5 *WHY*: the RATIONALE 4
1.2.6 *HOW*: the METHODOLOGY 5
1.2.7 LIMITATIONS 6
1.3 STRUCTURE of the THESIS 7
CHAPTER 2: 'PROBLEM'-RELATED CONCEPTS IN TRANSLATION STUDIES 8
2.1 TRANSLATION PROBLEMS 8
2.2 PROBLEMS as DIFFICULTY & UNCERTAINTY 10
2.3 PRODUCT-ORIENTED APPROACH 13
2.3.1 PROBLEMS as ERRORS 13
2.4 SUBJECT-ORIENTED APPROACH 14
2.4.1 DATA ELICITATION BASED ON Q&A 14
2.5 PROCESS-ORIENTED APPROACH 16
2.5.1 STUDYING PROBLEMS with CONCURRENT VERBALIZATIONS 16
2.5.2 PAUSES AS MANIFESTATIONS OF PROBLEMS 18
2.6 TRANSLATION as PROBLEM SOLVING 22
2.7 CLASSIFICATION of PROBLEMS 24
2.8 UNITS in TRANSLATION RESEARCH 26
2.8.1 TRANSLATION UNIT 26
2.8.2 PROBLEM UNIT 29
2.8.3 ATTENTION UNIT 30
2.8.4 COGNITIVE UNIT 31
2.9 RELATIONSHIPS AMONG UNITS 32
2.10 KEY CONCEPTS 35
CHAPTER 3: OVERVIEW OF CONCORDANCING TOOLS 36
3.1 TYPES of CONCORDANCING TOOLS 36
3.1.1 CONCORDANCERS in CORPUS STUDIES and ACADEMIA 36
3.1.2 CONCORDANCERS for the TRANSLATION PROFESSION 38
3.2 CONCORDANCERS in the TRANSLATION INDUSTRY 41
3.2.1 OFF-LINE CONCORDANCERS 41
3.2.2 ONLINE CONCORDANCERS 44
3.2.2.1 TAUS SEARCH 44
3.2.2.2 MYMEMORY 46
3.2.2.3 LINGUEE 47
3.2.2.4 GLOSBE 49

3.2.2.5 TRADOOIT 50
3.2.2.6 WEBITEXT 51
3.2.2.7 TRANSSEARCH 53
3.2.2.8 OTHER CONCORDANCERS 55
3.2.3 INTRANET-BASED CONCORDANCERS 56
3.2.3.1 EURAMIS 57
3.2.3.2 The EURAMIS CONCORDANCER 58
3.2.3.3 QUEST 62
3.2.4 CONCORDANCE USERS' PROFILES 64
3.3 KEY CONCEPTS 67
CHAPTER 4: CONCORDANCE SEARCHES, TRANSLATION PROBLEMS AND INFORMATION NEEDS 68
4.1 INTERNAL and EXTERNAL SUPPORT 68
4.2 CONCORDANCE SEARCHES & TRANSLATION PROBLEMS 71
4.3 CONCORDANCE SEARCHES in PROCESS RESEARCH 72
4.4 TRANSLATION PROBLEMS as INFORMATION NEEDS 74
4.5 CONCORDANCE SEARCHES vs. WEB SEARCH LOGS 78
4.6 STRUCTURE of a CONCORDANCE SEARCH 81
4.7 KEY CONCEPTS 84
CHAPTER 5: RESEARCH DESIGN 86
5.1 RESEARCH ENVIRONMENT 87
5.1.1 EUROPEAN COMMISSION 91
5.1.2 EUROPEAN PARLIAMENT 92
5.1.3 LANGUAGE POLICY for the EU TRANSLATORS 92
5.2 RESEARCH PARTICIPANTS 93
5.3 EXPERTISE 94
5.4 DATASET for the STUDY 96
5.4.1 TIME SPAN 97
5.4.2 DISTRIBUTION of SEARCHES by SOURCE LANGUAGE 98
5.4.3 DISTRIBUTION of SEARCHES by TARGET LANGUAGE 99
5.4.4 DISTRIBUTION of SEARCHES by INSTITUTION 101
5.5 DATASET PRE-PROCESSING 104
5.5.1 The FINAL DATASET 106
5.5.2 LANGUAGE FAMILIES 107
5.5.3 LANGUAGE 'AGE' 111
5.6 SEARCH SESSIONS 113
5.6.1 STANDARD DEVIATION and COEFFICIENT of VARIATION 116
5.6.2 LEVELS of ANALYSIS 118
5.7 KEY CONCEPTS 118
CHAPTER 6: ANALYSIS OF THE 'SEARCH STRATEGY' COMPONENT 119
6.1 STRING LENGTH 119
6.1.1 QUERY REFINEMENT CATEGORIES 126
6.1.1.1 RESUBMISSION 127
6.1.1.2 FORMAL CHANGES 127
6.1.1.3 REDUCTION 127
6.1.1.4 EXPANSION 127
6.1.1.5 REPLACEMENT 128
6.1.1.6 MIXED STRATEGY 128
6.1.2 CATEGORY DISTRIBUTION 129
6.1.2.1 MANUAL CHECK on FINNISH 132
6.1.2.2 REDUCTION vs. EXPANSION 138
6.2 TOOL SETTINGS 139
6.2.1 AUTOMATIC METADATA 139
6.2.1.1 DATE and TIME STAMP 139
6.2.1.2 EXECUTION TIME and RESULTS 140
6.2.2 SEARCH SETTINGS 143

6.2.2.1 INTERFACE and SEARCH MODE 143
6.2.2.2 SEARCH METHOD and DIRECTIONALITY 146
6.2.3 DOCUMENT-BASED FILTERS 147
6.2.3.1 DATABASE FILTER 147
6.2.3.2 REQUESTING SERVICE and DOCUMENT TYPE 149
6.2.3.3 YEAR(S) 150
6.2.4 FILTERS OVERVIEW 152
6.3 KEY CONCEPTS 154
CHAPTER 7: ANALYSIS OF THE 'PROBLEM UNIT' COMPONENT 155
7.1 STRING LENGTH 156
7.1.1 CUT-OFF LENGTH 157
7.1.1.1 STRING CATEGORIES 161
7.1.1.2 NAMED ENTITIES 162
7.1.1.3 OTHER STRING TYPES 163
7.1.1.4 N-GRAMS and CATEGORY DISTRIBUTION 164
7.1.1.5 CORE STRINGS 169
7.2 STRING CONTENT 172
7.2.1 EUROVOC 175
7.2.1.1 STRUCTURE of EUROVOC 176
7.2.2 METHODOLOGY 178
7.2.2.1 PRE-PROCESSING of the DESCRIPTOR LIST 178
7.2.2.2 PRELIMINARY TESTS 179
7.2.2.3 EDITING of the DESCRIPTORS 181
7.2.3 DOMAIN ANALYSIS TEST-PHASE 182
7.3 LINGUSTIC FORM of CONCORDANCE SEARCH STRINGS 188
7.3.1 LINGUISTIC CATEGORIES of WEB QUERIES 188
7.3.1.1 POS-TAGGING 189
7.3.1.2 VARIATION ACROSS STRINGS in SEARCH SESSIONS 193
7.3.2 PROBLEM CATEGORIES in EMPIRICAL STUDIES of TRANSLATION 199
7.3.2.1 COMPOUNDS 202
7.3.2.2 COLLOCATIONS 204
7.3.2.3 LANGUAGE CHAINS and PHRASEOLOGY 206
7.3.2.4 FORMULAIC SEQUENCES 208
7.3.2.5 TERMS 209
7.3.2.6 MULTI-WORD UNITS 210
7.3.3 The BILINGUAL MENTAL LEXICON 213
7.3.3.1 The TOT STATE 214
7.3.4 IMPLICIT vs. EXPLICIT INFORMATION NEEDS 217
7.3.5 TRANSLATION PROBLEM SCENARIOS 220
7.4 KEY CONCEPTS 226
CHAPTER 8: CONCLUSIONS 227
8.1 [RQ1] The LANGUAGE PAIR 228
8.2 [RQ2] The PROBLEM UNIT 229
8.3 [RQ3] The SEARCH STRATEGY 230
8.4 INFORMATION NEEDS 231
8.5 FUTURE AVENUES of RESEARCH 233
REFERENCES 235
APPENDIX 256
APPENDIX A 257
APPENDIX B 262
APPENDIX C 266
APPENDIX D 276
APPENDIX E 282

LIST OF TABLES

TABLE 1. PROPOSED TAXONOMY AND HIERARCHY OF LABELS FOR THE UNITS OF TRANSLATION ACTIVITY. 34
TABLE 2. SUMMARY OF RELEVANT FINDINGS OF THE TRANSSEARCH USER QUESTIONNAIRE, AS FOUND IN MACKLOVITCH ET AL. (2000: 1204-5). 54
TABLE 3. DISTRIBUTION OF EURAMIS USERS ACROSS THE EU INSTITUTIONS. 57
TABLE 4. TRANSLATING STAFF, QUEST TOTAL AND ACTIVE USERS BOTH IN 2010 AND IN SEPTEMBER 2010, TOGETHER WITH ESTIMATE FOR QUERIES DIVIDED PER INSTITUTION (SEE TABLE 3 IN SECTION 3.2.3.1 FOR ABBREVIATIONS). 88
TABLE 5. DISTRIBUTION OF SEARCHES PER TARGET LANGUAGE IN ASCENDING ORDER (740,000 DATASET). 105
TABLE 6. TARGET LANGUAGE DISTRIBUTION FOR THE FINAL DATASET (724,000) IN ASCENDING ORDER. 107
TABLE 7. LANGUAGE FAMILY DISTRIBUTION FOR THE LANGUAGES INVOLVED IN THE STUDY. 108
TABLE 8. DISTRIBUTION OF TARGET LANGUAGES ACCORDING TO THEIR RELATIVE AGE AS OFFICIAL EU LANGUAGES WITH RESPECT TO THE 2004 ENLARGEMENT. 111
TABLE 9. TARGET LANGUAGE DISTRIBUTION FOR THE FINAL DATASET (724,000) IN ASCENDING ORDER AND DISTRIBUTION OF LANGUAGE FAMILY AND RELATIVE AGE OF THE LANGUAGE IN EU TERMS. 112
TABLE 10. SIZE OF EURAMIS TM DATABASES PER EACH TARGET LANGUAGE IN TERMS OF NUMBER OF STORED SEGMENTS AS OF 4TH JANUARY 2010 (11 TMS FROM EURAMIS INCLUDED). 113
TABLE 11. A SMALL EXCERPT OF SEARCH LOGS FROM THE POLISH SUBSET SHOWING INSTANCES OF QUERIES. 114
TABLE 12. OVERVIEW OF THE DISTRIBUTION OF SEARCH SESSIONS AND SPOT SEARCHES ACROSS ALL LANGUAGES. 116
TABLE 13. COMPARISON OF PERCENTAGE DISTRIBUTION OF QUERY LENGTH BETWEEN THE TRANSSEARCH AND THE EURAMIS DATASETS. 120
TABLE 14. DISTRIBUTION OF STRING AVERAGE LENGTH ACROSS LANGUAGES FOR THE WHOLE DATASET OF 724,000 (BOTH TYPES AND TOKENS) AND FOR SEARCH SESSIONS AND SPOT SEARCHES. 122
TABLE 15. TAXONOMY OF REFORMULATION STRATEGIES FOR QUERY LOGS ACCORDING TO HUANG AND EFTHIMIADIS (2009). 126
TABLE 16. VERBATIM EXAMPLES OF QUERIES FOR EACH IDENTIFIED CATEGORY AND SUB-CATEGORY IN A SEARCH SESSION. 129
TABLE 17. LIST OF CATEGORY CODES EMPLOYED TO REFER TO A MACRO-CATEGORY AND EACH OF ITS SUB-TYPES. 130
TABLE 18. DISTRIBUTION OF SESSIONS CATEGORIES ACROSS ALL LANGUAGES AFTER THE AUTOMATIC LABELING. 130
TABLE 19. OVERVIEW OF MANUALLY IDENTIFIED CATEGORIES AND SUB-CATEGORIES (WITH EXAMPLES) AND EXEMPLIFICATION OF THE ABSTRACT STRUCTURE OF EACH SESSION. 133
TABLE 20. EXAMPLES OF SESSIONS LABELED E11. 135
TABLE 21. RESULTS FROM MANUAL CATEGORIZATION VS. RESULTS FROM PHP SCRIPT. 136
TABLE 22. PERCENTAGE DISTRIBUTION OF FAILED AND SUCCESSFUL SEARCHES FOR EACH OF THE THREE SUBSETS (724,000, SESSION AND SPOT) WITH A FURTHER BREAKDOWN FOR SUCCESSFUL SEARCHES. 142
TABLE 23. PERCENTAGE DISTRIBUTION OF SEARCH MODES FOR EACH GROUP OF STRINGS. 144
TABLE 24. DISTRIBUTION OF SEARCH METHOD ACCORDING TO SEARCH MODE FOR THE MAIN DATASET (724,000). 147
TABLE 25. OVERVIEW OF AVAILABLE TMS IN EURAMIS. THE FIRST HALF OF THE TABLE LISTS THE INTERINSTITUTIONAL MEMORIES, THE BOTTOM HALF LISTS TMS ACCESSIBLE TO A SINGLE INSTITUTION. 148
TABLE 26. DISTRIBUTION OF THE MOST POPULAR SINGLE AND MULTIPLE SELECTIONS OF YEARS. 150
TABLE 27. DISTRIBUTION OF NON-DEFAULT ADVANCED FILTERS FOR EACH SUBSET OF SEARCHES. PERCENTAGES WERE CALCULATED AS AN OVERALL AVERAGE: FIRST, AVERAGES FOR EACH LANGUAGE WERE OBTAINED AND THEN NUMBERS WERE NORMALIZED ON THE TOTAL NUMBER OF ADVANCED SEARCHES (EXCEPT FOR THE FIRST LINE). 153
TABLE 28. COMPARISON OF TOP 10 MOST SEARCHED BI-GRAMS AND TRI-GRAMS IN TRANSSEARCH (MACKLOVITCH ET AL. 2008: 415) AND EURAMIS. 160
TABLE 29. TOP 10 MOST FREQUENT 8- AND 11-GRAMS WITH ABSOLUTE FREQUENCY COUNTS (742,000 DATASET). 161
TABLE 30. LIST OF TRIGGER WORDS FOR THE IDENTIFICATION OF NAMED ENTITIES, GROUPED BY SUB-CATEGORY. 162
TABLE 31. OVERVIEW OF THE DISTRIBUTION OF THE MAIN CATEGORIES FOR NAMED ENTITY RECOGNITION FOR EACH SUB-SET AND RANGE VALUES FOR FREQUENCY COUNTS FOR THE 742,033 DATASET. 164
TABLE 32. DISTRIBUTION OF THE THREE CATEGORIES OF N-GRAMS REPRESENTING THREE DIFFERENT APPROACHES TO CONCORDANCE SEARCHES FOR EACH OF THE THREE ANALYZED DATASETS. 165
TABLE 33. POSSIBLE REALIZATIONS OF ONE SINGLE NAMED ENTITY, I.E. THE TITLE OF COMMISSIONER ASHTON, CALCULATED ON THE MAIN DATASET OF 724,000 STRINGS (ALL SEARCHES LOWERCASED). 167

TABLE 34. TOP 30 MOST SEARCHED STRINGS IN THE MAIN DATASET (724,000) WITH ABSOLUTE FREQUENCY COUNTS. ... 169
TABLE 35. DISTRIBUTION OF COLLOCATIONAL WINDOWS RELATED FREQUENCIES FOR TWO SAMPLE STRINGS EXTRACTED FROM THE TOP 100 MOST FREQUENT STRINGS IN THE OVERALL DATABASE (724,000). ... 170
TABLE 36. EXAMPLES OF FOUND TOPICAL CATEGORIZATIONS OF WEB QUERIES (CATEGORIES ARE OFTEN NON-MUTUALLY EXCLUSIVE). ... 173
TABLE 37. FIELDS AND HEADINGS OF THE THREE CANDIDATE EXTERNAL TAXONOMIES CONSIDERED FOR THE STUDY. ... 175
TABLE 38. DISTRIBUTION OF LGP AND LSP STRINGS FOR EACH LANGUAGE PAIR AT EACH OF THE THREE LEVELS NORMALIZED BY THE TOTAL NUMBER OF SEARCHES FOR EACH LANGUAGE. ... 182
TABLE 39. LIST OF DOMAINS AND CODES EMPLOYED IN THE DESCRIPTORS LIST (IN ASCENDING DESCRIPTOR ORDER). ... 184
TABLE 40. EXAMPLE OF POS-TAGGED STRINGS. FOUND TAGS (TO BE FOUND AFTER THE '/'): CC (COORDINATE CONJ.); DT (DETERMINER); IN (PREPOSITION/SUBORD. CONJ,); JJ (ADJECTIVE); MD (MODAL); NN (NOUN, SINGULAR OR MASS); NNS (NOUN PLURAL); NP (PROPER NOUN, SINGULAR); NPS (PROPER NOUN, PLURAL); POS (POSSESSIVE ENDING); VB (VERB, BASE FORM); VBG (VERB, GERUND OR PRESENT PARTICIPLE). ... 190
TABLE 41. AGGREGATE DISTRIBUTION OF THE MAIN CATEGORIES OF POS-TAGS FOR CA. 605,000 STRINGS. ... 191
TABLE 42. MOST FREQUENT POS PATTERNS IN THE N-GRAMS STRING CORPUS. ONLY THE TOP 20 POS COMBINATIONS OUT OF SOME 60,000 FOUND ARE SHOWN. ... 192
TABLE 43. CLASSIFICATION OF QUERIES IN SPOT SEARCHES GROUP AND SESSION CATEGORY A1 WITH VERBATIM EXAMPLES FROM THE LOGS. ... 194
TABLE 44. VERBATIM EXAMPLES OF ADDITIONS AND DELETION OF ELEMENTS IN NOUN PHRASES FOR THE TRIM DYNAMIC CATEGORY (C1, C2). ... 196
TABLE 45. VERBATIM EXAMPLES OF ADDITIONS AND DELETION OF ELEMENTS IN NOUN PHRASES FOR THE EXPANSION DYNAMIC CATEGORY (D1, D2). ... 196
TABLE 46. DELTA STRINGS FOR VP AND PREPOSITIONAL PHRASES. ... 197
TABLE 47. EXAMPLES OF ADDITION AND DELETION OF COORDINATE PHRASES OR CLAUSES FOR EACH DYNAMIC CATEGORY. 198
TABLE 48. CLASSIFICATION OF TRANSLATION PROBLEMS ACCORDING TO DÉSILETS ET AL. (2009). ... 200
TABLE 49. SIMPLIFIED CATEGORIZATION OF TYPES OF TRANSLATION PROBLEMS AS FOUND IN DÉSILETS, FARLEY ET AL. (2008). ... 201
TABLE 50. MOST FREQUENT COMBINATIONS FOR THE "/IN []* /IN" PATTERN. DATASET: 605,000 STRINGS BETWEEN 2 AND 11 WORDS. TOTAL NUMBER OF STRINGS MATCHING THE PATTERN: 2086). [DT=DETERMINER, NN(S)=NOUN, VBN=PAST PARTICIPLE, IN=PREPOSITION]. ... 218
TABLE 51. SUMMARY OF PROBLEM SCENARIOS WITH THEIR MOST LIKELY FLOWCHART PATH. ITEMS IN BRACKETS ARE THEORETICALLY POSSIBLE BUT LESS LIKELY TO OCCUR IN PRACTICE. ... 225

LIST OF FIGURES

Figure 1. TAUS Search interface with (language-dependent) advanced filters activated. 45
Figure 2. Results display in the TAUS Search Web interface. 46
Figure 3. MyMemory search interface. 46
Figure 4. MyMemory results page 47
Figure 5. Linguee main search interface. 48
Figure 6. Linugee results page. 48
Figure 7. Main Glosbe search interface. 49
Figure 8. Glosbe results page. 49
Figure 9. TradooIT search interface. 50
Figure 10. TradooIT results page 51
Figure 11. WeBiText Search interface with advanced search options activated. 52
Figure 12. WeBiText results page. 52
Figure 13. Euramis architecture (DGT 2010c). Items can be found in the top half of the green, orange and pink sections are illustrated at length in the text. 58
Figure 14. Euramis concordance Simple and Advanced interface as of 2012. 59
Figure 15. A closer look at the Euramis Advanced interface. 60
Figure 16. Euramis Results page (identical same for Simple and Advanced mode). 60
Figure 17. Example of the structure of the Euramis translation memory (DGT 2010c). 61
Figure 18. Quest main search interface. 63
Figure 19. Two interfaces of Quest. On the left, the list of resources available. Items in green are available for the selected languages. On the right, the Quest result page showing Eur-Lex as the active resource. 64
Figure 20. Quest result page with the Euramis concordance as the active resource. 64
Figure 21. Search-strategy steps in translation-problem solving using external support. 76
Figure 22. The classic model for IR (Broder 2002: 4). 76
Figure 23. Classic model for IR augmented for Web searching (Broder 2002: 4). 79
Figure 24. Model for external support usage in translation (adapted from Broder 2002: 4). 79
Figure 25. Breakdown of a concordance search with respect to the variables pertaining to the search strategy: string length and tool settings. 82
Figure 26. Breakdown of a Problem Unit into its various levels of analysis. 82
Figure 27. Complete breakdown of a concordance search (collated version from Figure 25 and Figure 26). 83
Figure 28. Distribution of queries submitted via Quest per requesting institution (September 2010). 89
Figure 29. Distribution of searches per institution (ALL>ALL) viz. Distribution of translating staff according to official statistics for 2010. 90
Figure 30. Distribution of total searches (970,000) per source language. 98
Figure 31. Breakdown of pages translated per source languages at DGT in 2008 (DGT 2009c: 7). 99
Figure 32. Breakdown of pages translated per target languages at DGT in 2008 (DGT 2009c: 7). 100
Figure 33. Distribution of total searches (963,000) per target language, with one single target language selected per search. 101
Figure 34. Distribution of total searches (ALL>ALL) according to submitting institution. 102
Figure 35. Distribution of requesting institution for EN>ALL (749,500 queries). 102
Figure 36. Distribution of requesting institution for FR>ALL (109,336 queries). 103
Figure 37. Graphical representation of the distribution of searches per target language (740,000). 105
Figure 38. Distribution of searches grouped by language family across each institution, normalized by the total number of searches for each institution (724,000). 108
Figure 39 Distribution of target languages within the institution COJ. 109
Figure 40. Distribution of target languages within two institutions: EP and COUNCIL. 110
Figure 41. List of European languages and year in which each language became an 'official language' of the EU. Note the major increase by 9 languages in 2004 (EC 2008: 3). 111

FIGURE 42. BREAKDOWN OF A CONCORDANCE SEARCH WITH RESPECT TO THE VARIABLES PERTAINING TO THE SEARCH STRATEGY: STRING LENGTH AND TOOL SETTINGS. 119
FIGURE 43. DISTRIBUTION OF SEARCHES (724,000) ACCORDING TO THE NUMBER OF WORDS IN A SEARCH STRING..... 120
FIGURE 44. GRAPHICAL REPRESENTATION OF THE DISTRIBUTION OF AVERAGE STRING LENGTH (IN WORDS, Y AXIS) ACROSS ALL LANGUAGES (X AXIS) FOR BOTH TOKENS AND TYPES 123
FIGURE 45. GRAPHICAL REPRESENTATION OF THE DISTRIBUTION OF AVERAGE STRING LENGTH (IN WORDS) ACROSS ALL LANGUAGES FOR BOTH SEARCH SESSION AND SPOT SEARCHES (TOKEN COUNT). 124
FIGURE 46. DISTRIBUTION OF AVERAGE STRING LENGTH PER INSTITUTION OVER THE WHOLE DATASET AND THE SESSION AND SPOT SUBSETS, RESPECTIVELY..... 125
FIGURE 47. DISTRIBUTION OF REQUESTS EACH DAY OF THE SELECTED MONTH (SEPTEMBER 2010). 140
FIGURE 48. DISTRIBUTION OF SEARCHES ON AN HOURLY BASIS. 140
FIGURE 49. DISTRIBUTION OF UNSUCCESSFUL SEARCHES (I.E. THOSE PRODUCING ZERO RESULTS) ACROSS LANGUAGE PAIRS, NORMALIZED BY THE TOTAL NUMBER OF STRINGS PER LANGUAGE SUBSET. 143
FIGURE 50. DISTRIBUTION OF SEARCH MODE (SIMPLE AND ADVANCED) ACROSS THE TARGET LANGUAGES (NORMALIZED BY TOTAL STRINGS PER LANGUAGE SUBSET). 145
FIGURE 51. DISTRIBUTION OF UNSUCCESSFUL SEARCHES (ZERO RESULTS) BETWEEN SIMPLE AND ADVANCED MODE FOR EACH TARGET LANGUAGE, NORMALIZED BY THE TOTAL NUMBER OF SEARCHES PER TARGET LANGUAGE. 145
FIGURE 52. DISTRIBUTION OF UNSUCCESSFUL SEARCHES (ZERO RESULTS) ACROSS DIFFERENT SUBMITTING INSTITUTIONS, NORMALIZED BY THE TOTAL NUMBER OF SEARCHES FOR EACH INSTITUTION. 146
FIGURE 53. DISTRIBUTION OF THE YEAR FILTER COMPARED WITH THE DISTRIBUTION OF SEARCHES WHERE "2010" WAS SELECTED. RESULTS WERE NORMALIZED BY THE TOTAL NUMBER OF ADVANCED SEARCHES PER LANGUAGE AND SHOW THAT 2010 WAS SELECTED IN THE VAST MAJORITY OF THE SEARCHES. 151
FIGURE 54. DISTRIBUTION OF THE MAX RESULTS FILTER ACROSS TARGET LANGUAGES, NORMALIZED BY THE TOTAL NUMBER OF ADVANCED SEARCHES PER LANGUAGE. 152
FIGURE 55. BREAKDOWN OF A PROBLEM UNIT INTO ITS VARIOUS LEVELS OF ANALYSIS. 155
FIGURE 56. CONTINUUM REPRESENTING THE MOST LIKELY SEARCH APPROACH AS SEARCH STRING LENGTH INCREASES. 157
FIGURE 57. DISTRIBUTION OF TYPE/TOKEN RATIO FOR EACH N-GRAM GROUP FROM 1- TO 20 (742,000 DATASET)..... 158
FIGURE 58. HIERARCHICAL STRUCTURE OF FIELDS AND MICRO-THESAURI (LEFT) AND BREAKDOWN OF THE HIERARCHICAL STRUCTURE AT MICRO-THESAURUS LEVEL AND BELOW (RIGHT). 177
FIGURE 59. EXAMPLE OF A TERMINOLOGICAL LIST IN EUROVOC. 177
FIGURE 60. AVERAGE DISTRIBUTION PER DOMAIN CALCULATED FOR EACH OF THE THREE LEVEL OF ANALYSIS (MAIN, SESSION AND SPOT), NORMALIZED BY THE TOTAL NUMBER OF LSP STRINGS (724,000). 184
FIGURE 61. DISTRIBUTION OF DOMAINS AND ZERO RESULTS (SL SAMPLE TAKEN FROM THE 724,000 SET). 185
FIGURE 62. DISTRIBUTION OF DOMAINS ACCORDING TO LANGUAGE FAMILIES. THE DATASET USED AMOUNTED TO SOME 510,000 QUERIES AND WAS NORMALIZED USING THE TOTAL AMOUNT OF LSP STRINGS. 185
FIGURE 63. DISTRIBUTION OF DOMAINS ACCORDING TO THE LANGUAGE AGE CLUSTERING (510,000 NORMALIZED BY TOTAL LSP STRINGS)..... 186
FIGURE 64. JOINT FREQUENCY DISTRIBUTION OF SEARCH MODE AND DESCRIPTORS DOMAINS, NORMALIZED BY THE TOTAL STRINGS IN EACH DOMAIN..... 187
FIGURE 65. DIAGRAM SUMMARIZING THE MAIN POSSIBLE SCENARIOS BEHIND A TRANSLATION-RELATED INFORMATION NEED. THE YELLOW BOXES REFER TO THE TRADITIONAL DICHOTOMY IN TPR BETWEEN RECEPTION AND PRODUCTION PROBLEMS WHEREAS THE RED DIAMONDS SPECIFY THE HYPOTHESIZED LEVEL OF INFORMATION NEED. 221

LIST OF ABBREVIATIONS

A/ADJ/JJ	Adjective
AU	Attention Unit
BC	Bilingual Concordancer
BG	Bulgarian
CAT	Computer-Assisted Translation
CI	Contextual Inquiry
CNA	Choice-Network Analysis
COA/CDCE	Court of Auditors
COJ/CDJ	Court of Justice
COR/CDR	Committee of the Regions
CS	Czech
CTR	Click-through Rate
CV	Coefficient of Variation
DA	Danish
DE	German
DET/DT	Determiner
DGT	Directorate General for Translation
DTS	Descriptive Translation Studies
EBMT	Example-Based Machine Translation
EC/CMS	European Commission
ECB	European Central Bank
EESC/CES	European Economic and Social Committee
EIB	European Investment Bank
EL	Greek
EP/PE	European Parliament
ES	Spanish
ES	External Support
ET	Estonian
EU	European Union
FI	Finnish
FR	French
GA	Gaelic
HCI	Human-Computer Interaction
HR	Croatian
HU	Hungarian
IN	Preposition
IR	Information Retrieval
IS	Internal Support
IT	Italian
KI	Knowledge Item
KS	Knowledge State
LGP	Language for General Purposes
LSP	Language for Special Purposes
LT	Lithuanian
LV	Latvian
MD	Modal verb
MT	Maltese
MT	Machine Translation
MWU	Multi-Word Unit
N/NN(s)/NP	Noun(s)/ Noun Phrase
NE	Named Entity/-ties
NL	Dutch
NLP	Natural Language Processing
P/PP	Preposition/Prep. Phrase
PL	Polish
PN	Proper Nouns
POS	Part-of-Speech
POS	Possessive
PT	Portuguese
PU	Problem Unit
Q&A	Questions & Answers
RO	Romanian
RVP	Retrospective Verbal Protocols
SD	Standard Deviation
SK	Slovak
SL	Slovenian
SL	Source Language
ST	Source Text
SV	Swedish
TAPs	Think-Aloud Protocols
TC/CDT	Translation Centre
TL	Target Language
TLA	Transaction Log Analysis
TM	Translation Memory
TOT	Tip-of-the-Tongue
TPR	Translation Process Research
TR	Turkish
TS	Translation Studies
TT	Target Text
TU	Translation Unit
UAD	User Activity Data
VB/VBG/VBN	Verb
XML	Extensible Markup Language

CHAPTER 1: INTRODUCTION

The present study aims to contribute to the research on translation as a process as well as the development of more targeted forms of translation support. The analysis focuses on a specific tool generally used by professional translators which can provide researchers with large volumes of authentic data about the difficulties encountered by translators while doing their work. The tool is a multilingual concordancer, which is also well known among scholars doing research in the field of corpus linguistics. The concordancer used as a translation aid can log users' activity and keep track of all the searches translators submit in the system. In particular, concordance searches can be seen as manifestations of translation problems, which are the main focus of this empirical investigation. According to Lörscher (1991a: 92),

> [t]he concept of 'translation problem' which manifests itself empirically has hardly been paid any attention by translation theory. The reason for this may be the fact that translation theory has had a strong theoretical, speculative, and thus non-empirical orientation [...] Although taxonomies or typologies of translation difficulties are sometimes given in the relevant literature [...], they are theoretical constructs which are based on single, individual, and largely unsystematic observations or, what is more likely, are hypothetically derived from a comparison of source- and target-language phenomena in a contrastive-linguistic way. Only very recently, with the development of a new empirical and process-oriented paradigm within translation studies [...], has the concept of 'translation problem' — and with it, that of 'translation strategy' — gained crucial importance.

Though this statement should be adjusted to account for the developments in translation process research that have taken place since the late 1990s with the introduction of keystroke logging, eye-tracking and screen recording, it still seems to apply to those empirical studies focusing on professional translators in their normal working environment.

According to Holmes' map of Translation Studies ([1972]2004), translation process research traditionally belongs to the *Pure* branch and more specifically to the sub-branch of Descriptive Translation Studies (DTS). A few years ago, the *Applied* branch was updated by Quah (2006) to include the most important technological developments. She identified four sub-branches, namely "translation training", "translation technology", "translation policy" and "translation criticism". As far as "translation technology" is concerned, a further break down into automatic translation tool (i.e. machine translation) and CAT tools is provided. The CAT tool group is further sub-divided into translation, linguistic and localization tools with additional branches departing from each group. Concordancers, in particular, are listed under the "linguistic" and more precisely "language-independent" sub-branching. The present study aims at looking at a specific aspect of the translation process that involves the use of concordancers, thus combining the applied with the descriptive branches. This is not the first study attempting to bring together translation tools and the study of the translation process (e.g. Dragsted 2004) but it is undoubtedly one of the very few to focus specifically on the concordancer and attempt a large-scale analysis in terms of language combinations. One of the goals is to gain further insights into how translators interact with this specific resource with a view to improving existing forms of translation support:

> [T]here is [...] considerable scope for research into the use of translation tools by translators, which may in turn lead to further improvement in the tools themselves but can also be expected to continue to our knowledge of the process of human translation in the modern idiom (Quah 2006: 196).

Recently, Chesterman (2011) has built on Toury's (1995: 249) distinction between the *translation act* and the *translation event*, the former being the actual individual translation task and the latter comprising the broader sociological framework within which the translation act takes place. Building further on this distinction, the present study will concentrate on the translation act and the analysis will have a very specific context, i.e. the translation services of the European Union. Research on the translation act can choose between at least two major approaches. The product-oriented approach is normally corpus-based and can investigate the features and properties of translated texts with a particular focus on translation errors or variability. Alternatively, the process-oriented approach is generally experimental and either investigates expert performance as opposed to student performance or focuses on psycholinguistic aspects of the translation process using a specific group of subjects. Rather than conducting a full-blown experiment, this empirical analysis draws data and information from a special kind of corpus and builds on existing findings from experiments and studies in the field of translation process research (e.g. Krings 1986; Lörscher 1991; Dragsted 2004) with a special focus on problem-solving as well as gaze and pause patterns in text comprehension and production.

1.1 AIMS, HYPOTHESES AND RESEARCH QUESTIONS

The present study aims at gaining further insights into the way professional translators operate in their daily work by looking at concordance searches. A concordancer enables users to retrieve from a repository text segments that match the text used as input for the search and obtain a translated version of that segment. Using concordance searches for research means resorting to a relatively unexplored data type in an attempt to "focus research and development efforts on addressing problems actually encountered in the workplace instead of introducing technology based on assumed needs" (Karamanis *et al.* 2010). Hopefully, findings from this study will be useful to tool developers to better target user needs and fine-tune existing forms of support or develop new ones.

This research project sets out to study translation problems in a systematic way across the largest possible number of language pairs under comparable conditions. To this purpose the translation services of the European Union provide the ideal test bed because their internal structure and daily practices and workflow allow for comparability of the data without necessarily limiting their volume.

The first research question to be addressed will be:

[RQ1] How do translation problems vary across different language pairs?

This question presupposes that there are in fact differences across different language pairs in terms of translation problems, thereby suggesting that the language pair might affect the problems translators encounter. This assumption derives from the expectation that translators will encounter different problems according to the specific language pair they are working with (Göpferich & Jääskeläinen 2009: 182). More specifically, translation between languages belonging to the same language family is expected to be more straightforward than translation between languages that have little in common. If English were taken as source language, translation into Germanic languages should theoretically be less demanding. As a consequence, the problems encountered when

translating into Germanic languages should present different features from those experienced when translating into languages belonging to, say, the Finno-Ugric family. Based on this assumption, the amount of "language-specific" problems can be expected to vary according to the language pair considered. In addition to language-specific problems, a number of "global factors" can be expected, i.e. recurring patterns that are not language-dependent. This initial assumption about the existence of language-specific problems as well as global problems will need to be further developed before addressing the main research question, which focuses specifically on the language-specific group.

Global problems are worthwhile being investigated on a large scale even when the language variable does not play a major role. To this purpose, a more general approach will be taken when investigating the following research question:

[RQ2] What are the types of searches translators submit into a concordancer?

This question first of all considers all searches submitted into a concordancer and focuses in particular on finding global recurring patterns in those searches. This involves categorizing the search strings. Seen from another perspective, this question can still relate to the first one in that the global analysis of searches might still highlight differences from language to language. At the same time, a concordance search is not only characterized by the searched-for text but can be looked at from the perspective of user behavior when accessing a concordancer, which generates the third research question:

[RQ3] How do translators interact with a concordancer as a translation aid?

This last question aims at studying the interaction between users and the concordancer, which falls into the more general field of Human-Computer Interaction. On the one hand, all users are expected to use the concordancer in a similar way, but, on the other hand, the previous hypotheses about language-related differences could imply that there are different types of interactions based on the types of submitted searches.

The use of computer-based translation aids has only been partially addressed by research in the process of translation, in that most of the existing studies have focused on dictionaries and only recently the Internet has been included in the analysis. Moreover, findings about the use of translation resources have rarely been employed to triangulate data collected via keystroke logging and eye-tracking.

In the field of translation process research, evidence from concordance searches could provide further insights into the way translators work and the types of challenges they face while translating. In particular, the concordancer could be seen as an additional source of data on the process of translation — one that could also be used to support or refute generalizations on translation problems across different language pairs.

As for the translation industry in general, a cross-linguistic large-scale investigation of translation problems could provide additional data to further develop and/or fine-tune existing forms of translation support and, in general, increase awareness about the actual needs of professional translators working between many language pairs.

1.2 SCOPE OF THE STUDY

Data provided by concordance searches are quite unusual in Translation Studies and no set methodology has been developed to systematically deal with this kind of data nor has their potential been fully investigated. Possibly, the best way to outline the analysis is to frame it using Rudyard Kipling's (1902) approach:

> I keep six honest serving-men / (They taught me all I knew); / Their names are What and Why and When /And How and Where and Who […].

In the following paragraphs, the research project will be outlined and contextualized before moving on to the overall structure of the study.

1.2.1 *WHAT*: THE DATA

The data for this study is a collection of authentic concordance searches. They have been collected ex-post from a real working environment and amount to almost one million searches. Because the focus is primarily on the concordancer as a form of translation support, generalizations are advisable. However, because the type of data is closely linked to the actual concordancer chosen for the analysis, no generalizations should be attempted until the peculiarities of the research environment have been detailed.

1.2.2 *WHERE*: THE DATA SOURCE

The real working environment referred to in the previous paragraph consists of the (internal) translation services of the European Union. The data were collected from an internally developed tool which is available to staff translators working in the translation units of eight different EU institutions, including the European Commission, the European Parliament and the European Council. The environment is well-defined and controlled, so that a number of important assumptions for the analysis can be made quite safely. Clearly, such a peculiar environment is useful for running an exploratory study and develop a methodology for the analysis but future studies will need to address the issue of better defining the research environment, particularly in the case of an experiment.

1.2.3 *WHO*: THE TRANSLATORS

This analysis is based on authentic data collected from a real working environment. The prototypical users of the concordancer under examination are professional translators working internally for the EU translation services and can range from senior officials (translators) to temporary staff and translation trainees. For this analysis, no comparisons are drawn between semi-professional and professional translators because the queries will be considered as coming from professional translators, who make up the vast majority of potential users.

1.2.4 *WHEN*: THE TIME FRAME

The data cover one full month of searches submitted in 2010. One month was deemed a long enough span to provide a balanced overview of the activity in the concordancer while keeping the dataset at a manageable size. Unfortunately, no comparisons can be drawn with data from a different time span, tool or environment but final results will at least serve as baseline for future analyses. A more detailed explanation of the reason behind the choice of September 2010 will be provided in the relevant chapter (Section 5.4.1).

1.2.5 *WHY*: THE RATIONALE

Experimental studies in translation process research have traditionally only considered one or two language pairs and usually in one directionality only. The most common

language combinations are closely related to the geographical area of the researcher(s) and their linguistic background. Nonetheless, English is generally one of the languages, with the main other languages found in the relevant literature being (in no particular order) German, Danish, Portuguese, French and Finnish. Clearly, there is a need for greater language-coverage in experiments on translation process research.

This study on the use that *professional* translators make of translation resources can provide some of the results scholars advocated when they pointed out the lack of empirical studies on the routines or strategies in the use of translation aids by professional translators:

> […] we should probably accept that, at least in the area of strategies of using dictionaries further studies of non-translators cannot provide us with much more useful information, and that we should now focus on studies of professional translators using various methodologies. […] we probably need to re-examine anecdotal (written from experience), theoretical and empirical studies of strategies of use of dictionaries and other reference sources by professionals and design empirical studies to obtain supporting evidence (Ronowicz *et al.* 2005: 592)

In particular, compared to Translation Memories, the concordancer as a form of translation support has not received much attention from the research community (other than tool developers). However, existing evidence shows that concordancers provide one of the main forms of support when used in conjunction with Translation Memories.

> While manual concordancing may indeed help to reuse phraseology, TM developers have long aimed to provide automated concordancing to reuse database content at the sub-segmental, phrase level, which is said to be the level at which repetition happens most often. This feature is currently being researched by many TM developers, drawing from statistical algorithms which may include some kind of example-based MT, and marketing the feature with different names […]. We do not yet have empirical studies on how beneficial these algorithms may be, but this is an area in which progress is expected; after all, there have been advances in natural language processing much more sophisticated than those currently present in TM, and the task of speeding up translation is gaining urgency (Garcia 2012: 455-6).

More recent forms of on-line based concordancers offer stand-alone support for translators and their number has increased in the past few years. The concordancer as a translation aid has become more visible but not enough information is available on how users go about using it.

1.2.6 *HOW*: THE METHODOLOGY

The present study aims to be an empirical investigation based on authentic data and will take a predominantly descriptive approach because of the peculiar nature of the dataset used with respect to traditional data types in empirical Translation Studies. Concordance searches can be examined both from a quantitative and a qualitative perspective. Because of the data used, the analysis is to some extent exploratory in nature and cannot rely on well-established and tested methodologies in the field of Translation Studies. The relatively new type of data comes with additional challenges such as the size and nature of the dataset and has imposed an interdisciplinary approach drawing on a number of methodologies from related fields. These include studies in process research (Krings 1986; Lörscher 1991), search log analysis (Jansen 2006), psycholinguistics (Ecke 2009) as

well as corpus-based studies (Sinclair 1991) and cognitive studies applied to translation (Shreve & Angelone 2010).

The language pair will be considered the main independent variable for most of the analyses so that results can be compared cross-linguistically using descriptive statistics. For parts of the analysis some computational methods will be applied using customized scripts, parts-of-speech tagging and some clustering techniques, which will be detailed in due course.

1.2.7 LIMITATIONS

To some extent, the present study can be considered corpus-based, in that the collection of searches can be said to represent a corpus in its own right. Corpus-driven studies, however, have some methodological limitations, as pointed out by Alves *et al.* (2010: 111):

> In corpus-based translation studies, there is often a wide methodological gap between high-level hypotheses about translations and the low-level linguistics/textual indicators drawn from the corpora to support the hypotheses. In other words, the connection or link between extracted linguistic phenomena and the hypothetical construct is very indirect and more complex than it often assumed.

Inevitably, a large-scale analysis runs the risk of over-simplifying the phenomena under scrutiny and for this reason it is advisable to turn to other disciplines such as search log analysis for suggestions on how to handle unfamiliar data. However, search log analysis can be just as problematic:

> [...] one implication of [the identified translators' behavior] is that one has to be careful about using log analysis to infer the needs of translators, because the queries contained in such logs may be strongly conditioned by the users' knowledge of that tool's strengths and limitations (Désilets *et al.* 2009).

The present discussion relies exclusively on problems that have been addressed by means of an external resource, in this particular case a multilingual concordancer. One possible consequence is that only a partial view of the actual range of problems that translators face in their daily work is obtained. There are probably other types of problems that occur at higher textual levels than the sentence or sub-sentence, such as syntactic and cohesive problems and problems related to the logical progression or sense of the text. For these other problems, translators have to rely heavily on their own cognitive resources or turn to colleagues for help, meaning that such problems cannot be systematically dealt with using a computer-based translation support tool.

Search logs have the advantage of coming in large quantities but at the same time they "may be strongly biased towards translation problems which are particularly appropriate for that one tool", and indeed professional translators seem well aware of the strengths and weaknesses of each tool when they use them (Désilets, Brunette *et al.* 2008: 343).

The EU translation services are an ideal test bed for overcoming traditional practical problems that prevent a wide-scale cross-linguistic analysis. On the other hand, such a "controlled" environment may produce only partial results because of the peculiarities of the EU institutions as opposed to the average working environment in the translation industry.

1.3 STRUCTURE OF THE THESIS

A number of questions have been raised in this introductory chapter about the nature and features of search logs as manifestations of translation problems. However, some theoretical premises are necessary before analysing the dataset. For this reason, Chapter 2 will cover the notion of 'translation problem' by taking different perspectives within the overall framework of translation process studies. Problems will be examined focusing on the product, the process and the subjects and according to the various methodological approaches taken to examine them over time. In the final part of the chapter, a brief review of the concept of unit in translation will be carried out to try and produce a synthesis of the main recurring concepts in the literature and reorganize them into a hierarchical structure.

Chapter 3 will deal with the tool used in this study: the concordancer. First, an overview will be provided on the literature about concordancers both from an academic and a professional perspective. The focus will then shift to the translation industry in particular, and the available types of concordancers will be briefly reviewed. A more accurate account of existing research will be provided for those concordancers that have been the object of scholarly work. The last part of the chapter will present in greater detail the concordancer from which the searches have been collected, which is an internal tool at that translation services of the European Union.

After the theoretical account on the concept of 'translation problems' and presenting the tool used in this study, Chapter 4 will provide the missing link between the two and will present concordance search logs as manifestations of translation problems. In addition, it will draw a comparison between concordance searches and Web searches and will suggest that the general idea of translation problem could be in fact reworded as an "information need". Before moving on with the actual analysis, the structure of a concordance search will be presented by breaking it down into main components and sub-components, which will represent the backbone of the chapters to come.

The research environment, data collection and pre-processing will be detailed in Chapter 5, where the study will be contextualized within the EU translation services. A number of preliminary analyses on several variables will be carried out and the main framework of analysis will be sketched. Most analyses will be carried out in parallel at three different levels for each language pair, namely the main language subset, the search session subset and the spot search subset.

Based on the concordance structure outlined in Chapter 4, Chapter 6 and 7 will each address one main component and the relevant sub-components and attempt to answer the research questions. More specifically, the component "Search Strategy" will be dealt with in Chapter 6, which reports on the main quantitative analysis in the study. In Chapter 7, the more qualitative part of the study will be presented, which deals with the component "Problem Unit". Finally, the very last sections of Chapter 7 will present a categorization of translation problem scenarios as information needs. In Chapter 8 the main conclusions will be drawn and a summary of the main findings of this study will be provided together with an illustration of possible future developments.

CHAPTER 2: 'PROBLEM'-RELATED CONCEPTS IN TRANSLATION STUDIES

Translation Studies has often dealt with the notion of translation problems. This has been done at different stages in the development of the discipline, for various purposes and using different methodologies. This chapter aims to provide an overview of the notion of "translation problems" as discussed in the literature and with reference to the related concepts of "translation difficulty" and "uncertainty". The chapter will also introduce other topics of direct interest to the present study, such as types of problem indicators, problem classification and the concept of "unit" in translation.

Think-Aloud Protocols and other research methodologies have often looked at the problem-solving aspect of translation. Translation problems can be challenging for researchers because they are generally identified by inference from micro-level data (e.g. fixations) or macro-level indicators (e.g. sighs). Before examining the various perspectives from which translation problems have been investigated, the concept itself needs to be defined.

2.1 TRANSLATION PROBLEMS

With regard to the notion of translation problem, one of the first things to be noted is the lack of a common definition, as recognized by a number of authors. For example, Presas (1996: 9) points out that part of the problem is the usage of the word "problem" in its 'general' and 'colloquial' senses:

> [...] els corrents teòrics [...] no s'ocupen en profunditat del tema "problema de traducció", gairebé sempre emprenen el terme en la seua accepció col·loquial, i per tant no n'ha estat desenvolupat un concepte que sigui comunament acceptat.

As a consequence,

> [...] nous ne disposons pas de définition du problème de traduction qui obtienne un certain consensus, ni d'une classification des problèmes de traduction qui ait été validée empiriquement (Hurtado Albir 2001: 280 in Gil-Bardají 2010: 279).

The lack of consensus can also be seen as deriving form the different senses attached to the word "problem" in Translation Studies, as pointed out by Toury (2002). More specifically, Toury has identified three different senses (and usages) of the concept. The first sense ("PROBLEM$_1$" for Toury) is centered on the source text and addresses issues of translatability without considering the context of a translation act. The second sense ("PROBLEM$_2$") is linked to the actual translation events, i.e. "individual translation acts situated in a particular time and space" (Palumbo 2008: 54) and takes a product-oriented perspective that highlights problems in retrospection. Finally, the third sense ("PROBLEM$_3$") can be considered process-oriented because it looks at the translation act in its unfolding. However, problems of the third type can only be investigated if the translation process has left (temporary) "traces comprising more than just the end-product" (Toury 2002: 65), such as interim replacements or reflections. The several definitions of the notion of translation problem that have been proposed over time can be placed individually at each of the three levels identified by Toury (2002), which he terms

(i) source-oriented and prospective; (ii) target-oriented and retrospective; and (iii) process-oriented and processual. One of the most famous and more widely used definitions is the one proposed by Nord (1991), who distinguishes between translation *problems* (in the sense of PROBLEM$_1$) and translation *difficulties* (PROBLEM$_2$ and PROBLEM$_3$):

> A translation problem is an *objective* problem which every translator (irrespective of his level of competence and of the technical conditions of his work) has to solve during a particular translation task (Nord 1991: 151; emphasis added).

The concept of problem as a shared and *common* incident is contrasted with that of translation "difficulty", which is a *subjective* concept and is related to the translator and his/her specific working conditions (Nord 1991: 151). Unfortunately, the definition does not lend itself to a straightforward implementation because, as Nord further explains, the same textual item can be seen both as a problem and as a difficulty:

> Ein Übersetzungsproblem, das für einen Anfänger eine nicht zu bewältigende Schwierigkeit darstellt, weil er noch nicht gelernt hat, wie es zu lösen ist, bleibt ein Übersetzungsproblem, auch, wenn er als langjähriger Berufsübersetzer keine Schwierigkeiten mehr damit hat. Ebenso kann ein anderes Übersetzungsproblem, das der erfahrene Ubersetzer mit den entsprechenden Hilfsmitteln [...] leicht löst, wieder Schwierigkeiten beraten, wenn es ohne Rückgriff auf die Hilfsmittel [...] bewältigt werden muss (2011: 255).

The trouble with implementing these definitions of problems was also pointed out by Lachat (2003 in Muñoz Martín 2009a: 30), who rejected the difference between problem and difficulty because translation problems "were found to be usually complex and ill-defined" and she found that the definitions lacked "psychological reality". This is maybe why, as already pointed out by Palumbo (2008: 48), it would seem that the concepts of "problem" and "difficulty" are often used interchangeably, reflecting "the uncertainty of researchers as to what translation problems are and how they should be categorized".[1]

Another well-known definition of translation problem (in Toury's third sense) was put forward by Lörscher, who pointed out the lack of attention the concept received in translation theory, particularly from an empirical perspective. According to Lörscher, a problem consists of "all those (linguistic) problems which a subject is faced with when performing a translation" (1991a: 94):

> [...] a translation problem is considered to occur when a subject realizes that, at a given point in time, s/he is unable to transfer or to transfer adequately a source-language text segment into the target-language (Lörscher 1991a: 80).

While Lörscher's definition does not seem to restrict the translation problem to a particular size or unit, Gonzáles Davies and Scott Tennent (2005: 164) explicitly add potential problem locations, i.e. global and local:

> A translation problem can be defined as a (verbal or nonverbal) segment that can be present either in a text segment (micro level) or in the text as a whole (macro level) and that compels the student / translator to make a conscious decision to apply a motivated translation strategy, procedure and solution from amongst a range of options.

[1] For a more in-depth discussion about the concepts of "problems" and "difficulties", see Palumbo (2008: 47ff.).

They consider translation problems to be linked with conscious decision-making for problem-solving, as Kiraly (1995: 105) did before them:

> Translation problems emerge from the intuitive workspace when automatic processing does not produce tentative translation elements. These problems are considered in the controlled processing center and a strategy is chosen and implemented in an attempt to deal with them.

This view assumes that alternating sequences of automatic and controlled processing occur while translating, the latter emerging whenever a problem is encountered. In this sense, Kiraly's definition brings to mind the "stop-and-go" structure of the translation process, originally suggested by Krings (1986a: 116), who described it as

> eine alternierende Auftreten von Phasen des raschen, relativ ungehinderten Fortschreitens und Phasen des Blockiertseins durch eine nur durch besondere übersetzerische Anstrengungen zu überwindende Schwierigkeit.

As a consequence, a clear interruption in the translation flow was taken to represent some kind of encountered difficulty. Once this dichotomy in the process had been established, the next step was to label stretches of continuous translation "non-problematic" as opposed to the "problematic" ones. Problems were identified with a non-productive stretch where no translation occurred (interruption) due to either SL comprehension or TT production.

More recent definitions consider translation problems as "difficulties encountered by the subjects when carrying out a translation task" (PACTE 2011a: 37) and more specifically "any source language word or expression which presents a difficulty for a human translator (not machine) during the process of translation" (Désilets *et al.* 2009). A similar translator-oriented view is adopted by Enríquez Raído, for whom "a central discussion inevitably revolves around what constitutes a translation *problem* and from *whose* perspective" (2011: 151; emphasis in the original). In her study,

> "problems" constitute those particular source text items that the research participants explicitly identified as problematic for translation in the online search reports, and as manifested in, and inferred from the participants' recorded translation processes and their resulting products (Enríquez Raído 2011: 151).

Most definitions of translation problems have been formulated from a translator's perspective. However, a machine-oriented view is also possible. In the case of Machine Translation, a translation problem would become an issue of "translation mismatches", i.e. a source expression that is wrongly matched with one of several target language equivalents (Prahl & Petzolt 1997: 125). In order to operationalize the definition and use it on a machine, a clear distinction between problems and non-problems should be made available. Humans are more flexible (and possibly unpredictable) in this regard because the problematic item is generally subject-dependent and difficulties do not necessarily occur in the same textual locations.

2.2 PROBLEMS AS DIFFICULTY & UNCERTAINTY

Difficulty is an intrinsic property of a translation task. In addition to her definition of (subjective) "translation difficulty", Nord (1991: 168) sets out the parameters to determine the degree of difficulty presented by a translation task. In the practice of translation, such parameters boil down to the particular features of the source text-in-situation, assuming that professional translators are fully proficient in language, subject

matter and transfer procedures. The discussion of the concept of difficulty can be further broken down into other dimensions of difficulty that Reiss (1974: 5 in Nord 1991: 165-6) calls (i) linguistic difficulties, (ii) factual difficulties and (iii) technical difficulties. In the field of translation teaching, a linguistic classification of difficulties has been proposed by Wilss (1982: 161), who distinguishes between (i) transfer-specific translation difficulties (TD); (ii) translator specific TD; (iii) text-type-specific TD; (iv) single-text-specific TD and eventually provides a definition of translation difficulty based on the concept of equivalence:

> TD occur whenever a lexical or syntactic one-to-one correspondence between SLT and TLT cannot be practised, because literal translation would inevitably entail a negative transfer (Wilss 1982: 164).

An alternative view on difficulties is offered by Campbell (1999: 34), who considers three distinct perspectives of analysis, namely the source text, the translation task (i.e. external factors such as time pressure) and the translator's competence. His analysis centers on a cognitive approach to translation difficulties that first distinguishes between on-line and off-line translation mode[2], the former roughly matching Jakobsen's "drafting" phase[3], the latter matching "orientation" and "revision", and then goes on to examine on-line translation difficulties in greater detail. Text difficulties in on-line translation can be characterized from a cognitive perspective (e.g. grammatical and lexical items whose processing consumes more working memory capacity) as well as a language processing perspective (e.g. availability and activation of source and target lemmas in the mental lexicon). These two variables have been operationalized by Campbell and later on by Dragsted (2012) to indicate difficult items in the source text by considering the amount of cognitive processing (i.e. mean number of alternate renditions by a group of translators for the same SL item) and the amount of editing in the target text (e.g. additions and deletions). Although the focus was on the source text, difficulty was first measured in the target text and then mapped back onto the source. A more straightforward source-text oriented analysis of source text difficulty is presented in Jensen (2009) and Hvelplund (2011: 88ff.), who used three quantitative indicators to measure difficulty, namely measurements of readability, calculations of word frequency and calculations of the number of occurrences of non-literal expressions (idioms, metaphors and metonyms).

On the other hand, a subject-oriented cognitive approach to translation difficulty can be found in Dragsted (2004: 58), according to whom "translational problems only exist to the extent that the translator experiences a problem", as if to say that translation problems are not to be established a priori. The main variable affecting the perception of difficulty is the translator's competence. However, some source texts will nonetheless be intrinsically more difficult and cause more problems to translators as revealed by reduced speed and segment size in the translation process.

In his study on metacognitive activity in translation, Angelone (2010) supports the view of a linear translation process where problem-solving translation sequences will alternate

---

[2] "On-line" mode means that "the translator works on a document or part of a document at a more or less a single pass" whereas "off-line" involves e.g. pre-reading, re-reading, checking words (Campbell 1999: 36).

[3] According to Jakobsen (2002: 191-193), a translation event (/act) can be subdivided in the three phases of orientation, drafting and revision. More specifically, the orientation phase "is the time delay between the appearance of the source text on the screen and the typing of the first text production key", the drafting phase "runs from the first text production keystroke until the first typing of the final punctuation mark" and the (end) revision phase mainly consists in "monitoring [the] existing text".

with unproblematic ones. Difficulties can arise at any stage of comprehension, transfer and production, impeding the translation activity, and are likely followed by uncertainty. "Uncertainty" is another concept that can be found in the literature as closely relating to translation difficulty, which Angelone (2010: 18) defines as

> a cognitive state of indecision that may be marked by a distinct class of behaviors occurring during the translation process [...], [it] is associated with an observable interruption in the natural flow of translation and is generally related to a discrete *problem nexus* in the translation. A nexus is the confluence of a given textual property and level (lexis, term, collocation, phrase, syntax, sentence, macro-level feature) intersecting with some sort of deficit in the translator's cognitive resources.

Because uncertainty behaviors are visible, a number of "diagnostic behaviors" marking the interruptions in the translation process can be studied empirically. They are physiological as well as verbal indicators about the interactions with the translation interface and include extended pauses in TT production, deletions, revisions, cursor repositioning, information retrieval behavior, eye movements, pupil size variations, increased brain activity, galvanic skin responses and direct and indirect verbal articulations. These behavioral markers for uncertainty can be said to overlap with translation problem indicators such as long pauses or fixations and (in)direct articulation, as suggested by Angelone (2010: 22):

> Ideally, each occurrence of uncertainty is demarcated by a problem recognition indicator marking the activation of uncertainty management and the beginning of a problem solving bundle.

According to Tirkkonen-Condit (2000: 123), uncertainty management in the translation process manifests itself in identifiable patterns which she investigated by means of verbal protocols. Such patterns can be seen as "conscious strategies for reducing uncertainty by solving the problems of comprehension, transfer or production" (Angelone & Shreve 2011: 109), i.e. an example of strategic sub-competence in the PACTE's model of translation competence (2011b: 326).

The PACTE group (2011a: 37) has provided a list of indicators for the identification of translation problems which include the coefficient of perception of the overall difficulty of the text and a number of indicators relating to "Rich Points". The concept of Rich Point was borrowed from Agar (1991: 176), who defined it as a "particular place in one language that makes it so difficult to connect with another". Rich Points are source-text elements identified by PACTE during pilot tests and include "linguistically, culturally and functionally challenging points, i.e. terms which are expected to result in extensive processing activities" (Jääskeläinen 2011: 25), and are complemented by "Points of Interest", i.e. "presumably non-challenging points where the choice of translation variants is limited and little conscious thought expected" (ibid.). In particular, Rich Points in the source text can be assigned to five categories of translation problems: (i) linguistic problems, (ii) textual problems, (iii) extralinguistic problems, i.e. relating to cultural difference or subject domain knowledge, (iv) problems of intentionality, i.e. difficulties in understanding ST information and (v) problems relating to the translation brief and/or the target-text reader (PACTE 2011b: 327).

Having established some overlaps between the underlying concepts of terms such as "problem", "difficulty" and "uncertainty", the present review will now turn to illustrating some common approaches employed in the study or description of translation problems, i.e. product-oriented, subject-oriented and process-oriented approaches.

2.3 PRODUCT-ORIENTED APPROACH

A product-oriented perspective on translation problems can be taken by considering, on the one hand, some of Kring's primary and secondary problem indicators, such as alternative tentative translation equivalents and gaps in the TT resulting from a participant not knowing how to translate certain ST units and, on the other, Campbell's indicators. Campbell (1999) interpreted source text difficulty as a function of the processing effort measured by looking at the mean number of alternate renditions of the same item by different translators (i.e. variability in the target text) as well as the amount of editing by the same translator e.g. in terms of additions and deletions during the drafting phase. Difficulty was therefore related to the complexity of choices available to the translator (Campbell 2000: 38). His investigation also introduced the concept of Choice Network Analysis (CNA; Campbell 2000), a research method aimed at inferring mental processes from the analysis of translations of the same source texts performed by multiple subjects. CNA can be "useful for estimating the relative difficulty of parts of source texts", where difficulty is closely related to the number of nodes and branches in the network of available TL choices (2000: 38). Building on Campbell, Dragsted (2012) conducted a study to correlate observable indicators of difficulty in translation product data (target text variation and target text modifications) and process data (gaze time, regressions/refixations and pause length) and concluded that "target text variation is a reliable predictor of difficulty indicators observable in process data" (Dragsted 2012: 95).

In translation pedagogy in particular, translation problems have been discussed both *prospectively* (i.e. predicted on the basis of a linguistic analysis of the source text) and *retrospectively* through error analysis of the target language text (Krings 1986b: 266), the latter to be covered in the next section.

2.3.1 PROBLEMS AS ERRORS

Uncertainty management has been assumed to reflect translation quality[4] in that potential correlations exist between the efficacy of uncertainty management and the number of errors in a translation and can be revealed by error analysis (Angelone & Shreve 2011). When tested, error analysis highlighted that the professional translator did not produce any errors as opposed to a bilingual subject and two students with different mother tongues. Due to the very small subject sample, no firm conclusions can be established on the basis of Angelone and Shreve's study. Nonetheless, the professional translator's profile suggested a possible correlation between the absence of errors and a higher frequency of problem recognition behaviors. Therefore, problem recognition was perceived as the most important element in uncertainty management and as a way of limiting and reducing the number of errors.

Problem awareness is also considered as a prerequisite for strategic competence in translation. Similarly to the previously mentioned findings, in PACTE (2011a) problem awareness is directly linked to the absence of translation errors, the number of translation problems solved and the number of translation errors reflected upon.

The concept of translation difficulty (and translation problem) is therefore closely linked to that of translation error (Palumbo 2008: 5). Translation errors in turn are the main

[4] In the case of Angelone and Shreve (2011) "quality" is to be interpreted as the absence of errors. Quality evaluation was performed by grading translations using a modified version of the error encoding framework adopted by the American Translators Association.

object of translation evaluation and "can first of all be seen as problems which the translator was not able to solve" (Palumbo 2008: 47). In such error-oriented approaches to translation problems, the main focus is on the target text, to be compared not only to the original source language text but also to readers' expectations. Whatever the touchstone, the translation product remains the central element based on which translation problems are examined.

2.4 SUBJECT-ORIENTED APPROACH

Most, if not all, empirical studies in translation need at least one subject to make data collection possible and in this sense the subjects (sometimes referred to as "research participants") play a central role. Some researchers have shifted the focus from the product to the subject, who is the one actually experiencing the problem:

> [...] a translation problem is considered to occur when a subject realizes that, at a given point in time, s/he is unable to transfer or to transfer adequately a source-language text segment into the target-language (Lörscher 1991a: 180).

Data elicitation in this case can take many different forms, the most obvious being a translation task. A number of different data elicitation techniques have been used which vary from direct interaction with the researcher in the case of e.g. interviews to more sophisticated monitoring systems such as eye-tracking. This section is concerned with those studies that presuppose a direct interaction with the subjects or where the subject is "actively" participating in data collection.

Verbalizations in the wide sense have been one of the favored methodologies for analysis. They draw on psycholinguistics studies and have been used from the early days of translation process research. Verbalizations can be grouped into two main categories: concurrent verbalization (i.e. Think-Aloud Protocols or TAPs; e.g. Krings 1986a,b; Gerloff 1987; Jääskeläinen 1987) and retrospective verbalization (e.g. Dragsted 2004; Enríquez Raído 2011), which are mostly used as an additional data elicitation method for purposes of data triangulation.

2.4.1 DATA ELICITATION BASED ON Q&A

Q&A as a data elicitation method relies on the direct interaction between the researcher and the subject and its most frequent realizations are interviews or questionnaires. Unlike TAPs, it generally does not involve multitasking on the part of the subject. An example of written Q&A is the questionnaire on translation problems used by the PACTE group (2011a: 56), aimed at studying translation competence and its acquisition. It consisted of four questions and inquired e.g. about the overall degree of difficulty of the text perceived by the participants (from very easy to very difficult). Eventually, the translator was asked to provide five instances of problems encountered while translating and for each problem, four additional questions were asked. The collected answers became the main data source for the identification of translation problems, though at times researchers had to interpret vague or confusing answers in order to categorize them. Overall,

> [i]t was observed that it was difficult for subjects to explain the difficulties they encountered given the procedural cognitive processes involved in translation and the automatisation of all expert knowledge (PACTE 2011b: 334).

A more articulated questionnaire with open-ended and closed questions was submitted by Enríquez Raído to her students (2011: 523-4). In the problem-related questions, the

interviewee was asked to first order a short list of actions (e.g. Try to understand the meaning of the word from the context; Consult a bilingual dictionary; Consult a monolingual dictionary; Try to express the same idea in as many ways as possible in the target language) according to their likelihood and then answer true or false statements (e.g. "The main problems encountered when translating non-specialized texts are vocabulary problems").

Questionnaires together with interviews were also one of the three main data sources for a study on the challenges posed by multilingual lawmaking within the EU (DGT 2010a). The questionnaires were filled by translators and lawyer linguists working at the European Commission, Council and European Parliament, as well as by Member State coordinators. In the present review, only the questionnaire for the translation departments will be considered. It consisted of 13 questions which did not explicitly ask about translation problems but touched upon scenarios linked to translation problems, asking for example whether translators were in contact with experts for clarifications of uncertain terminology and what approaches translators took in the case of e.g. outdated or foreign language expressions.

Questionnaires, surveys and polls can be considered a non-relational type of data elicitation because data collection does not require the physical presence of the researcher. However, they cannot be considered completely non-mediated in that the structure and wording of the questionnaire can reveal the researcher's intent. Alternatively, relational forms of Q&A can be adopted where the researcher directly interacts with the subject by asking questions or using cues to trigger responses. This can be the case of Contextual Inquiry and field studies in general (e.g. Drugan 2004; Désilets *et al.* 2009; Karamanis *et al.* 2010), where researchers observe the subjects in their authentic working environment and interact with them whenever needed. Sometimes, actual interviews (structured, semi-structured, tailored or informal) are carried out. For example, Désilets *et al.* (2008a,b) used interviews in the context of two translation tasks, a natural task and a controlled one, whilst transcripts provided qualitative and quantitative data to produce a list of translation problems of professional translators. Questions about problems are also found in Dragsted's retrospective interviews to both professional and student subjects (i.e. "Did you find the text problematic?"; 2004: 294ff.). A more recent example of interviews can be found in Enríquez Raído (2011), where participants' Web search behavior, rather than translation problems, was the main focus of attention.

In the case of Q&A sessions, data interpretation is more straightforward because researchers do not need to interpret the answers inferentially, provided the questions were formulated appropriately. For this reason, this form of data elicitation possibly produces the most direct type of results, if not the most reliable, and can be used as an additional reference to confirm results about translation problems, whenever possible and appropriate, as pointed out by Hvelplund (2011: 225):

> Retrospective interviews and questionnaire data, for example, would have been helpful in explaining some of the more surprising results. For instance, they could have explained why source text complexity did not appear to have an effect on translators' allocation of cognitive resources. Retrospective data might have revealed that the translators did not experience any difference between the experimental texts with respect to difficulty.

Because experimental results depend heavily on the subjects, comparisons of results should only be attempted when studies present comparable experimental conditions and involve the same categories of subjects (e.g. student vs. professionals or novice vs. experts). Unfortunately, this is not always possible. For example, early studies in

translation process research relied on advanced language learners while recent studies have become much stricter in the selection criteria for subjects (e.g. Martín-Mor 2011: 104ff.). Other studies have focused on the didactics of translation and the acquisition of translation competence and therefore focus on the student population. In principle, the present study is concerned with professional translators only and shall not attempt to compare students' and professionals' performances, whilst the overview in this chapter is concerned not so much on the results of each empirical study but rather on the different theoretical and empirical approaches as well as the methodologies used for studying translation problems.

2.5 PROCESS-ORIENTED APPROACH

Research on the process of translation started in the mid-1980s and is still very active with new technologies and methodological approaches being introduced or borrowed from other disciplines. Overall, three main stages in translation process research can be identified. The first covers the decade between the mid-1980s and the mid 1990s, when verbal protocols were the main approach chosen to tap into the translators' black box. In the late 1990s, a new era began with the introduction of keystroke logging to study patterns of target text production at a very minute level. This phase is still very much ongoing but from the mid-2000s a further branching in translation process research was started when the eye-tracking methodology was applied to the study of translation. With all these different data types available, the practice of triangulating translation process data has established itself:

> Assuming that navigating through uncharted waters requires several location points to establish one's position, and taking examples from the Social Sciences, [researchers] focus on the need to apply several instruments of data gathering and analysis in their attempts to throw light on the nature of the process of translation, and mainly on issues related to their inferential behavior, intersubjectivity, competence, segmentation, time pressure, dictionary use, and the novice-expert interface (Alves 2003: vii).

The following sub-sections present approaches and findings from process studies that relied on experiments, rather than interviews, to elicit process data. None of the experiments would have been possible without research participants. The role of the subjects is still paramount but the type of interactions with the researchers differ considerably from the previously described scenario in that the analysis is much more focused on the unfolding of a translation assignment rather than on a global (or punctual) assessment of a task.

2.5.1 STUDYING PROBLEMS WITH CONCURRENT VERBALIZATIONS

Concurrent verbalizations, also known as Think Aloud Protocols (TAPs), have been widely used in translation process studies since the mid-1980s. The main purpose of this approach is to gain insights into the cognitive processing in a person's brain while s/he is translating but also to find out about other translation-related questions, such as the size of the translation unit. Despite being a form of verbalization, TAPs differ from retrospective interviews because they are elicited simultaneously while performing the main experimental task, as opposed to post-task elicitation. A lot has been written about TAPs as a research method (e.g. Fraser 1996; Bernardini 2001; Jääskeläinen 2002) and their practical implications, including criticisms about the validity of the results which Enríquez Raído (2011: 100) summarizes in terms of completeness of the reports, their

effects on cognitive processes[5] and the ecological validity of the experimental setting. Similar feedback also comes directly from the participants. One example is provided by Lauffer (2002: 65) describing the outcome of her experiment with professional translators:

> I asked why he had not verbalized and he replied that he had found it very difficult and that having to think about talking while translating took him away from the actual work. Interestingly, the third translator who felt more comfortable thinking aloud once said "I'm not talking at all now. I could talk more but that would slow me down and it doesn't feel normal."

However, TAPs have also provided useful insights into the processes of translation at a time when no alternative methods were available. Today, introspective methods are still used in research projects (e.g. the TransComp project; Göpferich 2010) but instead of being the main source of data elicitation, they serve as complements to other data types for triangulation purposes.

In the case of TAPs, the experimental setup and in particular the choice of subjects is very diversified and existing labels to characterize subject groups, such as "professional translators", "expert translators", "novice", "advanced student", "semi-professional", could be discussed at length[6]. As extensive reviews on existing TAP studies are already available (Enríquez Raído 2011; Jääskeläinen 2011), no attempt will be made here to produce an additional overview of the existing literature. As with the previous sections, the focus still remains on the use of TAPs as a method for the analysis of translation problems as they unfold.

In his pioneering study, Krings (1986a: 188) reported that 90% of his subjects' verbalizations were connected with problematic items. This can be justified by the ensuing notion that subjects can only report on what is in their focus of attention:

> [I]t is only information stored in the short-term memory which can be externalized, and which, by definition, is in the individual's focus of attention. There is no way of getting information [read: verbalizations] about mental processes which are not given cognitive attention (Lörscher 1991b: 75).

In particular, Krings identified 454 single translation problems from the think-aloud protocols, which proved to be the most important single feature of the translation process (Krings 1987: 168-9). Similarly, one of Lörscher's (1986: 279) categories of analysis was the verbalization of a translational problem, i.e. "[a] subject verbalizes a problem connected with the translation of an SL text segment into TL. Translational problems are verbalized either before or during the search for their solution". Both Lörscher and Krings found hardly any thinking-aloud during unconscious (i.e. automatic) phases of the translation process and Jääskeläinen and Tirkkonen-Condit (1991) found that the verbalized problems changed according to the level of translation competence. Later experiments (e.g. Lauffer 2002; Buchweitz & Alves 2006) showed that the non-vocalized information (or at least part of it) could be elicited using retrospective interviews while showing a playback of the task previously recorded via a screen-recording software. In particular,

> [i]f a segment of the text (a word, a phrase, or even a full sentence) is instantiated in the verbalization of the participants, it represents a possible problem faced

[5] See Jääskeläinen (1999: 151-158); Jakobsen (2003); Krings (2001).

[6] For a comprehensive discussion on the concept of expertise in translation and its operational definitions, see Jääskeläinen (2010).

> during translation that required more decisive action, and as such is burned on the memory of participants and recalled in the think-alouds (Buchweitz & Alves 2006: 249).

Ultimately, implementing TAPs to study translation problems has often meant looking for Krings' primary and secondary indicators in the verbal protocols and task recordings of experiments (1986: 121). Krings' primary indicators are (1) explicit or implicit utterances by means of which the participants manifest a translation problem; (2) consultation of a source of reference; and (3) gaps in the TT resulting from a problem with certain ST units. Secondary indicators are (i) competing translation equivalents; (ii) changes in the target text manuscript; (iii) underlinings in the source text; (iv) negative evaluations of the target version; (v) metaproblematisation; (vi) unfilled pauses of more than 3 seconds; (vii) paralinguistic indicators; (viii) lack of a primary equivalent association. Broadly speaking, these indicators can be grouped according to two focal points: the translator (i.e. the subject) and the target text (i.e. the product). In his analysis, Krings classified an item as translation problem if at least one primary or two secondary indicators were found.

Other studies that used TAPs to investigate translation problems focused on subtitling (Kovačič 2000) and the characterization of expertise (Jarvella *et al*. 2002). Both studies used TAPs in combination with other methodologies and produced an *ad hoc* classification of problems, one specific for subtitling problems, the other distinguishing four degrees of problem difficulty. The assessment of the degree of difficulty was based on a "combination of pauses, repetitions of source and target text, returns made to problems, the quantity of verbalization produced relating to a problem [and] verbal cues" (Jarvella *et al.* 2002: 189). Eventually, both studies reported a limited amount of verbalizations and results were obtained by combining concurrent verbalizations with the other methodologies, such as key-logging and retrospective interviews. A comparative study between concurrent verbalization (TAPs) and cue-based retrospective verbal protocols (RVPs) highlighted that

> explicit information about the use of resources and reflections about strategies and dealing with translation problems might be more accessible with RVPs (Ehrensberger-Dow & Künzli 2010: 130).

In particular, Ehrensberger-Dow and Künzli found RVPs (to some extent comparable to Contextual Inquiry) more suited to investigate the process of professional translators in their normal work setting, whereas Hvelplund (2011: 225) noted that retrospective interviews and questionnaire data would help explaining some surprising results of eye-tracking and key-logging.

In sum, TAPs have provided useful insights into some aspects of the translation process but are probably not the best methodology to systematically study translation problems. A complementary approach that can produce more easily quantifiable data is provided by pause analysis.

2.5.2 PAUSES AS MANIFESTATIONS OF PROBLEMS

Pauses play an important role in the study of translation process using TAPs. Unfilled pauses, for instance, were used by Krings (1986a,b) as secondary indicators for a translation problem. Similarly, Lörscher (1986: 279) used pauses to possibly identify one of his categories of analysis, i.e. the realization of a translation problem:

> To the analyst [the realization of a translation problem] is very often perceivable by a pause and/or hesitation during the production of the TL text. […] Of course, not every pause or hesitation necessarily indicates a translational problem. The

> mental organization of SL test segments as well as problems in SL text reception can lead to pauses and hesitation.

Before Krings' primary and secondary indicators, Færch (1984: 60ff. in Lörscher 1991: 63) distinguished between implicit and explicit (strategy) indicators, pauses being listed in the implicit group as opposed to a verbalization that constituted an explicit type of indicator. One caveat with the existing indicators was pointed out by Lörscher, who warned that

> [i]ndicators are only constituted by the analyst's interpretive approach to the data. [...][They] are not strategy or problem indicators *by themselves* but *potential* indicators. In the process of interpretive reconstruction they can be interpreted as being strategy or problem indicators by the analyst (1991: 65; emphasis in the original).

After keystroke logging was introduced in translation process research with Translog in 1999[7], a new perspective on translation problems became possible. Translog logs the typing activity of translators and provides researchers with a new data type that, once analyzed and interpreted, gives new empirical insights into many aspects of the translation task (e.g. the comparison between experts and novices) and the translation process alike, such as revision, memory constraints and shifts in attention. For the purpose of the present overview, the focus will be on segmentation, and in particular on pauses. Building on Krings (1986a,b), pauses have been linked to the employment of translation strategies to solve translation problems, so that the slower the production of the TT due to the number and length of pauses, the more problematic the text is assumed to be (Dragsted 2005: 50). In particular, Dragsted's experiment compared translation students and professional translators translating texts of various levels of difficulty. Both groups were slowed down considerably when translating the more difficult text, due to longer pauses. Dragsted (2004, 2005) used pause patterns to study the segmentation of the source text in Translation Units (TU), whose length was expected to vary according to the level of difficulty encountered, i.e. the more difficult the source text, the shorter the TUs. The difficult text often had segment length reduced to one word, which "shows that the presence of a difficult item in the source text and the retrieval of the information required to solve the problem, took up practically all the WM [working memory] capacity" (Dragsted 2005: 57-8), possibly because the information processing system can only focus on one thing at a time when solving a problem (Newell & Simon 1972: 89 in Dragsted 2005: 50). In a study on directionality and recursiveness (i.e. "on-line" revision of a text[8]), translation into an L2 resulted overall more difficult (i.e. more time-consuming, resulting in a greater number of segments and requiring more revision) while segment and TAP analysis highlighted that the focus was frequently on words (Buchweitz & Alves 2006). Jakobsen (2005) studied the segmentation patterns of expert translators by analyzing the occurrences and distribution of pauses and suggested how to use this approach to identify the kind of cognitive activity going on at any given moment. Irrespective of the problem and the cognitive rhythm (which is linked to the level of expertise), target text production was again slowed down when translators struggled, though the specific type of problem (e.g. reception problems, L2-problems) would affect the delay in different ways (Jakobsen 2005: 181, 187). In other words, "the main obstacle to fluent translation is frequently to do with a local, e.g. semantic, problem occurring unpredictably" (Carl *et al.* 2008a: 116).

---

[7] For an overview of existing studies conducted using Translog, see Schou *et al.* (2009).

[8] For the concepts of "on-line" and "off-line" see Section 2.2, note 2.

Although translation process research made a considerable step forward after the introduction of Translog, there were still several pending questions regarding the type of processing taking place during pauses, e.g. reading, interpreting the ST, considering target text alternatives or checking the TT (Mees 2009: 28). This is when eye-tracking as an additional type of data elicitation methodology was introduced to complement data from key-logging. Pause patterns could be identified by considering the number and the duration of fixations which, in turn, were linked to ongoing cognitive processing due to the eye-mind assumption, i.e. "there is no applicable lag between what is being fixated and what is being processed" (Just & Carpenter 1980: 331 in Hvelplund 2011: 68). In addition to global data on fixations, measurements of pupil dilation can be added as an additional variable to measure cognitive effort (e.g. Pavlović & Jensen 2009; Hvelplund 2011).

With both keystroke logging and eye-tracking being used jointly in experiments, a new method for investigating human translation was established, which produced increasingly larger volumes of data related to keyboard activity and eye movements (Carl *et al.* 2008b: 21), termed User Activity Data (UAD). More generally, the notion of UAD is taken to

> subsume any kind of data which is consulted or generated by a translator during (or in context with) a translation session. […] UAD relates spatial, i.e. textual, product data with temporal process data (Carl & Jakobsen 2009: 126).

In this new UAD perspective, translation pauses were not just characterized by absence of keyboard activity, as in "pure" keystroke logging, but also by the exclusive occurrence of fixations (Carl 2009a). The underlying idea is to "link basic translation concepts i.e. major building blocks of mental representation, with patterns of UAD to detect factors which contribute to the problems which translators face during their work" (Carl *et al.* 2008b: 26). In particular, researchers hope to find a way to determine the specific cause of a translation pause, e.g. unknown terminology or a more complicated understanding and/or translating problem (Carl 2009) by triangulating micro-level data (e.g. UAD) and macro-level data (e.g. verbal protocols).

A critical element in the analysis of pauses is the cut-off length to determine process boundaries, i.e. pause duration to identify interruptions in the typing or verbalization flow and (long) fixations. The chosen time delay will mark the processing units to be analyzed. Pause cut-off length can vary between one second (Krings 2001: 210), one/two seconds (Dragsted 2004: 103)[9], three seconds (Krings 1986: 121) or five seconds (Buchweitz & Alves 2006: 249, Alves & Vale 2009: 257) and will affect the nature of the segmentation unit and the way statistics are calculated. However, this cut-off length cannot be easily defined, though

> [t]here is agreement among most researchers that considering very short interruptions as pauses would lead to the identification of automatic processes, corrections of typos or other instances of on-line text production in which no conscious problem-solving and/or decision-making takes place (Alves & Vale 2009: 255).

Once pauses have been mapped, researchers can try to establish what type of pause each instance represents. Alves and Vale (2009: 257) distinguish between four main types of

[9] In her study, Dragsted compared students and professionals and found an approximate pause length of 1.5 seconds for all subjects. However, the pause unit value was eventually adjusted according to the characteristics of the subjects (e.g. typing speed and processing speed) to account for inter-group variations: a pause unit value of 1 second was used for the group of professionals, whereas for the students, a pause of two seconds applied (2004: 100-103).

pauses: (i) planning, (ii) searching for a translation alternative, (iii) assessment of the previous production and (iv) beginning of a new reading phase. A comparative study of the writing processes for translation and monolingual text production presents three main interpretations for pauses in normal text production, i.e. indication of mental organization, problem-solving or the beginning of a new cognitive unit (Immonen 2006: 315), which clearly overlap with the types listed for translation. However, "[a]s the production of a written text is a highly complex activity and the outcome of numerous cognitive processes, it is difficult to determine any one cause to be responsible for a certain pause" (Immonen & Mäkisalo 2010: 46). A systematic pause analysis has been attempted to gain further insights into the role of pauses in translation which had generally only been used for segmentation purposes (Immonen & Mäkisalo 2010: 49). Pauses were first examined based on their location and duration at the boundaries of linguistic units[10] (Immonen 2006). Next, the correlation between syntactic units (e.g. type and function of phrases, clauses) and pause time distribution was investigated (Immonen & Mäkisalo 2010). Surprisingly, pause time distribution changed according to the size and nature of the linguistic units being processed (2010: 57). In particular, pause processing in translation was intensified for smaller units and reduced for larger units compared to monolingual text production. This was explained by considering that at textual level not much processing is required because the "paragraph and sentence structure [...] is often copied from the source text" (Immonen & Mäkisalo 2010: 60). In translation, the focus is often on the lexical level which seems confirmed by the finding that pauses preceding noun phrases were generally longer than pauses before verb phrases. In addition, translation of noun phrases seems characterized by a form of writing strategy[11] where writing begins before processing is completed which results in longer phrase medial pauses. Further research showed that processing in translation cannot be predicted by analyzing monolingual text production. The two writing processes differ the most at the level of syntactic processing, bearing in mind considerable inter-subjective variation. The most common pattern in translation was the processing of words as units of their own, separated from phrases and clauses that constituted a separate unit. In other words, the processing of phrases differs significantly from the processing of words but not so much from the processing of clauses, suggesting that in translation the processing of short phrases and words is emphasized, possibly due to the search for equivalence (Immonen 2011: 244-6, 251). One special case are compound words that, being multiword constructions, could in fact be considered between words and phrases and become another element on which the emphasis of word processing is placed (Immonen 2011: 250-1).

Key-logging and eye-tracking (possibly complemented by retrospective verbal protocols) have become the new standard in translation process research. These methodologies de-compose a translation event (a complex event) into simple minimal events such as reading and writing logged at the level of milliseconds; thanks to UAD, the evolution of a translation task can be tracked over time. Other methodologies, such as TAPs, interviews and screen recording focus more on the macro-level of the translation task. Triangulation (Alves 2003), a de facto paradigm in current translation process research, has brought together these two levels of analysis, given better insights into several aspects of the translation process and created a sounder basis for discussing experimental findings.

---

[10] Location is here intended to mean (in ascending order): character, syllable, compound word, word, phrase, subordinate clause, main clause, sentence and paragraph.

[11] The other strategy in written text production consists in pausing "long enough to process the intended portion of text before starting to write" (2010: 60). These strategies are two ways to control working memory overload (Immonen 2011: 236).

2.6 TRANSLATION AS PROBLEM SOLVING

In the previous sections, several references were made to translation as a problem-solving activity where translation strategies are employed more or less systematically. This section aims to present a more focused account of the problem-solving nature of the translation task.

The view of translation as a problem-solving and decision-making activity is widely accepted in Translation Studies but perspectives on the concept of "strategy" vary greatly (Palumbo 2009: 131-2) and can be used in the sense of either conscious (overt tactics) or unconscious procedures (mental procedures) (Séguinot 1991: 82). This overview takes a descriptive approach toward the concept of translation strategy and considers it in a narrower perspective, i.e. linked to the concept of "problem":

> A translation strategy thus becomes a procedure or method used to solve a particular kind of problem posed by the text to be translated or linked to the translation task (Palumbo 2009: 132).

From this definition, translation strategies could be said to refer to conscious procedures of problem solving. However, such a conclusion is probably too far-fetched, as testified by Krings' (1986: 268b) definition of translation strategies as "*potentially* conscious plans for solving a translation problem" (emphasis added) and considering Kiraly's (1995: 105) contention that "[s]trategies do not solve translation problems; they are merely plans carried out in an *attempt* to solve problems" (emphasis added). A point to clear up about the conscious character of translation strategies is the relationship between task automatization and problem solving. If strategies for handling translation problems are conscious, then they should not occur in the automatized stretches of a translation task. Krings' observation seems to speak in favor of conscious strategies: "Strategies emerge as soon as the translation cannot be carried out automatically" (1986: 268b).

Different kinds of strategies are used for different kinds of problems related to the translation task. In a broader view, problems can include "detecting properties of the source and target audiences, determining the extent of the translation brief, designing the structure of the translated document, etc." (Sharoff 2006). The most frequent type of problem is however linked to the choice of the appropriate target language version for rendering the source words and concepts, which brings to mind Pym's binary and minimalist view of translation competence (2003: 489). In particular, problem-solving activity is understood as "information processing aimed at solving a specific and identifiable translation problem" (Jensen & Jakobsen 2000: 111). Problem-solving strategies have been categorized in many ways (e.g. search strategies, creativity strategies and textual strategies; Chesterman & Wagner 2002: 57) and pertaining to different levels, e.g. global and local strategies (Jääskeläinen 1993), but strategy concepts are often flawed because the underlying notion of problem is too vague:

> What the authors consider to be a problem often is not made explicit. Nonetheless, an implicit use of the term 'problem' in a colloquial sense can be detected (Lörscher 1991a: 79).

Experimental studies highlighted varying translation behaviors with regard to strategies. In particular, Sharoff *et al.* (2006: 743) found that for the majority of problems, translators preferred very different translation solutions. Translators were found to switch strategies based on the interplay between memory constraints and the difficulties encountered in the source text (Séguinot 1989: 33-34). For example, Krings (1986b: 268ff.) identified five main strategies involved in problem solving:

comprehension strategies, strategies of equivalent retrieval, strategies of equivalent monitoring, strategies of decision-making and strategies of reduction. For the purpose of the present review, only the "comprehension strategies" will be considered, which originate from a comprehension problem. The two subtypes are (i) inferencing and (ii) the use of reference books[12]. The latter comprises further sub-types of strategies for the use of reference books, a very frequent one being a word look-up in a bilingual dictionary and a subsequent check for appropriateness in a monolingual dictionary. Inferencing strategies, on the other hand, were used when no support was available or the search was unsuccessful. Interestingly, inferencing strategies turned out to be non translation-specific and matched inferencing used in ordinary text comprehension (Krings 1986b: 270).

A more precise framing of the span of a translation strategy is proposed by Lörscher (2005: 599), who contends that

> [...] translation strategies have their starting-point in the realization of a problem by a subject, and their termination in a (possibly preliminary) solution to the problem or in the subject's realization of the insolubility of the problem at the given point in time.

The concept of translation strategies as solution to a problem is in line with the view put forward by Chesterman (1997: 89) and endorsed by Tirkkonen-Condit (2000). The starting point of a strategy as well as the translation process is one or more translation problems or obstacles that interfere with the natural progress of the translation. If the process of translation can be interrupted by obstacles and problems, there must be also sequences of a different kind in a translation that Mondahl and Jensen (1996: 102) call *spontaneous sequences* as opposed to *problem sequences*. The former are characterized by automatized uninterrupted behavior, the latter by stops and hesitations. Mondahl and Jensen derive problem indicators from TAPs collected from adult language learners working into their L2. Spontaneous sequences did not contain any (overt) problem signs such as verbalizations, stops and interruptions or any sign of problem-solving strategies. On the other hand, problem sequences could be identified not just through direct verbalization but also through a series of secondary indicators, i.e. competing translation suggestions for the same SL part, underlinings of ST elements, dissatisfaction with the chosen translation, corrections and pauses.

Once the starting and end points of a strategy have been established, a categorization of any additional activity taking place in between can be useful to map the translation process. Lörscher (1991a: 107ff., 2005: 599-600) identified two hierarchical levels: the lower level containing the *elements of translation strategies,* i.e. the smallest discrete problem-solving steps, and the second level capturing the *manifestations of translation strategies*. No matter how complex a translation strategy may be, it can be broken down into *basic structures* that in turn can grow into expanded structures and complex structures. *Expanded structures* occur when one element of a strategy is repeated whereas *complex translation structures* originate from the union of a basic an expanded strategy. Five types of basic structures are listed: (i) recognition of a problem, (ii) searching for a solution, (iii) verbalization of the problem, (iv) searching and verbalizing together and (v) a splitting-up structure (i.e. when a problem is too complex to be solved as a whole so the subject splits it smaller parts to be addressed separately). Some of these structures are in

[12] The examples provided obviously refer to the main forms of translation support available at the time of Krings' study. However, the type of resource can be easily updated to match current forms of support.

fact not relevant to the present analysis because no verbal protocols have been recorded. However a search instance in a concordancer can clearly represent an instance of type (ii) which necessarily follows type (i) for the very fact that the translator left the editing environment to consult a translation resource; type (v) is also plausible but the question remains as to whether it can be successfully recognized.

In a more cognitive perspective, Dragsted (2004: 55ff.) has developed an approach to problem-solving that used the concept of subject-dependent "problem space", i.e. "a person's internal (mental) representation of a problem, and the place where problem-solving activity takes place" (2004: 55). The ability to find solutions to a problem is linked to the "intelligence" (and expertise) of the problem solver which determines the goal-directed activity. The activities going on in the problem space are "a selective search through the possible TL items and structures [...] until the most appropriate solution [...] has been found" (2004: 55). When discussing human problem solving, Newell and Simon (1972: 88 in Dragsted 2004: 55-6) presented it as a serial task: first a representation of the problem is produced, then a method for solving it is selected and finally the method is applied. If no solution is found, the process repeats itself with the limitation that only one method is activated at a time. As all attention is geared towards the problem solving activity, in translation "[i]t can thus further be expected that the number of items in a translation unit will be limited to comprising only the problematic item" (Dragsted 2004: 56).

Comparisons between professional and non-professional translators have highlighted that the former engage in more problem-solving activities (Jääskeläinen 1999; Hvelplund 2011: 23) but the amount of problem-solving activity turned out to be affected by time constraints (Jensen & Jakobsen 2000). In addition to the approaches discussed here, other methods have been employed to identify problem-solving activities. By combining eye-tracking and keystroke logging, the *eye-key span* measure has been introduced which refers to the time span between looking at a ST word and producing its TT equivalent (fixation-to-production span) (Dragsted & Hansen 2008: 10,19). Problem words have been connected with longer eye-key spans and this measure has therefore been claimed to be an indicator of problem-solving activity and greater processing effort (Hvelplund 2011: 29). Screen recording (e.g. Göpferich 2009; Enríquez Raído 2011; Martín-Mor 2011) have allowed researchers to monitor any operation carried out on-screen by the subjects and obtain detailed information on translators' problem-solving strategies, among others.

2.7 CLASSIFICATION OF PROBLEMS

Closely connected to strategies is problem classification, which attempts to fit instances of translation problems (and/or difficulties) into typologies. According to Lörscher,

> [taxonomies or typologies of translation difficulties] are theoretical constructs which are based on single, individual, and largely unsystematic observations or, what is more likely, are hypothetically derived from a comparison of source- and target-language phenomena in a contrastive-linguistic way (1991a: 92).

Problem classifications are often employed in the didactics of translation e.g. to teach translation strategies. The distinction found in Nord (1991: 158ff.) between problems and difficulties (see Section 2.1) goes down to a further degree of detail to include pragmatic problems, cultural translation problems, linguistic problems and text-specific problems. Connected with the above-mentioned rich points, different types of translation problems have also been considered by the PACTE group (2009: 213): linguistic problems, textual

problems, extralinguistic problems, problems of intentionality and problems relating to the translation brief and/or the target text reader. These echo the list of problems mentioned by Sharoff (2006, see Section 2.6) and before him, by Jääskeläinen (1987), who classified problems into comprehension problems, monitoring problems (i.e. when searching for translation alternatives), editing problems and planning problems (i.e. style of the TT and requirements of the translation brief).

The alternation of problematic and non-problematic sequences seems to underlie the approach taken by Prahl and Petzolt (1997), who have distinguished between potential and actual translation problems. They attempted to develop clear criteria to tell problems and non-problems apart, with a view to finding common problem categories between human translators and machine translation. In their words,

> a potential translation problem becomes an actual translation problem when there is an information deficit at a certain point in time – without taking into account whether the translation is aware of this deficit information or not (1997: 25).

They proposed the following criteria to identify an actual translation problem: (i) a decision must be made during the translation process (ii) but there is an information deficit (iii) at a specific moment within the translation problem (iv) in a certain context. However, a systematic application of such criteria to both human and machine translation remains difficult.

One of the most widely adopted categorizations employed in translation process research distinguishes between reception problems (also receptive problems, comprehension problems, translation problems or L2-problems) and production problems. Problematic elements in the source text are not only caused by the inherent complexity of the text itself (*Rezeptionsprobleme*, i.e. problems related to the interpretation of the ST) but can also be due to difficulties in finding the appropriate target language equivalent for that particular source text element (*Wiedergabeprobleme*, i.e. problems related to the production of the target text) (e.g. Gerloff 1986: 252; Krings 1986a: 144-152; Mondahl & Jensen 1996: 102). In addition to these two main categories, a third "hybrid" category is sometimes mentioned to account for dubious instances. Krings (1986a: 144-152) called it reception-production problems whereas Mondahl and Jensen (1996) state that a problem can be both receptive and productive. Despite being an interesting classification, it proves difficult to implement without specific operational criteria and often the data and TAPs themselves do not qualify as indicators for a clear classification (Lörscher 1991a: 95-6). Lörscher (1991a: 201-217) grouped translation problems into three categories, i.e. lexical, syntactic and lexico-syntactic. Lexical problems, i.e. "single lexemes of the SL text for which the subject has no corresponding TL lexemes available" (1991a: 202), turned out to be the most populated category in his experiment (ca. 70%). The second category (syntactic problems, ca. 8%) was concerned with the syntactic arrangement of the lexemes, while the third group (lexico-syntactic problems, ca. 22%) either included both levels or was used when no distinction between the two could be made (1991a: 203). The main problem with these tripartite categorizations is the presence of the third "hybrid" group, not helpful for a systematic analysis. Lörscher further distinguished between problematic and non-problematic phases at the level of the source text, the former requiring the application of a problem-solving approach.

The source text has also been the focus of Campbell's analysis on difficulty (1999). Source-text related difficulties were attributed to word class, distributed meaning in a group of words, complex noun phrases, abstractness, frequency and familiarity of the subjects with source text expressions. As an alternative, a binary classification of problematic items in

the source text has been carried out by Désilets and colleagues (2009), who distinguished between Language for Special Purposes problems and Language for General Purposes problems according to the linguistic items that translators looked up in translation support tools.

Further details about problem classifications and operational categories will be provided in Section 7.3.2. Before moving on with the analysis, one last element needs to be taken into account when dealing with translation problems, i.e. the concept of "problem unit", which will be detailed in the following section.

2.8 UNITS IN TRANSLATION RESEARCH

When discussing translation (and the process of translation, in particular) researchers have often resorted to the concept of "unit" in an attempt to establish a clear and measurable constituent for more systematic analyses. The most discussed unit is without doubt the translation unit, which has attracted the interest of the researchers for over fifty years. More recently, other types of units have been introduced in translation studies to keep pace with the new developments in research methodologies and technological advances. This section aims at providing an overview of the main types of units found in the literature in addition to the translation unit, namely the cognitive unit, the attention unit and the problem unit.

2.8.1 TRANSLATION UNIT

The translation unit (TU) is by far the most popular type of unit discussed in the literature. Generally speaking, a TU represents the structure resulting from the cognitive processing and segmentation of the text while translating. A few extensive literature reviews have already been written about the different understandings of the concept and its evolution in time (Dragsted 2004; Alves & Vale 2009; Enríquez Raído 2011).

Dragsted (2004: 11ff.) first discussed the translation unit from a linguistic perspective and grouped scholars according to the suggested size of a translation unit, which is by now established as a flexible item with no set size. However, some scholars used to limit the scope of TUs to a given length in a somewhat normative perspective based on the concept of equivalence in translation. Dragsted focused on three main approaches proposing different sizes and scopes of TUs. The first approach considered the TU below sentence level[13] as the smallest unit of meaning expressed by a flexible unit (from morpheme to sentence). The second level considered the TU at clause or sentence level[14] as the smallest unit of analysis with an independent meaning. Finally, the TU can be found at the level of paragraph or text[15] with a greater focus on structure and style. In the second part of her literature review, she looked at the TU from a cognitive perspective[16], reviewing data-driven approaches to the definition of translation unit. Empirical data came largely from experiments on the translation process using TAPs.

[13] Dragsted specifically considered Vinay & Darbelnet (1995), Catford (1965), Newmark (1988), Barkhudarov (1993), Sager (1993) and Toury (1995).
[14] In this respect the following scholars were listed: Bell (1991), Hewson & Martin (1991), Lou (1999), Zhu (1999).
[15] For this approach, see Nida (1964, 1969), Bassnett-McGuire (1991).
[16] This is the view generally chosen in translation process research and promoted by Dechert & Sandrock (1986), Gerloff (1986), Lörscher (1986, 1991, 1996), Krings (1986a/b), Königs (1987), Kiraly (1995) and Jääskeläinen & Tirkkonen-Condit (1991), whom Dragsted discussed in detail.

Alves and Vale (2009: 254) identified three groups of researchers that focused on different aspects of the TUs. The fist group includes researchers who attempted to distinguish between problematic and non-problematic TUs and respective processing by translators[17]. The second group includes those who have tried to establish the different size of TUs processed by professional and non-professional translators[18]. The third group is made up of those who have attempted to determine the general size and nature of translation units[19]. A similar structure is found in Enríquez Raído's literature review (2011: 38ff.). She first discussed TUs from the perspective of empirical TAP studies focusing on translation units (and translation strategies) and then divided researchers into two main groups, i.e. those who considered TU size, level of analysis and the amount of processing involved,[20] and those who considered the (non)problematic nature of TUs[21].

Alves and Vale (2009: 253ff.) also presented other studies focusing on the concept of translation unit from a process-oriented perspective and briefly discussed Malmkjær's distinction between product-oriented and process-oriented TUs (2006). Product-based TUs are "pairs of ST and TT segments which can be identified by mapping the units of the TT onto the units of the ST" (Alves & Vale 2009: 254), which resembles Carl's understanding of "alignment units", i.e. correspondences between the source and the target text (static product data; Carl 2009b: 227). Alignment involves finding correspondences not just at word level within the same segment, but also between continuous or discontinuous sequences of different length and even across sentence boundaries. In this static perspective, product-oriented translation units do not change in time and can be directly observed. Process-oriented TUs, on the other hand, are dynamic units that change in time and that do not have a 1:1 correspondence between ST and TT elements. This is one of the reasons why "we do not exactly know what TUs actually are" (Carl 2009: 228) and "the concept of translation unit [...] cannot be easily framed" (Simard & Macklovitch 2005: 71). More importantly, the process-oriented view of TU relies greatly on experimental findings such as that the TU is "identified on the basis of cognitive processes observable (indirectly) in a set of data" (Dragsted 2004: 32).

Some researchers (e.g. Malmkjær 2006 and Alves 2006) agree on a definition of TU which highlights two main elements (i.e. the TU is mapped onto the source text and coincides with the translator's focus of attention). Translation units, therefore, can be understood as

> [...] segments of the source text, independent of specific size or form, to which, at any given moment, the translator's focus of attention is directed. It is a segment in constant transformation that changes according to the translator's cognitive and processing needs (Alves & Gonçalves 2003: 10).

In other words, "[o] foco de atenção e consciência é o fator direcionador e delimitador da unidade de tradução" (Alves 2006: 38). Similarly, Malmkjær defines a TU as "the stretch of the source text that the translator keeps in mind at any one time, in order to produce translation equivalents in the text he or she is creating" (2006: 92). Even though TUs are understood in terms of the ST, the translator's focus of attention (a central element for the

---

[17] Lörscher (1986), Krings (1986), Königs (1987), Kiraly (1995), Jääskeläinen and Tirkkonen-Condit (1991).
[18] Lörscher (1986), Kiraly (1995), Jääskeläinen and Tirkkonen-Condit (1991).
[19] Gerloff (1986), Lörscher (1986), Krings (1986) and Kiraly (1995).
[20] Dechert & Sandrock (1986), Gerloff (1986, 1987, 1988), Lörscher (1986, 1991, 1996), Jääskeläinen (1987, 1990), Séguinot (1989), Kiraly (1995), Jensen (1999), Lorenzo (1999), Jakobsen (2003), Malmkjær (2006).
[21] Krings (1986), Lörscher (1986, 1991), Jääskeläinen (1993), Kiraly (1995), Mondahl (1995), Mondahl and Jensen (1996), Dragsted (2004).

identification of a TU) is generally analyzed by looking at continuous text production (i.e. the TT) and in particular at production segments delimited by non-productive intervals (i.e. pauses) (Alves & Vale 2009: 254). Researchers point out that there is no direct correspondence but only a correlation between the TT segments and TUs because

> [a] text production segment is a text extract observable in TTs or along the text production process while units are only momentarily perceptible and identified in time by pause intervals as they catch the translator's focus of attention (Alves & Vale 2009: 254-5).

The other element of general consensus among researchers is the dynamic nature of TUs, which not only reflects the size of the processed unit but also the translator's personal past history and his/her general mental capacities (Malmkjær 2006). Experimental findings highlighted both intra- and intersubjective variability (e.g. between novice and professional translators[22]), in that "[…] translators navigate between different linguistic units and levels during translation" (Alves *et al.* 2010: 121). Dynamicity is derived to a great extent from the focus on time, which is a central element in Alves' definition of TU: "TUs are ST segments […], which attract the translator's focus of attention at a given time in the translation process" (Alves & Vale 2009: 254). The concept of translation focus is not entirely new: Bennett (1994: 13) used it a decade earlier to indicate "the section of text which the translator focusses on at any one time" and maintained that

> [e]ach UT [unit of translation] is part of a larger unit, and so on up till the entire text is reached, but no translator can work with any text other than the very shortest as an undivided UT, for reasons of memory limitations if nothing else.

Indeed, the dynamic approach is generally linked to considerations on cognitive load, working and long-term memory, which are reflected by pauses in text production. Alves and Vale (2009) mark the beginning of a translation unit with a pause in keystroke data; the unit then continues in a production phase until a new pause interrupts it. The translation process does not unfold in a linear fashion and previous units are re-used, revised, edited, deleted and so forth. These operations enable researchers to identify macro- and micro-translation units. In fact, translation units are not necessarily just seen as minimal units, but each TU "may be modelled as a macro-unit, a composite of constituent micro-units" (Alves *et al.* 2010: 123).

Alves and Vale (2009: 257) provide the following definition for micro- and macro-TUs:

> A micro TU is defined as the flow of continuous TT production – which may incorporate the continuous reading of ST and TT segments – separated by pauses during the translation process as registered by key-logging and/or eye- tracking software. It can be correlated to a ST segment that attracts the translator's focus of attention at a given moment.
>
> A macro TU, in turn, is defined as a collection of micro TUs that comprises all the interim text productions that follow the translator's focus on the same ST segment from the first tentative rendering to the final output that appears in the TT.

In operational terms, micro TUs include all changes implemented online (i.e. while translating) and found between two pauses of a given length. On the other hand, a macro TU is not necessarily time-bound as far as pause length is concerned and includes both online and end-revision operations. Researchers were specifically interested in studying

[22] e.g. Dragsted (2004); Alves & Liparini Campos (2009a,b).

the distribution of translation units between the drafting and revision phases. Another study that distinguished between micro and macro (attention) units was carried out by Hvelplund (2011), who looked at the distribution of units between the reading and writing processes (see Section 2.8.3 below).

The interplay between reading and writing phases became prominent in relation to the concept of the eye-key span proposed by Dragsted and Hansen (2008). They first defined segments by identifying their boundaries as pauses of an arbitrary length (e.g. at least 1.5 seconds) and then looked at the type of processing (comprehension and/or production) within each segment. Three scenarios emerged: (i) segments concerned solely with ST reading/comprehension (ST segments); (ii) segments containing only TT production (TT segments) and (iii) segments where both reading and writing activities took place (ST/TT segment) (2008: 16). The existence of instances for each scenario raised questions on the pause-defined understanding of the translation unit (or *translation segment*, as the authors tend to call it). In other words, the issue arose whether the notion of translation unit could be applied to all three scenarios or should involve target text production (or the mixed type). Dragsted and Hansen proposed a distinction between orientation (ST segments) and production (TT production and mixed type) (2008: 24). With respect to the previous study based solely on keystroke logging (Dragsted 2004), additional data could be collected in this later study (Dragsted & Hansen 2008) after introducing eye-tracking for data collection. It turned out that

> the input material (ST reading) associated with the corresponding TT output is rarely found within the same (pause-defined) segment, but rather in the previous segment, raising the question of whether the pause criterion for segmentation is operational (Dragsted & Hansen 2008: 25).

A closer analysis revealed that the coordination between reading and writing takes place across segment boundaries and that "the pause seems to signal a coordination effort, i.e. *a transition from SL comprehension mode to TL production mode*, rather than a transition from one ST/TT unit of meaning to the next" (Dragsted & Hansen 2008: 25; emphasis in the original). The scholars came to the conclusion that a pause-defined segmentation may not be the ideal solution and that a segment could rather be seen as "a TT string and the matching ST input often divided by a pause" (Dragsted & Hansen 2008: 25), which seems to indicate a "run-up" to production of the target segment (p. 27). This view of pauses as "run-up" to production will be central for the discussion of translation problems as information needs in Chapter 4.

2.8.2 PROBLEM UNIT

As opposed to the translation unit, the concept of problem unit has been sparingly used and usually more as an intuitive entity rather than an operationalized one. This is possibly due to the different approaches to the study of translation problems that sometimes do not lend themselves to a categorization of this kind.

Kiraly (1995: 86) explicitly distinguished two types of translation units:

> (a) units that were problems and required cognitive attention and the application of conscious or potentially conscious strategies and (b) units whose solutions came from intuition and spontaneous association, apparently without the intervention of problem-solving strategies.

He conducted a case study and reported that all subjects experienced both types of units and the mean number of problem units was higher than non-problem units with some

intersubjective variation. Kiraly's distinction can be directly linked to the conception of translation as alternation of problem and non-problem sequences, though for some reason the word "unit" has been generally avoided in the literature.

A problem unit intuitively represents a portion of the source text that corresponds to a problematic item for the translator. The question as to what relationship, if any, exists between a translation unit and a problem unit has not been explicitly discussed in the literature but will be tentatively addressed in Section 2.9 below.

2.8.3 ATTENTION UNIT

Attention unit is a concept originally borrowed form cognitive psychology and then applied to translation process studies particularly in relation to eye-tracking, to mean "visual attention, i.e. recordable eye fixations" (Dragsted 2010: 47). If all gaze activity allocated to a specific area is aggregated, an indication of the cognitive processing pertaining said area can be obtained (Hvelplund 2011: 73). Before eye-tracking was introduced in Translation Studies, the concept could not be linked directly to gaze activity and TAP data were used instead (Jääskeläinen 1999, in Hvelplund 2011: 73).

An attention unit thus represents "a unit of problem-solving activity, or an instance of *marked* processing which interrupts the smooth unproblematic flow of the translation process" (2011: 73; emphasis in the original). Moreover, Jääskeläinen (1987, 1990) found out that the majority of the attention units were related to target text production rather than source text processing which differs from the general mapping of units of attention onto the source text. This difference could be tentatively explained by considering the methodology. In TAP, translators tend to verbalize the problems they encounter in terms of the TT because the translation task itself triggers the verbalization of the problem. Keystroke logging uses target text data that are then mapped back onto the source text while eye-tracking can virtually consider both levels simultaneously. These consistent methodological differences may, to a certain extent, be at the root of different understandings of attention (and translation) units. A more recent perspective, maintains that

> [b]y concentrating our attention on a particular ST segment which attracts the translator's focus of attention [i.e. a TU], we can see it evolve into correlated text production segments (micro TUs) which together form a macro TU, i.e., a combination of processing steps in on-line text production, until it appears as a TT segment, a definite solution, at the end of the translation process (Alves & Vale 2009: 257).

Based on this definition, a "unit" can be defined as

> a segment of the source text independent of specific size or form to which, at a given moment, the translator's focus of attention is directed. It is a segment in constant transformation that changes according to the translator's cognitive and processing needs (Livbjerg & Mees 2003:129).

Hvelplund too considered attention units as markers for cognitive processing or problem solving. He operationalized an Attention Unit (AU) for his eye-tracking study and defined it as "a time measurement unit of uninterrupted cognitive processing, as indicated by eye movement data (fixations and saccades) and typing events" (2011: 73). As opposed to TAP-based AUs, attention units identified by eye-tracking have clearer boundaries and in Hvelplund's study, AUs are determined by three properties: (i) a task which is the focus of the translator's attention (*type* of AU), (ii) a specific attention shift (i.e. the *latest*) marking

the *beginning* of an AU and (iii) a specific attention shift (i.e. the *next*) marking the *end* of the AU[23]. In particular, AUs can be related to different sub-processes and be mapped onto the source text (ST reading and comprehension), the target text (TT reformulation and TT reading or typing) or may indicate parallel attention to both elements (ST reading/comprehension and TT reformulation/typing) and their duration is assumed to represent the translator's conscious response to the processing requirement of the task (Hvelplund 2011: 74,78).

Similarly to translation units, AUs have been subdivided into micro-AUs and macro-AUs. Micro-AUs were obtained by looking at neighboring data rows in the gaze and typing data whereas macro-AUs correspond to the three possible locations of an AU, i.e. source text, target text, parallel attention plus the case where not enough data is available (Hvelplund 2011: 116). Findings from Hvleplund's study may help dissipate doubts regarding the locus of the translator's attention. Building on Alves' assumption that TUs are source text segments and the focus of the translator's attention (Alves *et al.* 2010: 124), an inference can be drawn that all TUs are made up by source text elements but not all foci of the translator's attention are necessarily source text elements. In addition, Hvelplund found that there are differences in the distribution of AUs between professional translators and students (Hvelplund 2011: 138-9). The former focused longer on TT reformulation than ST comprehension, whereas students had more ST-comprehension related AUs than TT reformulation AUs, meaning that expertise (in addition to the specific data elicitation method) can affect experimental results.

2.8.4 COGNITIVE UNIT

The concept of cognitive unit has not been explicitly defined and is hardly to be found in the literature on process research. However, some researchers (e.g. Dragsted 2004: 42) discussed *cognitive segmentation,* which is dependent on the human memory system and presents translation units as a product of this segmentation.

The expression "cognitive unit" is mentioned in passing when discussing Translog data (e.g. Dragsed 2004: 143, Alves & Vale 2009: 255) to explain the presence of segments that were not justified by the syntactic structure of the text but nonetheless seemed to make sense for the translator. Moreover, cognitive units can be identified in real time by adding the time element to the linear representation of the text production.

When discussing average TU size, Dragsted (2004: 150) compared the length of processed segments with the size of "normal information processing units" and found out that TUs comprised approximately the same (or slightly fewer) items, irrespective of the level of expertise of the translator. *Information processing unit* can be used as an alternative label for cognitive units.

In this perspective, cognitive units can be understood as the result of cognitive segmentation, i.e. the object of cognitive processing at any given point in time. Cognitive segmentation was found to differ between professional translators and students suggesting that cognitive units would differ, too. As is the case with translation units, cognitive units do not need to be understood using a set size because they were shown to vary and they can also be related to a specific activity, such as problem recognition, solution proposal and solution evaluation defined by uncertainty management (Angelone 2010: 23). A statement in Sager (2001: 259) effectively summarizes the present understanding of a cognitive unit in relation to the translation task: "In its simplest form,

[23] For further details on the specific kind of attention shifts, see Hvelplund (2011: 73).

the focus in translation can be said to be on the linguistic representation of a cognitive unit; […]".

2.9 RELATIONSHIPS AMONG UNITS

This review of how units are understood and studied in translation research has highlighted many overlaps between labels and pointed to a number of synonymous expressions that sometimes add to the confusion. Given the considerable technological and methodological advances, particularly in the study of the translation process, it is perfectly understandable that the original theoretical concepts have changed and evolved. In turn, empirical studies need to propose new labels for newly discovered phenomena. All in all, there seems to be a relatively small set of recurring expressions that could be mapped so as to propose a basic framework of reference for translation process research and hopefully accommodate most existing approaches.

When studying the translation process, references have been made to existing research both in reading and writing, which are the two main (observable) activities underlying translation. In addition, some studies have specifically compared translation with "pure" reading and writing (e.g. Jakobsen & Jensen 2008 and Immonen & Mäkisalo 2010, respectively). On the basis of some of the previously discussed findings (e.g. ST segments and TT segments in Hvelplund 2011), it would therefore seem reasonable to consider reading and writing in addition to the translation activity.

Cognitive units represent the mental processes going on at any given time and could be placed at the highest level of the hierarchy. This is probably the hardest level to define because it cannot be physically accessed and it can be generally considered an unconscious type of processing because people are usually not aware of the fact they are reading something (they would rather focus on *what* they are reading) or the type of information processing going on while writing, let alone during a translation task. This level can be tapped into thanks to some kind of manifestations of ongoing cognitive processes, as already pointed out by Sager (2001: 259): "In its simplest form, the focus in translation can be said to be on the *linguistic representation* of a cognitive unit; […]"(emphasis added). Based on the reviewed literature, examples of manifestations could be eye movements and fixations for reading and typing behavior for writing.

When fixations occur, they are usually referred to as *attention units.* Because of the eye-mind assumption, the fixated items are taken to represent those elements the translator's mind is attending to, i.e. focusing on. As regards reading patterns, a number of factors can affect the typical duration of fixations (i.e. between 200 and 250 milliseconds; Jakobsen & Jensen 2008: 114), such as word familiarity, word predictability, word length and complexity and lexical and/or syntactic ambiguity (Dragsted & Hansen 2008: 19,20; Jakobsen & Jensen 2008: 103). On the other hand, one feature of written language production to have received scholarly attention is pause time distribution which Immonen and Mäkisalo presented in their study (2010: 46). In particular, researchers combined information about pause length (representing the writing process) and pause location in the text (the product) to infer the mental processes underlying the writing task, i.e. planning and decision-making. Moreover, the types of pause location were thought to be dependent on the type of text and the unit of language but there could also be other cognitive causes (e.g. physical or socio-psychological reasons) affecting the writing patterns. As translation consists of both tasks, these manifestations (fixations and pauses) can be expected to be applicable in the case of a translation task, as was empirically shown by e.g. Jakobsen & Jensen (2008) and Immonen (2006). Thus, fixations and pauses have

been used as manifestations of cognitive segmentation in reading and writing, respectively, i.e. they signal units of attention (or attention units). The concept of attention unit has been used in translation to refer to fixations (Hvelplund 2011) but also to target text production with pauses as boundaries (Alves & Vale 2009). The latter concept is perfectly justifiable if the activities of reading and writing are considered in isolation or rather in sequence. This view presupposes that the (written) translation process is *serial* (also referred to as *linear*, *sequential* or *vertical*), meaning that "full comprehension of the SL message must be achieved before any rendition of the message in the TL can begin" (Hvelplund 2011: 61). However, empirical evidence (Dragsted & Hansen 2008; Hvelplund 2011) and theoretical accounts (Séguinot 1989: 76) support the *parallel* view of the translation process (i.e. ST comprehension and TT production taking place at the same time) though some variables such as text type and translator's expertise may impact the overall processing (Dragsted 2004: 50). Dragsted and Hansen referred to these ST/TT segments as translation segments, which they also call translation units.

Translation units (or units of translation) can then be those instances where parallel reading and writing activities take place. *Attention units* can therefore refer to the general foci of attention that can be highlighted in reading and writing, thus comprising both instances of ST segments and TT segments in translation. Alternatively, attention units can be considered the "unmarked" type of unit found in general reading and writing tasks, whereas *any* unit identified in a translation task could be labeled translation unit, to distinguish it from the unmarked type for reading and writing. This alternative has been suggested in the event that the parallel processing perspective is perceived as too restrictive or when no UAD can be collected for both eye-tracking and key-logging. Translation units can represent a very dynamic situation because they could include a ST segment and a TT segment that do not directly match (i.e. where no alignment is possible). However, these units would still be manifestations of cognitive processing in translation just as attention units do for cognitive units in reading and writing. At this level, attention and translation units could be conscious or unconscious. An example of conscious unit of attention in reading might be an expression that the reader particularly enjoys and therefore s/he reads again or, in writing, an editing of the written text. However, chances are that in translation these units are dealt with unconsciously by default, e.g. the translator is not aware that while typing a TT segment s/he is actually reading the following ST segment. The question now arises as to whether or not there is a quantitative relationship to be established between the cognitive level and the attention level. On this regard, Dragsted explains that

> the TU size data suggest that all translators, regardless of expertise level, process segments comprising approx. the same or slightly fewer items than normal information processing units (2004: 150).

One hypothesis is therefore that segment size in translation (the translation unit) is basically comparable to the size of attention unit for reading and/or writing. At the level in which cognitive units are observable, no striking differences are found between the processes but the question remains as to how faithful the manifestation of these units is with respect to the actual processing because e.g. the eye-mind assumption does not always seem to hold (Hvelplund 2011: 68-9). Empirical observation has then highlighted that attention (and translation) units are not always the same. There are instances where "unusually" long pauses or fixations were found, e.g. in monolingual text production, pauses seem to occur mostly at sentence and paragraph boundaries, as if to indicate instances of global text planning (Immonen 2011: 250). At word level, a pause could be caused by a spelling problem in writing whereas a pause in a reading task could be triggered by anaphora resolution or a long and articulated word. In all these instances,

there appears to be one (or more) element(s) that somehow disrupt the normal flow and progression of attention units. Most likely, such disruptions are due to a problem (or difficulty) of some kind. Attention units that are delimited by indicators of greater cognitive effort such as "long"[24] pauses could be termed *problem units.* In translation, problem indicators have been well documented at the boundaries of segments (Dragsted 2004; Hvelplund 2011) and the resulting units could be labeled *translation problems* or *problematic translation units*, as a special case of translation unit. Seen from a different perspective, "'[p]roblems can here be seen to be a subset of 'attention units'" (Pavlović 2007: 30). As will be further discussed in Section 7.3.5, these translation problems can pertain to ST reading and comprehension as well as TT production. The question arises as to what relationship exists between translation and problem units. Available data unfortunately are not enough to give an answer to this question, though some hypotheses have been formulated: "[I]t can […] be expected that the number of items in a translation unit will be limited to comprising only the problematic item" (Dragsted 2004: 56). This means that if a text portion is considered a translation problem, the corresponding translation unit will be longer or, at most, equal to said unit. In terms of cognitive status, problem units and translation problems can be quite safely considered conscious units as abundantly shown by TAPs and maintained by several scholars: "Our definition of a 'problem' is very broad, meaning only that the unit in question was raised to the level of consciousness" (Livbjerg & Mees 2002: 161).

Furthermore, an additional level of analysis could be found within each overt level (i.e. attention/translation units and problem unit/translation problem) in terms of micro and macro structures, as already pointed out by Alves and Vale (2009) and Hvelplund (2011). In this view, Bernardini's statement seems to hold: "[A]ttention units are better defined in hierarchical rather than sequential terms, with smaller units being processed within larger units" (2001: 249). To better exemplify the proposed taxonomy, the main concepts, their hierarchical structure and the relevant cognitive status have been summarized in Table 1.

Table 1. Proposed taxonomy and hierarchy of labels for the units of translation activity.

LEVEL (top-down)	**ACTIVITY/TASK**			**COGNITIVE STATUS**
	READING	**TRANSLATION**	**WRITING**	
COGNITIVE UNIT[25]	Comprehension	Comprehension and/or Production[26]	Production	Unconscious
(manifested as)				
ATTENTION UNIT	Fixations (eye-tracking)	**TRANSLATION UNIT**[27] Fixations and/or Pauses	Pauses (keystroke logs)	(Un)conscious
(a special type is)				
PROBLEM UNIT	Long fixation	**TRANSLATION PROBLEM**[28]	Long non-productive pause	Conscious

[24] The value of "long" needs to be defined and agreed on within the research community, or at least it has to be clearly stated when presenting and discussing empirical studies.
[25] Alternative labels, *cognitive segment* and *information processing unit.*
[26] Assuming that translation is a PARALLEL process. The conjunction "and/or" is used to account for the scenario where translation units are taken to represent *any* processing unit in translation and those cases where no UAD for both eye-tracking and key-logging can be collected.
[27] Alternative labels, *unit of translation, translation segment.*

The concept of problem unit will be taken up again and further discussed in Section 4.6 to contextualize it in the present research, while Sections 7.3.2 and 7.3.5 will attempt to illustrate different types of translation problems based on the data used for this research.

2.10 KEY CONCEPTS

- There is still no full consensus among researchers about the nature and definition of translation problems, let alone their classification.
- Difficulty and uncertainty in translation are two closely related concepts to that of translation problems are sometimes used interchangeably.
- The notion of translation problem can be approached from different perspectives, the main ones being: product-oriented, subject-oriented and process-oriented.
- In the product-oriented approach, translation problems are examined by looking at translation errors, that is from a translation evaluation perspective which tends to be more common in translation pedagogy.
- In the subject-oriented approach, translation problems are elicited more or less explicitly and directly from the subjects by means of written or verbal Q&A sessions and/or retrospective verbalizations.
- Two main approaches fall within the process-oriented approach, namely concurrent verbalizations (TAPs) and pause analysis.
- Pause analysis used to be based on TAPs but technological advances turned it into a more quantitative analysis of keystroke and eye-tracking logs.
- Key-logging and/or eye-tracking (possibly complemented by retrospective verbal protocols) have become the new standard in translation process research. The data thus obtained are sometimes referred to as User Activity Data (UAD).
- Problems play a role in the very definition of translation, which can be seen as a problem-solving activity where decision-making and (possibly conscious problem-solving) translation strategies are applied in alternating problematic and non-problematic sequences.
- Over time, a number of different classifications for translation problems have been proposed, originating from various perspectives of analysis.
- A limited number of recurring concepts in the literature about translation was highlighted, namely translation unit, problem unit, attention unit and cognitive unit.
- Despite existing uncertainty about the nature of the translation unit, some researchers seem to agree on at least some features, i.e. it can be framed in terms of the source text and it has a dynamic nature, meaning no fixed size.
- A hierarchical relationship between the units has been proposed that takes into account at the same time the tasks of reading, writing and translating.

[28] Alternative labels, *problematic translation unit* and possibly *problem nexus.*

CHAPTER 3: OVERVIEW OF CONCORDANCING TOOLS

This chapter contains the second part of the literature review; it will cover in greater detail the topic of computer-assisted translation and, more specifically, the available types of concordancers. In the first part of the chapter, different types of concordancers will be presented together with the contexts in which they are used. The main section presents features and interfaces of a number of online concordancing tools and then moves on to examine the concordancer used for this study. This tool is used internally by EU staff translator, which implies that the framework for the analysis is well-defined and well-known. A few words on the internal organization and functioning of the EU translation services, with a special focus on the European Commission will be spent in Chapter 5.

3.1 TYPES OF CONCORDANCING TOOLS

In this section, concordancing tools will be presented and contextualized in different domains according to their context of use and purpose. On a terminological note, different classifications for translation technologies have been proposed over the years[29] that distinguish between a "tool" and a "resource".

> The word *tool* refers to computer programs that enable translators to carry out a series of functions or tasks with a set of data that they have prepared and, at the same time, allows a particular kind of results to be obtained. [...] By *resources* we refer to all sets of data that are organised in a particular manner and which can be looked up or used in the course of some phase of processing (Alcina 2008: 94; emphasis in the original).

Examples of the former are word processors, assisted translation software and terminology database management software, while the latter include dictionaries, corpora and other closed data sets. From this perspective, the concordancer can be seen to fall in both categories. Concordancers in local Translation Memory Systems can be seen as tools, whereas stand-alone concordancers become resources because they are used mainly as an external reference source. They enable users to query a repository of documents by entering a text string (the query). The program retrieves the matching items from the database and displays them to the user according to the specific interface and architecture of the system. Concordancers show the matching item in its surrounding context, the so-called KWIC (Key Word In Context). They can be monolingual if the texts to be analyzed are all in one language. Alternatively, concordancers can be bi- or multilingual and this is the type of tool discussed in the next sections.

3.1.1 CONCORDANCERS IN CORPUS STUDIES AND ACADEMIA

Traditionally, one of the first practical applications for concordancing software programs is corpus linguistics. A number of tools have been developed over the years in the research community to make full and better use of existing corpora. Concordancing

[29] See Alcina (2008) for an overview of the different classification approaches.

tools[30] in academia have long been known in corpus studies research and students in universities have been exposed to them since the late 1990s (Quah 2006: 113). For example, Multiconcord[31] is a multilingual parallel concordancing tool which has been mainly used for foreign language teaching (Gross 1998) and in general for classroom activities, though studies focusing on translation also exist (Scarpa 1999, 2000).

A probably more well-known parallel text concordancing software is ParaConc[32] (Barlow 1996). This tool is designed primarily for linguists and researchers to enable them to carry out contrastive language studies in the field of translation studies and/or to investigate specialized or technical information texts (Barlow 2004). It contains an alignment utility to perform semi-automatic alignment of the files, allows users to select different search modes and granularities and produces frequency statistics (Barlow 2002). Another corpus analysis software coeval with ParaConc is WordSmith Tools[33]. It offers word and keyword lists as well as a concordancing function, used in a number of studies in translation-oriented corpus studies (e.g. Olohan 2004)[34]. A slightly more recent addition has been the freeware software AntConc[35], a multi-purpose corpus analysis toolkit, originally designed for use in the classroom (Anthony 2004) but whose scope has been broadened to meet translators' needs when dealing with electronic text corpora e.g. to search for terminology in specialized domains.

The role of concordancing software in the realm of corpus linguistics resources has gained further ground in empirical studies, such that "the creation of concordances, i.e., formatted displays of all the occurrences of a particular type in a corpus, may be considered the most fundamental task" among the routine procedures of corpus analysis, such as counting, organizing and displaying the results (Wiechmann & Fuchs 2006: 107). Wiechmann and Fuchs (2006) conducted a comparative analysis of as many as ten concordancing software programs[36], both commercial and freeware, along three main dimensions: (i) functionality, (ii) performance and (iii) usability. The aim of the study was to provide an overview of the features and performance of search-and-retrieval software available at that time. For their study, the researchers used four datasets of different sizes, ranging from 100,000 words (a subset of the Brown corpus) to 100 million words (the British National Corpus in full). They found that "not a single one of the programs tested was able to perform all searches" (2006: 109) possibly because they used a test system that matched the standard office-computer, which was underpowered for some of the tasks performed. As a means of comparison, the same searches were also performed using a scripting language (Perl) and the command line and as a result all operations were carried out successfully, despite the limitations of the machine used. This result showed that "the queries are not impossible *per se* on [the] test system. Instead, the gap in

---

[30] As anticipated, reference will be made exclusively to bilingual (or multilingual) concordancers. For details about monolingual concordancers, refer to e.g. Bowker (2002: 53ff.) and Simard *et al.* (1993: 2-4).

[31] http://artsweb.bham.ac.uk/pking/multiconc/l_text.htm (last accessed: November 2012).

[32] http://www.paraconc.com/index.html (last accessed: November 2012).

[33] http://www.lexically.net/wordsmith/ (last accessed: November 2012).

[34] For a complete account of related works, refer to: http://www.lexically.net/wordsmith/corpus_linguistics_links/articles_using_wordsmith.htm (last accessed: November 2012).

[35] http://www.antlab.sci.waseda.ac.jp/antconc_index.html also including a list of related works (last accessed: November 2012).

[36] MonoConc Pro 2.2, WordSmith Tools 4, Concordance, Multi Language Corpus Tool, ConcApp 4, AntConc 1.3, Aconcorde, Simple Concordance Program, Concordancer for Windows 2.0 and TextSTAT 2.6.

performance [...] is probably best explained by the difference in functionality" (2006: 109) because the programs used a large amount of memory for the analysis while the script only executed each command sequentially without taking up all the memory but, in so doing, the scripts reduced the scope of the analysis to individual operations. For each system, suggestions were given as to which analysis or data type would be best suited for a given tool, based on its performance in the study.

Finally, another corpus analysis tool will be briefly mentioned that was developed in the 2000s following a cooperation between Italian universities. TaLTaC2 which stands for Trattamento Automatico Lessicale e Testuale per l'Analisi del Contenuto (Automatic lexical and text analysis tool for content analysis)[37]. It allows statistical analysis on any linguistic data type, can be combined with other linguistic and statistical programs, currently supports up to five languages and it has been employed in a number of studies in different fields, from linguistics and translation (e.g. Ondelli & Viale 2010) to text mining and computational linguistics[38].

In addition to their applications to language teaching and contrastive studies, concordancing tools have also been employed in translators' and interpreters' training, more specifically in the area of Language for Special Purposes and specialized translation (e.g. Gavioli 1999; Scarpa 2006). Further examples of applications of bilingual concordancers in academic settings for translators' training can be found in Bowker (2002: 19-20).

3.1.2 CONCORDANCERS FOR THE TRANSLATION PROFESSION

The brief overview of past and present research and teaching activities that use corpora and concordancing software in academia did not consider the usability of such concordancing tools for professional translators. Indeed, bilingual concordancers

> have long been known in fields such as language teaching or second-language learning [...] but it is only more recently that their potential as translation aids has been recognized (Bowker & Barlow 2004: 74).

Both WordSmith Tools and AntConc are not just used for corpus-based research but also perceived as resources for professional translators, so much so that they have been (and still are) regularly reviewed and compared in (online) journals about translation (e.g. Reppen 2001; Wilkinson 2011, 2012). However, this has not always been the case. Looking at the professional literature and magazines aimed at language professionals, Bowker and Barlow (2008: 8-9) noted that bilingual concordancers were hardly mentioned in the professional settings, suggesting that they were less widely known with respect to Translation Memories (TMs). In a purely chronological perspective, the development of the translator's workstation (and TMs) can still be considered a relatively recent phenomenon, which Hutchins (2005) attributed to two important developments, namely "the acquisition and management of terminology" and "the appearance of systems for storing large corpora of bilingual texts" with the possibility of extracting examples of translations of phrases and sentences. These have become known as Translation Memories and originated as a "by-product of research on statistical methods of MT [Machine Translation]" (Hutchins 2005). Macklovitch and Russell (2000: 412) have distinguished between two main usages of the label "Translation Memory". According to a

[37] http://www.taltac.it/it/taltac1.shtml (last accessed: November 2012).
[38] For a complete account of related works, refer to: http://www.taltac.it/it/materiali1.shtml (last accessed: November 2012).

narrow definition, TMs represent the type of translation support tool that automatically retrieves matches from a repository; the broader definition sees TMs "simply as an archive of past translations, structured in such way as to promote translation reuse". The concordancing function in TM systems allows translators to manually search within a TM and use it "as the reference tool it was originally envisaged to be" (Benito 2009). This seems to have been the most prominent (and useful) functionality in the adoption of Translation Memories. Because the narrow definition presupposes automatic look-up of translation matches, interactive bilingual concordancing tools would be excluded from the class of TMs (Macklovitch & Russell 2000: 412). A similar differentiation was proposed by Macklovitch *et al.* (2000: 1201) who used the concept of TM to refer to "an archive of existing translations, structured in such a way as to promote translation re-use" (see above), whereas "the widespread use of the term [...] has arisen as a result of the popularity of commercial translation support systems".

Irrespective of the scope of the definition, a common feature of bilingual concordancers and Translation Memories is information extraction from parallel corpora. The hidden potential of bilingual parallel corpora for translators had already been pointed out by Bowker and Barlow (2004: 70). They presented concordancers as the 'old' technology and reportedly "not well known in the translation industry outside of academic circles" (2004: 70). Because of sustained lack of investigations on concordancers and TMs, the authors took charge of the task and compared a bilingual concordancer (ParaConc) and a popular TM system (SDL Trados) (Bowker & Barlow 2004, 2008). Despite the fact that the same corpus data can be used with both concordancers and TMs, there is an underlying difference in the way the texts are stored in each system. A bilingual concordancer maintains the aligned units (at sentence level) of source and target texts within their surrounding text whereas a TM splits the source and target texts into segments of sentence-like length which are paired to form a Translation Unit (TU), though in the process "the very notion of a document is lost" (Macklovitch & Russell 2000: 415). Each TU is stored individually in the database and the original text cannot be retrieved as a whole, unless specific solutions are devised beforehand[39]. However, a direct comparison of the two systems as conducted by Bowker and Barlow seems somewhat unjustified as the two systems are concurring rather than mutually exclusive or complementary. Almost a decade earlier, Simard *et al.* pointed out that

> [a] translation memory has applications beyond the simple recovery of translated sentences: it can serve as the basis for a *bilingual concordance tool*, a program which finds occurrences of specified expressions or pairs of expressions, and displays them in their bilingual context (1993: 2, emphasis in the original).

The structure of a bilingual (or multilingual) concordancer is relatively simple compared to other translation tools. It only requires reference texts to be pre-processed so that the content can be quickly indexed and retrieved by the system. Reference documents usually come in the form of bi-texts or Translation Memories, which can be collectively (and loosely) defined as parallel-aligned corpora. All concordancing software works according to the same underlying principle: each search has to be manually entered and launched by the user as opposed to Translation Memory systems, whose matching algorithms run automatically and the user just needs to accept or reject the proposed solution.

---

[39] One exception is MultiTrans Prism, a tool that is able to incorporate the entire text and index the document as a corpus when a TM is created [http://www.multicorpora.com/ - Last accessed: December 2012]. Another attempt in this direction was the introduction of 'context match' in some TM systems. This corresponds to a 101% match because the TU also contains information about the preceding and following segments.

A comparison of "full-sentence repetitions processing" and bilingual concordancing was carried out in another study (Macklovitch *et al.* 2000: 1205). The authors stressed the greater flexibility of concordancing tools for searching and pointed out the differences between a concordancer, a translation tool and a terminology bank. If a term bank can count on evaluated and validated entries, a concordancing tool can help translators to solve "many problems that often aren't catalogued in either term banks or bilingual dictionaries" thanks to the wider topic coverage of a large corpus over more specialized TMs (2000: 1205). A major difference between the two systems pertains to the degree of automation for database retrieval. The TM carries out the comparison automatically, while concordancers require their users to manually enter the query in the system. Traditional TM systems match strings based on similarity, understood as number of shared characters (or string edit distance) but in order to deal with sub-sentential matching and retrieval, automation needs to be reduced and the user allowed "to manually select and submit a word or phrase to the bi-textual database via a [...] concordancing tool" (Macklovitch & Russell 2000: 414-5). The lack of automation in the concordancer was initially presented as a potential limitation (Bowker & Barlow 2004: 72), but is in fact a very important feature because the translator is always in control of the search process and deliberately looks up whatever unit is needed. Back in 1990, Pierre Isabelle had already suggested that "existing translations contain more solutions to more translation problems than any other existing resource" (in Macklovitch *et al.* 2000: 1205).

User behavior in relation to the use of linguistic resources and tools has been investigated through ethnographic studies carried out among Canadian professional translators by the National Research Council of Canada (Désilets *et al.* 2009; Désilets, Brunette *et al.* 2008; Désilets, Farley *et al.* 2008) using methodologies from ethnographic practice such as Contextual Inquiry (CI). CI is a one-to-one field interview conducted in the actual work environment of the subjects and the data are later analyzed using grounded theory[40]. A similar study was conducted in Ireland (Karamanis *et al.* 2010, 2011) focusing on real work practices and in particular on the areas of Human-Computer Interaction (HCI) and Computer Supported Cooperative Work (CSCW). The findings of these studies provide a solid basis for further investigation in the way translators use translation aids to solve translation problems.

In particular, concordance searching is usually referred to as a terminology- and/or phrase-oriented search (Benito 2009: 4) but previous studies (Bowker 2002: 88) highlighted that some translators record for personal reference (specialized) phrases or expressions that do not qualify as terms in the conventional sense. Even though this view is likely to be closer to the reality, a large-scale empirical research on authentic concordance searches looking systematically at search patterns is still lacking. Some earlier studies on concordancing can be found that focused primarily on idiomatic expressions and fixed formulae (Simard *et al*. 1993) but the Canadian research group has advocated the need for greater language coverage and analysis of problems relating to Language for General Purposes (Désilets *et al.* 2009) and has pointed out that, at least until 2008, current commercially available concordancers "have never been the object of

[40] Grounded Theory is an analytic framework for (qualitative) research developed by Anselm Strauss and Barney Glaser in the 1960s. According to this approach, theory is derived and emerges from a systematic analysis of data as opposed to theories that are already established but are not anchored in empirical research. In the early stages, the researcher does not generally have clear expectations and is expected to be flexible and open to multiple options while performing the analysis and building the theory.

scientific evaluation and publication" (Désilets, Farley *et al.* 2008). Moreover, further evidence of lack of attention towards concordancers can be found in the literature, according to which the "majority of concordancers are not widely advertised to professional translators" (Quah 2006: 113).

More recently, concordancers have been getting more exposure both in academia and among translation professionals, which makes their absence in studies on CAT tools all the more striking. The same can be said for specialized translation magazines and resources offering reviews of translation and localization software programs and/or new releases, but not even mentioning concordancers. This lack of information is made up for by blogs by users, professional translators and service owners presenting new products and services, carrying out their own comparative analyses and reviewing CAT tools and their features, which sometimes specifically target the concordancing function (Schiaffino 2011; Lossner 2012). Academic contributions present (bilingual) concordancers as a translation aid but they either only consider standalone corpus-based types of tools, like the ones described in Section 3.1.1 (Bowker 2002: 55ff.) or refer to the integrated concordancing function of TM systems (O'Brien *et al.* 2010). Furthermore, a labeling issue emerges as to how to refer to these resources, suggesting that awareness about these tools and resources is sometimes still low. The available concordancers are sometimes presented and referred to by their own developers or by users writing blog posts about them as "translation search engines", "terminology search tools" and "translation memory databases" (Gough 2012), adding to the confusion as to what these tools are and do.

3.2 CONCORDANCERS IN THE TRANSLATION INDUSTRY

On the basis of a more straightforward — albeit, due to Web-based TM systems, no longer water-tight — distinction, concordancers can be grouped into two main categories: "off-line concordancers", i.e. those integrated in the local TM system (or the traditional software for corpus studies) installed on the user's machine and "online (standalone) concordancing tools", which that give users access to data stored 'in the cloud'[41]. The next sub-sections will present each type of translation aid in greater detail.

3.2.1 OFF-LINE CONCORDANCERS

Traditional Translation Memory systems are installed locally on the user's machine and so are the databases they manage, i.e. Translation Memories. This type of system can be considered off-line in that there is no interaction with other users and only locally stored resources are used. The main difference between the concordancing function and the TM systems matching algorithm lies in the role of the translator. In a concordance search, the user is fully in control of the search operation, while in the other scenario the translator might be closer to an editor and/or evaluator of an existing translation automatically chosen by the system. However minor, this difference plays an important role in the way translators may relate to and perceive a tool because translators "appreciate being in control when using the TMs" and flexibility was found to be a requirement in HCI (Karamanis *et al.* 2011: 49). In a (manual) sub-segment match, the system displays all TUs containing the requested chunk and serves as "an extension of the translator's own memory" (Benito 2009).

[41] The concept of "the cloud" derives from "cloud computing", i.e. a way of providing services using a number of centralized resources stored on servers and accessed via the Internet.

When Translation Memory systems are reviewed and/or compared, not much attention is given to the concordancing feature. When the results of the well-known Translation Memory Survey (Lagoudaki 2006) were presented, hardly any mention of concordancing features or tools was made. In a retrospective overview of the history and evolution of TM systems, Garcia (2012: 453) points out that

> [o]utside th[e] area of technical translation/localisation, TM was of not much use for the individual freelancer. Except for updates, sentence repetition or similarity in natural language is scarce and *the usefulness of the memory database*, which came empty on purchase of the tool, *was small beyond concordancing* (emphasis added).

The potential usefulness of sub-segment matches has been emphasized for some time (Simard & Langlais 2000) and eventually TM systems have enabled it. Melby (2006) has suggested automatic sub-segment lookup, with additional language-specific morphological processing in the case of inflected languages[42], and display of matching target language units according to their likely relevance. Target text recombination could furthermore benefit from predictive typing mechanisms (Benito 2009) such as the one used in some CAT prototypes (Kohen & Haddow 2009). Relevance of results can also be measured using co-text (Melby 2006), as Simard and Langlais (2001) had done a few years earlier. They considered the TM as a large parallel corpus and computed the most likely target sub-segment on the basis of the context of the retrieved matching TUs. However, this approach proved to be computationally expensive, it suffered from inconsistencies in the translation of the same phrase and was not likely to perform well in detecting the correct boundaries in the translation (Benito 2009). The issue of "granularity" affects both TM and EBMT (Example-based MT) because the longer the units the lower the chance of finding an exact match, but at the same time a short unit increases the probability of ambiguity, i.e. multiple and conflicting matches, and quality inevitably suffers (Somers & Fernandez Diaz 2004: 9).

According to Melby (2006), there are some challenges in (manual) sub-segment matching that may speak in favor of automatic sub-segment lookup, namely (i) the potential overwhelming number of hits to evaluate and (ii) the possibility of not finding the chunk in the database thus wasting time in an unfruitful search. Not too long ago, SDL Trados (v. 6.5) added an auto-concordance function that would automatically look for sub-sentence matches of the segment in case no match was found and was later improved to the 'Auto-Suggest' function in SDL Trados Studio 2009 offering predictive suggestions to the typing translator (O'Brien *et al.* 2010: 187-8). However, this was far from being the first example of such a feature. The concept of sub-segment matches had already been defined by Bowker as falling "partway between fuzzy and term matching" (2002: 103) and in 2004 was considered "very desirable in TMSs [TM systems] but so far has only been a promise, with one notable exception" (Somers & Fernandez Diaz 2004: 16). This exception was the commercial TM system Déjà Vu, which in its Déjà Vu X release introduced the functionality 'assemble from portions'. Using EBMT technology, this functionality was able to self-repair fuzzy matches and replace them automatically with exact matches, provided the right conditions were met (2004: 17). This feature is still researched by TM developers (e.g. 'Advanced Leveraging' in Multicorpora, 'Auto-Suggest' in SDL Trados and 'Deep-Miner' in Déjà Vu X2), who use statistical algorithms and example-based machine translation for sub-segmental, phrase-level matching, "which is said to be the level at which repetition happens most often" (Garcia 2012: 456). Even professional translators

[42] A well-known limitation of many TM systems is their inability to deal with inflection and derivation (Macklovitch & Russell 2000; Bowker 2002: 106).

agree that "it had long been obvious that below the sentence is where the true linguistic treasure of translation memories was buried" and today, having moved past to complete segment matching,

> most major tools have found some sort of subsegmenting technology that provides a way to unearth automatically the real value that translation memories hold and to give us that material at our fingertips (Zetzsche 2012: 31).

The "usability" of sub-sentential segmentation has been investigated by Colomnias (2008 in O'Brien *et al.* 2010: 188) by means of traditional IR measures of recall and precision, though the analysis was carried out without consulting translators. To fill in this gap, O'Brien and colleagues set up an eye-tracking experiment to measure the usefulness of sub-segment matching and evaluate the design of the user interface. They looked at the number and duration of fixations on the Concordancing window as an Area of Interest (AOI) and measured productivity using average task time length with and without the concordancing feature enabled. Results were then triangulated with the outcome of a quality check and an opinion survey completed by the six subjects, finally suggesting that translators found concordancing useful and "even favoured [the information provided via the Concordance] over the longer segments displayed in the TM window" (2010: 189). Translators used the concordancer "primarily as a terminology and context-checker, as opposed to a productivity enhancing tool" and they also did "not wish to have [the Concordance window] turned on constantly" (2010: 187, 190). Moreover, most translators reported using the concordancer daily and closing it as soon as the search was completed.

A broader study on translation support tools was carried out in two Language Service Providers (LSPs) via Contextual Inquiry (Karamanis *et al.* 2011). The main tools translators used were Translation Memory systems and they generally included a concordance function to manually search the TM for a specific chunk of text:

> Translators were observed searching the Concordance frequently, mostly for sub-parts of the segment they were working on such as single word or a short phrase (i.e. a term) (Karamanis *et al.* 2011: 40).

However, using the concordancer did not necessarily imply that the translator was looking up an unknown item: as stated by an interviewee, "'[i]n most cases the translator is not really stuck as in they don't have a clue about what a term means'" (2011: 40). Another result was the occasional necessity to consult with other team members to ensure consistency. Researchers report about an incident where the translator needed to identify the most appropriate translation for a particular segment among the several versions available, i.e. a consistency check. In this particular context, the Concordance interface also provided information about the author of the translation in the metadata of the segment. Finally, translators were found to look up a phrase or a sub-part of the original sentence and in some cases a sub-segment search was performed even though the TM system provided the translator with a full sentence match (Karamanis *et al.* 2010).

In his doctoral dissertation, Simard (2003 in Somers & Fernandez Diaz 2004: 10) looked at the arbitrary sequences real users submitted to a bilingual concordancer and found that such 'chunks' were syntactically well formed. He also implemented a system to select the most useful results from the hits produced, which otherwise were too numerous, and found that his implementation was 15-30 times more effective than a sentence-based system.

One of the main issues with sub-segment matching is related to the pricing of a translation job (O'Brien *et al*. 2010: 187) because the more matches are found in a TM, the

lower the rate for the translator and further discounts will possibly be demanded for sub-segment matches. The likelihood of TM matches increases with the size of the TM. The current trend is towards integration and sometimes even replacement of locally stored resources with cloud-based ones, such as corpora, TMs or term bases in a translation environment that increasingly integrates more resources (both on- and off-line).

3.2.2 ONLINE CONCORDANCERS

Online standalone concordancers are less widely known in academia than among professional translators. Sometimes they are presented as "massive bilingual databases", "translation search engines", "terminology search tools", "translation memory databases" or "corpus-based tools" (Gough 2012). In fact, they are all Web-based multilingual concordancing tools which allow users to manually search large repositories of Translation Memories stored in the cloud. The fist attempts to build these massive databases date back to 2005 but in recent years, their number has noticeably increased, one reason being that "corporations now seem to be realizing that they have more to gain by pooling linguistic resources together" (Garcia 2009: 29) and understandably each company attempts to offer some specific functionality and diversify the features provided. These tools differ from traditional corpus analysis tools but also from standard TM systems, the main reason being the nature of the underlying translation repository. The repositories were first created to train statistical machine translation engines and were later made accessible to translators via a Web interface, a dedicated desktop application or directly from the TM system. In this last case, a plugin is installed in the TM system and the repository is queried as a regular TM. However, the likelihood of exact or fuzzy matches is not great and the user can try and use manual concordancing to obtain "interesting hits" (Garcia 2012: 455).

The next sections will provide a systematic review of the most renowned standalone online concordancers among the several tools that have been made available in recent years by presenting their interfaces and main features. Differently labeled tools are still grouped under the category of "concordance" because of the similarity in their use.

3.2.2.1 TAUS SEARCH

TAUS[43] (Translation Automation User Society) was formed in 2004 as a forum to exchange ideas and share experiences in the translation industry. In particular, the TAUS Data[44] promotes language data sharing among industries, governments and non-governmental organizations as well as research centers. Free and open access is provided to a whole language data repository through a search interface (TAUS Search) that receives over half million searches per month. Additional services are offered such as download/upload of words and Translation Memories. According to the reciprocity principle, different ratios for uploads and downloads of data are given to members and non-members and credits vary on the basis of the amount of data provided. About 40 languages are currently covered, with some additional locale-specific sub-grouping for a total of 2,135 language pairs and over 54 billion words. The interface of the TAUS Search service is shown in Figure 1 with the advanced search options activated.

[43] http://www.translationautomation.com/ [last accessed: December 2012].
[44] http://www.tausdata.org/index.php/home [last accessed: December 2012].

Figure 1. TAUS Search interface with (language-dependent) advanced filters activated.

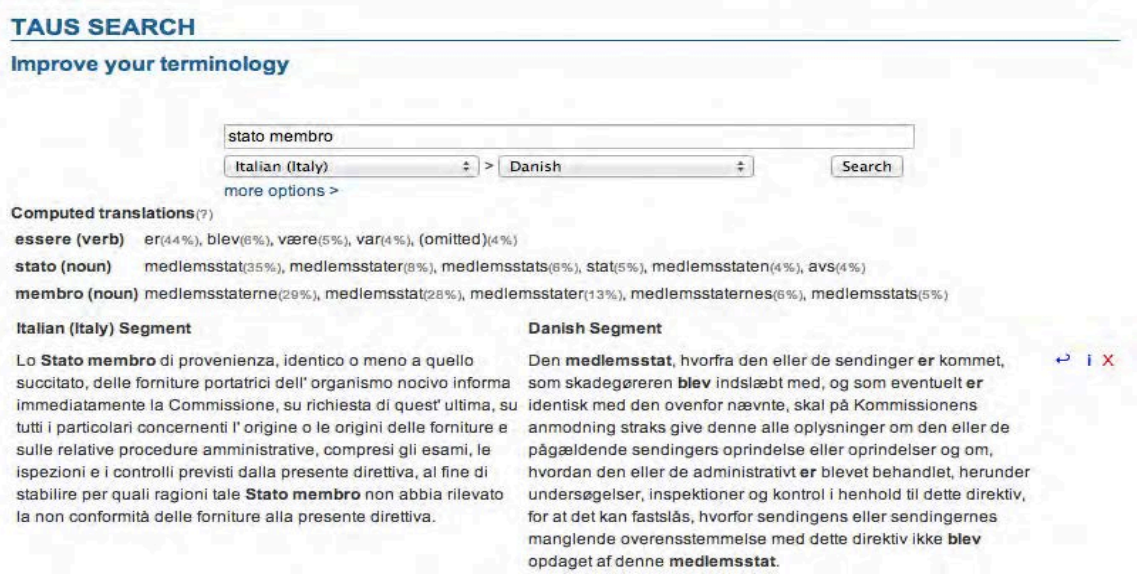
As can be seen from the title, the focus is on terminology (and phrases). The *Industry* filter enables users to select an industry domain e.g. automotive-manufacturing, legal services, computer software/hardware, telecommunications and energy, water and utilities. The *Owner* field contains a list of contributing users and companies to be used as filter, whereas the *Content type* distinguishes between instructions for use, support content, software/strings/documentation and sales/marketing materials, among others. Finally, the last two fields relate to linguistic filters and dis-/activate lemmatization and/or stemming and allow users to refine the search by choosing a specific part of speech among noun, verb or adjective. However, no customization features seem available. In addition to the Web page, a Search Widget is available for download to access the repository directly from the desktop. In September 2012, TAUS launched a common Translation Services API[45] to promote interoperability and standardization in the translation industry.

These features clearly suggest that this resource is mainly industry-oriented and indeed TAUS Search is aimed not only at translators, but also developers, support professionals and anyone who deals with industry-specific terminology. A general Web user with different needs and expectations may not find this resource particularly useful.

The system provides results in a two-column format with additional information on the side, i.e. segment metadata, a feedback button and the directionality specification as to whether or not the segment pair is direct, reverse or else a 'matrix' translation, i.e. a translation that is obtained by aligning two different pairs of segments using a common language (e.g. CS>DA obtained by linking the pairs CS>EN, EN>DA). Whenever possible, the system also provides computed translations for the entered words generated by a statistical engine (Figure 2).

[45] API stands for Application Programming Interface and is a protocol used as an interface to enable direct communication between software components and ensure interoperability between systems.

Figure 2. Results display in the TAUS Search Web interface.

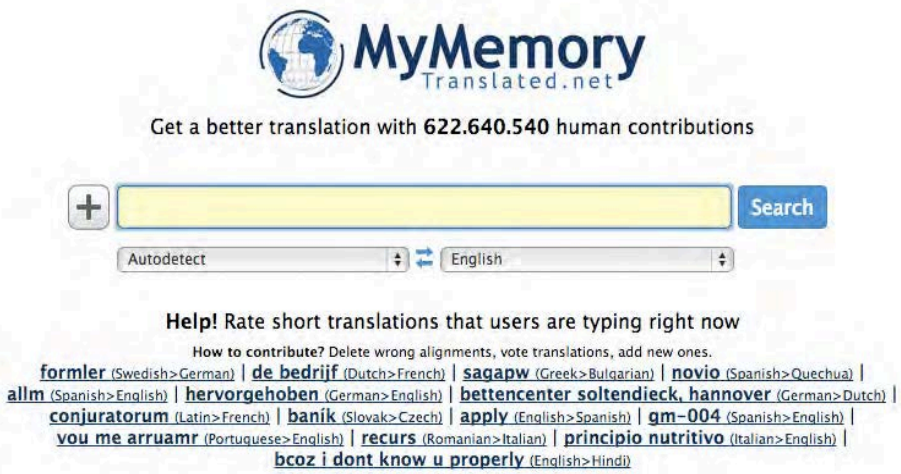


3.2.2.2 MYMEMORY

MyMemory[46] is presented as "the world's largest Translation Memory" and currently contains over 622 million TM segments covering over 150 languages. The TM repository has been populated with TMs from the European Union, United Nations, domain specific multilingual websites and is constantly updated thanks to the contributions of professional translators, Language Service Providers, customers and multilingual Web content. Memories can be downloaded for free in TMX format as well as uploaded or edited. A plug-in has been developed for use in conjunction with CAT tools, namely SDL Trados Studio 2009/2011 and MemoQ and a translation API has been available for a while.

The search interface is very simple and is offered in 9 different languages. The source language can be manually selected or automatically identified and a list of the latest searches offers quick links to submitted searches (Figure 3).

Figure 3. MyMemory search interface.



[46] http://mymemory.translated.net/ [last accessed: December 2012].

Results are displayed in a two-column format and an additional filter allows users to refine the search by subject (e.g. marketing, music, pharmaceuticals, aerospace, banking) (Figure 4). If the source text is not exactly matched in the database, a machine-translated version is shown. For existing matches, a reference is provided e.g. a user or a Website. At any time, a user can contribute to the TM by manually adding a translation, editing an existing translation or voting for the best translation.

Figure 4. MyMemory results page

Currently, MyMemory is employed in productivity studies with professional translators with a view to develop MateCat[47], a new open source Web-based CAT tool. Researchers are interested in measuring productivity gains while using a commercial CAT tool integrated with an MT engine and TM matches from MyMemory which the translator receives via the MyMemory plug-in (Federico *et al.* 2012).

3.2.2.3 LINGUEE

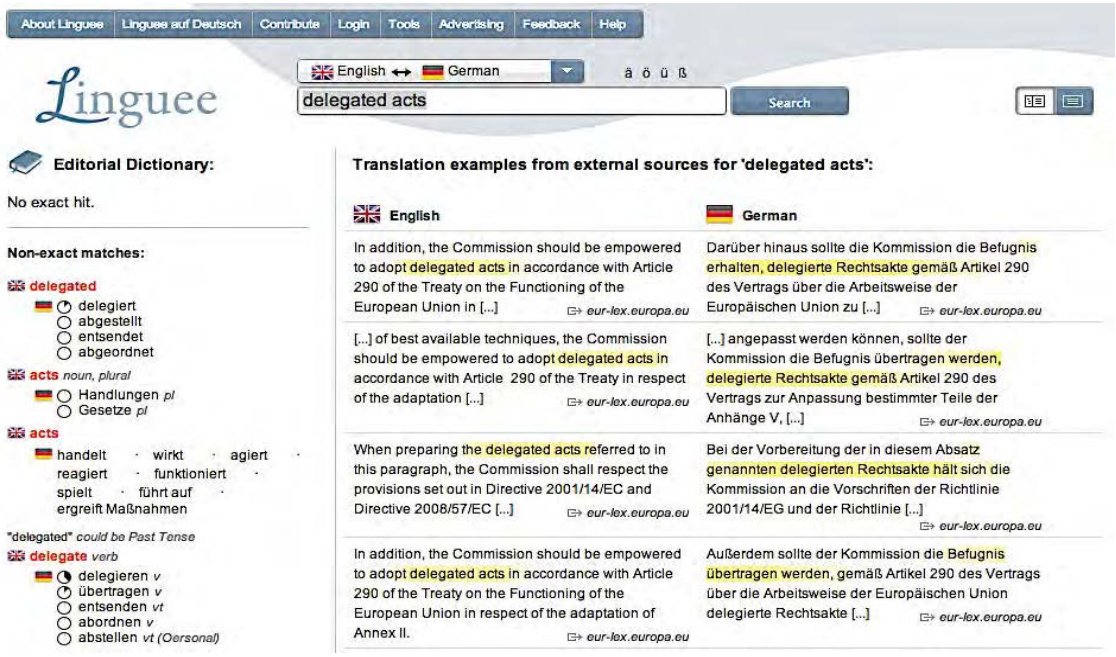
The Linguee[48] home page presents it as a "dictionary and translation search engine". Differently from the previous resources, it only covers a limited number of language combinations, i.e. translation between English and German, Spanish, French or Portuguese, as clearly shown in the main interface (Figure 5), though Chinese and Japanese are under development.

[47] http://www.matecat.com/ [last accessed: December 2012].
[48] http://www.linguee.com/ [last accessed: December 2012].

Figure 5. Linguee main search interface.

Results are displayed in two sections. Results from the editorial dictionary are displayed on the left whereas example sentences from other sources (e.g. bilingual texts) are shown on the right (Figure 6). Linguee is said to offer more advantages compared to online dictionaries with the examples providing more context and illustrating collocations or phrases. Great emphasis is placed on the dictionary component, which also provides information about the frequency of each translation. The dictionary component is manually checked and enhanced, whereas the bilingual examples are extracted via Web crawling (e.g. bilingual websites and reliable domains in addition to EU documents). When typing the query, the system suggests existing strings or previous searches and attempts to highlight the relevant portions both in the source and target language examples. When hovering with the mouse over the dictionary translations on the left hand side of the screen a relevant example is shown in context on the right hand side.

Figure 6. Linugee results page.



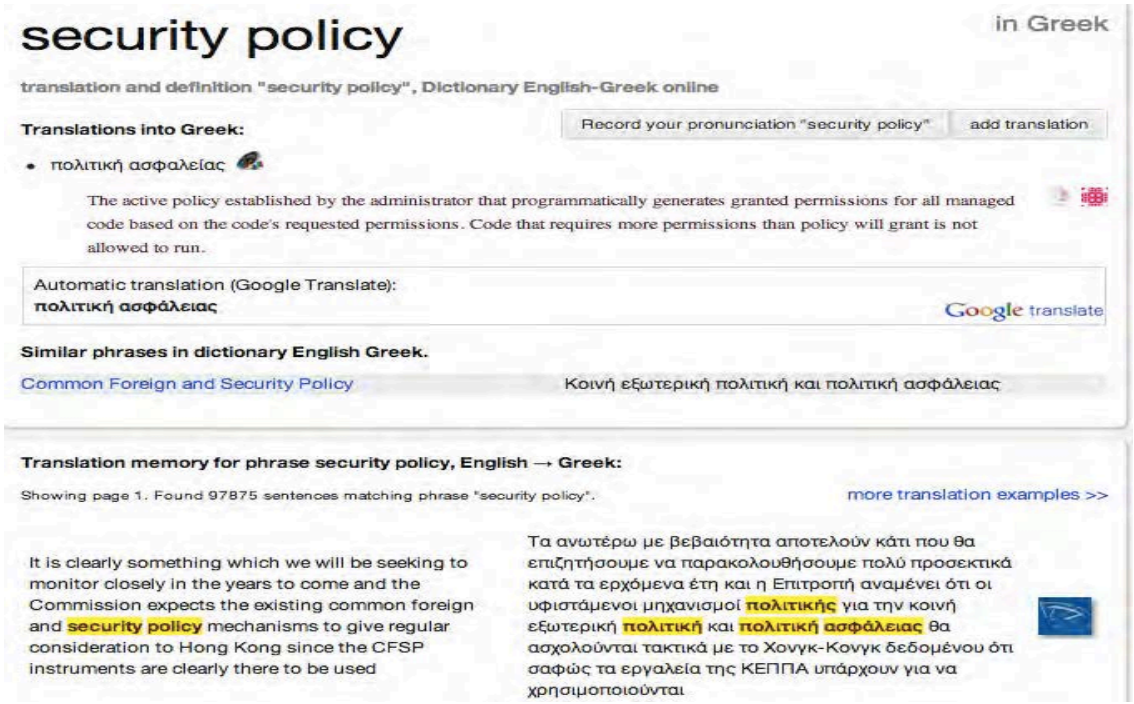
3.2.2.4 GLOSBE

Glosbe[49] is a "multilingual online dictionary" and provides free dictionaries "for almost every existing language" as well as a Translation Memory with over 1 billion sentences. Glosbe also promotes collaboration and resource sharing and offers a Glosbe API.
The interface (Figure 7) is very basic and is available in many different languages. The system stores the last few selected languages for easy access in case of multiple searches.

Figure 7. Main Glosbe search interface.

The left pane in the results page contains the alphabetical list of neighboring strings. The main results pane contains a definition of the searched expression and the target language version, with a machine translated version just below. The lower pane contains translation segments that contain further examples in context together with an indication of the source of the segment. Relevant text portions can be highlighted both in source and target segments (Figure 8).

Figure 8. Glosbe results page.



[49] http://glosbe.com/ [last accessed: December 2012].

3.2.2.5 TRADOOIT

TradooIT[50] is a bilingual concordancer currently covering three languages: English, French and Spanish. Its database contains over 260 million words and is described as "a computer-assisted translation suite" that comprises a Translation Memory, a term bank and a bilingual concordancer. The first two are built using public resources and the concordancer enables users to query them. Language professionals from all fields are the main target users of this service. The TM repository covers legislative texts, corporate documents, government and international organizations websites and movie subtitles, among others, while the term bank contains Wikipedia entries and well-known term banks. A lighter version of TradooIT (i.e. the bilingual concordancer only) was launched online in 2011 for the general public as well as language professionals but originally served a different purpose:

> TradooIT was created in a basement by a developer for his sweetheart, a freelance translator who found that current translation tools were poorly adapted to translators' needs. Among other things, she considered existing translation tools to be too slow, too invasive, too compartmentalized and too expensive (Okidoo 2012).

As far as the original version is concerned, TradooIT was a tailor-made tool for the needs of a professional translator and was later adopted and adapted by a translation agency and further customized.

The search interface closely resembles the Google search engine (Figure 9) whereas access to the Translation Memory requires user log-in. Both are offered only in English and French.

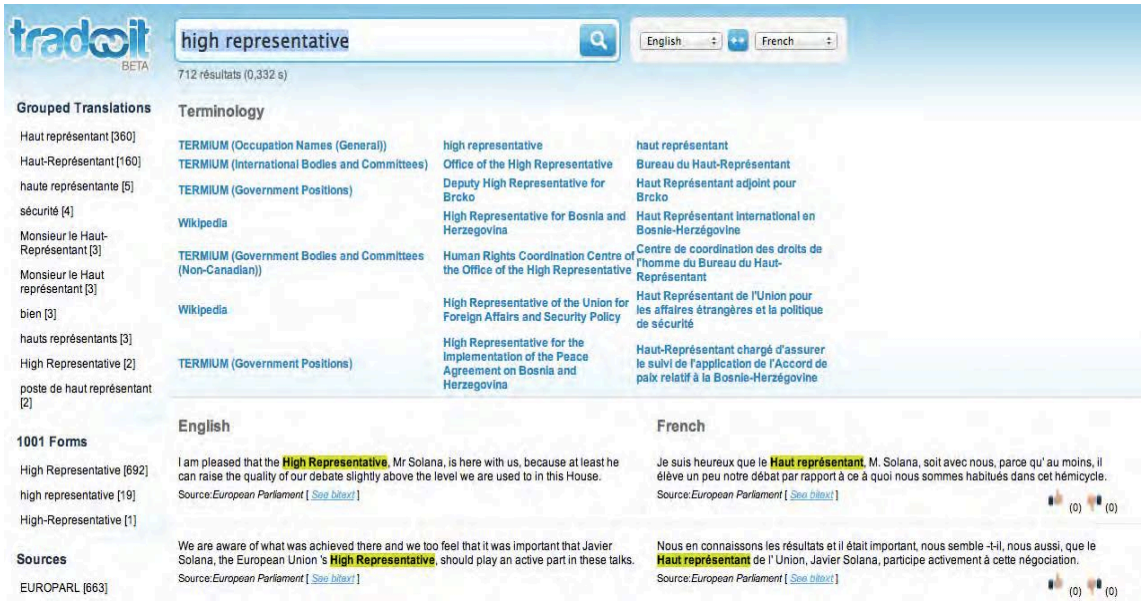
Figure 9. TradooIT search interface.

The concordancer displays results in table format with highlights both in the source and the target columns. Statistics on variants are also provided together with advanced filters to refine the search by e.g. form or source. Stemming is also available as well as variants for expressions. In the case of insufficient results, a smart feature prompts the translator for a different search that may return more results or presents him/her with a translation suggestion. The results page (Figure 10) is divided into three panes. The left pane contains possible translations together with frequency counts, all the variants in which the source strings appears and a list of the sources. The top pane contains the

[50] http://www.tradooit.com/ [last accessed: December 2012].

terminology entries paired with their source while the lower pane provides examples in context and highlights both the source and the target columns.

Figure 10. TradooIT results page.

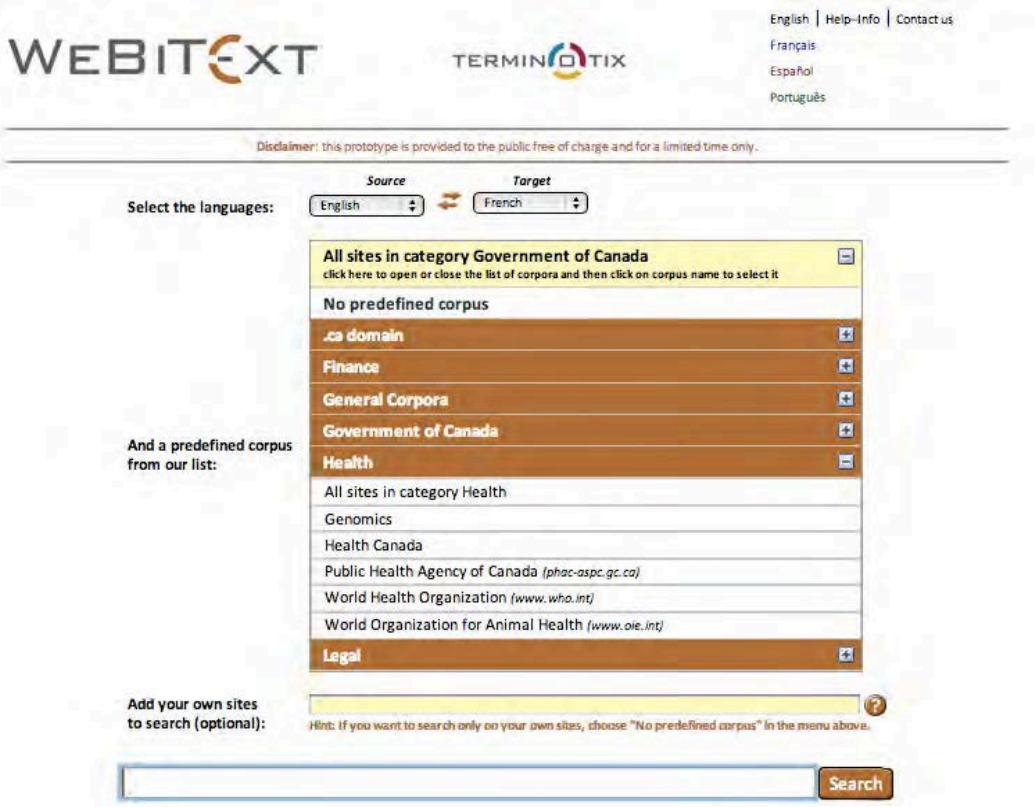


3.2.2.6 WEBITEXT

WeBiText[51] is a multilingual translation help tool that retrieves translations from bilingual corpora. The system is powered by Terminotix, a Canadian software company. The free version offered to the public is a prototype and is made available for a limited time only. The user can read the interface in four different languages. The WeBiText framework combines two search approaches into a hybrid system (Désilets *et al.* 2008: 5b), namely previously aligned content and online content aligned on the fly. Because of the different approach to searching taken by WeBiText, the search mask is slightly different from those of the previously described tools. There are 30 languages the user can choose from (including all 23 EU official languages) and two additional fields for corpus selection. The first corpus option covers the pre-processed available material, which includes a number of Web sites that were crawled and aligned in advance. The default option is "All sites in category Government of Canada" but additional resources are available in the predefined lists (Figure 11). Should the user prefer a specific Web site, the URL can be added in the specified box for on-demand search. In this latter case, the crawling is performed *ad hoc* but requires some extra time.

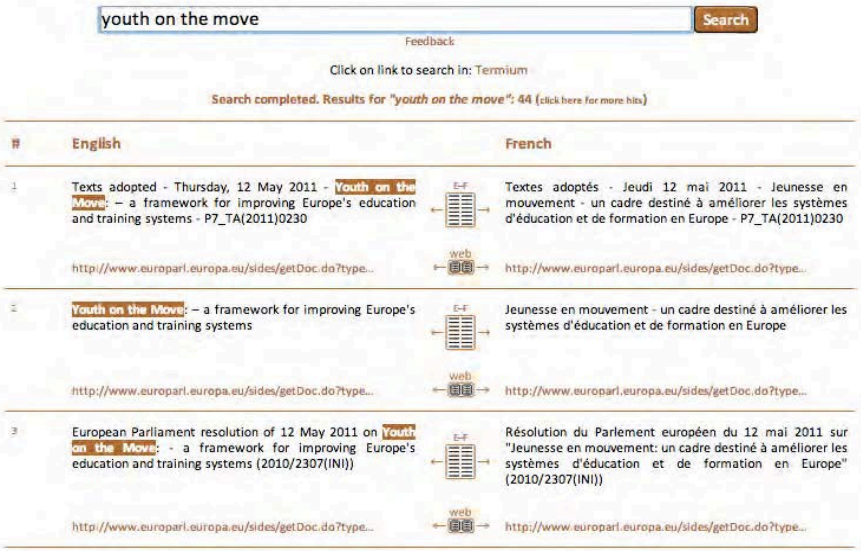
[51] http://webitext.ca/bin/webitext.cgi [last accessed: December 2012].

Figure 11. WeBiText Search interface with advanced search options activated.



Results are displayed in a two-column format with highlights on the source-text side while an additional link right below the search mask provides access to a term bank (Figure 12). In addition to the displayed results, the user can click on either icon in the middle to be taken to a new Web page. In one case, the full text is displayed in a similar format, i.e. segmented and aligned. In the other, the user is taken directly to the original Web site that is displayed in a split screen, each half containing one different language.

Figure 12. WeBiText results page.



WeBiText was already known in academia thanks to the research work by Désilets and colleagues (2008b) who investigated how a multi-purpose TM could be created from Web

mining and in particular "whether such a TM could add value over TMs built from other large, publicly available parallel corpora" (2008b: 1). English and French translators were interviewed and observed in their real working environment in Canada and a large TM built from crawling the Web pages of the Government of Canada was compared to a smaller general purpose TM built from the Canadian Hansard corpus. Results showed that the larger TM performed better in covering the observed translation problems, particularly those related to specialized terminology. At the time of the experiment, other languages were being tested using Web parallel corpora but results were less remarkable. Thanks to Web crawling, a TM can be populated much faster and become large enough to be useful for translators, but at the same time the data harvested are likely to be much noisier than those of existing corpora.

3.2.2.7 TRANSSEARCH

TransSearch[52] is a bilingual concordancer (EN to/from FR or ES), originally developed at the CITI, a Canadian research center, in the early 1990s[53] in cooperation with the University of Montreal and commercialized by Terminotix since 2003. Since then, it has been the object of a number of scientific studies and publications so much so that it can be considered one of the most researched commercial linguistic resources. In the past decade, research on the tool mainly focused on the analysis of the concordancer search logs in order to shed light on the human translation process, given that "users submit their queries in the natural course of their work, as they encounter translation difficulties" (Simard & Macklovitch 2005: 71).

To the public, TranSearch is presented as a database of past translations, a bilingual concordancer and a tool for professional translators, lexicographers, terminologists, linguists and (technical) writers who have to write in a language that is not their mother tongue. As of 2008, the system had about two thousand regular users, mostly Canadian translators submitting an average of 177,000 queries per month (between 2006 and 2007) (Macklovitch *et al.* 2008: 412).

Concordancing tools work according to some matching principles which are likely not known or understood by users. As a consequence,

> the way TransSearch interprets a query does not necessarily conform to a "naïve" user's intuition: while TransSearch looks for couples which satisfy a query, the user expects occurrences of a particular expression (Simard *et al.* 1993: 16).

This seems still to hold for some current concordancing tools because users, particularly non-translators, misjudge or misuse the systems due to differing expectations.

Originally, TransSearch was a bilingual concordance program that could extract concordance subsets from an underlying Translation Memory structure (TransBase) using a "fairly traditional search algorithm" and an abstract query language (Simard *et al.* 1993: 10). The TM repository contained two main databases, i.e. the Canadian Hansard with the parliamentary debates and the Canadian Court Rulings containing legal documents, but additional smaller databases were added over time to include translation

[52] http://www.tsrali.com/ soon to be discontinued in favor of http://tsrali3.com/Main.aspx?cc=true [last accessed: December 2012].

[53] For a detailed description of the original architecture of TransSearch, its underlying Translation Memory structure TransBase and any further technical details, see Simard *et al.* (1993) and Macklovitch *et al.* (2000).

memories in Spanish (i.e. the International Labour Organization database). As of 2008, the translation databases hosted 508 million words from debates from 1986 to 2007.

As the architecture of the system was designed to handle bilingual translations, multilingual translation could not be handled effectively. Results of a concordance search could be obtained in list mode, document mode or browse mode and both the source and target sides had highlights in the list mode display. In 1996, a Web-based version of TransSearch was opened to the public and in 2001 the tool was turned into a commercial service accessible to subscribers via the Internet. The bitexts were centrally stored on a server.

User activity in TransSearch has been logged since 1997 (Macklovitch *et al.* 2000: 1203). Based on the information retrieved from the logs, most searches come from Canada, France and Belgium but also from the UK and Russia, with educational and organization domains covering an even greater share. In 1999 researchers added a questionnaire to the Web site to elicit users' profiles and collect feedback. Results were discussed (Macklovitch *et al.* 2000: 1204) based on 119 voluntary responses about the users' profession and mother tongue, how they learned about the system, what they used it for, what they (dis)liked about it and the potential improvements they recommended. The most relevant findings for this study are summarized in Table 2. Apart form the questionnaire results, there did not seem to be any global statistics (demographics) about the users (Simard & Macklovitch 2005: 71), though technically these could be retrievable, particularly in the case of subscribers.

Table 2. Summary of relevant findings of the TransSearch user questionnaire, as found in Macklovitch et al. *(2000: 1204-5).*

Question/Topic	Answers	Percentage (n = 119)
User's Profession	Translator	51%
	Student	32%
	Linguists, Terminologists, Professional Writers	12%
Native Language	French	73%
	English	21%
Tool Usage	Find a translation solution or verify a translation[54]	75%
	Find monolingual information, such as collocations	10%
User Satisfaction	Very useful or indispensable	94%
	Adequate	8%
User Experience	Never consulted on-line help[55]	61%
Additional Domains Wished	Informatics, scientific & technical	45%
	Financial & economic	25%

If search results were logged in addition to queries, researchers could study both the chosen translation for a given query and the translators' strategies. However, this type of information raises even greater ethical issues than query logs alone as users were not personally consulted before their search data was analyzed (Simard & Macklovitch 2005: 76-77).

A number of additional studies have been carried out with TransSearch logs to investigate translation units and the nature of the text units with which translators work (Simard & Macklovitch 2005). In order to study the queries, researchers could only use the context

[54] For a finer-grained distinction between the two concepts, see Section 7.3.5.

[55] The authors believed that this way users were missing out on advanced features; as shown in the logs, 95% of the searches were considered "simple".

in which they appeared in the database, as the original source text was not accessible. Researchers first resubmitted a number of queries to TranSearch, automatically chunked the source-language sentences from the result page and compared each query with the newly obtained chunks. At the level of boundary matches, they checked whether or not the beginning (or end) of a query matched the beginning (or end) of a syntactic chunk in the source. The chunk ends matched in 85% of the instances whereas the beginnings were matched less often (55%), mostly in nominal chunks (NP) and less frequently with verb and prepositional phrases, which were the three main types of chunks identified. Modifiers, auxiliary verbs and determiners can all be easily left out in a search to shorten it or as a reflection of frequent Internet searching, while the lexical head on the right (i.e. chunk end) tends to be retained. To better track user behavior, researchers (Simard & Macklovitch 2005) developed the idea of an In-line TransSearch (ITS) functionality, an add-on that would integrate TransSearch into a word-processor and support complex queries such as 'bite+ the dust' (for inflection) and 'bite .. dust' (for ellipsis).

More recently, a more systematic analysis has been carried out on the log files collected over 6 years (Macklovitch *et al.* 2008) to study the main types of translation problems addressed with TransSearch. The most frequent searches for each length group were examined and a few "true terms[56]" were found. Contrary to researchers' expectations, almost no figurative expressions were identified. Instead, a high number of prepositional phrases including compound prepositions and complex prepositional groups was found together with predicate expressions followed by a preposition and single word queries, generally adverbials and adjectives (Macklovitch *et al.* 2008: 416-7).

Another area where much effort has been spent is that of "translation spotting", or "transpotting" i.e. "the task of identifying the word-tokens in a target-language (TL) translation that correspond to the word-tokens of a query in a source language (SL)" (Bourdaillet *et al.* 2009: 28). This has proven a very challenging task but was eventually implemented in the TS3 project. An enhanced version of TransSearch was developed to automatically highlight the corresponding translation to the query in the target versions[57] and display them in descending order of frequency after reducing nominal and verbal inflection and merging similar results. A "transpot" consists of "the target-word tokens automatically associated with a query in a given pair of units (sentences)" (Bourdaillet *et al.* 2009: 28) and frequency was found to be a good indicator of good translations with hapax items often corresponding to variants of the most frequent results (Bourdaillet *et al.* 2010: 253). Transpotting is actually one of the two main functionalities of the new TransSearch, i.e. bilingual concordancer and translation finder. As a bilingual concordancer, the focus is on the successful identification of the reference transpot in the target version. As a translation finder, the system will output a set of different translations for the same query (2010: 255-6).

3.2.2.8 OTHER CONCORDANCERS

Additional concordancing tools developed in academia are TotalRecall (Wu *et al.* 2003), a Web-based concordancer for English and Chinese mainly thought for second language learners, and Linear B (Callison-Burch *et al.* 2004) which is a tool meant to exploit

[56] "True terms" are here understood to mean "technical terminology". The small number of searches of this kind was partially justified by the limited number of domains covered by the TransSearch corpus as well as the availability of two large terminology banks in Canada, possibly better suited for specialized terminological searches.

[57] Further details on the transpotting algorithm and other technical details can be found in Huet *et al.* 2009a and 2009b and Bourdaillet *et al.* 2009 and Bourdaillet *et al.* 2010.

translation memories to build an MT system. As of 2010, Linear B covered Arabic, Chinese and seven European languages. Its interface is similar to a standard concordancer and so is the way it presents results. However, it also performs some additional processing of the query so that it produces the most likely translations of the source, ranked according to their probability.

In the present study, all listed stand-alone concordancers will collectively be referred to as concordancing tools irrespective of the specific label given to each tool by its developers, advertisers or users. Some labels could be misleading for the users because they sound restrictive (e.g. "terminology search tools"). In fact, these systems lack some of the traditional features of terminology tools, are memory-based rather than lexicon-based (Somers & Fernandez Diaz 2004: 10) and often assume a professional use by translators. To collectively refer to all these tools, a more general label such as "concordancers" seems more effective, at least until a sufficient number of empirical studies are systematically carried out about the actual use translators make of these tools, such as the questionnaire study conducted by Gough (2012).

Based on the previous definitions, most (if not all) tools share the same features of a traditional concordancer and for this reason they have been grouped and discussed together:

- They are all based on a large repository of aligned documents or TMs, i.e. types of aligned parallel corpora;
- Translators deliberately access this resource at their discretion, i.e. the tool does not come as an integral part to a workstation but rather can be added later if desired;
- Users traditionally submit concordance searches manually instead of being presented with an automatic translation;
- They are generally aimed at retrieving sub-segment matches, though in practice they are used for segments of various lengths.

3.2.3 INTRANET-BASED CONCORDANCERS

In the previous section, several examples of freely accessible Web-based concordancers were provided. There is however a third level at which concordancing tools can be found beside the off-line and Web-based types. They are concordancers found at the *intranet* level of businesses and particularly large corporations but also within international organizations such as the European Union. From a research-oriented perspective, the intranet offers many advantages as opposed to the Web, e.g. a finite number of potential users. Generally speaking, intranet solutions provide a more controlled environment which can be useful for empirical studies (see Section 5.1).

At EU level, one concordancing tool is available and used daily by EU translating staff in addition to a metasearch engine. They both provide an ideal test bed to conduct this exploratory study without incurring in the many challenges posed by regular tools on the Internet. In the next sub-sections, both the concordancing tool (Euramis) and the metasearch engine (Quest) will be presented in greater detail so as to outline the reference framework for the main study.

3.2.3.1 EURAMIS

EURAMIS stands for European Advanced Multilingual Information System. It was originally launched in 1995[58] at the European Commission and is currently available to other institutions, namely the Council, the Court of Auditors, the Court of Justice, the Committee of the Regions, the European Economic and Social Committee, the European Parliament, the Translation Centre for the Bodies of the European Union[59] and, since the end of 2012, the European Central Bank. Euramis consists of a series of centralized Web-based applications for document search and retrieval, including alignment, machine translation and a concordancing tool.

To access the system, users need to enter their log-in details so that correct access rights can be assigned, because each institution has a set number of customized resources so as to protect sensitive content and make storage and retrieval more efficient. Statistics can be collected about the total number of registered users per institution but numbers are only indicative (in fact they are much smaller) because they may include inactive users or people who accessed a service other than the concordancer (Navarre, personal communication). Table 3 summarizes the total number of registered users and the number of "active" users as of January 1st 2012, i.e. people who have used the application at least once during the previous six months. No further breakdown of user-related information will be carried out because user IDs have been removed in this study due to privacy reasons.

Table 3. Distribution of Euramis users across the EU institutions.

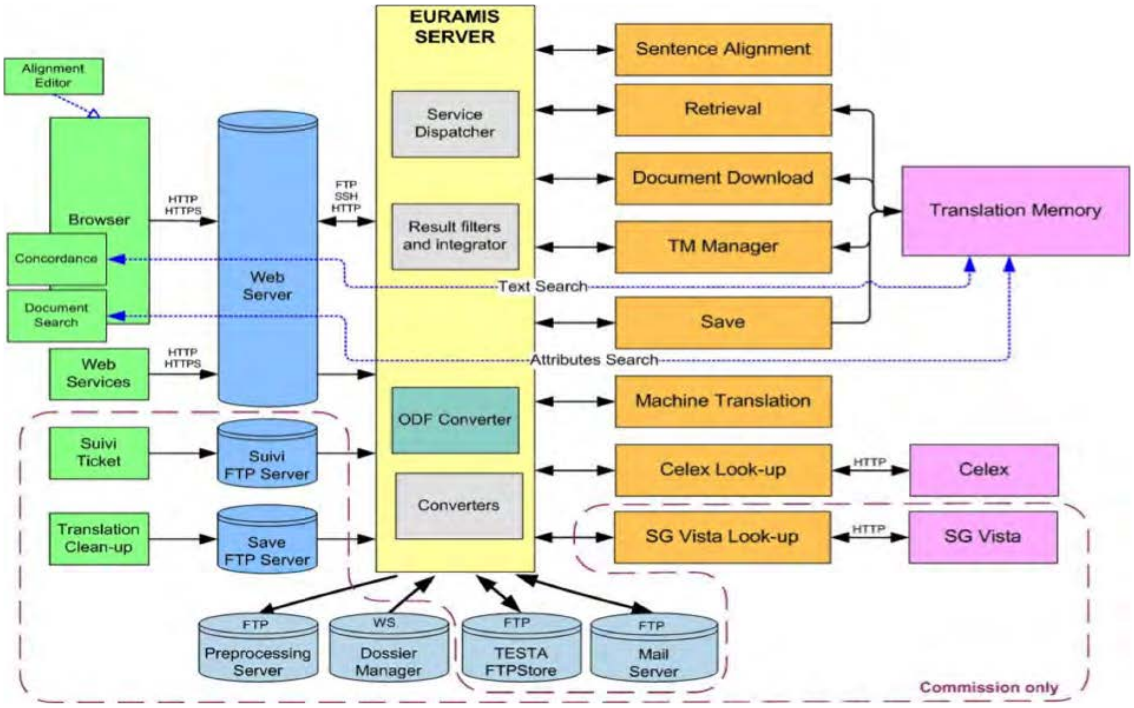
Institution	Registered Users	Active Users
European Commission (EC)	3493	2503
Committee of the Regions (COR)	181	86
Council (COUNCIL)	912	556
Court of Auditors (COA)	157	107
Court of Justice (COJ)	52	25
European Economic and Social Committee (EESC)	282	126
European Parliament (EP)	1296	857
Translation Centre (TC)	87	36

Every service in Euramis can be accessed from its tabbed interface. *Alignment* produces a bitext version of existing documents directly into the user's mailbox; *Translation Memory*, retrieves TMs from the central TM repository based on a given document; *Search* can lead to the *Document Search* page to retrieve a document using metadata; *Concordance* searches the Translation Memory on the fly; finally, another interface allows users to upload alignments to the central TM repository[60]. Euramis is generally referred to as a "central translation memory" (DGT 2009a: 10), "inter-institutional translation memory"

[58] For a brief history of the origins of Euramis until the late 1990s, see Leick (1998).
[59] About five years ago, there were talks for creating a public version of Euramis along the same lines of what of had been done for IATE but the costs were eventually found to be too high and the project was abandoned. A good deal of the basic material that Euramis uses is available in TM form via the Joint Research Centre (http://ipsc.jrc.ec.europa.eu/index.php?id=61, last accessed: December 2012). However no interface or tool to query it are provided.
[60] About 80% of the translations made by DGT are stored in Euramis but there are some confidentiality and file-format issues that prevent some documents from being uploaded. TM uploads can also be carried out automatically, as is the case for some institutions.

(DGT 2010c) and also "outil de traduction" together with the Translator's Workbench (Kowalska 2010). On other occasions, the Euramis concordancer is presented as a terminology search tool (Rusu 2009) with multilingual and multi-directional capability. Euramis at the EC also integrates machine translation for some language pairs. If the MT service is available for the chosen language pair, the translator is offered an automatic translation whenever the TM does not provide any suggestions.
Figure 13 represents the Euramis architecture. The Euramis Portal interfaces can be found at the very left of the chart, in the green section. The first six orange boxes list the available services discussed above, while the main TM repository is placed at the far right, in pink. The concordance application and consequently the retrieval service will be the main object of the present study[61]. The concordance interface will be presented in greater detail in the following sub-section.

Figure 13. Euramis architecture (DGT 2010c). Items can be found in the top half of the green, orange and pink sections are illustrated at length in the text.



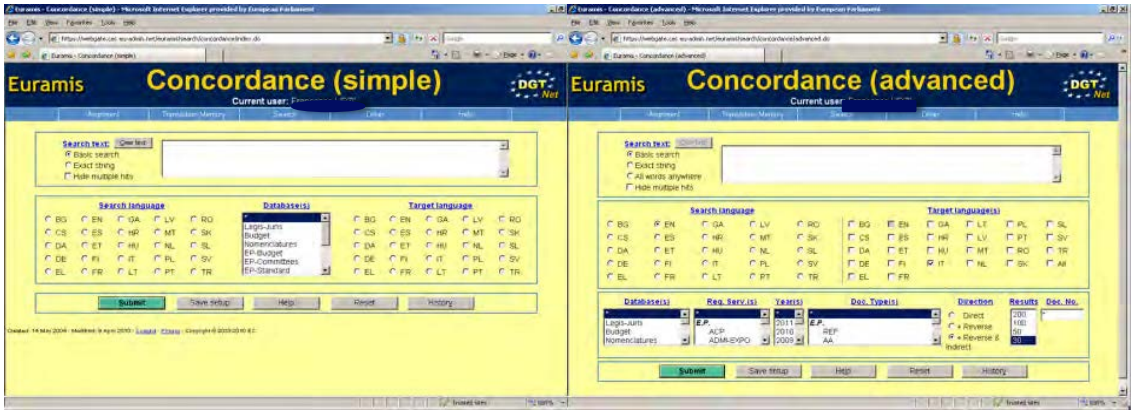
3.2.3.2 THE EURAMIS CONCORDANCER

The present analysis will mainly focus on the concordance application, which received about 30,000 queries per day at the time of data collection[62], and will consider both the interface and the retrieval service. The user can submit a query by opening the Euramis concordance page directly in the Web browser from the Search tab and typing or pasting the string. Alternatively, Euramis can be launched via Quest (see Sub-section 3.2.3.3) using a customized Word toolbar after highlighting the desired text portion.[63].

[61] A detailed description of the early Euramis algorithms for indexing and retrieval as well as the structure of the Translation Memory can be found in Blatt (1998a, 1998b).
[62] Currently, Euramis receives between 55,000 and 60,000 queries each working day.
[63] From direct observations of translators, Désilets *et al.* (2008: 3b) calculated that their subjects experienced approximately two problems for every 100 words they translated.

Two search interfaces are available: Simple and Advanced, as shown in Figure 14.

Figure 14. Euramis concordance Simple and Advanced interface as of 2012[64]*.*



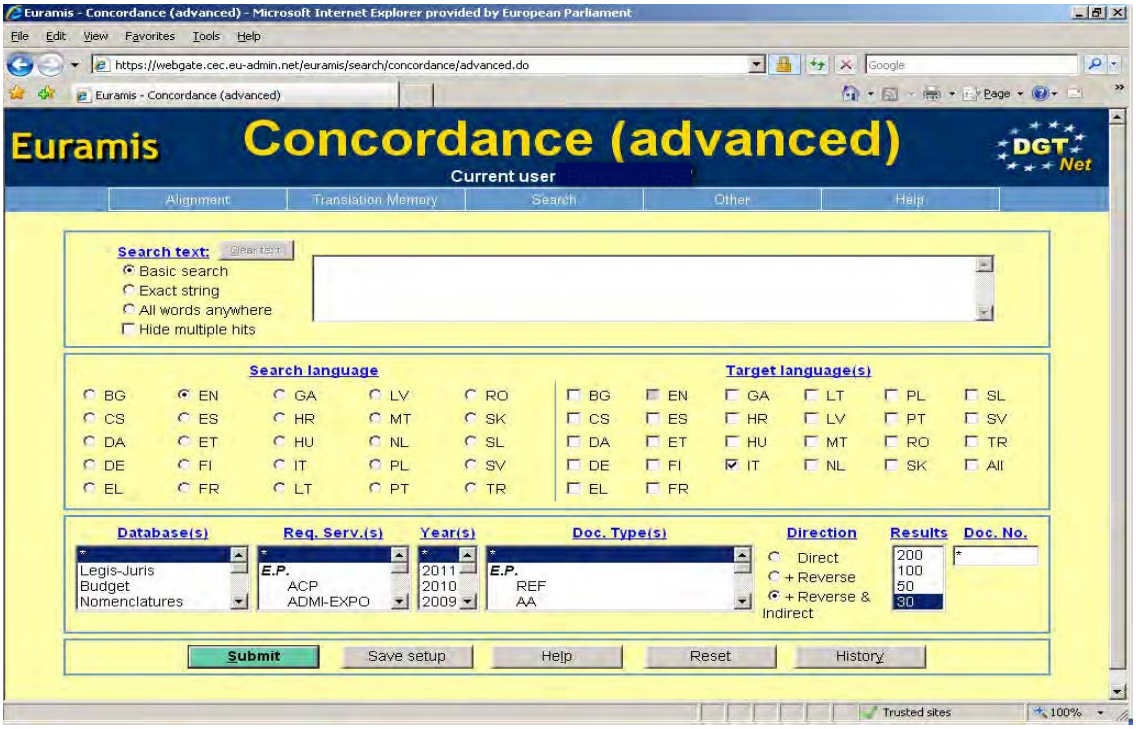
Starting from the top, there is a white box where users can enter the desired source text chunk, change the search mode by selecting one option in the *Search text* list and/or *Hide multiple hits*. *Basic Search* means that some grammatical words (the so-called "stop-words") in the string are ignored (in English and French only) while the other lexical items are searched (in the same order as presented); *Exact Search* means the string is searched verbatim while *All Words Anywhere* means that any segment that matches all these words found anywhere in the string is returned as a result. The central box in the Simple interface lists all source languages on the left, the available target languages on the right [65] and in the middle a list of available TMs in the middle, is shown. In the Advanced interface, multiple target languages can be selected in the same search episode while the database filter has been moved to the bottom part of the screen where a number of additional filters can be found: *Requesting DG, Year, Document Type, Document Number, Directionality* and *Maximum Number of Results*[66]. The lower bottom of the page contains some action buttons to submit the search, save or reset the current search setup, access the help features and go to the search history, i.e. a list of the most recent searches submitted to the system. This list contains some information regarding each search (some metadata and settings) and most importantly, the search strings. Each string is in fact a link that translators can click on to go back to results from a previous search with the same settings used. Figure 15 shows the Advanced interface in greater detail, which gives translators more scope for fine-tuning searches but takes considerably more time.

[64] The release of a new interface is foreseen for the middle of 2013. It will contain a single Web interface instead of a Simple and Advance one. Advanced settings will be shown or hidden, as desired by the user.

[65] Currently, the system supports 25 languages, i.e. 23 official EU languages plus HR (Croatian) and TR (Turkish), as can be seen in Figure 15. Chinese and Icelandic are planned. In general, the system can be customized to support any language as long as there is some/enough material to be uploaded.

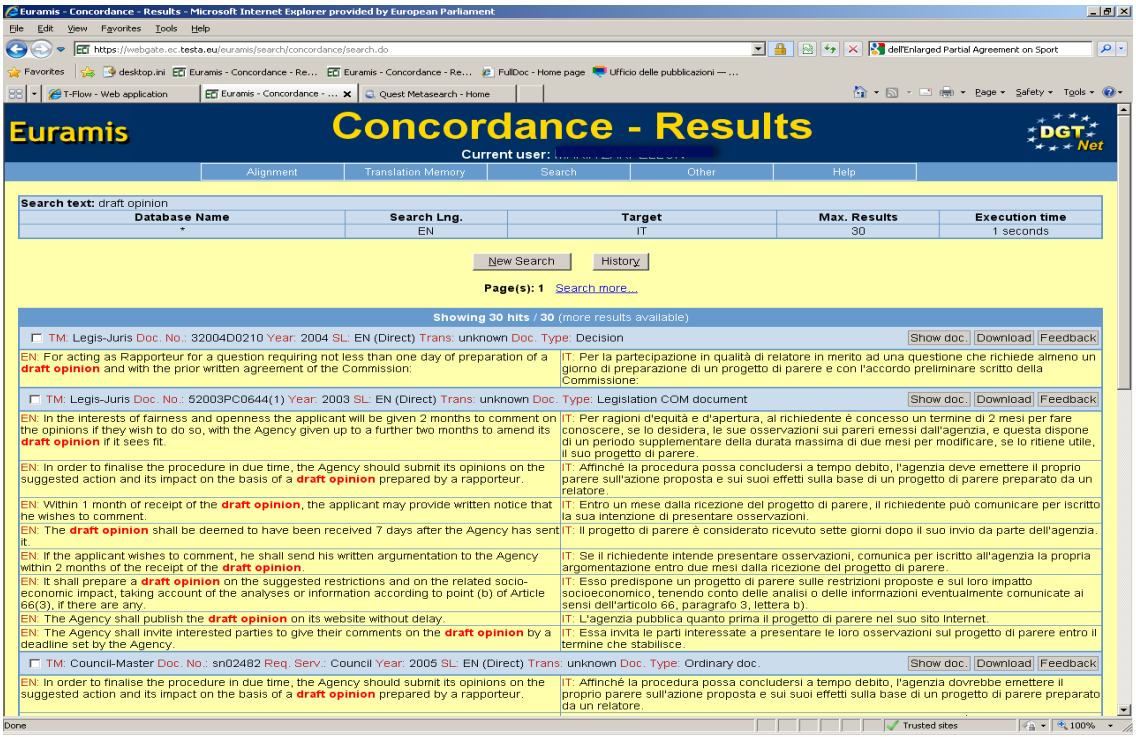
[66] The author of the translation (*Translator*) was originally shown as part of the segment metadata (Blatt 1998a: 86). After removing it for a while, translators would like this feature to be reintroduced.

Figure 15. A closer look at the Euramis Advanced interface.



Once a search is launched, the retrieved results are shown in the Results page (Figure 16).

Figure 16. Euramis Results page (identical same for Simple and Advanced mode).

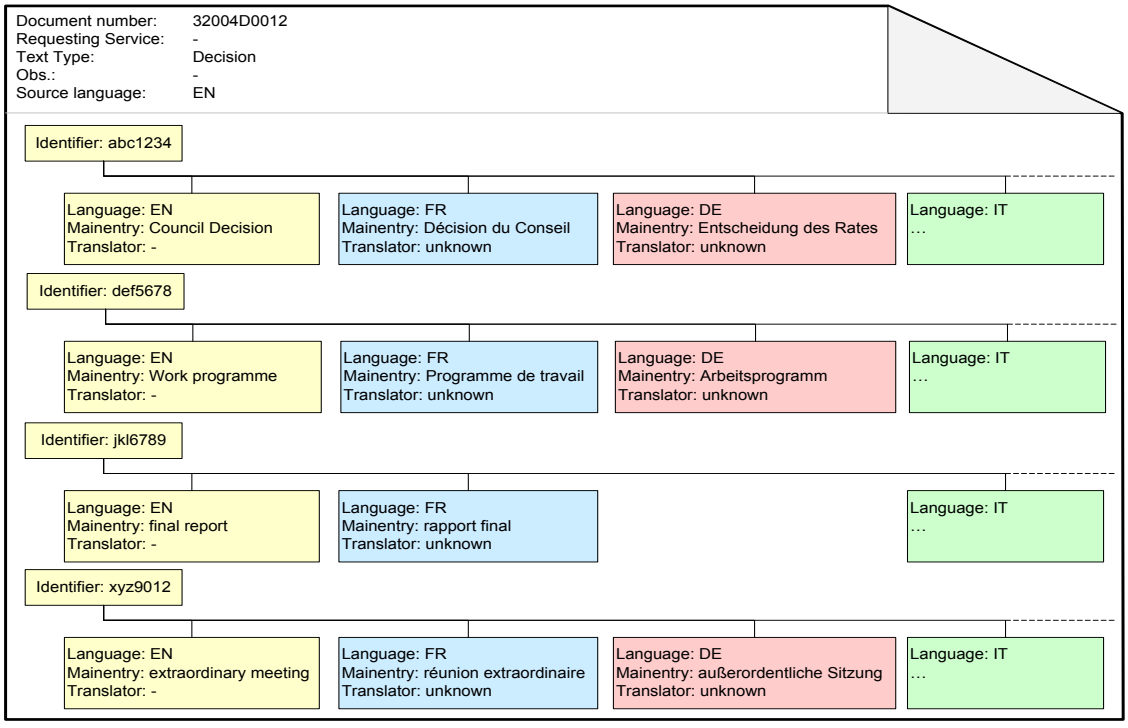


At the top of the Results page, there is a blue box containing a summary of the search setup and the search string. As previously seen with other concordancers, results are displayed in a two-column format, source text on the left, with red highlights, and target

text on the right. If the user is given the option to *Search More* (Figure 16), this means that the system has retrieved more results than the displayed ones.
Unlike previous tools, Euramis groups results from the same document under a common heading using the metadata of the document. In the heading, users can perform different operations, namely (i) open a new window to see the whole aligned bitext; (ii) download the document and (iii) send feedback in case of mistranslations or misalignments. If multiple target languages are selected, the multi-language target segments are clustered in the same target cell with the corresponding language code. The order in which results are displayed matches the order in which segment pairs have been uploaded in the database[67] and there may be instances where some sentences are missing from a document because the system does not allow duplicate source sentences within the same document. This seemingly strange behavior goes back to the very notion of "document" in Euramis, in that the system is designed as a centralized Translation Memory and not as a document repository or management system.

The underlying architecture is a SQL-based database and has a specific structure for storing text segments using the document metadata. The segment identifier is composed by a segment ID together with a language code that matches the language of the segment. Segment matching is performed not by comparing segments but by retrieving matching metadata between source and target sentences and using the language code to identify the corresponding translation(s) for the source segment. Figure 17 gives an example of the TM structure in Euramis.

Figure 17. Example of the structure of the Euramis Translation Memory (DGT 2010c).



The first segment is identified by the ID 'abc1234' and the Translation Unit (source + target) is obtained by matching the source segment ID with the corresponding target language code. This is probably the reason why some searches are unsuccessful, i.e. no

[67] In the future, a new feature, i.e. anti-chronological results display, will be added to the database software.

common ID can be found for all submitted words. For example, if a user searches for 'final report' (EN>FR), the system will match 'jkl6789 EN' with 'jkl6789 FR' but will fail if the language pair was EN>DE. The first step is to looks for segments that include the search words in the selected search language. There is no limitation to the amount of text that can be entered in the white search box. However, the system will only consider the first 230 characters of the searched-for string. Secondly, for each of the retrieved segments the system checks whether a translation in the target language(s) exists by browsing the segment IDs. As a consequence, a search that produces many results in the first step (e.g. common words) but has few entries in the target language (e.g. in case of an unusual language pair) takes much longer to be processed than a search that produces many results in the second step because in the latter case, the maximum number of results to be displayed is reached much faster.

In order to provide the history of the past searches, the system saves the queries in log files, which have now been collected for about six years. One log stores the following information about each search: date and time stamp, user ID, institution code, source language, target language(s), search mode, search interface, searched TMs, search method, execution time, number of results, requesting DG, year, document type, document number, directionality and maximum number of results. TransSearch logs contain similar information such as date, time, submitted query, number of hits produced, but also the IP address of the machine[68] (i.e. the source of the query), how results were displayed and details on the source of the query (Macklovitch *et al.* 2008: 413; Macklovitch *et al.* 2000: 1204; Simard & Macklovitch 2005: 70). While in TransSearch, users are identified using the IP of the machine whereas in Euramis login details are used. Before Euramis data could be analyzed for the purposes of the present study, user information had to be removed in order not to infringe on users' privacy. As a result, Euramis searches could only be studied according to the requesting institution because information about individual translators was lost.

3.2.3.3 QUEST

Quest is a meta-search engine developed in the early 2000s by the European Commission to speed up the search process and provide simultaneous lookups into several databases and online resources (both general and language-specific). A new inter-institutional version of Quest (Quest II) was released in 2009. Once again, the tool seems centered around terminological searches. When a user launches a search in Quest, the system forwards the query to each active resource as if a human user had accessed each resource separately and then consolidates results in one single Web interface while saving search logs (*History*) for future reference.

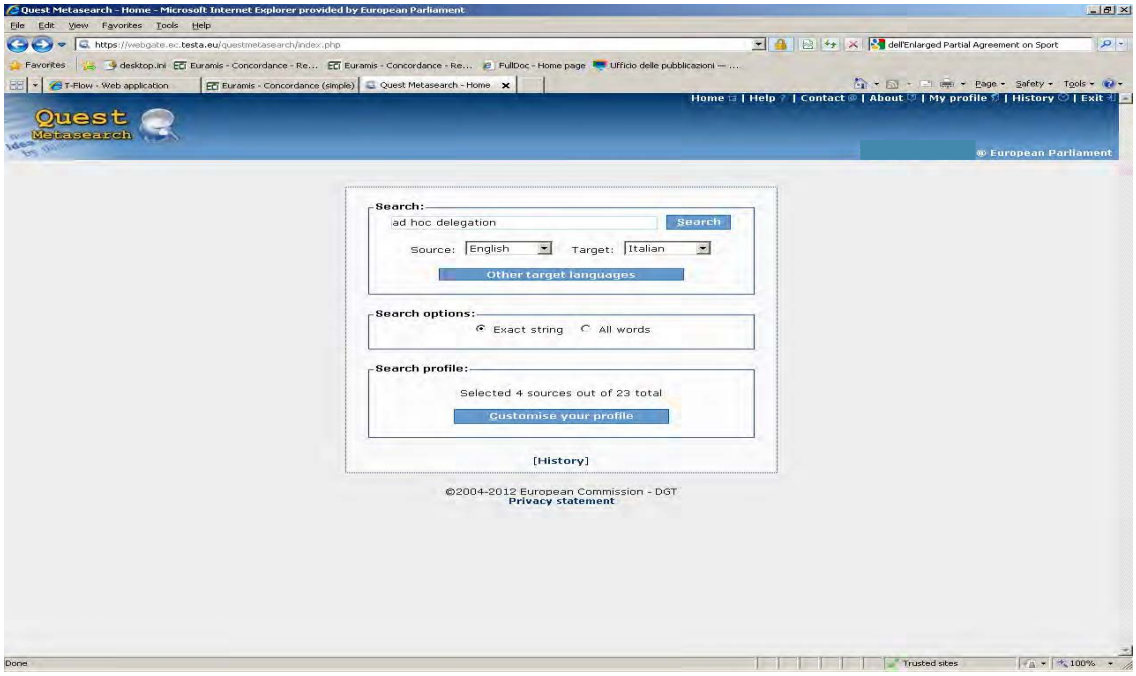
Similarly to Euramis, Quest can be accessed from a Web interface or directly from the Word toolbar. A Word macro can interact with the Trados Workbench and e.g. retrieve the language pair currently activated[69]. Compared to Euramis, Quest has a much simpler interface where users only select source and target languages and choose between two search modes: *Exact String* (default) and *All Words* (which corresponds to the mode *All*

[68] This information allows researches to keep track of the rough number of users (i.e. one IP is thought to represent one user) as well as to cluster potential users according to their organization and/or country.

[69] Until recently, EP trainees did not have access to the Commission's ECAS authentication system which is used to access Quest. This means that they were a group of users who did not have access to Quest at the time of data collection for the present study.

Words Anywhere in Euramis) (Figure 18). A customizable advanced search feature for Quest is foreseen but has not yet been implemented.

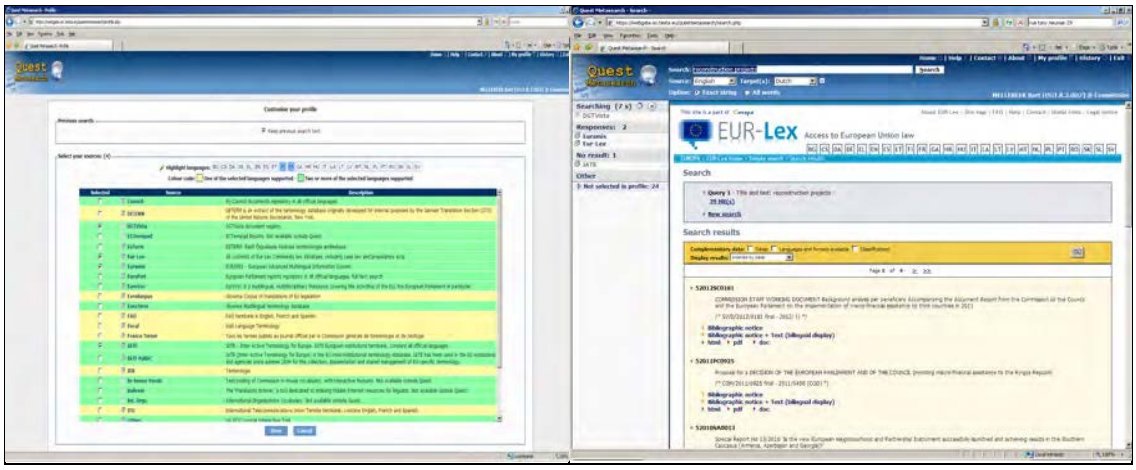
Figure 18. Quest main search interface.



The bottom box refers to the *Search profile*, i.e. the resources selected by that the user. Each user can select up to twelve different resources to be stored in the default user profile (Figure 19, left), though users are generally asked to select up to 4 or 5 resources. The number of resources available depends first of all on the institution but also on the language combination. The three main resources are the Euramis concordancer, IATE and Eur-Lex. IATE is the terminological database of the EU institutions, with multilingual versions of terms in many different domains. There are two versions of IATE: a public one which can be accessed from the IATE website[70] and an internal version that can only be reached via the EU intranet. Eurlex[71] (previously known as *Celex*) is a publicly accessible online repository where published EU legislation is stored. Whenever results are obtained from a resource, they are displayed in the Quest frame so that users can quickly switch resource or change search settings (Figure 19, right).

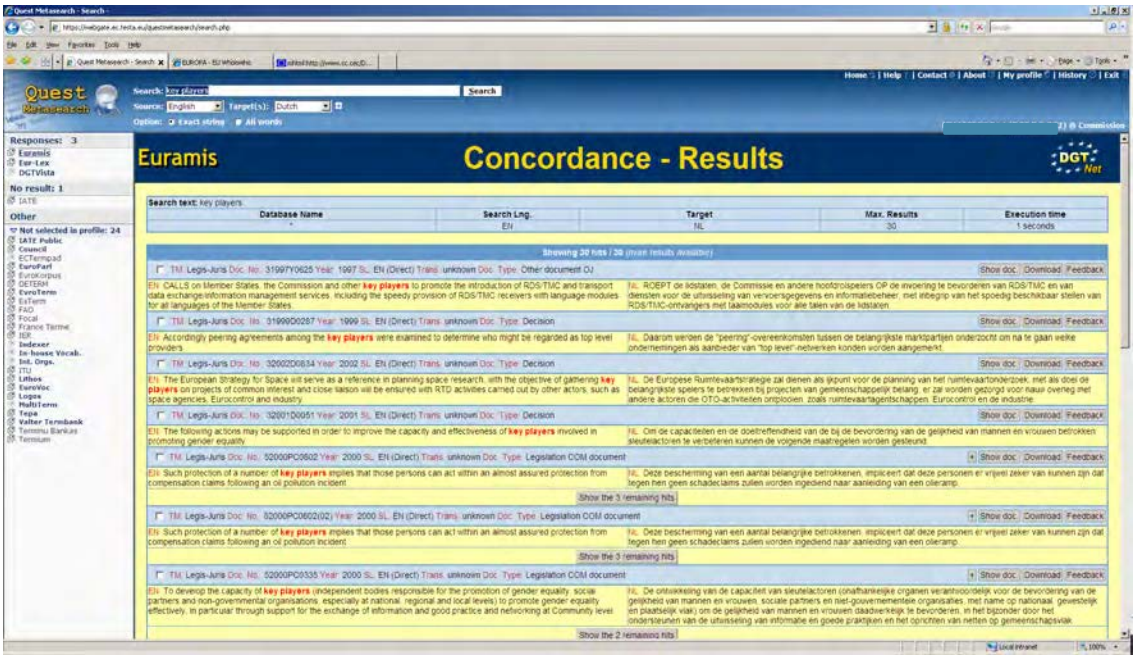
[70] http://iate.europa.eu/ [last accessed: July 2012].
[71] http://eur-lex.europa.eu/en/index.htm [last accessed: December 2012].

Figure 19. Two interfaces of Quest. On the left, the list of resources available. Items in green are available for the selected languages. On the right, the Quest result page showing Eur-Lex as the active resource.



The result page comes in a two-pane format (Figure 19 – right and Figure 20). The left pane lists all responsive resources at the top. The first resource to respond is automatically displayed as a nested webpage in the larger right-hand pane; Euramis is very often the first resource to respond. By clicking on the relevant link, the user can change the displayed resource.

Figure 20. Quest result page with the Euramis concordance as the active resource.



3.2.4 CONCORDANCE USERS' PROFILES

Whenever Euramis is selected as a resource (i.e. for most searches), Quest will act as a standard Euramis user and submit the search using the Euramis Simple interface and default settings. Because the number of Quest users is on the rise[72], the number of refined

[72] As of July 2012, over 70% of the Euramis searches were submitted by Quest vs. about 63% in September 2010.

searches is bound to decrease further because users seem to prefer pooled resources where they can compare results from different sources at once. Quest is possibly the main channel through which Euramis is accessed. All requests from Quest are logged under the same user in Euramis (i.e. 'Quest') but there are in fact virtually hundreds of translators submitting searches via Quest. Supposing that the Euramis user names could be retained in this study, Quest as user would prevent any further considerations on the *individual* behavior of translators because no further breakdown would be possible. The issues related to user anonymity (see Sub-section 3.2.3.1) is one of the reasons why the present analysis is more concerned with trends than individual preferences.

However, general user profiles could still be useful to obtain a more accurate idea of potential users of Quest and Euramis, and by extension, professional profiles for the detailed on-line resources presented earlier. Information about user profiles was collected during informal interviews with actual users of Euramis both at the EC and the EP and on the basis of the author's personal experience as a direct user of the Euramis concordancer at the EP.

According to the general understanding, Euramis is used to look up terms in context but also for document retrieval and download, as demonstrated by the fact that a number of retrieved documents were also downloaded. For professional translators, Euramis is a useful support at virtually any of the three stages of translation, i.e. orientation, drafting and revision (Jakobsen 2002: 191), though searches in the drafting phase will outnumber the others, as suggested by Lörscher (1996: 30), who found that professional translators often manifested difficulties, hence the need for some form of support, *while* they were rendering a unit into the TL, whereas foreign language students realized them *before* they started translating. Finding documents about the same topic using keyword searches was also mentioned as a type of support offered by the system. Very often, EU translators are required to quote existing legal texts rather than translate them from scratch (DGT 2010c) and this may account for a large number of searches for document retrieval purposes[73]. With this latter approach in mind, Euramis becomes more similar to a traditional Information Retrieval system.

Although Euramis is chiefly used by translators, not all translators use it consistently. The choice of the concordancer by EU translators varies on a case-by-case basis. Some users are not aware of all available tools while others find some tools not particularly convenient, despite the shortcuts, and prefer to access the resource directly from the webpage or use alternative internal resources, such as FullDoc at the EP (see Section 5.1.2). Some people tend to avoid the Euramis webpage in favor of a systematic use Quest by launching the searches directly from the text editor. However, the first choice is often the Trados concordancer followed by Quest and/or Euramis; if all fails, Internet-based resources such as Google or other concordancers such as Linguee are accessed because they also provide references to EU material. When evaluating retrieval, some users want to see as many results as possible but, at first, evaluate only the first dozen. If the search is unsuccessful, users tend to quickly turn to the Web (e.g. Google and Wikipedia). Using external resources such as Google and other concordancers is also helpful because they can provide the desired document in its final version, whereas internal resources may contain several versions of the document depending on the specific policy of each institution which may produce inconsistent results.

[73] Once a document is downloaded, it can be added to the local Translation Memory currently in use to increase leverage.

The list below summarizes the main scenarios for concordance searching by translators who are likely to access Euramis:

- when translating or revising;
- to find the exact translation of a given chunk in the existing legislation;
- to check unfamiliar or unknown expressions (e.g. because the source language is the translator's *n*th L2 and they feel less confident about the lexicon);
- to find the intended meaning and/or the most frequent translation when a source word has several potential translations;
- to check whether the term has been translated and how to find the established translation in a given context[74];
- to produce consistent translations with other parallel target versions;
- to double check word usage when a translation is not convincing;
- and for document retrieval purposes.

Three main translation aids have emerged from this overview: Translation Memory systems, the concordancers and the Web browsers. They all require the user to interact with the system and often to submit a query. Undeniably, there are links between the systems, as the same query could easily be submitted to all three tools. However, the direct interaction between the user and each system is likely to be different. The interaction with a TM system presupposes distinct actions from those required by a concordancer, let alone a Web browser. The question arises as to what extent user interaction differs across the tools and how much the habit of using one tool can affect the interaction with another. For example, Web searching has become a de facto standard in today's society and is not just used for translation or professional purposes (though in this case some specific search strategies may apply). It could be argued that users have become so accustomed to interacting with a search engine that they might reproduce this behavior with tools whose appearance closely resembles Web browsers (see the examples of concordance interfaces provided earlier). Before Web browsers, other forms of information retrieval existed, such as library catalogs, and even early Web browsers[75] worked according to a different interaction principle with the system. In addition, translation technologies such as TM and MT have forced "a sub-segment mode of processing, which potentially leads to cognitive friction" because translators seem naturally inclined to process larger textual units (O'Brien 2012: 118).

Computer literacy is also likely to play a role. Every new tool presupposes a learning curve which can be linked to the Law of Least Effort and traditional concepts in Human Computer Interactions such as human factors and (cognitive) ergonomics (O'Brien 2012: 103). For a number of reasons (e.g. time pressure) users are not willing to spend too much time learning about new (advanced) features which would explain the success of largely intuitive and easy-to-use tools. At the same time, one may wonder whether the underlying architecture of these systems is so different that each tool requires a distinct approach. In the light of these observations, a more detailed comparison between information retrieval and concordance searching will be carried out in Chapter 4.

[74] Sometimes translators may use the concordancer in reverse mode, i.e. they search their own translation in the other directionality to check the context of usage in the source language.
[75] For a detailed description of the functioning of early Web browsers, see Austermühl (2001: 52ff.).

3.3 KEY CONCEPTS

- Concordancers are traditionally used in academia and especially in corpus studies;
- Concordancers applied to translation memories and bilingual parallel resources originate from research in machine translation;
- There are several types of concordancers and some are widely used by professionals but they generally do not attract researchers' attention;
- In the translation industry, concordancers can be found at three levels: off-line and online; the latter branch into Web-based and intranet-based tools.
- Off-line concordancers are traditionally found in Translation Memory systems; Web-based concordancers are generally commercial standalone tools whose number has noticeably increased in the past few years; intranet-based concordancers can be found at corporate level or within international organizations, such as the EU;
- Special attention was devoted to TransSearch, an online bilingual concordancer which has been the object of a number of research studies on concordance searches and user interaction;
- Two intranet-based EU concordancers (Euramis and Quest) have been thoroughly analyzed as they are the source of the search logs studied in the present work;
- Concordance users profiles at the EU have been sketched and broader questions about user interaction with different types of language resources have been put forward.

CHAPTER 4: CONCORDANCE SEARCHES, TRANSLATION PROBLEMS AND INFORMATION NEEDS

In Chapter 2, different perspectives taken on the concept of translation problem were discussed and the use of reference books was mentioned as a primary indicator of a problem. Chapter 3 dealt at length with the features of one specific type of translation support tool, i.e. the concordancer, and special attention was devoted to the functioning of the concordancer used in the present analysis. In this chapter, the concordancer as a standalone resource will be added to the list of reference material available to translators and concordance searches will be considered as an additional type of user activity data (UAD) that could be successfully triangulated with other data types to gain further insights into the translation process.

4.1 INTERNAL AND EXTERNAL SUPPORT

When discussing translation problems (see Chapter 2), a number of problem-solving behaviors were identified, ranging from conscious translation strategies to the consultation of reference material. Problem-solving strategies can broadly fall into two main categories, linked to the use of internal or external support (Alves 1997). *Internal support* comprises the translator's world knowledge and in general his/her cognitive resources whereas *external support* refers to the use of documentation sources (e.g. dictionaries, reference material and online resources) available to the translator. A similar distinction is found in Ronowicz *et al.* (2005: 583), who claimed that external resources (particularly dictionaries) are used when neither internal mechanism (the Frequent Lexis Store[76] and the Lexical Search Mechanism) provides an answer to the problem.

The distribution of internal and external support in the process of translation has been studied empirically using translation protocols, key-logging and direct observation (e.g. PACTE 2005; Alves & Liparini Campos 2009a,b). In one study by the PACTE group (2005), a number of different sequences of actions were observed during translation that could be grouped in five main categories:

1. Simple internal support (IS);
2. Internal support dominant combined with external support (ISD);
3. Balanced interaction between internal and external support (IS-ES);
4. External support dominant combined with internal support (ESD);
5. Simple external support (ES).

The first category occurred when a "Definitive Solution" was found without any external consultation, whereas category five was assigned to those instances where a bilingual dictionary was used as external support. Categories two and four contained instances of "complex consultations", i.e. a chain of consultations or a complex type of search like a

[76] The Frequent Lexis Store is a concept developed by Bell (1991: 47) when describing the translation process and is defined as "an instant look-up facility for lexical items both 'words' and 'idioms'". It is a component of the brain's syntactic analyzer (Ronowicz *et al.* 2005: 581).

targeted Internet search. The dominant type of support was considered the one that eventually provided the solution. For the third category, three further subcategories were identified (generally involving the use of dictionaries): (i) Consultation to confirm a decision; (ii) Consultation using a dictionary but solution rejected; (iii) Internal support plus consultation and solution accepted (PACTE 2005: 615). Category distributions were then combined with the acceptability of the chosen solutions for the likely problematic items in the text (i.e. the "rich points"; see Section 2.2).

PACTE's classification was later reformulated to accommodate a different classification that distinguished between support for orientation and support for revision (Machado 2007 and Batista 2007, in Alves & Liparini Campos 2009a: 78-9). PACTE's five categories were first re-organized by removing the IS-ES type and renaming the remaining categories SIS, DIS, DES and SES[77], respectively. Because of the new added dimensions of orientation and revision, eight new categories[78] were listed according to the phase of translation at which each of the four forms of support can occur and their distribution was tested in a translation task with and without Translation Memory systems. The study by Alves and Liparini Campos (2009a) also tested translators' behavior during orientation and revision by applying Machado's and Batista's categories and recording translator's activity. A screen-recording software was used in conjunction with TM systems whereas Translog was used for the task without TM systems. Overall, internal support was more frequent both for orientation and revision with slight variations between simple and dominant categories according to the task type, one involving the use of TM systems, the other leaving them out. However interesting, such categorization cannot be applied "as is" to the present analysis because no information is available about the stage at which the concordance search was launched (orientation, drafting or revision), though intuition and existing evidence point to the drafting phase as the most likely.

Alves and Liparini Campos considered every (single) consultation of external resources and concluded that "[m]ost instances of external support involve web searches or dictionary look-ups to find translation alternatives for specific terms" (2009b: 203). This statement suggests that one of the main reasons for turning to external support was in fact not so much the need to understand a term or find out its meaning but rather to find "translation alternatives". Further evidence showed that the use of external support was not affected by the use of TM systems, though the nature of the search changed according to the test scenario. When no TM system was available, external support comprised Web searches for parallel texts and dictionary look-ups, otherwise the Trados Concordancer was the most frequently used type of external support (2009b: 204).

Overall, this group of studies is particularly interesting for the present analysis because these experiments explicitly consider external support both in isolation and in combination with internal support. Experiments in process research have at times allowed participants to use some kind of translation support, under different circumstances. Including support tools in an experiment was a deliberate choice on the part of the researcher. Early studies in the 1980s (e.g. Krings 1986a,b) only had the option

---

[77] Simple internal support (SIS); Internal support dominant combined with external support (DIS); External support dominant combined with internal support (DES); and Simple external support (SES).

[78] For the orientation phase: Simple internal support for orientation (SISO); Simple external support for orientation (SESO); Dominant internal support for orientation (DISO); and Dominant external support for orientation (DESO). For revision: Simple internal support for revision (SISR); Simple external support for revision (SESR); Dominant internal support for revision (DISR); and Dominant external support for revision (DESR).

of paper-based support, usually in the form of mono- and bilingual dictionaries or some other kind of printed reference work. Over time, forms of support have increased in number and types but researchers have in some cases decided to do without external support altogether so as to reduce the number of variables and focus on the purely cognitive task (Immonen 2006; Dragsted & Hansen 2008). Usually, experiments involving eye-tracking and key-logging tend not to allow dictionaries or other translation resources because of practical constraints during data elicitation. However, some empirical studies have specifically focused on the way translators interact with translation support in the form of paper-based resources (Künzli 2001; Livbjerg & Mees 2003) and Internet-based resources (Enríquez Raído 2011; PACTE 2011b). Some other studies have decided on having the Internet as sole external aid for translators during the experiments (Jakobsen 2003; Alves *et al.* 2010) whereas others allowed any support tool translators wished for (Buchweitz & Alves 2006). Ethnographic research has also contributed to shed light on the type of tools translators use, this time without experiments involved. Researchers have observed professional translators at work, taking notes of how they worked and what type of aids they were using (Désilets *et al.* 2009; Karamanis *et al.* 2010). No study, however, seems to have used an experimental design focused on the use of the concordancer and the only empirical evidence on the use of this resource is derived from verbal protocols (Buchweitz & Alves 2006; Karamanis *et al.* 2010).

Generally speaking, two main groups of external resources can be identified: *relational* and *non-relational* resources (Lu & Yuan 2011). Examples of relational resources involve a direct interaction with other people e.g. colleagues or experts, be it in the form of spoken or written communication, from direct dialogue to instant messaging. They provide "immediate and focused responses given specific questions" (2011: 136) and can be further broken down into synchronous and asynchronous forms of communication, i.e. with or without time delay between interactions. Non-relational resources (e.g. the corporate intranet, paper-based resources) do not involve direct interpersonal contact and can be thought of as the traditional type of resource translators use, such as dictionaries, glossaries or reference documents. Relational support is more likely to occur within a translation team than in individual projects and two main patterns of interaction can be identified. On the one hand, translators seek advice from their "peers" who have experience in the same or a similar project or topic, in particular when a team leader is available; on the other hand, junior translators tend to turn to more experienced colleagues, which implies a unidirectional form of information exchange. More importantly, "translators do not consult resources arbitrarily. Trusted resources are given priority over less trusted ones" (Karamanis *et al.* 2011: 43). Source trustworthiness plays a major role in deciding whether or not to use the proposed translation. In the case of online searches (i.e. when the source is less trusted), translators were found to place greater importance on the reason for choosing a translation over another (*why*) rather than the meaning of the word (*what*) (2011: 49). Vice versa, if the translator is using an internal TM, the metadata of the translation unit (e.g. date or author) may be sufficient to make an informed choice and the *why* is not really questioned.

When translation support is technology-based, the phenomenon under scrutiny is referred to as Human-Computer interaction (HCI)[79] (O'Brien 2012), a research area where field studies and Contextual Inquiry (CI) are often chosen as methods for data

[79] Should (external) computer-based translation support be interpreted in the broad sense of translation technologies such as TM systems, a number of additional studies can be mentioned that deal specifically with the relation between translators and TM systems (e.g. Dragsted 2004; O'Brien 2008; Alves & Liparini Campos 2009a,b).

elicitation. In a CI study (Désilets, Brunette *et al.* 2008; Karamanis *et al.* 2010, 2011), researches observe and interview users during their daily work. This approach has the advantage of being grounded in the end-user's world. TAPs studies have provided a wealth of data about the process of translation but were not very effective for investigating how technology can better support translators because "what [translators] *think* they need [...] often turns out to be different from what they *actually* need" (Désilets, Brunette *et al.* 2008; emphasis added). Conversely, available experiments have tended not to take into account *how* translators employ technology while working because they were focusing on different research questions. Désilets, Brunette and colleagues (2008) carried out their analysis using grounded theory in conjunction with pre-established categories such as user goals or recurrent workflows. Additional categories were found which were more specific for the translation task, such as those relating to translation problems[80]. In particular, they identified the following three categories: (i) problem-type (e.g. finding an equivalent or understanding the meaning of the ST), (ii) problem-solving approach (type of action performed) and (iii) employed linguistic resource(s).

In the Contextual Inquiry studies, researchers noted that available resources (e.g. glossaries or other reference material) were sometimes not enough and translators would then turn to online search (Karamanis *et al.* 2011: 43). In general, translators tended to be very cautious when evaluating results of online searches and sometimes multiple searches were carried out to verify the initial suggestion. Both the destination website and the number of results were taken as additional criteria to evaluate search results whereas the pool of resources shared by the translation team was generally considered trustworthy right away. Because translators seem to access resources according to their trustworthiness, the order in which external forms of support are accessed can provide additional insights into the reliability and/or usefulness of a specific resource over another. For example, Karamanis *et al.* (2011: 43) found that

> on many occasions the translators were observed doing a Concordance search as their first step for resolving a problem. Moreover, they would normally consult the TM and their references (provided that these were available) first and then search online. Trusted team members were asked to verify an inconclusive search and their own decisions.

4.2 CONCORDANCE SEARCHES & TRANSLATION PROBLEMS

Concordancers, as described in Chapter 3, can undoubtedly be considered a type of external support that translators use in addition to dictionaries and the Internet. For example, in an experimental study, the

> Trados Concordancer was considered to be a source of external support since it has to be looked up by the translator and works differently from the standard solutions provided by the TM (Alves & Liparini Campos 2009b: 201).

Evidence collected via Contextual Inquiry (Karamanis *et al.* 2010) showed that "translators interrupted editing in order to perform a Concordance search on several occasions". If the text editor is taken as an area of interest for the mapping of the translator's activity, a concordance search will be logged as an interruption in the typing (i.e. a pause) and is possibly preceded by a fixation of a text item. Long pauses and

[80] The researchers understood translation problems as words or expressions that caused difficulties to the translator and for which multiple resources had to be consulted.

fixations together with consultations of external support are all considered established problem indicators, hence concordance searches can be said to represent manifestations of translation problems.

Concordance searching was often found to be the first step translators took were facing a problem that could not be solved with internal support only, despite the fact that at times no answer could be obtained (Karamanis *et al.* 2010), in which case online searching was attempted. The concordancer remained nonetheless the main cause of interruption. While it can be quite safely assumed that all concordance searches are contained within a larger pause in target text production, not all pauses in the process represent an instance of a search using external support (the concordancer) because internal support, mixed support and virtually any other interruption can occur in a professional setting. Translators were reported using resources at their discretion whenever they felt appropriate. In the case of the concordancer, a user stated that "[at]'*[a]ny time I can highlight a portion of the sentence that I am translating and do a concordance search*'" (Karamanis *et al.* 2011: 44), which suggests that performing a concordance search is not perceived as labor-intensive or particularly time-consuming. In fact, a concordance search can be performed very quickly because the tool is integrated in the editing environment and the user only needs to highlight the desired source text portion and click on the appropriate shortcut (see Sub-section 3.2.3.2 and Karamanis *et al.* 2010). A bilingual concordancer may be less automated but keeps the translator in control of the search pattern. Even if concordancers "might appear to achieve less [than e.g. TM systems], they may be quicker to provide translators with results they can actually use, and [translators] are more likely to be more tolerant of unexpected situations" (Bowker & Barlow 2008: 12), particularly if they eventually learn "which types of patterns are likely to produce valuable information and which are likely to waste time" (2008: 13).

Moving now from the search act to the searched item, references can be made to the theoretical approach to translation problems. The chunk of source text chosen as the input for a concordance search can be said to represent the translator's focus of attention at a given moment, which justifies the view of a search query as a translation unit. However, these translation units are paired with some kind of external support, hence they fall in the category of problem units, the special kind of translation units discussed in Section 2.9. Indeed, "when working with a BC, the translator initiates the search and therefore only looks for passages for which he requires help" (Bowker & Barlow 2008: 11).

In sum, searches can be seen as manifestations of translation problems because they (i) represent items attracting the translator's attention at given moments, (ii) produce an interruption in the translation workflow (pause, very likely above the adopted cut-off length), (iii) involve using external support, (iv) are consciously performed and (v) ultimately aim at finding a target language version for a source text item. In this perspective, a concordance search can be regarded as a complex type of search because it involves many levels of analysis and a combination of reading, typing and strategic choices.

4.3 CONCORDANCE SEARCHES IN PROCESS RESEARCH

Concordance searches are saved and stored in the form of logs. In general, search logs (as opposed to other types of logged data) are records of the user requests to a system. Search logs as a data type have the advantage that source text segmentation is already provided and its identification does not rely on interpreting eye-tracking or keystroke

data. Instead, user searches can be used in the first place to better interpret micro-level data. This can be clearly seen from the example found in Alves *et al.* (2010: 132) where key logging alone would not have been enough to make hypotheses about the translator's cognitive activity. At some point in the experiment, a 45-second typing pause was recorded, after which a portion of a previous translation unit was typed. Only after triangulating key logs with eye-tracking data, researchers found out that during those 45 seconds, the translator used external support (an online dictionary) to look up a translation of a German verb. Using an external support involved a (much) longer pause than the threshold length set for the experiment but still the experiment showed that translators were quick at searching and making decisions.

In a previous experiment involving TAPs, the following scenario was reported where triangulation played a central role:

> The first translator started by reading the entire text very carefully. I assumed he would begin talking once he started translating. Several minutes later, he had still not said a word. Even though I knew not to interfere with the process, I quietly said to him "remember to think-aloud." He looked at me with great surprise and said "I am not thinking about anything" and went back to typing. Several minutes later, he was still not saying anything. Once again, I quietly asked him to think-aloud. He said he was just looking for a word in the dictionary, but did not give any details. Luckily, I had the camera to zoom in and could see that he was looking up the word écarter. He then continued to translate and did not say another word until he finished two hours later (Lauffer 2002: 64).

Situations of this kind could be virtually applied to many experimental scenarios just by replacing the type of aid available. Concordance searches do not require any verbalization on the part of the subject to be identified. They do not require any arbitrary cut-off length within the unit or in the pause time because each logged search has clear-cut boundaries, deliberately selected by the user[81]. In the above-mentioned experiments, knowing the searched items (beforehand) from a separate data source would have provided a good starting point to interpret process data (and in particular key-log and eye-tracking results) and analyze problematic items.

Generally speaking, the only thing that can be clearly measured is the fact that the translator left the editing environment to perform a search operation on a different tool. Presumably, there is a time frame[82] during which the translator considers whether or not s/he can immediately translate that source text portion using his/her own internal resources and then decide which translation aid is best suited to solve the problem[83]. This time frame is very likely to be extremely short but long enough to be recorded as a fixation of the source text. According to studies on eye-tracking and reading patterns, factors such as word familiarity, word predictability, word length and complexity, lexical and/or syntactic ambiguity affect fixation length (Jakobsen & Jensen 2008: 103). These

---

[81] This statement has in fact only a general validity because, as will be discussed in the following chapters, there are logs which can be grouped into sessions and consequently the identification of the "main" Problem Unit becomes trickier (see Section 5.6). However, each concordance search can also be considered a Problem Unit in isolation as each search would still represent an interruption in the target text production.

[82] This time frame may be considered as the "run-up" to production mentioned in Section 2.8.1.

[83] Following the "eye-mind assumption" (Just & Carpenter 1980: 330-1), "there is a high correlation between long fixation durations and effortful processing" (Jakobsen & Jensen 2008: 114).

factors have been documented specifically in a reading task but they are all the more likely to play a role in a translation task.

The main features of search logs is that they are Web-based and they are authentic, i.e. they can be collected without setting up an experiment and thus maintain a high ecological validity. Search log volumes can be of considerable size and can represent a large and often heterogeneous user group[84]. This can easily be controlled using standard IT procedures to filter specific computer IPs or isolate a geographic area. Obviously, the need to have a more or less controlled user group highly depends on the scope of the analysis but in general, large data volumes from a large pool of users tend to highlight (search) trends. Alternatively, generalizations have to be made from smaller datasets with a variable number of subjects (often not higher than 20 people) which can be problematic since problem perception has a strong intra-subjective component: not everyone perceives the same element as difficult (Séguinot 2000: 145).

In sum, in spite of possibly less controlled conditions during data collection, search logs provide a reliable segmentation of the source search segment and can lead to the identification of trends – as opposed to idiosyncratic behavior. This type of data can also produce (comparable) data in comparable quantities about e.g. different language combinations without the limitations that lab experiments usually impose. Search logs were chosen for this study to gain a better understanding of what translators are searching for in a systematic fashion. As a data type, search logs can be considered as a special case in that they represent a non-invasive method of data elicitation. By triangulating search logs with pauses and fixations, researchers may have a more objective basis to identify segmentation patterns and problem units by e.g. measuring the pause length immediately before a search. This could provide an empirical basis to identify the most appropriate cut-off length and possibly determine whether it should vary according to other parameters such as expertise or text difficulty.

4.4 TRANSLATION PROBLEMS AS INFORMATION NEEDS

In her study, Dragsted (2004) has looked at problems in cognitive terms as triggers of a goal-directed activity. Generally speaking, the ultimate goal in translation is to produce a target language version of a source text item but sometimes source text comprehension or target text production problems impede the goal. Solving these problems involves engaging short and long term memory. By looking at pause length, researchers can best estimate "how long it takes a subject to retrieve relevant information from the memory system, and […] to prepare each TL segment" (Dragsted 2004: 129).

Information retrieval is generally triggered by some kind of information need. An information need "traditionally denotes the start state for someone seeking information, which includes search using an [Information Retrieval] system" (Cole 2011: 1216). According to Canfora and Cerulo (2004: 178), there are three types of information need scenarios (for document-related searches): (i) known item information need, (ii) conscious information need and (iii) confused information need. In the first case, users try to locate or verify the existence of documents they know exist, while in the second scenario, users know about the subject matter but not about what documents they are

[84] Not all search logs, however, necessarily come from human beings, just as not 100% of the concordancer search logs always originate from professional translators. In the Web, robots (BOTs) exist whose function is to crawl the Web and navigate through links, thereby often performing search operations and queries in the freely accessible concordancers.

searching for. Finally, in the third scenario users know neither the documents nor the subject.

The question now arises as to whether translation problems can be generally thought of as information needs. An information need can be defined as "the gap between people's current information and information sufficiency threshold" (Lu & Yuan 2011: 134); it can be considered to be implicit in the user's mind (Canfora & Cerulo 2004: 178) and plays a central role when balancing cost and benefits for attaining the desired information. In particular, people are not necessarily motivated to expend too much effort on processing information, which means that sources likely to provide the result with the least effort would be selected more easily (Lu & Yuan 2011: 134). Given the points made in Section 4.2 about the frequency of use of the concordancer, this tool can be regarded as an information source requiring low effort on the part of the user. When accessing a concordancer, the user is basically trying to extract sufficient information from an external memory to fill the current knowledge gap which prevented the translation process to continue uninterrupted.

If translation problems can indeed be considered as knowledge gaps, then the task of problem solving can be regarded as an Information Retrieval (IR) task:

> Information Retrieval (IR) is the scientific discipline that deals with the analysis, design and implementation of computerized systems that address the representation, organization of, and access to large amounts of heterogeneous information encoded in digital format (Canfora & Cerulo 2004: 175).

This applies to most of today's forms of external support but in a metaphorical sense it can be used to refer to the instances of internal support too, where the translator attempts to retrieve some kind of information from his/her memory system, using the brain as a computer.

In classical IR, information needs are defined as "the perceived need for information that leads to someone using an information retrieval system" (Broder 2002: 3). The information need (the problem) is represented as a statement usually containing keywords or phrases, which involves a loss of information between the original information need (e.g. a phrase in context) and the way it is formulated (e.g. the phrase without context). This representation of a user information need is known as *query* and originates from a problem that the user is trying to solve (Canfora & Cerulo 2004: 175,8). Information seeking can be seen as closely related to problem solving in that both require the adoption of some kind of search strategy on the part of the user. In the field of translation,

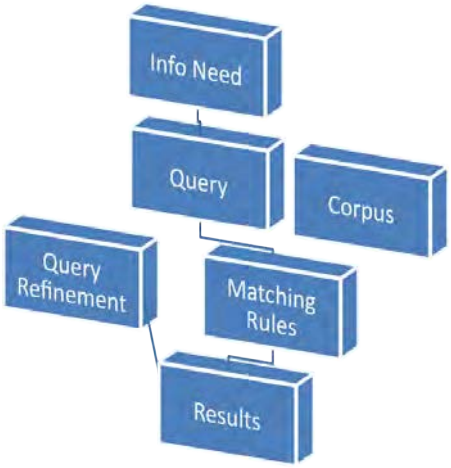
> translation strategies have their starting-point in the realization of a problem by a subject, and their termination in a (possibly preliminary) solution to the problem or in the subject's realization of the insolubility of the problem at the given point in time. [...] [F]urther verbal and/or mental activities can occur which can be interpreted as being strategy steps or elements of translation strategies (Lörscher 1991a: 96)

In this sense, translation strategies relate to forms of both internal and external support. If the focus shifts on the use of external support, the strategy steps in Figure 21 can be identified within the broader problem-solving strategy. This opens up new approaches to the task of analyzing translation problems and translation strategies.

Figure 21. Search-strategy steps in translation-problem solving using external support.

According to the classic model for IR (Figure 22), a search process can be broken down into a number of actions. An information need drives the user to formulating a query which is then submitted to the system. Matching rules enable the system to select the relevant documents from a larger collection (e.g. a corpus). After evaluating the results, a query refinement can be considered, where the query is reformulated to better balance information needs and results (Broder 2002: 4).

Figure 22. The classic model for IR (Broder 2002: 4).



The purpose of a search is retrieving some information which in the case of translation can take the form of an equivalent target language version, a definition, a word in a specific context, a synonym, background information and many more. However, retrieving at least one result from the system does not necessarily guarantee that the translation problem has been solved. Désilets and colleagues (2008b: 8) list typical situations where retrieving results does not come to a solution to the problem:

1. The translation provided was in the wrong sense or topical domain;
2. The target segment was not correctly aligned with the source in the repository;
3. The target language contained the same text as the source because the segment was not translated;
4. The problematic item in the source segment cannot be found in the target segment e.g. because it had purposely not been translated.

By considering translation problems as information needs, the word "problem" can be avoided altogether; Pavlović (2007: 30) noted that "the term 'problem' [...] might be laden with negative connotations". In order to avoid such connotations, some scholars have resorted to alternative labels such as "attention unit" (Jääskeläinen 1990) because "labeling all points at which the translation seems non-automatic in the same way has the disadvantage of investing the source text with the difficulty" (Séguinot 2000: 144 in Pavlović 2007: 30). During informal conversations with translators at the European Commission, the word "problem" was indeed hardly mentioned. There were references to things that looked "strange" or "checks" to be performed but there was no clear mention

of "problems". As far as professional translators are concerned, the notion of "problematic" should maybe be interpreted loosely, i.e. in the cognitive sense of "requiring more processing effort", as manifested by the fact that translators resorted to the concordancer.

If translation problems can be referred to as information needs, translation support tools could in turn be regarded as information sources and, in an even broader perspective, TM applications too can be seen from an IR perspective:

> When using a TMS [i.e. TM system], the translator is actually just searching for documents that might help in translating a given sentence. As for the query, it is constructed automatically by the TMS, from the SL sentence to be translated. The retrieval operation is carried out by matching this query as closely as possible. [...] the default strategy of existing TMS's is an extreme form of high-precision, low-recall search: return only the best matching document, and only if it displays a sufficient resemblance to the source sentence (Simard & Langlais 2001: 335).

The underlying assumption is that the user is not willing to browse through large volumes of results. However, the exact repetition of complete sentences within and across texts is a rare event in general language, the only exceptions being texts with specific content such as updates or localization-related texts.

The measures of Precision and Recall in IR derive from the assumption that any retrievable item can be considered (i) retrieved or not retrieved and (ii) relevant or not relevant (Lu & Yuan 2011: 134). Document relevance in translation depends on whether the translator deems that the "target-language (TL) segment constitutes an acceptable translation for the source-language (SL) sentence" (Simard & Langlais 2001: 335). In the case of a bilingual concordancer, solution recall can be understood as "the proportion of all relevant solutions that a tool is able to suggest" within the set limit of results to be displayed, whereas "solution precision is the proportion of [...] solutions proposed by a tool, which are in fact relevant for the problem at hand" (Désilets *et al.* 2008b: 7). In this type of search, higher recall can be obtained by making the search more general, i.e. by reducing the length of the string and/or by not applying any filters to the search. In this way, the search will produce more results but they might turn out less precise. Conversely, a search with high precision will be characterized by longer queries and/or a higher number of search filters, which will reduce retrieval effectiveness and limit the number of results. According to Simard and Langlais (2001: 336), the main purpose of a concordance search is to provide the translator with useful matches for the searched portion and not so much cover the whole source segment or recombine retrieved sub-matches automatically.

Source trustworthiness is certainly important but in the case of a information need, it does not seem to be a major issue because translators are good at dealing with noisy results provided that they have some results to evaluate in the first place. Therefore low precision would seem less harmful than no recall because "when translators use a tool to search for a solution to a particular problem, their main goal is to find *one acceptable solution* in the first 10 to 20 suggestions" (Désilets *et al.* 2008b: 7).

The selection of the resource to be queried first of all depends on the number of resources available (the fewer the resources, the more limited the choice) but it also depends on their specific features, in particular their quality and accessibility. Quality was hinted at in Section 4.1 while discussing trustworthiness because the two can be considered closely linked. Accessibility is a measure of how easily the resource can be accessed. For example, an Internet search engine can generally be accessed very easily but results might not be necessarily trustworthy in terms of quality. Ultimately, the information need will

determine which feature matters most for each search (Lu & Yuan 2011: 140-1). In particular, "[i]nstead of tending toward one source or the other, individuals skillfully adjust their information-seeking strategies according to their information needs" (Lu & Yuan 2011: 142) and

> they also like to be given a choice of different relevant solutions, so they can select the one that seems most appropriate for their current situation. Multiple suggestions are also often used as a source of inspiration for creating new solutions that are ideally suited to the situation at hand (Désilets *et al*. 2008a).

Lu and Yuan (2011: 135ff.) distinguish between three levels of information need (high, medium and low) which directly impact the trade-off between quality and accessibility (see Section 7.3.4).

4.5 CONCORDANCE SEARCHES VS. WEB SEARCH LOGS

Search logs in Translation Studies represent an underused and to some extent even unknown data type. Information Retrieval developed in the course of the 20th century and, in the late 1990s, it branched out to include Web searching. Queries submitted to search engines have been studied to find out what users are after and how they go about searching for it. The literature on Web log analysis has identified three main categories of Web queries (Broder 2002: 5-6; Rose & Levinson 2004: 14-15): (i) navigational, (ii) transactional (or resource) and (iii) informational queries. Navigational queries help the user reach a particular site, which can be either known or assumed to exist. Transactional queries are meant to reach a site where further actions will be required on the part of the user, such as shopping sites, downloading pages or Web-mediated services. Queries in the third category – informational queries – are aimed at finding (static) information on the Web for the main purpose of reading and are closest to classical IR. According to the distribution of queries found in several studies (Broder 2002: 8; Rose & Levinson 2004: 18; Baeza-Yates *et al.* 2006: 102; Jansen *et al.* 2008: 1262), the majority of Web queries are informational, which is also the category where the majority of translators' queries can be said to happen.

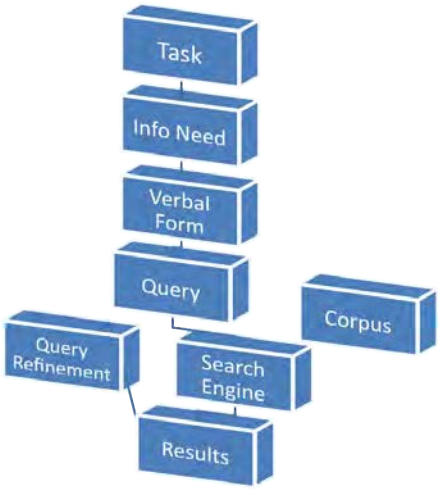
Broadly speaking, studies in the area of Web searching can be tentatively grouped into four main categories:

1. Studies on general Web searching as opposed to standard IR.
2. Studies on user behavior focusing on the goal of the searches, which can be identified in several ways (e.g. via the log itself, page results or clickthrough rates).
3. Studies on search sessions using time cut-offs to reconstruct the operations performed by one user when interacting with a Web search engine (e.g. quantitative and qualitative analysis of user sessions or query reformulation strategies).
4. Studies on queries in quantitative (e.g. statistics on length and frequencies) as well as qualitative terms (e.g. query clustering, taxonomies, topical categorization of queries and linguistic analysis).

Web searching is said to differ from traditional IR searching in that it consists of a "highly simplified type of searching by the broad public" whereas more sophisticated users and professionals would use other IR systems and submit more complex queries (Spink *et al.* 2001: 230). When it comes to modeling Web searching, however, the classical IR retrieval model (see Section 4.4) becomes slightly more articulated, as shown in Figure 23. An information need originates from a task and is verbalized (mentally) into a query that is

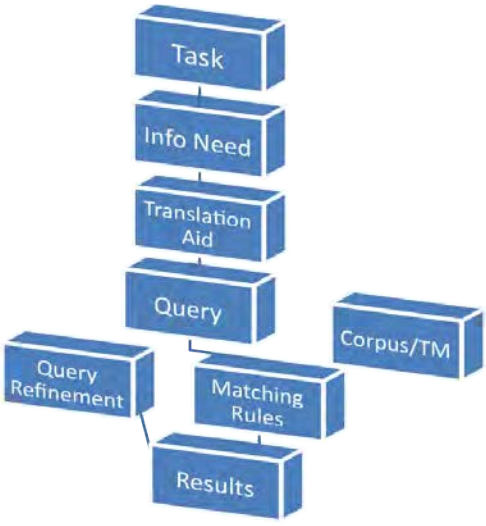
submitted to a search engine, which in turn extracts the relevant information from the available collection of resources and displays results (Broder 2002: 4).

Figure 23. Classic model for IR augmented for Web searching (Broder 2002: 4).



Building on the notion of translation problems as information needs, a similar diagram can be conceived for a translation task (Figure 24). In this case, the task from which an information need arises can be reduced to translating or revising a text. The information need can be handled in several ways, one of which is resorting to a translation aid as a form of external support. Nowadays, alternatives to Google as a Web search engine are scant so the selection of a Web search engine does not represent a challenge. However, the picture is much more complicated when it comes to translation and language resources in general. Contextual enquiries (Désilets *et al.* 2009) have shown that translators have at their disposal and use a wide range of resources which include terminology databases (private and public), online concordancers, translation memories and archives of previous, unaligned translations, general-purpose and specialized monolingual dictionaries, thesauri, the Web and specific websites, manuals of style, spell-checkers and newspaper articles relevant to the subject at hand. The list is by no means exhaustive and depends very much on the specific task, personal preferences, the working environment and the industry sector in which the translator operates (e.g. localization vs. legal translation).

Figure 24. Model for external support usage in translation (adapted from Broder 2002: 4).



As soon as the translator chooses the concordancer over existing and available tools a search strategy is in place. In the specific case of the EU translation services, the translator also chooses what interface s/he is going to use (the Web portal or Quest). During the query process, additional components come into play that are part of the strategic component, such as the use of filters or the amount of text to be searched. In the study by Karamanis *et al.* (2010), delays were sometimes observed when translators had to sift through the results and online searching was found to be more time-consuming than concordance searching because additional checks were required before accepting the suggested translation (Karamanis *et al.* 2010). In another study, manual search on Web search engines could last from a minimum of 30 seconds to 5 minutes. Despite being time consuming, subjects used it fairly frequently. In particular, of the 11 subjects, 7 used Google, 3 TransSearch and 3 a custom TM but Google was by far the most queried resource with 13 searches compared to 5 and 4 for the other resources respectively (Désilets *et al.* 2008b: 4)[85].

In the present study, the only translation aid to be considered is a multilingual concordancer, which means that the query is not "formulated" by typing keywords as in Web searches, but rather it is "copied" directly from the source text as a chunk of text. As opposed to Web searches, where there are virtually no limits to the amount of documents to be searched, the performance of a concordancer relies on the size of the underlying repository of aligned segments. The matching rules in the system retrieve the relevant units from this corpus of aligned segments, which can be referred to as a Translation Memory. The translator then evaluates results and, if necessary, refines the query.

In reviewing existing concordancing tools in Chapter 3, some comparisons were drawn between the concordancer and Web search engines, particularly at the User Interface level. Concordance searching shares some features not only with IR but also with Web searching. The boundaries between user categories and resources are increasingly blurred, as are the categories of users, now ranging from professional translators to general Internet users. This is also justified by the fact that most standalone concordancers are generally Web based and translators themselves resort to Web search engines to satisfy their information needs. However, search logs within Translation Studies are fundamentally different from generic Web search logs. Web searching can be triggered by a variety of purposes, whereas concordance searching has always an ultimate goal, i.e. the retrieval of a target language version. Irrespective of the specific features of each search, the ultimate user goal for the translator is to produce a target version of the text.

User behavior is an important component of the analysis because by inferring user behavior, search strategies and reformulation strategies can be identified. This is possibly where the greatest differences in the type of recorded data are to be found between Web searches and concordance logs. Search engines usually log information such as IP address and/or a cookie, which basically correspond to user ID. Often, click-through data and click-through rate (CTR) are also obtained. The former is a collection of user transactions using queries and their corresponding clicked links to discover correlations between the two e.g. for the purpose of clustering (Chuang & Chien 2002: 76). The latter is calculated by dividing the number of clicked links in the results by the total number of links obtained for an individual query. CTR is considered a measure of user satisfaction with the

[85] This result is possibly due to the fact that the TransSearch repository was not "customized" for the specific translation task at hand and its usability might have been limited in some cases.

retrieved result based on the submitted query and it is also the main source of revenue from advertising (Zhang *et al.* 2009: 1-2). Concordancer logs from Euramis do not contain either types of data and different metrics need to be found to (anonymously) identify users and user satisfaction. The time stamp is another important factor when studying the interaction between the user and the search engine. A temporal cut-off is sometimes used in addition to IP and cookies to identify search sessions, which can be plainly defined as "the entire set of queries by the same user over time" (Spink *et al.* 2001: 227).

Compared to more traditional data types, query logs are a rather poor source of data for individual events (Grimes *et al.* 2007: 1) in that they only show the recorded actions and not the intent behind the queries, which can be difficult or impossible to identify without context — usually missing form the logs (Huang & Efthimidias 2009: 84). In the case of translation-related searches, the scope of search logs is more constrained than in general Web searching and each interaction with the system could virtually be ascribed to a relatively small number of possible search intents (i.e. information needs) that will be addressed in Section 7.3.5. The existing distinction between "reception problems" and "production problems" (Krings 1986a: 144-152) will not be strictly enforced in the present analysis because any submitted query already implies that the problem was important enough to justify a search, irrespective of whether it was a reception or production problem. This is in line with Lörscher's approach (1991a: 94), which does not differentiate between reception and production problems in that the distinction "does not play any role *for the subjects*" (emphasis in the original). In Lörscher's study, translators would employ "problem solving strategies which are independent of the analytical artefacts of L2-competence problems and translation-competence problems" (1991a: 94). The difference, if any, which should emerge in the present study is the way translators evaluate the results of the concordance search, i.e. what elements they focus on more.

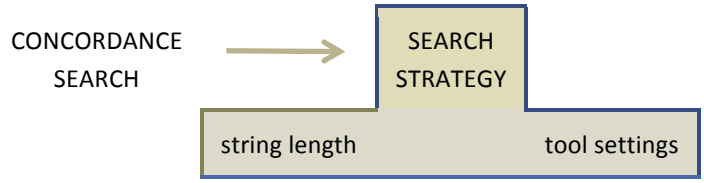
One obvious problem when studying search logs (in particular for Web searching) is the risk of using outdated technologies or even obsolete data and recent data can be hard to obtain. Researchers are aware of the rapid changes in technology but reassured that "[i]n contrast, people, their information needs, and behavior do not [change]" (Spink *et al.* 2001: 227), given that "the method of interaction has remained the same (i.e. enter query, retrieve results, scan results, view results, refine query as needed)" (Jansen *et al.* 2008: 1252).

4.6 STRUCTURE OF A CONCORDANCE SEARCH

Translation process research has developed a model for the use of translation support (PACTE 2005; Alves & Liparini Campos 2009a,b) and Web log analysis has developed a methodology to analyze search queries. By bringing them together, a methodological framework could be developed to analyze concordance search logs. Concordance search logs resemble very much a traditional query log but there are a few fundamental differences between the two. First and foremost, concordance search logs serve a different purpose from Web search logs. Web search log analysis, a branch of Transaction Log Analysis (TLA) has identified three main types of queries but in the case of a concordance search the only purpose is for the translator to find a target language version of a portion of a source text. This introduces the second main difference from traditional Web queries: a concordance search is normally not freely formulated by the user but rather it consists of a verbatim selection of the text to be translated, which implies that the form and structure of a concordance query is likely to be different from a traditional Web search.

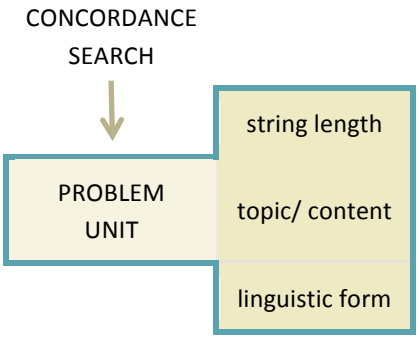
Earlier in this chapter (see Section 4.4), a concordance search was presented as a trade-off between recall and precision, i.e. between retrieval and relevance. This trade-off is accomplished by adjusting the search strategy (i.e. problem-solving strategy), which basically depends on two main components: string length and tool settings (Figure 25). The longer the string, the more precise the search will be but the less likely that the string will be found "as is" in the repository. Concordancers — just like Web search engines — can be manually set to include some more or less extended filtering options for the search (see Section 3.2). The more settings (i.e. filters) are applied to the search, the more stringent the matching rules will become. Therefore, applying filters is another way to restrict the search to the relevant results only but increases the risk of obtaining no results. The strategic component of a concordance search has a quite marked quantitative component, in that it requires string length to be measured and the distribution of tool settings to be analyzed. String length and tool settings will be systematically surveyed in Chapter 6.

Figure 25. Breakdown of a concordance search with respect to the variables pertaining to the search strategy: string length and tool settings.



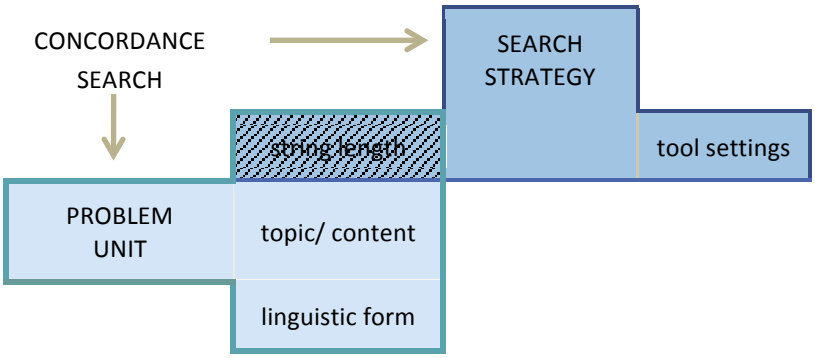
The central element of a concordance search is the query text that triggered the search and represents the underlying information need. The searched segment (i.e. the Problem Unit) deserves special attention and will be analyzed separately in Chapter 7. Kiraly (1995: 117ff.) distinguishes between different unit lengths for Problem Units (i.e. word and word string), whereas other studies focus on other properties of problematic items, such as word class or particular types of noun phrases (Campbell 1999), which can be referred to as the linguistic structure of the source text item. Other labels have also been used that include the category of abstractness (Campbell 1999) or distinguish between Language for Special Purposes (LSP) and Language for General Purposes (LGP) (Désilets 2009) thereby looking at the semantic level (i.e. content) of the string. In sum, length, linguistic form and semantic content of the query can all be considered possible aspects pertaining to the problematic unit, as summarized in Figure 26.

Figure 26. Breakdown of a Problem Unit into its various levels of analysis.



At this point, a concordance search has been broken down into a number of sub-components, grouped under two main headings: Search Strategy and Problem Unit. If the two representations are joined in one single chart, a comprehensive break-down of a concordance-search structure can be achieved, where the "string length" sub-component is shared between the two main components (Search Strategy and Problem Unit), as shown in Figure 27.

Figure 27. Complete breakdown of a concordance search (collated version from Figure 25 and Figure 26).



A concordance search derives from an information need that requires a search strategy to be employed to retrieve the desired information. By using a concordancer the translator is in control of the search process and can exactly define the search criteria. Unlike a standard Web query, translators using Euramis can profit from a controlled environment of texts pertaining to the specific domain of interest. These texts were most likely previously translated by the user him-/herself or by unit or department colleagues. Web queries are not ruled out as source of information but they are more likely to be chosen if the concordancer does not provide any usable information. The concordancer is less likely to be used to solve content related problems (i.e. Krings' reception problems) when an extremely efficient resource like a search engine is just a few clicks away. In fact, concordance query language is optimized for "linguistic", rather than "document-oriented" queries (e.g. those submitted to a traditional Web search engine):

> For example, when a [concordance] user specifies a sequence of words, their order is generally considered significant [...]. Even when this property does not hold, it is usually the case that the words which satisfy a query must appear within the same linguistic context, e.g. within the same sentence (Simard *et al.* 1993: 3).

The text of the query is the central part of the search. The text string can be analyzed using both a quantitative and a qualitative perspective, as shown by studies on Web search logs. The quantitative approach focuses on measurable and countable features of the string such as length, frequency counts and basic descriptive statistics such as query distributions. The qualitative analysis is more varied and includes clustering by topics, categorization and linguistic analysis.

In sum, a search in Euramis means that 1) a user (a translator) was translating or revising a document, or in any case performing an activity with a document involving two or more languages; 2) The translator encountered some problems that s/he could not (or did not want to) solve using internal support only; 3) The translator interrupted the activity (= paused) to elaborate a search strategy, e.g. selecting the appropriate tool; 4) The translator filled (part of) the pause by querying the Euramis concordancer to try and

solve the problem. 5) The translator evaluated the results and selected the most appropriate solution, if found.

The dataset for this study consists of concordance search logs. Given the above, every instance of a search represents a problem. The main analysis that can be carried out on this dataset is focused on the submitted searches from a descriptive perspective, i.e. an analysis of *what* translators looked up in Euramis. The dataset could also be analyzed from a different perspective, namely focusing on the rationale behind the searches (*why* translators submitted that particular search). However, no direct information is available to answer this question and results will mainly depend on the inferencing on the part of the researcher. Inferring the rationale of a problem by assigning it to some category is a common practice in translation (see Section 7.3.2) but there are virtually many possible causes for a problem and triggers for a search, which depend on a number of situational variables. For instance, Livbjerg and Mees (2003: 129) provided one example using the Danish word *'menneskerettighedskommissionen'* that can be considered problematic when translating into English due to possible issues of (i) spelling (capitalization), (ii) choice of words, (iii) grammatical form and/or (iv) the structure of the phrase. Deciding what rationale is applicable, if any, is not always a straightforward task.

Search logs have the advantage of coming in large quantities but at the same time they "may be strongly biased towards translation problems which are particularly appropriate for that one tool" (Désilets *et al.* 2008a: 343): when studying the logs from TransSearch, researchers found relatively few terminology problems compared to phraseology and idiomatic expressions. Professional translators seem well aware of strengths and weaknesses of each tool when they use them and thanks to Contextual Inquiry a more comprehensive picture of the range of problems encountered can be sketched (*ibid*.). The present analysis can only address some potential types of problems based on the available data and results from existing and relevant studies. However, direct triangulation of results is unfortunately not possible because no direct access to concordance users was possible when the queries were submitted.

4.7 KEY CONCEPTS

- Translation problems can be addressed by resorting to internal support (the translator's knowledge and cognitive resources), external support (documentation and reference material) or a combination of the two.
- External support can be grouped into relational and non-relational forms of support and each can presuppose synchronous or asynchronous forms of interactions.
- The present analysis focuses on a specific form of non-relational external support: the concordancer.
- Concordance searches can be considered manifestations of translation problems.
- Concordance search logs can be included in translation process research as an additional form of User Activity Data to be triangulated with data coming from key-logging, eye-tracking and verbal protocols.
- The notion of translation problem could be revised in terms of knowledge gap and therefore equated to an "information need".
- Both concordance searches and Web searches can be considered as forms of Information Retrieval; comparisons can be drawn between user behavior and search strategies in Web and concordance searching.
- A concordance search can be seen as a complex event that can be broken down into a Search Strategy component and a Problem Unit component.

- Each component can be further broken down into sub-components, namely tool settings and string length for the Search strategy component and string length (again), content and linguistic form for the Problem Unit.

CHAPTER 5: RESEARCH DESIGN

In the previous chapter, a concordance search was seen and described as a manifestation of a translation problem. In this chapter the dataset will be presented together with the main methodological approach to the analysis. As previously mentioned, no experiment was carried out for this analysis and the data were selected so that at least some variables could be controlled for the study to be scientifically acceptable. The first part of the chapter will deal with the ecological validity of the study and the controlled variables. The second part will present the data and explain all the pre-processing steps that were carried out on the strings before obtaining the final dataset for the main analysis. In the third part, the final dataset will be examined and some preliminary analyses will be carried out before the main analysis of the search subcomponents can be developed in the relevant chapters.

In Sections 3.2.3 and 3.2.4, the background for the study was sketched by presenting the translation tools that will be used to collect the data for the analysis. Interesting elements of this study are the high ecological validity that can be obtained by collecting authentic data on the daily work of translators and the wide coverage of language pairs. In this perspective, some theoretical conceptualizations will be derived from the analysis of recurring patterns, i.e. *trends*, rather than from the study of individual translators as such, provided they are clearly identifiable. Hypotheses and trends will be derived for each individual analytical step and will then be organized in categories leading to greater levels of abstraction, in line with some of the underpinnings of grounded theory[86] (Corbin & Strauss 1990, Strauss & Corbin 1994). An additional reference for approaching this exploratory study was found in the levels of data analysis suggested by Krings (2005: 354) that work best with structured data elicitation. The very first step in the (systematic) analysis of data is of a phenomenological nature, in that interesting examples are identified in the data and used as a starting point for the analysis that should then proceed in a systematic fashion for the whole dataset. This bottom-up approach should ideally be pursued towards increasing abstraction, starting from observing the data, moving on to classifying and quantifying them and identifying correlations and causality between variables to eventually formulate a theory. Immediately after the phenomenological analysis, individual phenomena should be labeled and categorized and then each category should be quantified so as to highlight phenomena that occur with higher frequency. In a traditional study on the process of translation, statistical measures for hypothesis testing would be employed to determine whether two phenomena are correlated to one another, i.e. whether there is a causal relationship between two variables from which the final step of theory generation can be reached. The present analysis, however, will largely be based on 20 individual subsets which means that there is an exponentially high number of possible correlations, making it very hard to perform systematic tests. Assuming that all the possible pairwise comparisons were performed, in order to test their statistical significance the p-value (usually used with a threshold of 0.05) would have to be lowered to account for multiple comparisons — the more the comparisons, the lower the p-value should be. By lowering the threshold, the test would

[86] Grounded theory is also used as theoretical basis of Transaction Log Analysis (TLA), i.e. Web search log analysis, in that the starting point are observations of the "real world" used to ground resulting theories or models (Jansen 2006: 409).

become more stringent, because the null hypothesis can only be rejected when a lower p-value threshold is reached. However, the lower the p-value, the harder it becomes to obtain significant results. This problem would occur at each analytical step where all target languages are separately analyzed. For all these reasons, statistical significance will not be tested.

Data collection for the present study did not rely on any structured or systematic approach, such as surveys, field observation or interviews. Some were nonetheless used to complement the analysis but they were rather informal in nature and did not directly pertain to the collected dataset. When translators use online forms of translation support, authentic user activity can easily be logged in a non-disruptive way and collected ex-post. This is the way in which the present dataset was obtained. Due to the highly exploratory nature of the study, a proper experimental setup was avoided, so as to reduce idiosyncratic components as much as possible and to focus on possible approaches to studying the available data. Despite the numerous methodological challenges that arose in the course of the analysis, this approach helped to eschew some of the renowned limitations of translation experiments, summarized by O'Brien (2009: 251) as relating to, among others, the research environment, research participants, ethics, data explosion and validity[87]. The following sections will each cover some of the main aspects by giving an account of the way in which such variables could be "controlled."

5.1 RESEARCH ENVIRONMENT

For the analysis to include some controlled variables, the research environment was limited to the level of the EU intranet. The users who have produced the data work internally at the eight European institutions (European Commission, European Parliament, Council of the European Union, Court of Auditors, Court of Justice, Committee of the Regions, Economic and Social Committee and Translation Centre for the Bodies of the European Union) that have access to Euramis and Quest. More specifically, they work for the translation services of these institutions and are generally referred to as translating staff. *Translating staff* may include any person actively participating in producing a translation because the translation services comprise not only permanent translators, but also assistants, lawyer linguists, terminologists, stagiaires, contracting agents and temporary staff (and possibly interpreters from the SCIC services) which amount to over 4,000 potential users across the eight institutions[88]. The total number of potential users who have access to the EU-intranet is obviously much higher but active users are very likely limited to the translating staff.

As detailed in Section 3.2.3, Quest is a predominantly linguistic tool whereas Euramis offers a range of services that are used by a potentially larger pool of users (e.g. pre-translation services and IT departments). For this reason, statistics referring exclusively to Euramis (see Table 3, Sub-section 3.2.3.1) might be considered less reliable because

[87] In a broader perspective, Krings (2005: 357) pointed out that translation process research would profit, among others, from a variety of methodological approaches, larger subject groups possibly involving translation professionals, real-life text types, greater language-coverage, more field studies and more frequent adoption of inferential statistics to test correlations and causality. This research project is an attempt to meet a number of these desiderata.

[88] According to official statistics, translating staff across EU institutions in 2010 was distributed as follows (Kowalska 2010; values in brackets refer to 2012): EC 1750 (1650); EP 760 (750); Council 650 (=); COA 100 (=); COJ 620 (=); EESC/COR joint service 350 (=); TC 110 (=); ECB 70 (50); EIB 30 (=).

they virtually include all users accessing the whole range of services in the Portal. Provided that only users accessing the concordancer could be counted, it would still be impossible to exactly determine how many unique users accessed the tool on one single day by simply looking at the global statistics. In September 2010, Euramis received a total of 971,322 searches of which 630,935 (or 65%) were submitted via Quest and 340,386 (35%) using the Euramis Web-Portal. The searches cover all 23 language combinations; hence they will be referred to as the ALL>ALL dataset. In the period under consideration, the busiest day for the Euramis concordancer was September 20th. The automatically generated statistics for that day reveal that 52,000 requests were submitted via the Euramis Web Portal and 919 different ('unique') active users were counted in the Euramis Portal, one being Quest. Of the 52,000 searches, 33,278 (64%) were submitted by Quest, which means that this "user" alone was responsible for almost two thirds of the requests of that day. Quest has its own user logging system and requires a separate user registration from the Euramis Portal. Statistics from the Quest engine show that 1,398 unique users performed 33,291 searches[89] in the chosen date, which means that the user 'Quest' in Euramis comprised almost 1,400 real unique users, though some might have logged in the Euramis Portal as well. The total Euramis user count for September 20th would therefore range from a minimum of 478 (=1,398-919-1) — if all users accessed both systems separately, i.e. using both accounts — to a maximum of 2,316 users (=1,398+919-1) — in the event that all users queried one system only. If the potential users were taken to be 4,300, translators working with Euramis and/or Quest on September 20th would range from 11% to 54% of the total. These numbers are only meant to provide an estimate of active users on a given day. More detailed statistics on Quest are provided in Table 4.

Table 4. Translating Staff, Quest total and active users both in 2010 and in September 2010, together with estimate for queries divided per institution (see Table 3 in Section 3.2.3.1 for abbreviations).

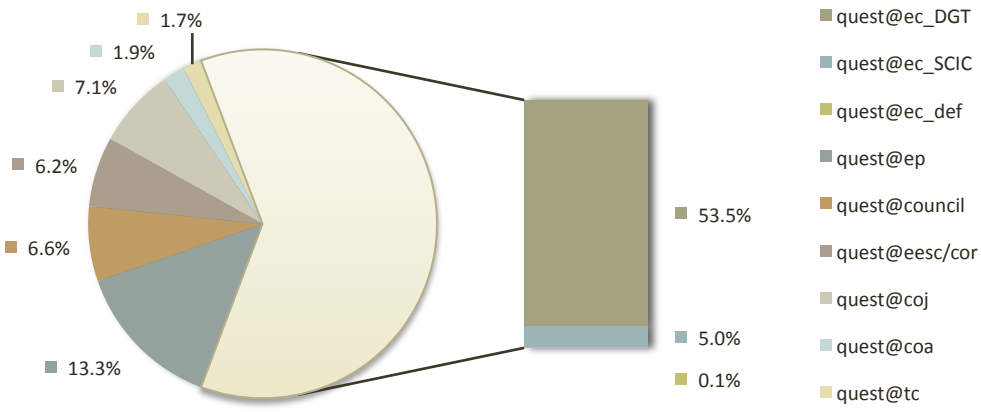
	EC	EP	COUNCIL	COA	EESC/COR	COJ	TC	TOTAL
Tot. Transl. Staff 2010[90]	1,750	760	650	100	350	620	110	4,340
Total Quest users 2010	2,518	595	530	134	576	428	134	4,915
Active Quest users 2010	2,032	405	352	106	432	300	109	3,736
% Active Quest users	80.70%	68.10%	66.40%	79.10%	75.00%	70.10%	81.30%	76.01%
Quest users Sept 2010	2,345	575	502	125	550	388	127	4,612
Active Quest users Sept 2010[91]	1,892	392	333	99	413	272	103	3,504
Total Quest queries Sept 2010	365,202	84,022	41,618	12,091	72,907	44,584	10,499	630,923[92]
Avg. queries per Active User - Sept 2010[93]	193	215	125	122	177	164	102	180

From Table 4, different percentages emerge with respect to the ratio of active Quest users in September 2010 and the total translating staff in 2010. In some institutions (EC,

[89] Compared with the statistics from Euramis (33,278), there is a delta of about a dozen strings, suggesting that there might have been instances where Euramis was not used as a resource in Quest.
[90] Interpreters and lawyer-linguists not included (Kowalska 2010).
[91] Estimate on the basis of available data.
[92] Again, compared with the previous statistic (630,935) there is a delta of a dozen strings which is possibly due to some minor differences in the two logging systems.
[93] Estimate calculated as the average per user (supposing equal distribution of queries).

EESC/COR) the number of active users exceeds the stated translating staff but this may be due to the fact that official statistics on translating staff do not include lawyer linguists and interpreters and possibly other people such as assistants[94]. In the remaining institutions, the number of active users is almost equal to or in some cases (EP, COUNCIL, COJ[95]) much smaller than the number of potential users. This indicates that the potential pool of users lies somewhere between 3,500 and 4,000, i.e. a considerable number of (virtual) subjects compared to other studies in the field. A closer look into Quest searches divided by institution provides a more accurate breakdown of resource usage (Figure 28).

Figure 28. Distribution of queries submitted via Quest per requesting institution (September 2010).



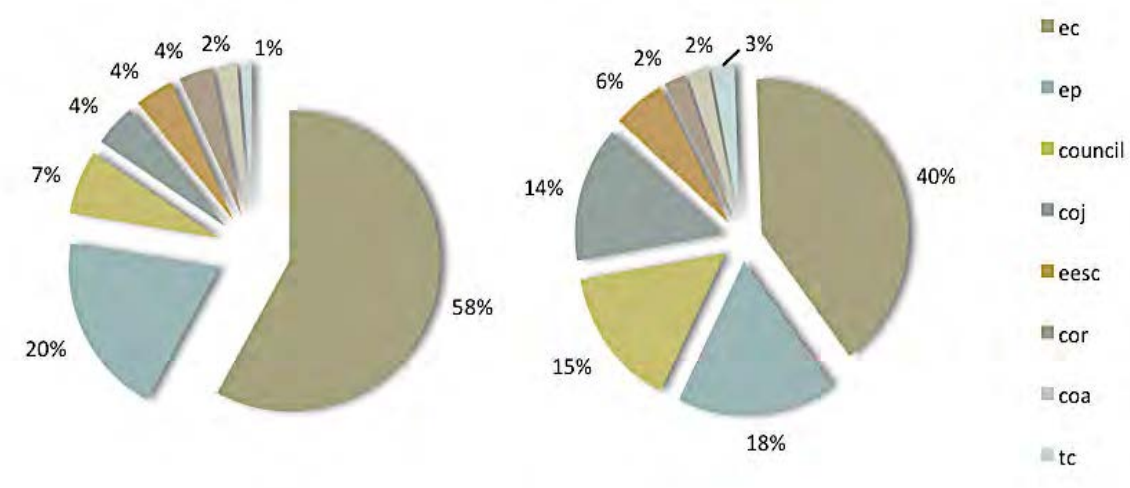
The main pie clearly shows that the majority of the queries in Quest in September 2010 were submitted by the EC (as suggested by the relative size of the user pool), followed by the EP and the Council as individual institutions. A further breakdown of the searches from the EC shows that the overwhelming majority comes in fact from DGT and a small percentage from the interpreting services (SCIC). This confirms the initial assumption that translating staff, most likely translators, are the main users of Quest and the Euramis concordancer, the only exception being former translators who have changed their career path and are now working in a different unit (0.1%) while still possibly using translation resources.

These statistics help defining the research environment within the translation services, which means that features that will be discussed in the following sections (i.e. directionality, translation domains, working conditions) are quite homogeneous, hence

[94] According to DGT statistics about 2010, there were 1750 translators and 600 "support staff" with 110 translators working at the English department (DGT 2010a). A breakdown of function groups in DGT in 2009 provides the following statistics for a total of 2336 employees: Senior Manager (6), Middle Managers (94), Translators (1413), Web translators (99), Field Offices (33), Administrators (147), Assistants (544) (EC 2010). As for the EP, the following statistics were available as of 1st January 2012. Out of a total of 1164 employees, there were 724 AD (i.e. official translators), 429 AST (i.e. assistants/secretaries) and 11 other (e.g. temporary agents). The overall number of translators usually ranges between 700 and 800 (Poulis, personal communication).
[95] According to the information provided by the Project Team, only a small percentage of translators at the COJ have in fact access to Euramis, which would explain the small number of users.

controlled for upstream as much as possible (i.e. at user level, as opposed to filtering search logs downstream, after data collection). To exemplify how relevant the statistics about the EU translation services are, Figure 29 compares two pie charts, one showing the distribution of searches in the whole dataset, the other the distribution of the translating staff across all institutions according to official data. As pointed out in Section 3.2.3, searches can be made to and from all available languages, meaning the whole dataset covers a total of 506 language combinations[96]. The initial dataset will be referred to as ALL>ALL and amounts to 971,321 searches. The first comparisons were carried out on the initial dataset with a view to identify more meaningful subsets and reduce the number of language pairs to analyze.

Figure 29. Distribution of searches per institution (ALL>ALL) viz. Distribution of translating staff according to official statistics for 2010.



By comparing the two charts, the amount of searches seems to positively correlate with the number of translators for most institutions with the exception of the Translation Centre (TC), which seems less active than expected with respect to the size of its translation services. The Commission (EC) shows greater activity in proportion to its size whereas the Council and the Court of Justice (COJ) seem to have only a small proportion of users accessing these resources. This argues in favor of the conclusion that translators are the ones who actually take advantage of these resources and consequently the query logs to be analyzed can be said to come for the most part – if not completely – from translators.

Information about the research environment, in this case an accurate distribution of the virtual subjects across institutions, allows for correct data interpretation and provides a sound basis for additional considerations. In view of the main analysis, these results suggest that the European Commission, the European Parliament and possibly the Council will be the main data sources. Given the above distributions, a closer look into the internal structure of the two most active requesters, EC and EP, is deemed appropriate to better sketch the research environment and complement the overview provided in Chapter 3.

[96] This is the total number of language combinations when only EU official languages are considered, thereby excluding Croatian (HR) and Turkish (TR) which were also present in the initial dataset.

5.1.1 EUROPEAN COMMISSION

The EC deserves a special analysis because it possibly has the most articulated organization and is one of the largest translation services worldwide. A whole Directorate General is dedicated to translation (DGT), which deals with a wide range of texts[97] and target readers and also counts a separate unit for Web publications. Generally speaking, the documents at the EC are divided into three main groups: incoming, outgoing and internal (DGT 2009b: 54-55; EC 2010: 47). A large part of the translation work at the EC is devoted to translating working documents, including regulations, directives and decisions and in many cases the texts are based on standard models; in 2009, 74% of the documents were made up of existing legal obligations such as implementation measures and monitoring reports; 10% consisted of non-core documents (whose translation depends on resources and costs); 8% were political priorities possibly creating new legal obligations and 8% were communication priorities of the EC (DGT 2010c). The documents produced by DGT can be further divided into two main categories: (i) legislation-related documents, i.e. texts using specific EU terminology and usually available in all 23 official languages, accounting for over two thirds of the total DGT workload; and (ii) communication documents aimed at EU citizens, to be adapted to the national context of each Member State (EC 2010: 47,48) and coming in the form of printed materials or website content.

Several studies have presented the EU translation services and their internal workflow, with a special focus on the European Commission (EC) which offers the largest translation service of all. Some of these studies were conducted by academics and researchers giving an external perspective (Dollerup 2001; Drugan 2004) while other studies originated from within the institutions and provided an inside perspective (Wagner *et al.* 2002; Cosmai 2007; Koskinen 2008). DGT offices are split between Brussels and Luxembourg and count a separate language department for each of the 23 EU official languages. The language departments are distributed into three of the six Directorates of DGT, called Translation Directorates – the other three being the Transversal Linguistic Services Directorate, the Resources Directorate and the Translation Strategy Directorate. Every language department counts some 60 translators who are further divided into three 'thematic' units each serving a number of Directorate Generals and dealing with specific topics (Peressini 2010): Justice, External Relations, Administration, Budget, Competition (Unit 01); Environment, Climate, Agriculture, Energy, Research, Economic and Financial Affairs, Internal Market, Information Society (Unit 02); Enterprise, Education and Culture, Trade, Health and Consumers, Employment (Unit 03). The enlargement of the EU brought about an urgent need for reorganizing the translation service of the EC, which went back to a language-based system integrating the pre-existing thematic organization.

[97] At the EC the following document types are translated: Legal acts and preparatory documents; International agreements; Policy statements; Commission decisions and communications (answers to be written and oral parliamentary questions); Publications (scripts and captions for films and other promotional material); Correspondence (internal administrative matters and staff information, correspondence with ministries, firms, interest groups and individuals); Speeches (speeches and speaking notes); Minutes; Reports (technical studies, financial reports); Working Documents (briefings and press releases); Web Pages and publications of every size and format on a huge range of topics for opinion-formers and the general public (DGT 2009b, DGT 2010c).

Such internal organization allows translators to specialize in some domains and the number of staff per unit depends on the demand for translations[98]. Different sizes and degrees of specialization contribute to making workflow management more challenging (Drugan 2004: 18): people use diverse translation tools in a different manner and staff works at least in two sites, not to mention teleworkers. This is the main reason behind the centralization and standardization of some pre- and post-translation tasks.

5.1.2 EUROPEAN PARLIAMENT

The structure of the translation division at the European Parliament is more straightforward: Directorate A deals with support and technological services for translation (ITS services), and Directorate B ("Translation and Terminology") is divided into 23 language units, one for each official language plus one Terminology Unit. The language regime at the EP is more balanced: there are no procedural languages, as is the case with English, French and German at the Commission, although they are used as *pivot* languages, English being the most common. In 2009, about 80% of the source documents at DGT were drafted in English by non-native speakers (DGT 2009b: 56). Translations are made from and into all official languages, at least for some document types. There is also a Pre-Translation Unit (PreTrad) dealing with more technical tasks and the management of translation resources and support. Of the three main institutions (EC, EP and COUNCIL), the Parliament has the largest document re-use, the highest level of multilingualism and the shortest deadlines.

The workflow at the Parliament is also quite articulated and differs from that of the Commission one reason being the specific features of documents to be translated (e.g. multilingual documents in the case of amendments). A detailed breakdown of some sample workflows at the EP can be found in Poulis (2009) and Hands (2012). While still working with Euramis, Quest and other resources developed by the European Commission, translators at Parliament also have a number of internally developed tools and resources available, the main ones being Twist, Shout!, FullDoc and Cat4Trad. Twist stands for Translator's Workbench Integrated Suite and is a standalone application providing a customized interface for the Trados workbench and used specifically for translation and TM handling. Shout! is a Web application acting as an interface with Euramis for EP users who do not need to use the main Euramis Portal. The main reason for developing this interface was improve integration between internal EP databases and resources and the Commission's resources. FullDoc is an intranet-based tool for document search and retrieval and works similar to the Euramis concordancer but it only accesses EP texts and can also display results from on-going translations. Finally, Cat4Trad is an *ad hoc* developed XML-based CAT tool, specifically designed to handle multilingual documents (i.e. amendments) using Translation Memories thereby complementing the functionalities of the existing CAT tool (Trados).

5.1.3 LANGUAGE POLICY FOR THE EU TRANSLATORS

One very important aspect of translation at the European institutions regards the language policy and justifies the division in language units with translators sharing the same native language: Translators translate into their mother tongue. However, the

[98] English, French and German are an exception: they are the procedural languages of the EC with a higher workload and consequently more staff (about 120 translators). Their language departments are divided into six units and not three like all other languages.

uncompromising statements in Dollerup (2001: 31; emphasis added): "[...] translations are *always* made into the translators' mother tongue [...]", Wagner *et al.* (2002: 32; emphasis added): "Translators translate *exclusively* into their mother tongue" and Cosmai (2007: 93; emphasis added) "[...] l'attività di traduzione comunitaria è *sempre* passiva, si svolge cioè dalla lingua straniera verso quella materna – o indicata come tale – del traduttore" need to be slightly revised for the sake of precision. In recent times, DGT has introduced a "two-way translation" option (anticipated by Wagner *et al.* 2002: 110). On a voluntary basis, translators can also translate documents for internal information only (usually newspaper articles or letters) that are not intended for publication, generally into English or French (EC 2010: 48, DGT 2009b: 54) and this is reflected in the new statement that "[translators] translate out of several languages, but *almost always* into their mother tongue" (DGT 2009a: 2; emphasis added), or rather "into the language they regard as their main language, generally their mother tongue" (DGT 2009c: 9). Although there are some openings to *retour* translation (i.e. out of the main language, generally into English) particularly as far as the new Member States are concerned, almost all institutions have rejected this option (Cosmai 2007: 190) and rather use *pivot* language versions (English, French and German) of the documents.

This translation policy is particularly relevant to the present study and the analysis of the research environment because it implies that EU translators are native speakers of the language they translate into, differently from the TransSearch scenario studied by Macklovitch *et al.* (2000) where concordance users could be translators into either French or English and it was not possible to determine from the search what the directionality of the translation was. In the case of Euramis and Quest, the selected target language in a query will be a clear indicator of the language unit the translator belongs to and consequently of his/her native language because there is one main directionality in translations for the EU. The following section will deal in greater detail with the average user profile of the "research participants".

5.2 RESEARCH PARTICIPANTS

The overview on the research environment suggests that the average profile of a concordance user is a translator working into his/her native language and that each language unit of an institution will count approximately the same number of translators, with a few notable exceptions.

Given the large pool of translators, considerable differences in terms of age groups, computer skills and expertise should be expected adding to the well-established inter-subjective differences in translation styles and needs established in the literature (see Chapter 2). This user pool can comprise subjects ranging from a senior translator with over 20 years' experience translating for the EU to a newly arrived trainee who is completing his/her university degree in translation. In terms of traditional experimental design, such internal variability can be considered a disadvantage for data interpretation. However, the present study also aims to provide a baseline for the use of a concordancer by the average user so that service providers of commercially and freely available concordancers may benefit, too. For this reason, the examined searches need to account for the variety of potential tool users, which becomes even greater in the case of freely available concordancing tools accessed by users who are not proficient in a language and where search directionality is unknown. For the purposes of this research, a more precise study on the actual profiles of users was deemed unnecessary. The aim of the analysis is to obtain overall trends from as large a volume of data as possible as opposed to generalizing results from data obtained from a small set of controlled users where

idiosyncratic behavior may be quite marked. A previous study on translation and multilingualism (DGT 2010a) was partly based on replies to three different questionnaires, one addressed to translation units of the EC, the EP and the Council, one to lawyer-linguists and one to the Member States' central administration. According to the aforementioned statistics on translating staff, some 3,000 potential respondents might have been expected for questionnaires one and two together. However, only "a total of 45 replies from those involved in the multilingual drafting process could be analyzed" (2010a: 40), which would provides only a partial snapshot of the overall picture.

When it comes to establishing user profiles, Web users accessing concordancers can be expected to have any possible background because freely available online concordancers do not restrict access to the tool. Filtering users (e.g. professional translators vs. general Web users) on the basis of existing logs becomes nearly impossible. However, the EU working environment provides a way of controlling user access to the tool early on in the search process so that this exploratory study focused on professional translators can be carried out. Irrespective of the individual background, EU staff translators need to comply with some general regulations and selection procedures before they can start working for the EU, which helps to provide partly controlled conditions for the study. To become an official translator, a candidate has to complete a special selection procedure, as laid out in the document "Staff Regulations of officials and the conditions of employment of other servants of the European Communities" adopted in 2004 (the first version dating back to 1968). Since July 2002, all selection procedures have been implemented by EPSO (European Personnel Selection Office) but translators can also be selected and employed as temporary staff, contract agents and trainees to work internally at the institutions (Cosmai 2007: 93,4). Free-lance translators are another important group of contributors, but they work remotely and "do not benefit from all the facilities available in Euramis" (Drugan 2004: 13) and more generally in the intranet. Irrespective of the contract type, a more or less stringent selection procedure is carried out either by EPSO or (a commission of) officials from within a translation unit (e.g. in the case of translation trainees). The wide variety of subjects and expertise levels offers an interesting spectrum of the translation profession covering the whole range of potential users, from the novice to the expert translator.

Search logs are a data type that may raise issues of anonymity, confidentiality and intellectual property. Simard and Macklovitch (2005: 77) rightly point out that "whatever knowledge we can extract from [TranSearch] log-files certainly cannot be seen as 'voluntary contribution' from the community". In order to make data available to the scientific community, user anonymity must be preserved. Encryption of the communication channels may be necessary and data ownership, particularly in the cases where surrounding context is also stored (see Sub-section 3.2.2.7), may be thorny if data collection takes place wholesale. In the case of Euramis, however, the researcher did not have direct access to the queries, as in the TransSearch study. Permission was asked to use and analyze the data and, once granted, the searches were anonymized before release.

5.3 EXPERTISE

In translation process studies, the question is often raised as to what constitutes an expert translator and where the dividing line between categories should be drawn.

One general definition of expertise is

> the property of a person who performs an operation or a set of operations in a limited domain with exceptional results when compared to others capable of performing the same operation. *Expertise* generally implies useful and large amounts of knowledge and fluent action, and it may depend on abstractions such as individual mental models, rather than on knowledge alone (Muñoz Martin 2009a: 25; emphasis in the original).

Muñoz Martin goes on to provide an operational definition of "expertise" specifically for translation which, however, uses some slippery concepts such as "expert translator":

> Translation *expertise* [is defined] as the *capabilities* which underlie the performance of human *expert* translators, including extensive domain knowledge, but crucially also heuristic rules that simplify and improve approaches to *problem* solving, metaknowledge and metacognition, and compiled forms of *behaviour* which afford great economy in skilled performance (2009: 25; emphasis in the original).

Jääskeläinen (2010) provides a comprehensive overview on the evolution of the concept of expertise and the problems related to generalizing results of empirical research involving "naïve translators" (i.e. traditionally language learners). Translation students are generally labeled "novice translators" or "semi-professionals", whereas "professional translators" populate the top-level and are loosely defined as "somebody who is a practicing translator and not a student" (O'Brien 2009: 254). This raises the problem of defining the concepts of "expertise", "professionalism" and possibly "specialization". Recent studies have used a combination of requirements to label a translator as a professional: over 3 years of experience and over 50% of personal income coming from translation jobs (Martín Mor 2011: 106) or a translation degree or certification and at least 3 years of experience in the relevant language pair (Hvelplund 2011: 85).

Aside from the definition of expertise, approaches to translation work may also vary according to the source text format and individual preferences having to do with diverging working methods and acquired skills, which in turn depend on the type of aid a translator uses for drafting the text (ranging from translation using Trados Workbench, to Word or dictation), as Drugan's field study clearly shows (2004: 5-6). Irrespective of the preferred working method, the Euramis concordancer is always available and accessible as an external resource to all EU translators. It can be assumed that a good share of Euramis searches come from expert translators and the study can be said to focus primarily (but not exclusively) on professional translators[99]. While the average number of participants in traditional translation process studies has been on an average of 8 "translators" for experiments involving TAPs to about a dozen in more recent studies involving eye tracking and key-logging, the potential number of participants in this study could be as high as 4,000.

From a methodological point of view, it is always necessary to specify "which view is adopted [...] to allow for comparison of research results, [...] the definition of expertise or professionalism used and the relevant background information on the subjects" (Jääskeläinen 2010: 223). In this research, there are no "subjects" as traditionally understood in empirical research because neither an experiment nor the correlated recruiting were performed. However, the controlled environment of the EU allows to

[99] For example, at the European Parliament, there are about 700 official translators and some 50 trainees (2-3 per language unit) that rotate every three to six months. At the time of data collection, EP trainees still did not have access to the Quest search engine, which means that they are not included among the Quest users for September 2010.

sketch average profiles of the participants while guaranteeing their anonymity at all stages, which is one of the requirements for ethical approval (O'Brien 2009: 259).

The present study focuses on real-working conditions to try and investigate "real scenarios" and increase validity (O'Brien 2009: 262) while sidestepping the problem of discerning "the extent to which findings in a laboratory can be extended to other environments" (Muñoz Martín 2009a: 26), a problem raised by several authors in the field who pointed out potential limitations in the generalization of results from experiments due to e.g. limited number of participants, time constraints to run the experiment, nature of the text used and selection criteria for the subjects (e.g. Muñoz Martin 2009b; O'Brien 2009). A possible objection to the subjects in this study may reside in the lack of a clear-cut distinction between the professional and trainee translators owing to the impossibility, with the available data, to further tell the searches apart. This objection can at least partly be countered by the fact that sometimes trainees at the EU institutions are in fact translation professionals who decided to take on a different job for a while. Moreover, a study conducted by Lörscher with language learners and professional translators has shown that "in spite of the differences, professional and non-professional translation processes have many features in common [...]" and suggested "[...] that the two kinds of mental processes are similar" and "[f]rom the point of view of the strategies detected, the mental processes of the two kinds of translators did not reveal significant differences" (2005: 604ff.). In Lörscher's study (2005), the two groups have rather shown differences in the quantitative aspects of the translation process, i.e. the type and frequency of the different translation strategies, the size of the translation units (larger among professionals, smaller among students) and the degree of control over the TL output both in terms of grammar and style.

5.4 DATASET FOR THE STUDY

The dataset for this study covers a month's worth of user searches and was collected ex-post, i.e. at the end of September 2010. The searches together can be said to form a special kind of multilingual corpus of search logs covering all 23 EU official languages. Following the definition of Bowker and Pearson (2002: 9), "[a] corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria," the keywords being 'authentic', 'electronic', 'large' and 'specific criteria'. On this basis, the dataset under consideration can be equated to a corpus and, following the corpus typology in Laviosa (2002: 34ff.), it would fall in the categories "sample", "synchronic", "general and terminological", "multilingual"[100] and "written" corpus. However, there is an additional word in the Bowker and Pearson's definition that requires attention, i.e. *texts.* The dataset for this study is in fact a collection of text *strings*, not whole texts. This may prove problematic for employing established corpus-analysis methods and standard corpus-analysis tools. In a translation process research perspective, a corpus of translated texts

> only constitutes post-hoc data, as the product is only the end point of the process(es) that created it. Experimental data from psycholinguistic tests, on the other end, allows better and more direct access to cognitive processing, yielding copious process data, but the product data obtained in such artificial settings is on a much smaller scale than that available in corpora. A combination of both product- and process-oriented methods should provide

[100] This only pertains to the initial dataset. As will be detailed in Section 5.5, the actual dataset used in the analysis has been reduced to a monolingual corpus.

substantively greater empirical process and product evidence [...]" (Alves *et al.* 2010: 111).

A corpus of search strings provides insights into the translation process by framing the translator's activity in a collection of snapshots that can be joined into a novel type of corpus. This data type alone is however not sufficient to provide conclusive information about the translation process but should rather be seen as an additional data type to be used for triangulation[101] so as to complement and/or verify findings based on other methods of data collection, e.g. keystroke logging, eye-tracking and product analysis (e.g. Alves 2003). Similar to keystroke, large volumes of concordance search logs can be collected in a reasonably short time span, which may expose the study to the data explosion problem. If generalizations of research findings are usually the problematic step, in the case of data explosion researchers acquire a rich set of data that may become overwhelming if there is only one researcher to carry out the analysis or if processes cannot be automatized (O'Brien 2009: 260-1). Any methodological inaccuracy resulting from such superabundance of data can negatively impact the validity of the research and its results.

After presenting the backdrop of this study, a few additional considerations about the dataset are necessary before delving into the analysis. In order to control for data explosion, some pre-processing of the initial dataset was carried out. The following sections present in greater detail the initial dataset and the evaluations that guided the corpus pre-processing stages.

5.4.1 TIME SPAN

The dataset was collected in October 2010 and covers searches submitted from September 1st to September 30th 2010. September 2010 was chosen so as to have a period of time that included the most common working scenarios for translators and one with a reasonable balance between peak activity and holiday time. More accurate results could have been obtained by considering longer periods (two months or more) and averaging or comparing the results. However, the data volume for one month was challenging enough and the methodology had still to be developed, so the comparison phase has been postponed to a future replication of the study. The main motivations for choosing September 2010 over other months are listed below in greater detail.

- The summer months had to be excluded because EU activity is lower than usual and tends to rely more on trainees due to annual leaves of staff translators.
- In September 2010 there were no official holidays, which allows for a full log of 20 working days without potential disturbing factors.
- September seemed to strike a good balance between holiday time and activity peaks. Many people are still on holiday in the first week (which would represent activity in the summer months) and gradually resume work in the first couple of weeks of the month. The second half is characterized by normal (or peak) activity.
- In Section 5.1, the EC and the EP were found to be the two largest users of Euramis, so they will likely affect the general search trends. Generally speaking, the workload at the EC is fairly constant, so it was important to avoid unusual situations as much as possible. As for the EP, activity usually peaks around session week (i.e. a week-long plenary session in Strasbourg, generally followed by a

[101] Triangulation involves "a weaving of results" (O'Brien 2009: 260) attained using several instruments of data gathering, ranging from qualitative to quantitative approaches, as well as different data processing and analyses (cf. also Alves 2003: vii).

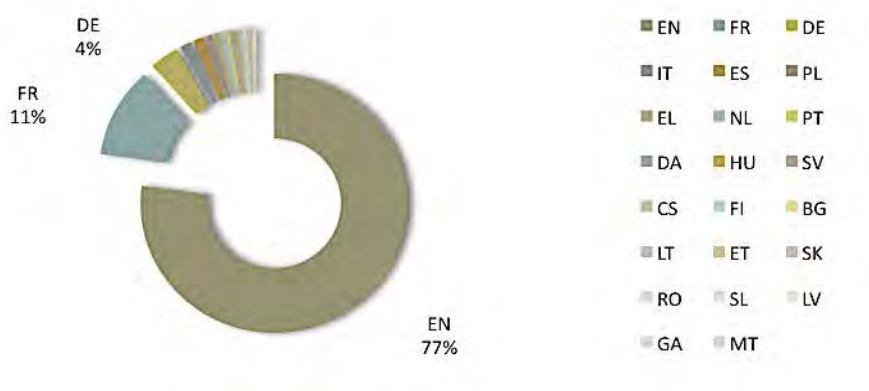
shorter session in Brussels). In September 2010, two Strasbourg sessions were held to make up for the missed August one, implying a heavier workload for the EP.

- The members of the Barroso Commission (2010-2014) took office in early February 2010. A few months were allowed for adjustments to policies and workflows, which ruled out spring months.
- The Treaty of Lisbon came into force on 1 December 2009. Procedural and terminological changes were introduced by the Treaty and a sufficiently broad time span was deemed necessary for the required adjustments to be fully implemented.
- This round of data collection was made in early October 2010. September was the most recent fully available month of logs. Choosing the most recent available logs was advisable in view of the overall length of the research project (i.e. at least two more years) so that the dataset would not become obsolete.
- Originally, a different dataset was analyzed that contained logs from March 2009. At that time, the choice was justified by similar criteria such as the absence of official holidays. In addition, the new Barroso Commission was designated in 2009 and the end of the term of office of the MEPs was approaching. The EP elections had been scheduled for June 2009 and parliamentary activity was likely to be feverish in the previous months. Logs from the second half of the year would have had to be discarded because of the adjustment time for the newly elected EP. Unfortunately, this initial dataset had to be discarded after some time due to a discovered technical glitch that did not record the distribution of searches between the two interfaces (Euramis and Quest), which is a relevant piece of information for the analysis.

5.4.2 DISTRIBUTION OF SEARCHES BY SOURCE LANGUAGE

The initial dataset covered all 23 EU official languages. The first analysis will look at the distribution of searches according to the selected source language (Figure 30). The results will provide information about the relative weight of each language and possibly the most representative source language(s).

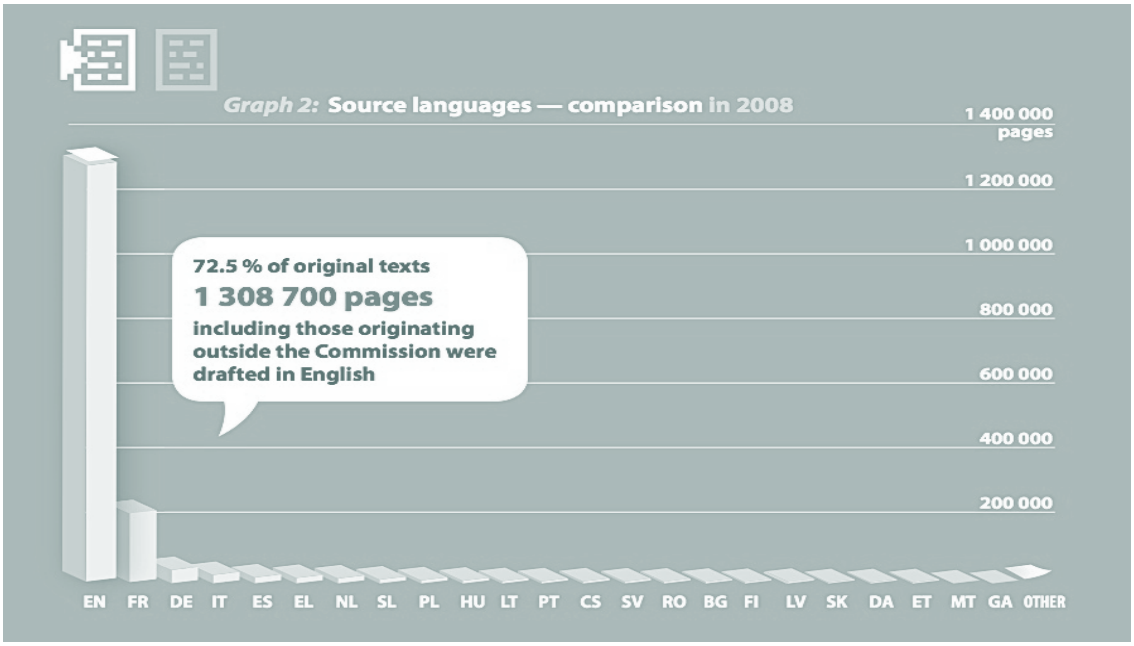
Figure 30. Distribution of total searches (970,000) per source language.



Pre-enlargement statistics about the main source languages of texts translated at DGT in 2003 report that 58.9% of documents were drafted in English, 28.1% in French, 3.8% in German and 8.9% in other languages (Drugan 2004: 9). More recent DGT data from 2008 provide a different overall picture: 72.5% of original texts have English as the source

language, 11.8% were drafted in French, 2.7% in German and 13% in other languages (DGT 2009c: 6). As shown in Figure 31, the overwhelming majority of searches are submitted in English. The other two procedural languages of the EC (French and German) rank second and third respectively, with all remaining languages together accounting for only 8% of the total searches. The figures for Euramis searches closely resemble the 2009 estimates. A finer-grained distribution of source languages (in terms of the number of drafted pages; Figure 31) provides a justification for the very low scores for most source languages in Euramis.

Figure 31. Breakdown of pages translated per source languages at DGT in 2008 (DGT 2009c: 7).

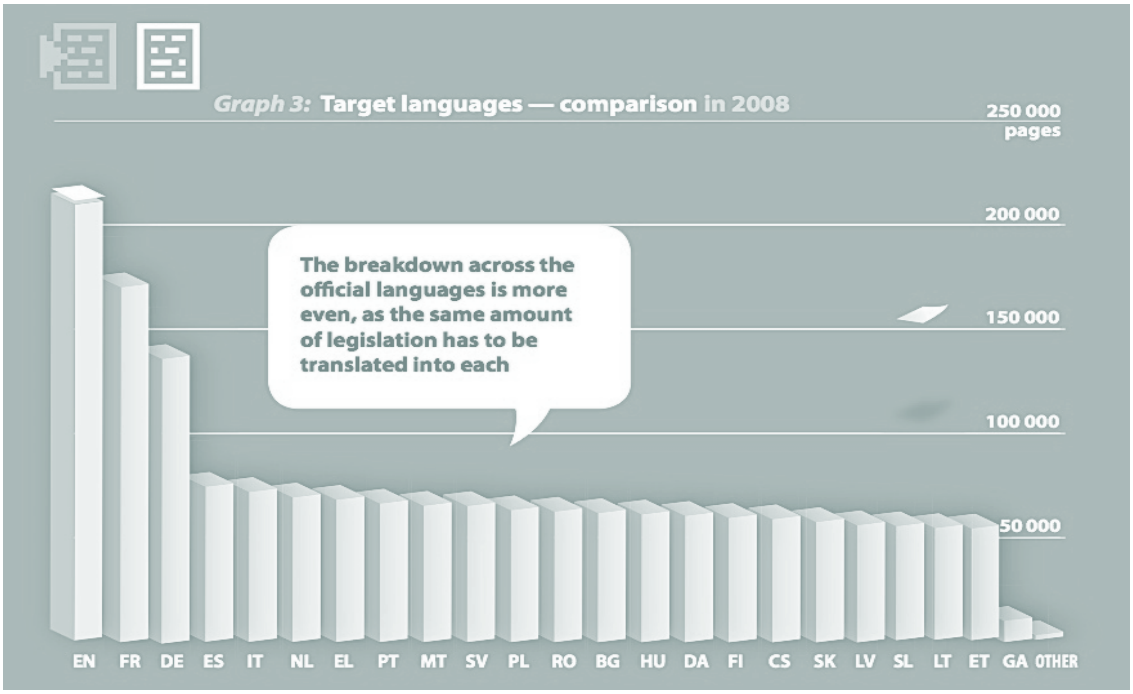


The massive gap between English as a source language and all remaining languages makes any potential comparison unsuccessful because of the irreconcilable size difference with any other source language subset. There simply is not enough data to work with other source languages, unless the English subset is dramatically reduced. For this reason, English was selected as the sole source language, thereby limiting the whole analysis to a single directionality only, i.e. translation from English into the native language of the translator (any of the remaining 22 languages). The whole dataset of searches will only consist of English queries and was consequently brought down to 749,500 searches.

5.4.3 DISTRIBUTION OF SEARCHES BY TARGET LANGUAGE

The analysis of source languages strongly indicated that English should be chosen as the source language for the study. However, an additional check should be made in order to verify the potential data loss in terms of target-language distribution when English is chosen as the only source language. As translation output per translated pages shows (Figure 32), the distribution of target languages is higher and slightly less unbalanced with respect to source languages. Here, too, English ranks as the top target language but for the present analysis it shall not be taken into account in that it cannot be used as source and target language at the same time.

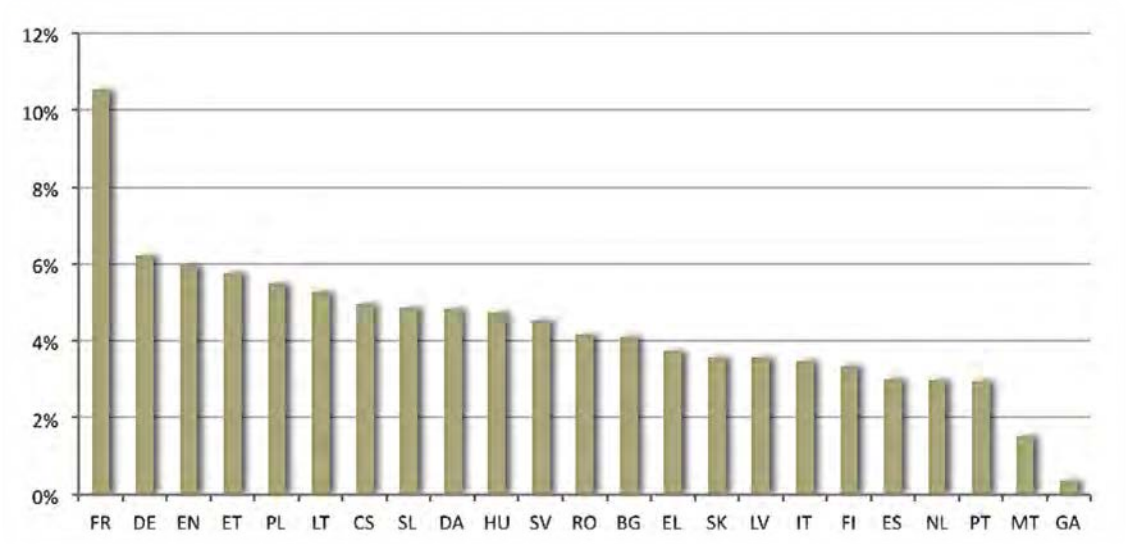
Figure 32. Breakdown of pages translated per target languages at DGT in 2008 (DGT 2009c: 7).



Because of the close correspondence found between concordance searches and the number of drafted pages in terms of source language distribution, a similar pattern can be expected for target languages. Unfortunately, this check cannot be performed directly on the available data. The advanced search mode in Euramis has a feature that enables the selection of multiple target languages. This causes a proliferation of target language combinations that brings the total from 24, i.e. the maximum number of individual target languages (non EU languages included), to 341 different combinations. In order to study target language distribution, the query corpus has to be pre-processed to remove all instances where multiple target languages were selected. Whenever multiple target languages are selected, there is no way to tell which of the selected languages the translator was working into and, by extension, what his/her native language was. This is an important step to ensure the validity of the assumption that the selected target language would match the translator's native language.

After performing this clean-up operation, the total number of strings was brought down to 963,439 from the original 970,000 searches. This subset will be labeled ALL>ALL_1tgt. Figure 33 summarizes the distribution of the searches by target language for the ALL>ALL_1tgt dataset.

Figure 33. Distribution of total searches (963,000) per target language, with one single target language selected per search.

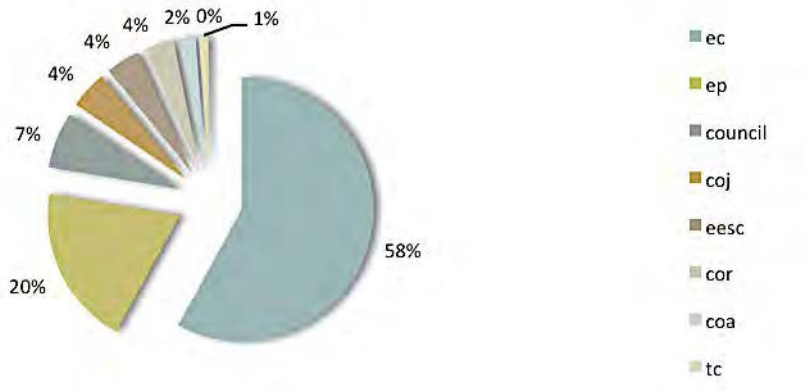


As anticipated, the distribution of target languages is at first sight very similar to the official statistics for the number of translated pages. However, a few notable differences need to be pointed out. The most searched language is French, followed by German at a considerable distance. English only comes third and then the order in which languages appear is very different from the one found in Figure 32. This indicates that there is no clear relationship between the number of pages translated and the number of searches that were performed. It should be noted that the two charts are in fact not directly comparable because one refers to the yearly production of DGT in 2008 and the other to a month's worth of searches from several institutions in 2010. In this overview, large numbers and approximate statistics are used, hence the focus is on potential trends rather than exact findings. Some further explanations for the distributions of source and target languages can be obtained from the analysis of the searches according to the requesting institution.

5.4.4 DISTRIBUTION OF SEARCHES BY INSTITUTION

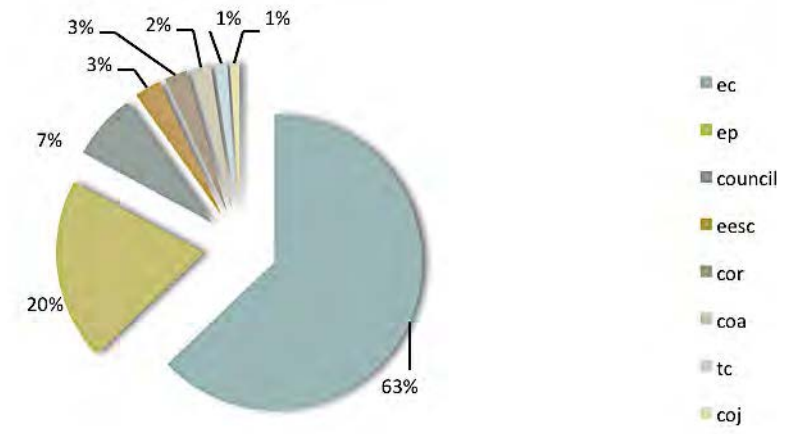
As previously explained (see Section 3.2.3), Euramis and Quest can be accessed by translators working internally at eight different EU institutions having translation units of different sizes. Figure 34 provides the distribution of all searches (ALL>ALL) according to the submitting institution, which was partly discussed in Section 5.1.

Figure 34. Distribution of total searches (ALL>ALL) according to submitting institution.



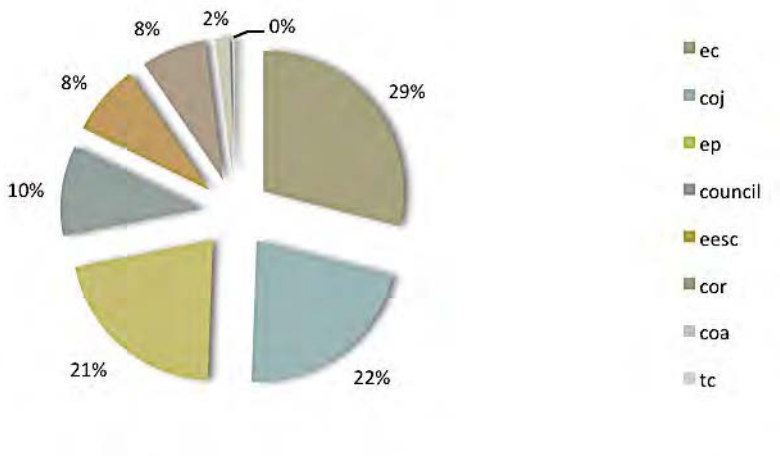
The majority of the searches come from the Commission and in particular from DGT. The EC, the EP and the Council are the top three requesting institutions. The analysis of source languages has already established that English should be the only source language used in the study. An additional check will be performed to ensure that the distribution of searches per institution is not affected by the selection of just one source language. Figure 35 details the distribution of the searches considering English as the only source language (EN>ALL) per EU body. The total searches in this case amount to 749,500 queries covering 77.16% of the total searches.

Figure 35. Distribution of requesting institution for EN>ALL (749,500 queries).



A striking similarity can be noted between the pie charts for the EN>ALL distribution (Figure 35) and the ALL>ALL distribution (Figure 34). The same ranking can be found for all institutions but one: the Court of Justice (COJ). In the overall (ALL>ALL) distribution, COJ ranked fourth, but came last in the EN>ALL distribution. In order to account for this phenomenon, a comparable distribution will be examined where the sole source language is French, i.e. the second most requested source language, which accounts for 11.26% of the searches, or 109,336 FR>ALL queries (Figure 36).

Figure 36. Distribution of requesting institution for FR>ALL (109,336 queries).



In this case, the distribution looks quite different from the previous two. COUNCIL, EESC and COR have all increased by a few percentage points. At a closer look, the ranking is the same as before in all cases except for the Court of Justice. The COJ percentage moved from 4% in the ALL>ALL distribution to 22% for FR>ALL, while searches coming from the EC have halved (29%). This apparently odd behavior of the Court of Justice can be easily explained by looking at its internal structure and work practices. Within the Court of Justice, the Translation Directorate General is the largest service, employing 46% of the total staff in 2008 (i.e. 876 people), according to internal statistics[102]. All language combinations for the official languages have to be covered and the translation volume exceeds 700,000 pages per year; all texts are of a legal nature with highly technical content. The Directorate General is further divided into two Directorates, one comprising the language units for BG, CE, DA, DE, ET, EN, FI, LT, MT, NL, PL and SV; the other dealing with ES, EL, FR, IT, LV, HU, PT, RO, SK and SL. The French language unit is the largest because all legal documents have to be available in that language, which is the main working language of the institution (Cosmai 2007: 83).

In practical terms, choosing English as the sole source language penalizes some institutions (in this case, the Court of Justice) where the distribution of source languages is more diversified. However, this loss can be put into perspective by considering total counts: the volume of searches from English is about 7 times the amount of searches from French. Had the total sums for EN>ALL and FR>ALL been comparable, searches from the COJ would have amounted to some 165,000 searches (i.e. 22% of 749,500) but in fact searches from French "only" accounted for 24,000 queries (i.e. 22% of 109,336), which is only about three times bigger than the subset of COJ searches from English, i.e. 7,500 (1% of 749,500). This means that all other source languages, which are proportionately more frequent in institutions other than the EC, would still be very much underrepresented in the whole dataset. Source languages other than English can thus be removed without incurring in major data loss, as can be easily seen by comparing the pie charts for ALL>ALL (Figure 34) and EN>ALL (Figure 35).

[102] Statistics obtained from http://curia.europa.eu/jcms/jcms/Jo2_10742/direction-generale-de-la-traduction [last accessed: December 2012].

5.5 DATASET PRE-PROCESSING

Once English was selected as the sole source language, the new main dataset became EN>ALL and amounted to 749,500 searches. This dataset was obtained by filtering the initial ALL>ALL dataset by source language but no further operations were carried out on the target language side. Multiple target languages had been previously identified as problematic for the analysis (see Section 5.4.3) so they had to be removed before the strings could be analyzed[103]. In addition to the cleanup of multiple target languages, additional pre-processing operations on the EN>ALL dataset were deemed necessary to remove other forms of noise as much as possible. Additional sources of noise were non-official EU target languages and mismatches between the selected source language and the actual language of the search. The preprocessing and cleanup operations can be summarized as follows:

1. REMOVAL of MULTIPLE TLs

   Because of the assumption that translators are working into the target language they select for the search (i.e. their native language), multiple target languages make it impossible to determine the relevant language for the statistics. Searches were removed that contained the following elements in the *Tgtlang* field:

   a. "*" [i.e. ALL available languages were selected at once]

   b. COMMA and/or WHITE SPACE [i.e. there was more than one element in the field, hence multiple languages]

2. REMOVAL of INDIVIDUAL TLs
   In a few occasions, some non-official EU languages were selected and in some other cases source and target languages were made to match, which turned the search operation into a monolingual query. Given the nature of the material in the repository (i.e. parallel aligned segments in different language pairs), a monolingual search would not be useful. In this case, searches were removed that matched the following filtering criteria in the *Tgtlang* field:

   a. *TGTLANG* = HR (Croatian) or TR (Turkish) [i.e. non-official EU languages, as of 2010]

   b. *TGTLANG* = EN [i.e. SL same as TL]

After this pre-processing, the total number of strings was brought down to 743,611 and the dataset was renamed EN>ALL_1tgt. An additional manual cleaning was performed to remove the searches where the selected and actual source language did not match, starting with the longest text strings. After a first round of manual cleaning, the EN>ALL_1tgt dataset was brought down to 742,033 searches and a second round further reduced it to 740,000 strings. This new EN>ALL_1tgt dataset included English as source language and 22 of the 23 official EU languages as targets (as English only counts as SL). The overall distribution of searches for the 740,000 dataset is summarized in Table 5

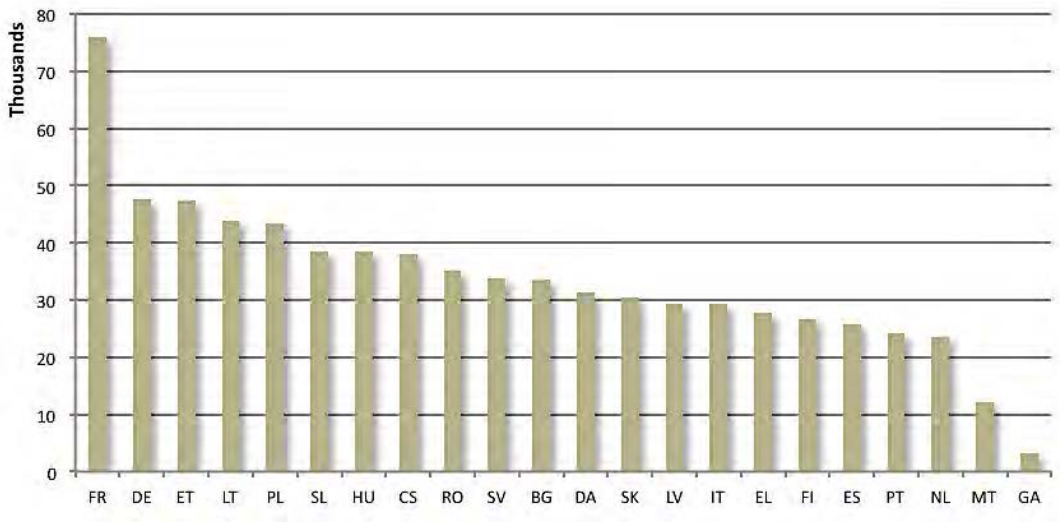
[103] Multiple target languages can be interesting as a separate subset. Recurrent language combinations that translators select in the searches could be analyzed to reflect either the working languages and language combinations of the translator or possible criteria for finding additional support by using multiple languages (i.e. language similarity, even without working knowledge of one or more of the selected languages). However, for the purposes of the present study, this type of search does not provide any useful information.

whereas the chart in Figure 37 contains a graphical representation of the same distribution per target language. The chart closely resembles the distribution found in Figure 33 and confirms that the overall distribution of the searches is not particularly affected by pre-processing the corpus.

Table 5. Distribution of searches per target language in ascending order (740,000 dataset).

TL	Count	% (of total)	TL	Count	% (of total)
GA	3,246	0.44%	BG	33,508	4.53%
MT	12,211	1.65%	SV	33,826	4.57%
NL	23,594	3.19%	RO	35,075	4.74%
PT	24,173	3.27%	CS	38,064	5.14%
ES	25,880	3.50%	HU	38,512	5.20%
FI	26,765	3.62%	SL	38,527	5.21%
EL	27,812	3.76%	PL	43,431	5.87%
IT	29,270	3.96%	LT	43,942	5.94%
LV	29,407	3.97%	ET	47,403	6.41%
SK	30,422	4.11%	DE	47,617	6.43%
DA	31,266	4.23%	FR	76,049	10.28%
Total	**740,000**	**100.00%**			

Figure 37. Graphical representation of the distribution of searches per target language (740,000).



Two outliers can be immediately identified in Figure 37: French and Gaelic. The former is about 1.5 times as large as the second most popular target language, while the latter is about four times smaller that the one before last (Maltese), which in turn was almost half the size of the previous language (Dutch). Clearly, individual language subsets are not directly comparable in absolute terms. There is a delta of over 10 percentage points between the first and the last language, which amounts to over 70,000 searches. The delta could be reduced considerably (by about 3 percentage points, i.e. some 24,000) if the target languages considered ranged from DE to NL. The gap between the first and the last language in the smaller group would still be relatively high but the delta between two adjacent languages is proportionately very small, allowing for some straightforward

comparisons. French and German have the two largest subsets, which comes as no surprise due to the large volumes of translations into these two procedural languages for DGT.

The third largest subset is Estonian, which is an unexpected outcome. Intuitively, this can be related to two phenomena: either ET translators have large volumes to translate or ET translators are simply very active users of Euramis submitting on average more queries than translators of other non-procedural languages. According to information collected at DGT, the workload for each official language is usually approximately the same. The second hypothesis seems more plausible: the target languages in the top half of the list can be considered more active users of Euramis.

If comparisons across languages are carried out using simple frequency counts for the analyzed phenomena, languages with larger datasets are likely to have higher frequency counts for the phenomenon under analysis. This is why it is important to avoid as much as possible discussing data based on simple frequency counts when the language variable is considered. To ensure that even the relatively small delta between each language pair is taken into account, the vast majority of inter-language comparisons will be carried out using percentage values, with counts normalized by the total number of strings for each language, if not otherwise specified in the course of the analysis.

5.5.1 THE FINAL DATASET

Preliminary analyses and small pilot studies showed that despite the prediction that languages with smaller datasets would be penalized, Gaelic and Maltese behaved very oddly. Their very small datasets, compared to those of other languages, were probably the cause of the odd results, with ratios placing them consistently either at the very top or at the very bottom of the ranking. On the other hand, French is not perceived as problematic because a larger dataset is not considered as likely to negatively impact French with respect to the other languages. Gaelic and Maltese were eventually removed from the final dataset for the analysis because the delta was too marked and there was no way to increase the size of the respective subsets. Without GA and MT, the main dataset was further reduced to 724,500[104] searches and 20 target languages; this will be referred to as the overall (or main) dataset (i.e. EN>ALL_1tgt_20). In the course of the study, additional *ad hoc* subsets may be generated or interim versions of the final dataset used. In either case, the applicable dataset will be detailed in the description of the relevant analysis.

Table 6 contains the final distribution of languages in the final dataset and clearly shows that percentage values are higher compared with the previous table for the 740,000 dataset (Table 5). However, the delta between German and Dutch has remained at 3.3% after removing GA and MT, showing how little an impact these two languages had in statistical terms.

[104] To be precise, there is an additional delta of 43 strings between the original 22-language dataset and the new 20-language dataset. This is due to the additional deletion of 43 noisy strings (where the SL was not English) from the 20-language dataset which were randomly selected across several languages. This was also meant to make the total count of the 20-language dataset a round number to make calculations more convenient. Given the extremely small delta, overall percentages were not affected.

Table 6. Target language distribution for the final dataset (724,000) in ascending order.

TgtLang	Count	% on tot	TgtLang	Count	% on tot
NL	23,570	3.3%	SV	33,826	4.7%
PT	24,170	3.3%	RO	35,075	4.8%
ES	25,880	3.6%	CS	38,064	5.3%
FI	26,765	3.7%	HU	38,510	5.3%
EL	27,810	3.8%	SL	38,520	5.3%
IT	29,270	4.0%	PL	43,431	6.0%
LV	29,407	4.1%	LT	43,942	6.1%
SK	30,420	4.2%	ET	47,400	6.5%
DA	31,266	4.3%	DE	47,617	6.6%
BG	33,508	4.6%	FR	76,049	10.5%
Total	724,500	100%			

One of the research questions (see Section 1.1) mentions a possible differentiation of translation problems based on a specific language pair, i.e. different languages would experience different types of translation problems because of some recurring language specific transfer issues. In particular, languages belonging to the same language family (e.g. Germanic) may highlight similar problems in a translation from English and these may differ from problems experienced by another language family (e.g. Romance). Consequently, grouping target languages by their language family may highlight some recurring search patterns to compare translation problems more easily.

5.5.2 LANGUAGE FAMILIES

Contrastive studies usually provide interesting information to build hypotheses on potential difficulties in translation. The smaller the linguistic distance, the more straightforward the translation task can be generally expected to be. This was also noted in a study conducted by the European Commission within the MT@EC project (Eisele & Lavecchia 2011: 5-7). In May 2011, a so-called "Maturity Check" was conducted to assess the quality of machine-translated sentences from English into all 22 EU official languages. The study highlighted some of the main problems shared by languages belonging to the same language family. Inflection was considered problematic for Romance languages; Germanic languages shared the problem of translating "composita" (compounds), whereas the remaining families shared the problem of inflection in addition to strong agglutination for the Finno-Ugric family. These problems were all identified in the output of machine-translated texts, which is not directly comparable to the difficulties experienced by human translators. This nonetheless suggests that there may still be some common difficulties for human translators that pertain to a specific language family. Table 7 groups the target languages of the final dataset according to their respective language families[105].

[105] The Semitic and Celtic families are missing because the respective languages, i.e. Maltese and Gaelic, have been left out in a previous stage.

Table 7. Language family distribution for the languages involved in the study.

Language Family	Target Language(s)
Baltic	LT, LV
Finno-Ugric	ET, FI, HU
Germanic/Scandinavian	DE, DA, NL, SV, (EN)
Hellenic	EL
Romance	ES, FR, IT, PT, RO
Slavic	BG, CS, PL, SK, SL

The distribution into language families is clearly uneven. Some families (e.g. Slavic) count as many as five elements, whereas the Hellenic family only has one member. While this may not be a problem per se, an uneven distribution makes studying phenomena across languages more difficult because results will inevitably be evened out in the case of larger families. In the case of Greek, on the other hand, the behavior of the language family and the single language will coincide.

The phenomenon can be exemplified with a cross tabulation of the main variable 'target language' grouped by language family and the submitting institution, which can be assumed to deal with different text types, at least to some extent:

> A cross tabulation is a joint frequency distribution of cases according to two or more classificatory variables. The display of the distribution of cases by their position on two or more variables is the chief component of contingency table analysis and is indeed the most commonly used analytical method in the social sciences. (Nie *et al.* 1975: 218).

A cross tabulation would highlight the distribution of searches per language family according to the submitting institution, thus looking at the overall distribution instead of individual languages as in Section 5.4.4, when only EN and FR were analyzed as source languages. Keeping all 20 languages isolated within each institution would have made data interpretation very challenging while language families seem to serve the purpose of studying common phenomena.

Figure 38. Distribution of searches grouped by language family across each institution, normalized by the total number of searches for each institution (724,000).

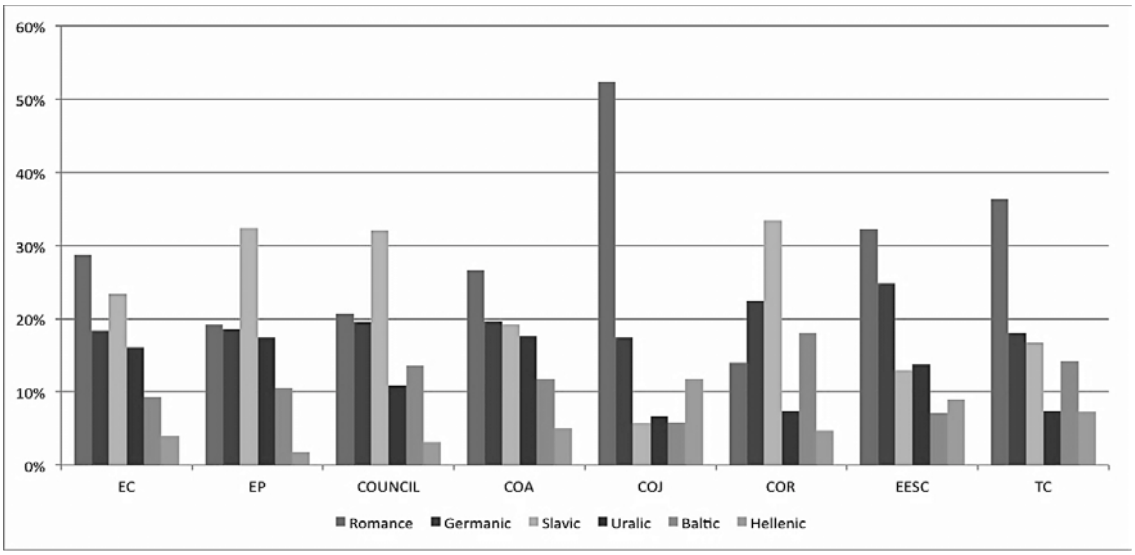
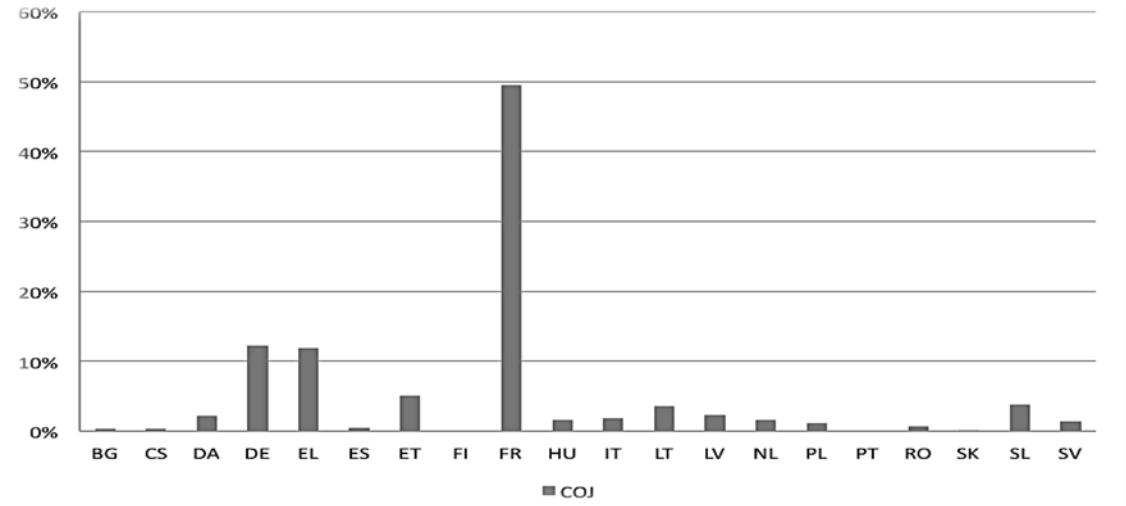


Figure 38 clearly shows that the most searched-for language families are either Romance or Slavic languages. The gap between the first and the second language family is most striking at the COJ, but also TC, COUNCIL and EP register a quite steep fall after the top family group. Greek is quite popular at the COJ but only accounts for a very small percentage in the other institutions. Uralic languages seem quite popular at COA, EP and EC. Baltic languages are the most searched-for at COR, TC and COUNCIL. Germanic languages seem most problematic at EESC and COR. Slavic languages are by far the top group at COR, COUNCIL and EP, whereas Romance languages at COJ and TC. The most balanced distributions are possibly found at COA and EC.

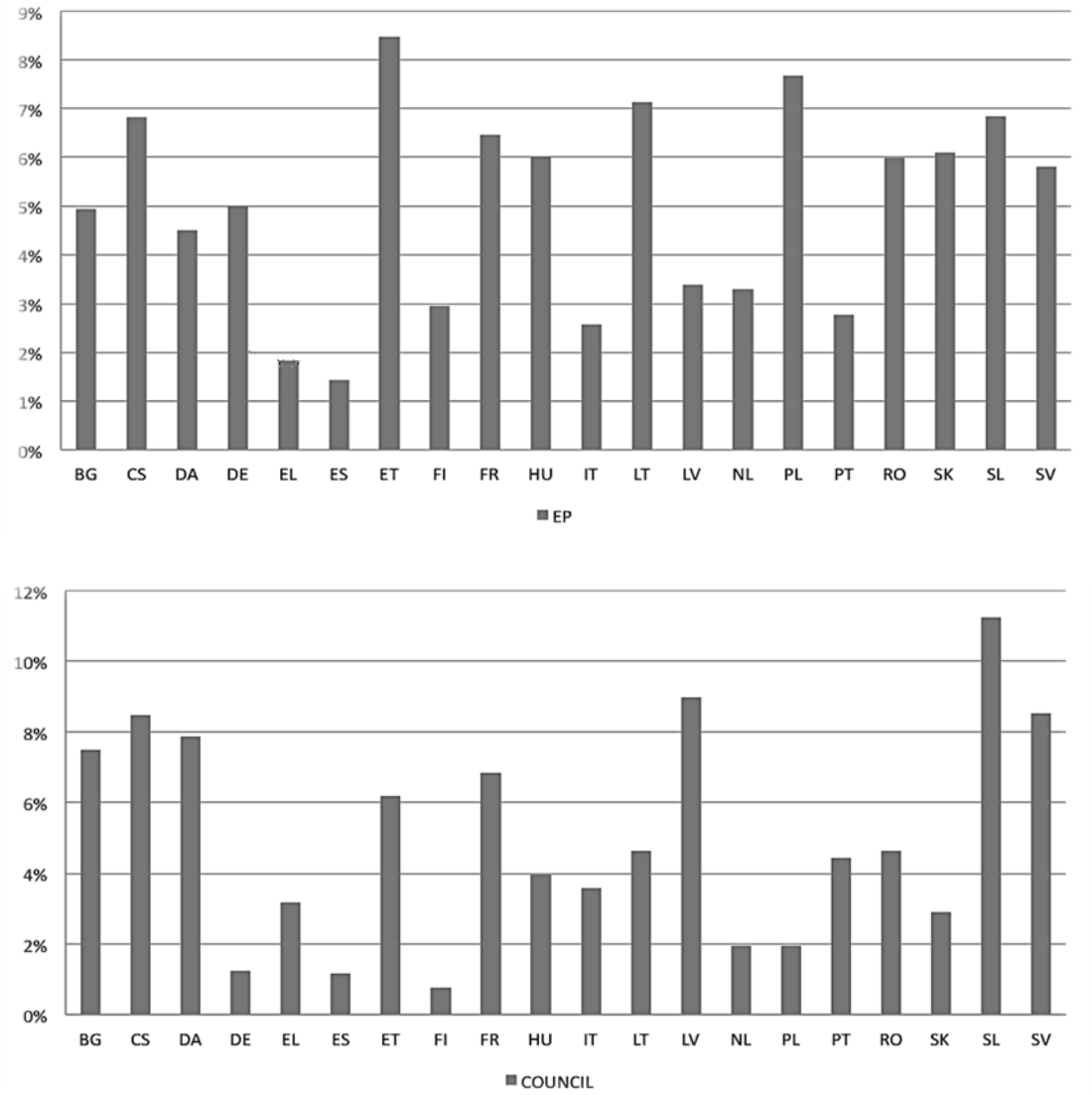
The most popular language families at each institution are either Romance or Slavic languages and each comprises four to five languages. A closer look into the individual distribution of all 20 languages within one institution may help shed light on the effects that the language family grouping may have on the analysis. Building on the previous findings for the Court of Justice, it can be hypothesized that French, as the main working language, plays a major role not just as a source but also as target language. The most popular language family at COJ is the Romance family, so French will likely be the most searched- for target language and Figure 39 confirms this assumption. Romance languages other than French are hardly searched for whereas Greek (EL) stands out as it did in the previous chart. Stating that Romance languages were (equally) problematic for translators at COJ would have been misleading because in fact French is almost the only Romance language that is searched for.

Figure 39 Distribution of target languages within the institution COJ.



Analyzing COJ was relatively easy because there was enough background information to make initial assumptions and data interpretation was rather straightforward. A similar test can be made with other institutions and language families to try and see whether similar results in Figure 38 would in fact match the actual distribution of individual languages. EP and COUNCIL have similar percentages of searches into Slavic languages. The distribution of individual languages for both institutions is shown in Figure 40.

Figure 40. Distribution of target languages within two institutions: EP and COUNCIL.



The most popular Slavic language at EP is Polish. This is not the case at the COUNCIL, where Slovenian is the most searched-for target language. Slovenian (SL) ranks also quite high at the EP, just like Czech (CS) which is also the second most searched Slavic language at the Council. A considerable difference can be found for Slovak, very common at the EP and less so at the Council. Apart from the Slavic family, Estonian turns out to be the number one language at the EP but this was not evident from the aggregated chart where Uralic languages ranked fourth and the aggregated results for Slavic languages took precedence. This small analysis confirmed that there are some imbalances within the language families in terms of language distribution. While language families prove to be a very effective way to graphically represent language distribution in one single chart for a general impression, caution should be exercised when discussing the results because distributions for the same language family are uneven across institutions.

5.5.3 LANGUAGE 'AGE'

Preliminary analyses conducted on the old March 2009 dataset suggested that there may be an additional criterion for grouping the target languages: their relative age as official EU languages. A European language becomes an official EU language when its country is accepted as a new Member State. In 2004, a major enlargement took place and the EU acquired 10 new Member States. Some adjustments to the language policy of each institution were introduced so as to cope with a major increase in language combinations. Figure 41 shows the increase in the number of official languages over time. 2004 is taken as a watershed because of the impact the enlargement had on the internal organization of the institutions. Ten of the languages under examination are found in the pre-2004 group (i.e. the "old" languages) whereas the other half belongs to the post-2004 group (i.e. the "new" languages), as shown in Table 8.

Figure 41. List of European languages and year in which each language became an 'official language' of the EU. Note the major increase by 9 languages in 2004 (EC 2008: 3).

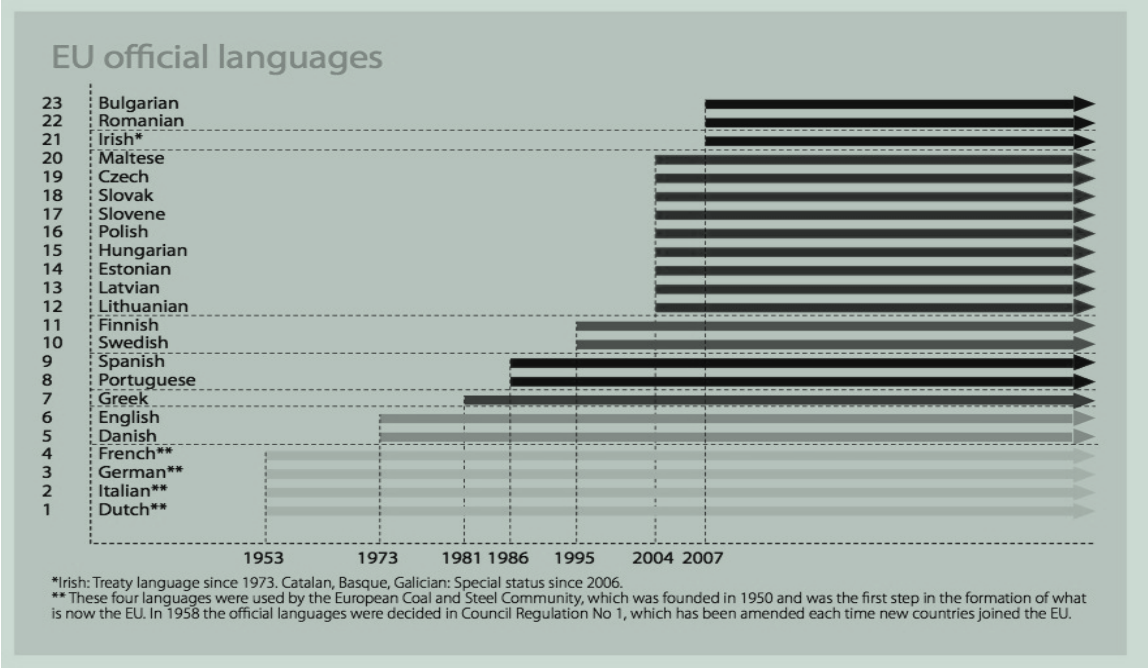


Table 8. Distribution of target languages according to their relative age as official EU languages with respect to the 2004 enlargement.

Pre-2004, or "old", languages	DA, DE, EL, ES, FI, FR, IT, NL, PT, SV
Post-2004, or "new", languages	BG, CS, ET, HU, LT, LV, PL, RO, SK, SL

In some cases, all languages from the same family belong to the same age group; in other cases the situation is mixed. When results are discussed, languages that fall within the same age group are harder to analyze because it cannot be clearly established which aspect (age or family) plays a bigger role in the results, and boundaries between the two features are often hard to draw. Table 9 provides an overview of the distribution of language family and age using the same ascending ranking as the one for subset size.

Table 9. Target language distribution for the final dataset (724,000) in ascending order and distribution of language family and relative age of the language in EU terms.

TgtLang	Count	% on tot	Family	2004
NL	23,570	3.3%	Germanic	pre
PT	24,170	3.3%	Romance	pre
ES	25,880	3.6%	Romance	pre
FI	26,765	3.7%	Finno-Ugric	pre
EL	27,810	3.8%	Hellenic	pre
IT	29,270	4.0%	Romance	pre
LV	29,407	4.1%	Baltic	post
SK	30,420	4.2%	Slavic	post
DA	31,266	4.3%	Germanic	pre
BG	33,508	4.6%	Slavic	post
SV	33,826	4.7%	Germanic	pre
RO	35,075	4.8%	Romance	post
CS	38,064	5.3%	Slavic	post
HU	38,510	5.3%	Finno-Ugric	post
SL	38,520	5.3%	Slavic	post
PL	43,431	6.0%	Slavic	post
LT	43,942	6.1%	Baltic	post
ET	47,400	6.5%	Finno-Ugric	post
DE	47,617	6.6%	Germanic	pre
FR	76,049	10.5%	Romance	pre
Total	724,500	100%		

Aside from FR, Romance languages tend to appear in the top half of the list (where subsets of searches are smaller), whereas Slavic languages seem to occupy the lower half (where subsets are larger). The remaining families (Germanic, Finno-Ugric and Baltic) seem evenly distributed, possibly as a consequence of the distribution of languages according to the age criterion. As previously noted, languages belonging to the same family in some cases also belong to the same age group (e.g. Slavic and Germanic languages). When considering the age column, most of the lower half of the ranking is populated by "new" languages (i.e. post-2004), whereas most of the "old" languages (i.e. pre-2004) appear in the top half where datasets are smaller. This suggests that older languages tend to submit concordance searches less frequently than post-2004 languages. This finding can be due to a number of factors. First, new languages can be said to encounter more problem than older languages, which is understandable because younger generations of translators are not as experienced as translators of pre-2004 languages. At the same time, however, younger translators may be more familiar with computerized translation tools and tend to submit queries more systematically than translators who had a different translation training and/or have different approaches to the translation task. If all languages are instead assumed to experience a comparable number of problems, this clustering might be explained by that fact that translators into "old" languages turn to resources other than Euramis whereas "new" languages may see Euramis as their main resource. A third explanation would look at the size of the translation memories for each language in Euramis. Given the age difference, one can assume that TM repositories for "old" languages are (much) bigger than those for new

languages, hence the pre-translation phase and the project-related translation memories produce better results in terms of text coverage than they do for new languages, and some pre-2004 material is likely not to be available in the "new" languages. This third hypothesis would also find its justification within the research field of NLP and more specifically Machine Translation, where a number of the new languages would probably still be considered as "under-resourced" languages. However plausible, the third hypothesis has to be discarded after examining the size of the Euramis database per target language (Table 10).

Table 10. Size of Euramis TM databases per each target language in terms of number of stored segments as of 4th January 2010 (11 TMs from Euramis included).

BG	CS	DA	DE	EL	ES	ET	FI	FR	GA	HU
6,894,617	6,771,911	7,344,720	6,810,395	5,262,780	7,485,710	8,181,431	8,201,125	7,267,604	306,487	7,275,624
IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
6,110,942	8,199,356	6,289,393	6,110,631	9,200,427	7,127,518	13,046,404	6,226,266	7,244,807	8,920,486	9,345,120
Total TGT segments			161,539,702			**Total SRC segments**			38,323,579	

After looking at Table 10, removing Gaelic was a justified choice given the huge gap in database size compared to all other languages. Despite an average database size, Maltese had also to be discarded because of the lack of a sufficient number of searches. The hypothesis that there may be a relation between the amount of searches per language (size of the subset) and the size of the corresponding Euramis TM might hold true only for Dutch and Portuguese, which ranked last and one-before-last respectively for amount of Euramis searches but came at the top with respect to database size. A large TM increases the chances of finding matches either in the pre-translation phase or during the translation with the CAT tool, thus reducing the need for manual search in the concordancer. However, there does not seem to be any clear relation between TM size and the amount of searches: for instance, ET has a very populated database but is the third most searched-for language in Euramis after French and German.

This overview served the purpose of presenting the dataset and explaining each step of the pre-processing that was carried out on the original dataset in order to obtain the final dataset of 724,000 searches that will be analyzed in the following chapters. From now on this final 724,000 dataset will be referred to as *main* or *whole* dataset and will be abbreviated as EN>ALL, unless otherwise specified. Similarly, any reference to *'overall* XY' by default refers to the 724,000 dataset.

5.6 SEARCH SESSIONS

While pre-processing the dataset, large volumes of searches were scanned through to the point where some recurring search patterns could be identified. They can be summarized in four main scenarios:

1. A one-time search with or without settings/filter selection;
2. A repeated search where the exact same query is resubmitted more than once;
3. A repeated search with the same search string but with different settings/filters;
4. A repeated search with a different search string, with or without some changes in the settings.

A repeated search means that more time was devoted to finding a solution to a translation problem, suggesting that there was one information need underlying the repeated

searches. One information need is expected to trigger a search episode, i.e. an interaction between the user and the system, which can consist of one or more individual searches. Web search log analysis has used the concept of *query reformulations* or *query refinements* (Huang & Efthimiadis 2009: 77-78) to account for this phenomenon, which involves "a modification to a search query that addresses the same information need". Modified queries can be grouped together to form a search session, which is defined as "a series of queries submitted by a user and related interactions during an episode of interaction between the user and the Web search engine around a single topic" (Jansen *et al.* 2009: 1361). As briefly explained in Section 4.5, Web log analysis uses some metadata (i.e. user ID, IP address, cookies, time stamps) to identify search sessions. Euramis does not log the same metadata as Web search engines but logs all searches using the user log-in details, which in the present study had to be removed to ensure anonymity (see Section 3.2.4). In the following sections, operational criteria will be developed to try and identify search sessions in the concordance search logs using available data, in compliance with the above definition. To better exemplify the type of regularities in the searches, some examples have been provided in Table 11.

Table 11. A small excerpt of search logs from the Polish subset showing instances of queries.

Line	Date	Time	Inst[106]	Quest	TgtLang	Results	Query
1	9/1/2010	10:51:31,176	cdt	TRUE	PL	30	Staff Policy Plan
2	9/1/2010	10:51:31,527	cdce	TRUE	PL	0	sworn declaration that
3	9/1/2010	10:51:32,066	ep	FALSE	PL	8	Territorial scope of the Regulation
4	9/1/2010	10:51:39,332	cdce	TRUE	PL	9	sworn declaration
5	9/1/2010	10:52:04,017	consil	TRUE	PL	30	enshrines
6	9/1/2010	10:52:24,756	cms	TRUE	PL	4	Wireless infrastructure
7	9/1/2010	10:52:26,377	cms	TRUE	PL	30	effective supervision
8	9/1/2010	10:52:29,191	cms	TRUE	PL	11	modern foreign languages
9	9/1/2010	10:52:46,985	ep	FALSE	PL	3	Material scope of the Regulation
10	9/1/2010	10:52:54,549	cms	TRUE	PL	1	modern foreign language
11	9/1/2010	10:53:04,463	cms	TRUE	PL	3	modern language
12	9/1/2010	10:53:10,036	ep	TRUE	PL	0	Bachelor of Engineering science
13	9/1/2010	10:53:13,176	cms	TRUE	PL	7	modern languages
14	9/1/2010	10:53:18,938	ep	TRUE	PL	30	Bachelor
15	9/1/2010	10:53:35,881	cdr	TRUE	PL	30	state budget
16	9/1/2010	10:53:37,436	cdr	TRUE	PL	30	state budget

Only logs from the PL subset are displayed because the target language is used to separate the logs into operational subsets, but similar patterns can also be found in the main dataset. Displayed logs refer to the first day of the month and are chronologically ordered according to the time stamp, i.e. they appear in the order in which they were received by the system. At a closer look, the 16 queries were all submitted within a time span of 2 minutes. The third column contains the institution code and in this case represents 6 out of 8 institutions. When the *Query* column is considered, recurrences in the searches seem to emerge which can be of two kinds, the string is exactly the same (e.g. "state budget") or different to some degree, e.g. "modern (foreign) language(s)" and "sworn declaration

[106] Institution abbreviations in the logs are often based on the French version. More specifically, cms = EC, cdr = COR, cdt = TC, cdce = COA.

(that)." These instances are not exactly the same but can nonetheless be considered to be similar due to a certain overlap in the words they contain. In a short string containing e.g. two words, the minimum overlap would need to be one word[107]. However, only in one case (lines 10 and 11) were the strings consecutive; in all remaining instances there were other strings in between as many users were using the system at the same time and if more target languages were considered, the strings in question would likely be even further apart. Intuition tells that these strings do belong together but proximity in the list cannot obviously be a sufficient criterion for isolating sessions. If metadata for recurrent strings are compared, the strings "modern (foreign) languages" turn out to be all submitted on the same day from the EC whereas "sworn declaration" came from the Court of Auditors. If time stamps are included, it emerges that the queries in the first group were submitted within about 30 seconds, whereas the latter within less than 10 seconds. In sum, target language, date and time, institution code and query text (provided there is at least one word in common) are pieces of information that can be used to extract search sessions. Search sessions are assumed to be handled by the same user to satisfy one information need, i.e. to solve one problem. In order to isolate a search session, the following four conditions have to be simultaneously met:

1. The searches must come from the same institution
2. The searches must be submitted on the same day
3. The searches must be submitted within a two-minute time span[108]
4. The search string must have at least one word in common with the next string or the one after that (excluding stop words such as 'a', 'and', 'by', 'for', 'of', 'the', 'to')[109]

A customized script in PHP[110] was employed to apply these extraction rules on each language subset. Two output files were generated for each language: one containing all strings that matched the criteria, labelled *search session*; the other containing all the remaining strings that could not be grouped into a search session and labeled *spot searches* — a one-time search event. However, not all strings contained in the spot file were necessarily one-time searches. There were instances of search sessions that could not be identified because one or more criteria were not met, e.g. the strings were too far apart from each other. After re-running the PHP script on a Spot file, additional sessions could be retrieved. The difference was however considered negligible and no further systematic re-runs of the script were attempted. Statistics were generated for each language subset considering the ratios of sessions and spot searches, as summarized in Table 12.

---

[107] A comparable example would be "Bachelor (of Engineering science)" in lines 12 and 14 where "Bachelor" is the only word in common.

[108] The time span of two minutes was arbitrarily chosen after studying the logs because it was felt that it was large enough to accommodate not only most of the actual sessions, which are usually terminated within less than a minute, but also sessions that lasted longer but whose logs were not necessarily stored one after the other without intervening strings. Longer spans would run the risk to increase noise in the results. This time span also falls within the range identified by Désilets *et al.* (2008b: 3) i.e. 30 seconds up to 5 minutes, which represents the time it can take a translator to retrieve a matching segment (and its translation).

[109] The fourth criterion can be seen as a proximity rule with some additional matching constraints. It nonetheless allows for the identification of long sessions because every string will still be considered, making it possible to extract sessions longer than three strings. Some stop words were ignored in an attempt to reduce noise and false positives.

[110] The author is indebted to Antonio Farina, Head of Engineering at Translated s.r.l. (Rome, Italy), for his help in writing the script.

Table 12. Overview of the distribution of search Sessions and Spot searches across all languages.

Lang	Main Subset	Session Sub.	Ratio Session	# Sessions	Av.Sess.Lgth	Spot Subset	Ratio Spot
BG	33508	15638	46.67%	6029	2.59	17870	53.33%
CS	38064	13180	34.63%	5530	2.38	24884	65.37%
DA	31266	9937	31.78%	4276	2.32	21329	68.22%
DE	47617	14766	31.01%	6406	2.31	32851	68.99%
EL	27812	9409	33.83%	4016	2.34	18401	66.16%
ES	25880	9610	37.13%	4014	2.39	16270	62.87%
ET	47403	16472	34.75%	6977	2.36	30928	65.24%
FI	26765	9465	35.36%	4060	2.33	17300	64.64%
FR	76049	19454	25.58%	8582	2.27	56595	74.42%
HU	38512	14246	36.99%	5892	2.42	24264	63.00%
IT	29270	10722	36.63%	4504	2.38	18548	63.37%
LT	43942	18151	41.31%	7462	2.43	25791	58.69%
LV	29407	11306	38.45%	4740	2.39	18101	61.55%
NL	23594	8226	34.86%	3441	2.39	15344	65.03%
PL	43431	14676	33.79%	6243	2.35	28755	66.21%
PT	24173	9894	40.93%	4025	2.46	14276	59.06%
RO	35075	13246	37.76%	5500	2.41	21829	62.24%
SK	30422	11988	39.41%	4971	2.41	18432	60.59%
SL	38527	13439	34.88%	5661	2.37	25081	65.10%
SV	33826	12149	35.92%	5071	2.40	21677	64.08%
TOTAL	724543[111]	255974	35.33%	107400	2.38	468526	64.67%
MEAN	36227.15	12798.7	36.08%	5370	2.39	23426.3	63.91%
ST_DEV	11933.44	3100.7	4.33%	1326.9	0.067	9363.6	4.33%
CV	0.329	0.242	0.120	0.247	0.028	0.400	0.068

The first thing to be noticed from Table 12 is the generally even distribution of queries across languages between the two groups in percentage terms (see *Ratio Session* and *Ratio Spot*). This is confirmed by the low value of both standard deviation (SD) and the coefficient of variation (CV).

5.6.1 STANDARD DEVIATION AND COEFFICIENT OF VARIATION

Standard deviation is one of the three main measures of variability together with range and variance (Oakes 1998: 6) and is obtained by extracting the square root of variance, which measures the distance of every data point from the mean (one of the measures of central tendency for a dataset). In this case, there are two measures of standard deviation in the ratio column (4.33%) but they both derive from the same variable because percentage values for each language for sessions and spot searches are complementary, hence SD value is the same. As a rule of thumb, "the smaller the SD in relation to the mean, the less dispersed the data is, that is, the closer individual values are to the mean"

[111] See note 104 of this chapter for an explanation for the "extra" 43 strings.

(Rasinger 2008: 129). SD for sessions is 4.33% against a mean of 36% for sessions and 64% for spot searches and in both cases it seems quite low. Both mean and SD — which is based on the mean — are heavily influenced by outliers (e.g. BG and FR in the case of the session ratio). It was previously noted (see the Introduction to this Chapter) that statistical significance will not be tested and differences will have to be discussed using different means. A useful measure to interpret the relative magnitude of SD is to calculate the coefficient of variation (CV).

> The coefficient of variation measures the variability of a series of numbers independently of the unit of measurement used for these numbers. [...] The coefficient of variation can be used to compare distributions obtained with different units (Abdi 2010: 1).

In TABLE 12, the CV is computed by dividing the SD by the mean and can range from 0 to $\sqrt{N-1}$, where N is a finite sample of non-negative numbers with a real zero (Abdi 2010: 2). There is no distinction between the CV derived from whole (or decimal) numbers and percentage values. A high CV reflects inconsistencies with the group, i.e. variability, suggesting that the phenomenon can be considered language-specific to some extent. If the CV is close to zero, low relative variability is found across the languages such that the phenomenon under observation might be considered a general trend common to most or all languages. In the case of sessions and even more the case of spot searches, CV is very close to 0 suggesting that the general trend prevails.

Data obtained from the string counts for sessions and spot searches (columns 3 and 7) have been provided in the table but will not be taken into account in that they are too dependent on the initial size of each language subset (column 2). The column labeled "# Sessions" provides the count of the number of sessions found in each language subset (whereas "Session Sub." provides the total number of strings in the session subset). After dividing these two values, "Average Session Length" is obtained, i.e. how many searches a session contains on average. This is an alternative approach to measuring "Session Duration" (Zhang *et al.* 2009: 8) defined as "the period from the time of the first interactions and the time of the last interaction by a searcher interacting with a search engine," because the temporal cutoff was arbitrarily chosen as a prerequisite for the identification of sessions. Results suggest that the vast majority of sessions consist in fact of two searches and a small percentage consists of longer interactions. CV is again very low, suggesting very little variability in session length across languages. The two outliers are BG and FR. Bulgarian has the highest percentage of sessions as well as the longest session duration. On the other hand, French has the lowest values for both categories but has, consequently, the highest number of searches in the spot category. The latter result may be a side effect of the much larger volume of searches for FR with respect to the other languages, which means that individual searches (possibly from many more users) have a higher relative weight than search sessions. This trend seems to be confirmed by findings in Web searching, according to which about 28% of the approximate 2 billion Internet searches submitted daily to a search engine[112] are modifications to the previous query (Huang & Efthimiadis 2009). Generally speaking, spot searches seem twice more frequent than sessions and possible reasons for these skewed distributions will be discussed in Sub-section 6.2.1.2.

[112] Statistics from 2006-2007, based on the year of publication of the reference works.

5.6.2 LEVELS OF ANALYSIS

In this section, the concept of "search session" has been introduced, adding another dimension to the analysis of concordance searches. The components and sub-components of a concordance search have been outlined in the previous chapter (see Section 4.6) to frame every individual search string. This static dimension of concordance searching can be used to describe spot searches, i.e. one-time search events. If a burst-shooting mode is applied, the snapshots will capture the unfolding of the search process. This is the case of search sessions, which represent the dynamic dimension of a concordance search and should be analyzed as a unit.

The static and dynamic search dimensions seem to point to the levels of analysis found in Web search log analysis, namely (i) session level, (ii) query level and (iii) term level (Jansen 2006: 417ff). Session level looks at the repeated interactions between the user and the system in a limited time span, query level considers an individual string whereas term level addresses a string of characters delimited by e.g. a space, in this sense comparable to a token in a corpus.

From now on, most analyses will consider search sessions and spot searches separately to account for both dimensions. Whenever relevant, an analysis at the term level will be added. The next chapter will deal specifically with the search strategy component by looking at the search sessions in greater detail and these will be further categorized according to the types of query refinements, among others.

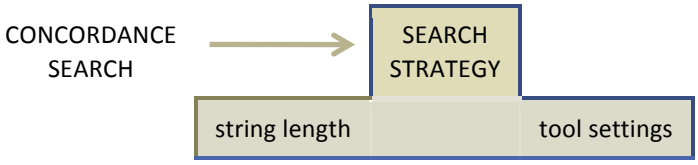
5.7 KEY CONCEPTS

- The analysis focuses on ecological validity by using data from "real scenarios".
- Data collection took place under partly controlled conditions (e.g. working environment and translation directionality).
- The aim is not so much studying individual behavior, but rather making generalizations and highlighting trends.
- The dataset consists of a special kind of corpus (i.e. no whole texts, only strings).
- English has been selected as sole source language without incurring in major data loss.
- The number of target languages was reduced to 20 because MT and GA had a excessively small dataset that caused these two languages to behave oddly.
- The dataset reduction phases can be summarized as follows: 970k ALL>ALL → 963k ALL>ALL_1tgt → 750k EN>ALL → 744k EN>ALL_1tgt → 740k EN>ALL_1tgt → 724k EN>ALL_1tgt_20 (i.e. the final dataset).
- The EC and the EP will be the main institutions to be examined because of the volume of the searches they produce.
- Language Families and Language Age have been chosen as grouping criteria but eventually no clear-cut distinction among languages was possible.
- A number of criteria for the identification and extraction of search sessions vs. spot searches have been identified and used.
- Thanks to the identification of spot searches and search sessions, two additional dimensions have been added to the analysis, i.e. static and dynamic, respectively.

CHAPTER 6: ANALYSIS OF THE 'SEARCH STRATEGY' COMPONENT

An overview of the various components of a concordancer search log was presented in Chapter 4 and the methodology for extracting sessions was detailed in Chapter 5. The present chapter will deal in greater detail with the search strategy component of a concordance search, which is made up of the string length and the tool settings. Figure 42 shows again the diagram discussed in Section 4.6, which summarizes the relevant components for this analysis, to be separately analyzed in the following sections.

Figure 42. Breakdown of a concordance search with respect to the variables pertaining to the search strategy: string length and tool settings.



6.1 STRING LENGTH

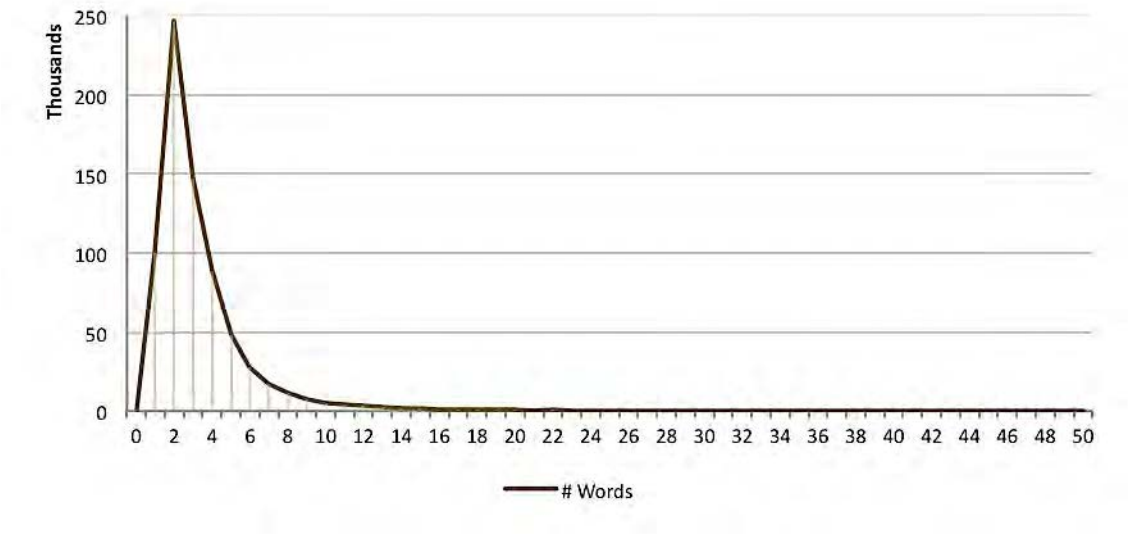
String length is one attribute of queries and search logs that is also shared by concordance searches. Query length is defined by Zhang *et al.* (2009: 7) as "the number of terms in the query" (including stop words), where a *term* is to be understood as "a series of characters separated by white space or other separator". As previously noted (see Section 5.6.2), a term often equals a word, but there may be instances where e.g. numbers, URLs or codes fall within the scope of a term, as clearly stated by Spink *et al.* (2001: 227):

> [A term is] any unbroken string of alphanumeric characters entered by the user. Terms included words, abbreviations, numbers, and logical operators [...]. URLs and e-mail addresses were treated as single terms.

When string length is being measured, the number of terms is counted without exceptions and the same was done in the case of Euramis searches. A customized Perl script[113] was used to extract basic frequency counts of the number of words (terms) per string across the whole dataset. Overall distribution (EN>ALL) of string length is reproduced in Figure 43.

[113] The author is indebted to Denis Navarre, IT Project Officer at the European Commission, for his help in writing the script.

Figure 43. Distribution of searches (724,000) according to the number of words in a search string.



A left skewed distribution towards the short end clearly emerges from the chart. The right tail is very long, ranging from about 20 words to a maximum length of about 200, but frequencies for such strings are low, meaning that very long strings are much less common than short ones. More than half of the strings have between two and three words (34% and 20%, respectively) and the vast majority range between one and four words.

Results are in line with those obtained for TransSearch (see Sub-section 3.2.2.7). In an early study (Simard & Macklovitch 2005), researchers looked at about 29,000 queries submitted in one week to this bilingual concordancer and then increased the dataset to 7.2 million queries covering a six-year period (Macklovitch *et al.* 2008). Table 13 provides the percentage distribution according to string length (in number of words[114]) for TransSearch next to the percentage distribution for Euramis.

Table 13. Comparison of percentage distribution of query length between the TransSearch and the Euramis datasets.

Query Length	TransSearch ('08) (6 years /7.2m)	TransSearch ('05)[115] (1 week /~ 29k)	Euramis (1 month /724k)
Single-word queries	13.2%	13.2%	13.83%
Two-word queries	39.6%	40.9%	34.02%
Three-word queries	27.7%	27.3%	20.33%
Four-word queries	13.0%	12.9%	12.27%
Five-word queries	4.3%	3.4%	6.66%
Six-word queries & above	2.2%	2.3%	12.90%
Total	**100%**	**100%**	**100%**

The most frequent queries have two words, followed by single-word queries and three-word queries. Irrespective of the considerable size difference between the three datasets, distributions and percentages are remarkably similar. The only difference worth noticing

[114] Assuming the same understanding of the concept of "word" in all studies.
[115] Percentage values calculated manually on the basis of the distribution chart with frequency counts in Simard and Macklovitch (2005: 72).

is in the last line where queries with six or more words are included. TransSearch queries account for slightly over 2% showing a decreasing trend from the previous tier, whereas in the case of Euramis they account for almost 13%, almost twice as many as the five-word group. This could partly explain why percentages in the 2- to 4-word groups are smaller for Euramis than for TransSearch. In the case of Euramis, strings longer than four words have a greater weight, suggesting that translators tend to submit a fair number of (very) long queries, as opposed to the more varied user group of TransSearch that focuses almost exclusively on very short queries. Researchers studying TransSearch interpreted their results as follows:

> […] either *TransSearch* users rarely encountered "lexical" problems (i.e. concerning a single, isolated word), or they turn to other resources when that happens, e.g. dictionaries, glossaries, thesauri, terminological banks, etc. What is clear, however, is that multi-word translation problems are the number one motivation for using *TransSearch* (Simard & Macklovitch 2005: 72).

For Euramis, the explanation might be a little different. Because Euramis offers a range of services to its users, there may be some overlaps between different services. The concordancing function in Euramis is sometimes used to retrieve documents, as explained in Section 3.2.4. This can be done by querying the repository for a few specific terms or expressions for that document or by inputting a larger chunk of text, as if it were a TM sentence match. If results are returned, then the user can download the document and e.g. integrate it into the locally stored TM. Unfortunately, the available data do not make a clear distinction between the two approaches and statistics cannot be produced.

The results in Table 13 are also in line with results from Web studies, where mean query length was calculated to be about 2 to 3 terms per query (e.g. Silverstein *et al.* 1999: 8; Spink *et al.* 2001: 230; Johnson *et al.* 2006; Arampatzis & Kamps 2008; Jansen *et al.* 2009: 1365)[116]. For TransSearch, this matching value may be a result of trial and error after its users noted that submitting strings longer than six words considerably reduced the chances of getting results (Macklovitch *et al.* 2008: 414). There might well be other possible explanations, but for the time being reference will be made chiefly to the study by Azzopardi (2009) stating that "the communication with [the Information Retrieval, IR] system appears to be the most efficient […] when two to five query terms are used. […]" (2009: 560). Despite shorter queries being most efficient for communicating with the system due to the Principle of Least Effort[117], the best total retrieval performance in Azzopardi's study was obtained when the query length was 30. On a smaller scale, multi-word units (e.g. bigrams and trigrams) "convey more specific meaning than single word features and therefore should be more effective in targeting relevant web search results" in that they make the intended sense clear (Johnson *et al.* 2006). However, the most commonly occurring bigrams and trigrams in English convey little meaning because of the high number of stop words.

---

[116] Unlike traditional IR searching, where the mean number of search terms varies from 7 to 15 (Jansen, Spink & Saracevic 2000: 217).

[117] The principle states that "a person attempting to apply a tool to a job does so in order to minimize the expected effort in using the tool for the given job", hence "in communicating with the IR system, the user wants to expend minimum effort in explaining their information need to the system, whereas the system wants to expend minimum effort in interpreting the query in order to return relevant documents. Consequently, the system would like longer and more precise queries, which uniquely identify relevant documents, whereas the user would like to submit short and vague queries" (Azzopardi 2009: 557).

Because frequency distribution for Euramis was calculated on the whole dataset of 724,000 strings, there was no correction for imbalances between languages (e.g. the French subset is double as large as the second largest subset). Distributions in Figure 43 run the risk of misrepresenting the overall trend by favoring the most populated languages. In order to correct for such size bias, mean query length was calculated for each language and results were normalized. A customized Python script[118] was developed to calculate average string length for every language subset (Table 14). Average length is usually counted in words (stop words included) but some decision-making was necessary to determine what was to be considered a word. Generally speaking, a word is delimited by two white spaces but the concept of "term" in Web log analysis is "a string separated by others by punctuation, white space, or a string of characters contained within square or curly brackets, or quotes" (Herskovitch *et al.* 2007: 214). There are also borderline cases such as the apostrophe 's' or the hyphen. The former can stand for a lexical item (verb 'to be' or 'to have') or a possessive form, while the latter can be used for compounding nouns or adjectives. For both Perl and Python scripts, only the hyphen was retained, whereas apostrophes were removed. Terms with slashes were split only if they contained letters but if they only consisted of numbers (e.g. document numbers), they were kept as one word.

Table 14. Distribution of string average length across languages for the whole dataset of 724,000 (both types and tokens) and for search sessions and spot searches.

(724k)	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
TOKEN	3.89	3.42	3.36	3.60	3.49	3.35	3.51	3.46	3.56	3.67	3.61	3.57	3.66	3.58	3.44	3.99	3.54	3.36	3.57	3.53
TYPE	3.06	2.88	2.96	3.14	3.18	2.93	2.99	3.12	3.14	3.09	3.16	2.83	3.13	3.25	2.87	3.56	2.88	2.73	3.05	3.01
	MEAN		SD		CV															
TOKEN	3.56		0.162		0.045															
TYPE	3.05		0.182		0.060															
(Tokens)	BG	CS	DA	DE	EL	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
SESSION	4.21	3.78	3.82	4.29	4.12	3.78	3.96	3.91	4.01	4.12	4.11	3.94	4.17	3.97	3.87	4.45	4.05	3.78	4.14	3.96
SPOT	3.61	3.24	3.15	3.29	3.17	3.09	3.27	3.21	3.41	3.40	3.32	3.31	3.34	3.38	3.22	3.67	3.23	3.09	3.27	3.30
	MEAN		SD		CV															
SESSION	4.02		0.18		0.045															
SPOT	3.30		0.15		0.045															

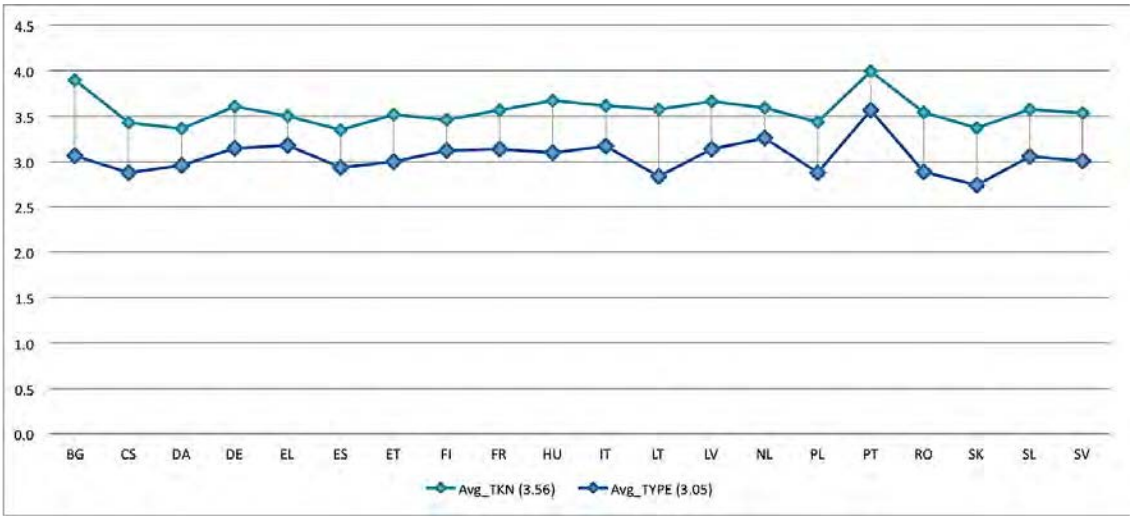
In the analysis of the overall dataset (724,000), a distinction was made between statistics including all searches in the dataset and statistics where repeated queries were removed. In the field of Web log analysis, researchers distinguish between *unique queries*, i.e. "all *differing* queries entered by one user in one session" and *repeat queries*, i.e. "all multiple occurrences of the *same* query" (Spink *et al.* 2001: 227; emphasis in the original). In a corpus-oriented perspective, the definitions seem to be comparable to the concepts of *tokens* and *types* at word-level, where the former represents "orthographic running word forms in the corpus" while the latter "refers to the number of different words used" (Olohan 2004: 80). The corpus of search strings may be analyzed by considering all strings separately (*string tokens*) or by looking at unique searches (*string types*). The same logic was applied by Lörscher (1991a: 207) in his analysis of translation problems:

[118] The author is indebted to Sabine Hunsiker, Linguistic Solutions Architect at Euroscript Deutschland, for her help in writing the script.

> In the terminology used in this study, a distinction is made between a *problem* and an *instance of a problem*. The distinction corresponds to that between *type* and *token*. The *different problems* of the corpus of investigation represent the *types*, the entire number of *instances of problems* corresponds to the number of *tokens* (emphasis in the original).

Table 14 provides two distinct distributions of the average length of the searches across languages in terms of tokens and types. The lower row in Table 14 shows the average length distribution for the token group between search sessions and spot searches. The total number of token strings was obviously 724,000, while total number of types (lowercased beforehand) amounted to almost half the total (i.e. just below 340,000). Mean values for types and tokens across languages taken separately differ by about 0.5, i.e. half a word. SD and CV become slightly higher in the case of types. The distribution for token and types is graphically represented in Figure 44 to better visualize differences and interpret results.

Figure 44. Graphical representation of the distribution of average string length (in words, Y axis) across all languages (X axis) for both tokens and types[119]

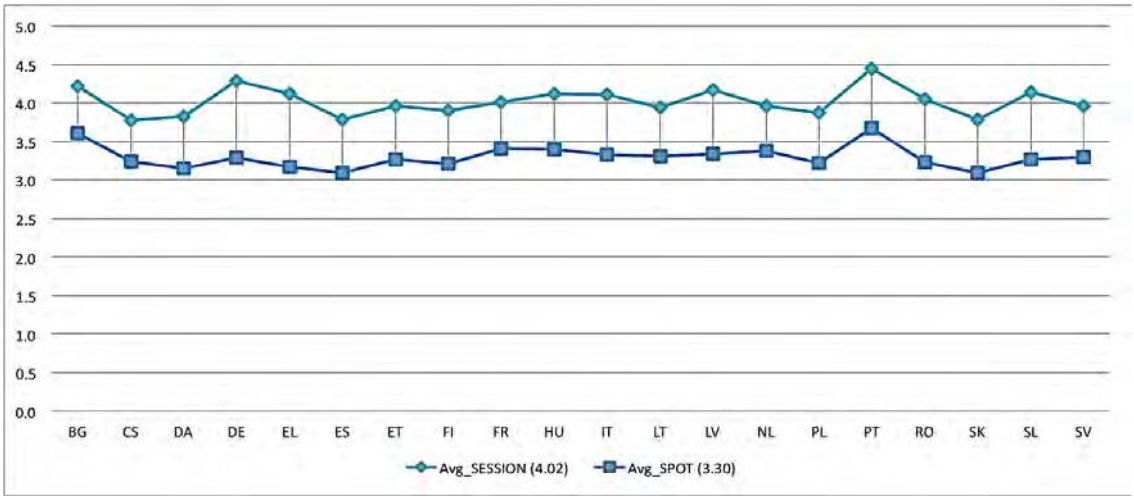


If all languages behaved in the same way, the distance between the same data point on the two lines (i.e. the high-low lines) would be of the same length. In fact, some lines are clearly longer than others, e.g. BG vs. FI and LT vs. EL, meaning that some languages have a higher number of repeated queries. If string types were to be studied instead of tokens, languages with longer high-low lines (e.g. BG and LT) would suffer from greater data loss. When types alone are considered, a tentative trend seems to emerge where 'old' languages submit longer queries (on average) than 'new' languages. This may suggest that old languages perceive an added value in Euramis (and/or Quest) for querying longer portions of source text. On the other hand, new languages use Euramis more often (cf. total subset size) and for shorter queries, a possible indication of different search strategies or different information needs. It could also imply that they tend to use Euramis for problems for which 'old' languages prefer to use other resources and/or that old languages tend to use Euramis more often for document retrieval than language queries. As will be shown later (see Section 6.1.2), string resubmission is expected to play a role in terms of the specific search strategy adopted (in other words, it is not accidental)

[119] Lines linking data points horizontally do not have any significance but are just meant to help visualizing the relative positions of each data point and compare relative distance between the two lines.

and for this reason the analysis is better carried out using token counts instead of using unique counts (i.e. types). As regards the distribution in the lower row of Table 14 between search sessions and spot searches, two clear trends can be identified as shown in Figure 45.

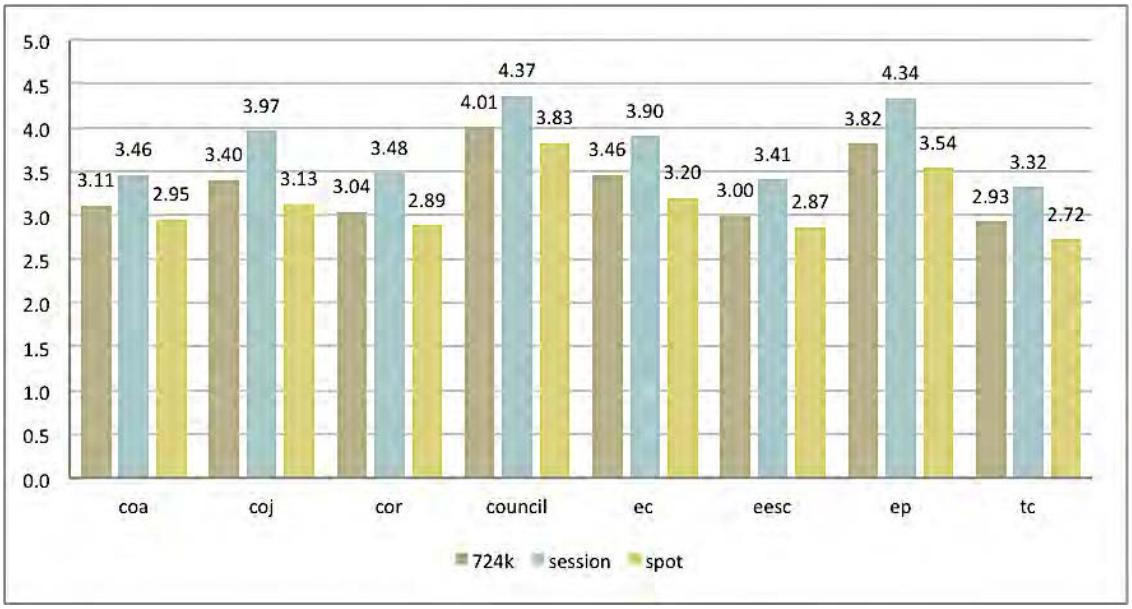
Figure 45. Graphical representation of the distribution of average string length (in words) across all languages for both search session and spot searches (token count).



Session and spot searches follow each a distinct trend with no intersections between the lines, i.e. in no subset the average length of spot searches is greater than average length of sessions. Mean length for sessions is greater than mean length for spots by over half a word (4.02 vs. 3.30 words). High-low lines are not of equal length, suggesting a different distribution across languages. Portuguese (PT) has the highest average both for sessions and spot, i.e. tends to submit (much) longer strings than other languages. German (DE) has the second highest average for sessions but for spot searches its average is close to the mean. Incidentally, German also has the longest high-low line (i.e. delta value), meaning that German translators tend to submit (very) long strings in a search session but submit (much) shorter queries in isolation. At the other end of the spectrum, Czech translators have the smallest difference in average length between sessions and spot searches, i.e. CS translators do not differentiate as much between the two types of searches. Search sessions having a higher average length may be explained by the fact that the (first) search in a session is an attempt to submit a query in a way that was not optimal for the retrieval effectiveness of the system, hence query reformulations. On the other hand, spot searches average close to the 2- to 3-word gravitation point which seems to provide a solution to the problem after the first attempt.

For each institution, search average length was also separately calculated for the total dataset and for sessions and spot searches so as to see whether different search habits could be highlighted (Figure 46).

Figure 46. Distribution of average string length per institution over the whole dataset and the session and spot subsets, respectively.



The first thing to be noted is the trend for each institution to have systematically higher session lengths than overall average length; average spot length was on the other hand only slightly below the overall average. Differences in average length between institutions are also to be found. Council translators submitted the longest queries, closely followed by the Parliament. On the other hand, the Translation Centre, EESC and COR used shorter queries, the difference possibly due to the specific text types at each institution. For each result, deviations from the mean should also be taken into consideration. For all three subsets, the Court of Justice had the highest value for SD (see Appendix B) because there were no queries coming from this institution for two languages (FI and PT), while other languages (e.g. IT and LV) reported the highest averages of all institutions put together, i.e. 7.62 and 5.50 respectively, with a peak of 11.75 for the IT session subset submitted by the COJ.

Consistent average lengths seem to be obtained from observing the data. These results could be compared to findings from other studies in slightly different areas which can be said to have come to similar results. One example is the study carried out by Dias *et al.* (1999) which will be further discussed in Sub-section 7.3.2.6. The authors extracted multi-word units from proceedings of the EP in four languages (EN, FR, IT and PT) and found that most of these units were made up of 2 to 4 words, independently of the language used. Dragsted (2004) also focused on length in terms of segmentation and translation units (see Section 2.8.1); in particular, she measured segment length in terms of words to try and find a cut-off length. She found that both professionals and non-professionals used most frequently 2- and 3-word segments as Translation Units.

> Hence, the distribution of 2-, 3- and 4-word segments as the most frequent segment size was identical for the two groups of subjects, and it can be concluded that segments of 2 and 3 words were the ones occurring most frequently in both groups (Dragsted 2004: 126-7).

Interesting parallels have emerged that could connect the performance of IR systems, translation and problem unit sizes and average query length in Web searching. The question arises as to how these levels can be related to one another in a more systematic way.

6.1.1 QUERY REFINEMENT CATEGORIES

Before delving into the analysis of tool settings, an additional study will be carried out at the level of search sessions where query refinement takes place. Refining a query means changing the search approach (i.e. the strategy) e.g. by changing string length (adding, removing, replacing portions of the original search string) or adjusting the filters and the settings. Ideally, a complete account of search sessions should consider both elements at the same time. However, this section will only consider changes in the text string mostly in terms of length; tool settings will be covered in the second half of the chapter.

After search sessions have been separated from spot searches, statistics for each group can be easily generated. As a rule, spot searches are one-time searches where parameters are selected only once (i.e. scenario n.1 in Section 5.6). In search sessions, on the other hand, different combinations are possible because each variable can be changed at every new search and each of the three remaining scenarios is equally possible. This analysis will focus specifically on the text string component and will make use of findings in Web search log analysis to develop a methodology for a finer-grained classification of the searches. A brief review of existing reformulation strategies for query logs was carried out by Huang and Efthimiadis, who noted that available taxonomies "are generally constructed by examining a small set of query logs. Some studies are out of date or incomplete. None have built an automatic classifier distinguishing reformulation strategies [...]" (2009: 78). The authors instead developed a matching rule for each strategy to identify the query reformulation types listed in their own taxonomy, combining categories found in previous work. They identified 13 reformulation categories[120] (Table 15).

Table 15. Taxonomy of reformulation strategies for query logs according to Huang and Efthimiadis (2009).

1	Word Reorder	8	Expand Acronym
2	Whitespace and Punctuation	9	Substring
3	Remove Words	10	Superstring
4	Add Words	11	Abbreviation
5	URL Stripping	12	Word Substitution (synonym, hyponym, hypernym, meronym, holonym)
6	Stemming		
7	Form Acronym	13	Spelling Correction

Some of the categories are specific for Web navigation (e.g. URL Stripping); and others require some additional resources for analytical purposes, such as Porter's stemming algorithm for category 6 (Stemming) or the Wordnet database to identify the semantic relation between two words (category 12 – Word Substitution). Given the extended proportions of this study and limited computational resources, the number of external resources to be employed had to be kept to a minimum. Such categories, if at all present, would have to be manually identified.

Preliminary observations of the language subsets suggested that similar categories could be employed in the case of concordance searches. Automatic categorization was deemed necessary because consistent manual categorization was not feasible without incurring

[120] For a more detailed discussion about taxonomy creation, individual reformulation strategies, matching rules and data volume, refer to Huang and Efthimiadis (2009).

difficulties such as inter-rater agreement, i.e. statistical measures to assess agreement between the ratings of two or more evaluators. Given the conditions for the extraction of sessions (see Section 5.6), at least one shared word between two strings in the same search session was to be expected and this could be used as a first anchor point to study query reformulation both in quantitative and qualitative terms. By observing the automatically extracted search sessions, a number of general trends were identified that were labeled as macro-categories, namely (A) resubmission, (B) formal changes, (C) reduction, (D) expansion, (E) replacement and (F) mixed strategy. Each macro-category may include a number of sub-categories which are detailed below and for which examples will be provided in Table 16. Because strings were lowercased to ensure consistent matching in the analysis, all examples provided will also be lowercase.

6.1.1.1 RESUBMISSION

With *Resubmission*, the system logs the same query without any changes made to the string. The user must have either re-submitted the same query manually or simply clicked on the "Show more" button in the Euramis result page, which the system interprets as a new query. This category also includes resubmissions where the text string was left untouched and the changes only affected the settings; at this stage no distinction is made between these two scenarios. The resubmission category was not considered in the reformulation strategies in Web logs (Huang & Efthimiadis 2009; see Table 15), but in that case "same queries" accounted for some 40% of the analyzed dataset.

6.1.1.2 FORMAL CHANGES

Formal Changes occur when casing is changed, punctuation is removed or the string locale is changed. In all three instances, there is no loss or addition of alphanumeric characters, at most the replacement of a letter. In this sense, this strategy can be seen as the closest to category A, Resubmission, and is likely to be included in scenario n. 2 in Table 15 due to lowercasing of strings in the main study.

6.1.1.3 REDUCTION

Reduction involves a shortening of the string (i.e. removal of words) and can be compared to the 'Remove Words' category in Web searching (Table 15); in the present study, however, a more detailed categorization of the types of deletion has been made. The sub-categories take into account the position or the specific type of deleted portion in the original search strings. Five possible sub-categories were identified:

1. Left trim (left-most part removed)
2. Right trim (right-most part removed)
3. Middle trim (central part removed)
4. Cross trim (both left and right)
5. Removal of plural (or genitive) 's'. It is often the case that a single letter is removed from the last word of the string and this generally is the plural 's' suffix. It has been included as a special kind of reduction that has only a minimal impact on the formal appearance of the string.

6.1.1.4 EXPANSION

Expansion is the opposite of reduction, i.e. text is added to the original string, and is the same as category 'Add Words' above. According to the point in the string where text is added, five sub-categories can be found:

1. Left expansion (text added at the beginning of the string)
2. Right expansion (text added at the end of the string)
3. Middle expansion
4. Cross expansion
5. Addition of the plural 's' suffix (rarer than the corresponding subcategory in Reduction)

6.1.1.5 REPLACEMENT

Replacement occurs when individual words in a string are changed. It involves neither deletion nor addition of words but rather word substitution, ranging from stemming and typo fixes to actual replacements with semantically related words. Given the variety of operations that can be performed, this category can be said to include strategies n. 6 (Stemming), 12 (Word Substitution) and 13 (Spelling Correction). More specifically, replacement comprises:

1. Tense change (usually from past/compound tenses or conjugated forms to the base infinitive form)
2. Paraphrase (possibly with a shift in grammatical category)
3. Synonym/Antonym
4. Word substitution involving a replacement without there being a direct semantic relationship between the two words. This partly falls into category n.12 (Word Substitution) above. Word substitution can sometimes involve unrelated search events (i.e. two independent spot searches) submitted at close distance from one another and possibly originating from the same source text.
5. Typo fix, i.e. category n. 13 (Spelling Correction)

The macro-category of *Replacement* is possibly the hardest to identify because it does not rely so much on quantitative measures (i.e. addition/deletion of content,) but rather on subjective judgments, e.g. on what can be considered as a synonymous (or antonymous) expression.

6.1.1.6 MIXED STRATEGY

Finally, there can be instances where search sessions are longer than two strings and the search strategy changes from one search to the other within the same session. There can also be instances where a short session of two strings employs more than one strategy at the same time. All these instances are grouped under the heading of *Mixed Strategy*, which was not clearly singled out for Huang and Efthimiadis' (2009) analysis of Web queries. Examples of mixed strategies are:

1. Expansion + Trim
2. Mixed Trim (first left, then right or vice versa)
3. Typo Fix + Expansion/Trim
4. Expansion/Trim + Resubmission
5. Typo Fix + Resubmission
6. Resubmission+ Replacement (usually, Word Substitution)

No further distinction was made for the internal order of the strategies nor the number of times each strategy was employed, otherwise the number of possible combinations would have grown exponentially and become unmanageable.

Table 16. Verbatim examples of queries for each identified category and sub-category in a search session.

CATEGORY	FIRST QUERY >>	LAST QUERY
RESUBMISSION	trade repository	trade repository
FORMAL CHANGES	- Intergovernmental	Intergovernmental
REDUCTION		
1. Left trim	government response to the consultation	to the consultation
2. Right trim	legally-binding measures	legally-binding
3. Middle trim	Consumer goods and services	Consumer services
4. Cross trim	geographic county of	county
5. Plural / Genitive 's'	Tuna purse seiners	Tuna purse seiner
EXPANSION		
1. Left expansion	over board	thrown over board
2. Right expansion	application process for	application process for the FP7
3. Middle expansion	convention human rights	convention on human rights
4. Cross expansion	performance	regional innovation performance index
5. Addition of plural 's'	official duty passport	official duty passports
REPLACEMENT		
1. Tense change	participated in this regulation	participates in this regulation
2. Paraphrasing	Failures in protection of	Failures to protect
3. Synonym / Antonym	troop allowances short-term segment	soldier allowances LONG-term segment
4. Word substitution	update the factual information	revise the factual information
5. Typo Fix	Hiigh-level Group on Gender Equality	High-level Group on Gender Equality

On the basis of the previous taxonomy, the category codes listed in Table 17 below will be employed to refer to a specific macro-category and one of its sub-types. There is also a possible additional category, which will be mentioned in passing. It was not explicitly listed because it contained 'false positive' sessions, i.e. groups of strings that were classified as sessions but in fact were not. This could be due to either a function word that did not appear in the list of stop words, or pure chance, i.e. two unrelated strings that meet all criteria. Moreover, there could be potential sessions that, due to synonymy or paraphrase, were not correctly identified and ended in the spot group. Although session extraction could not be further improved given the available means, false positives and false negatives are still expected to balance each other out in most cases.

6.1.2 CATEGORY DISTRIBUTION

After detailing the characteristics of each category, the existing sessions for each language had to be assigned to a given category. To make the process more efficient, each category and subcategory was associated with a specific alphanumeric code, as shown in Table 17.

Table 17. List of category codes employed to refer to a macro-category and each of its sub-types.

A	**RESUBMISSION**	**D**	**EXPANSION**
	A1. Repeated query		*D1. Left expansion*
	A2. Wildcards		*D2. Right expansion*
B	**FORMAL CHANGES**		*D3. Middle expansion*
	B1. Casing		*D4. Cross expansion*
	B2. Punctuation		*D5. Addition plural 's'*
	B3. Locale	**E**	**REPLACEMENT**
C	**REDUCTION**		*E1. Tense change*
	C1. Left trim		*E2. Paraphrase*
	C2. Right trim		*E3. Synonym/Antonym*
	C3. Middle trim		*E4. Word substitution*
	C4. Cross trim		*E5. Typo Fix*
	C5. Plural/Genitive 's'	**F**	**MIXED STRATEGY**

The matching of a search session with the corresponding search-strategy category was carried out with a PHP script, which automatically assigned the labels for some of the categories. Because of the relatively simple nature of the script, not all categories in the taxonomy could be automatically recognized. The labeling was limited to those categories that were quantitatively easy to identify, i.e. those words that were added or removed. Categories A1, C1, C2, D1 and D2 above met the criteria. If no difference between two consecutive strings in a session was found, the session was labeled A1 (Repeated Query); if words were removed from the beginning of the string, the session was labeled C1 (Left Trim) and if they were removed from the end, it became C2 (Right Trim). Conversely, if words were added at the beginning the session was labeled D1 (Left Expansion), if they were added at the end, the label was D2 (Right Expansion). All remaining categories (B and subcategories of C and D) were collectively labeled *Other* and left aside for manual evaluation. Table 18 summarizes the results of the automatic labeling of the sessions into the 5 selected categories plus the miscellaneous category *Other.*

Table 18. Distribution of sessions categories across all languages after the automatic labeling.

Lang Pair	**Search Strategy** (normalized by tot. # sessions)					
	Cat_A1	**Cat_C1**	**Cat_C2**	**Cat_D1**	**Cat_D2**	**Cat_Other**
BG	38.97%	15.82%	17.26%	1.02%	1.64%	25.26%
CS	21.37%	22.45%	20.43%	1.89%	2.45%	31.37%
DA	12.27%	24.08%	24.34%	2.31%	3.22%	33.74%
DE	17.17%	21.83%	21.90%	2.09%	2.76%	34.23%
EL	8.16%	26.71%	25.67%	2.01%	3.01%	34.41%
ES	15.27%	24.36%	22.94%	2.84%	3.13%	31.43%
ET	17.77%	25.31%	23.54%	1.47%	1.87%	30.01%
FI	9.13%	25.61%	23.05%	1.77%	2.46%	37.95%
FR	10.73%	25.94%	22.25%	1.75%	1.85%	37.45%
HU	20.48%	21.99%	20.60%	2.13%	2.57%	32.19%
IT	14.09%	23.71%	22.73%	2.35%	3.04%	34.05%
LT	32.61%	20.97%	18.42%	1.03%	1.32%	25.62%
LV	23.20%	23.37%	21.35%	1.41%	1.89%	28.75%
NL	10.69%	23.01%	20.22%	2.49%	3.37%	40.19%

PL	14.48%	24.93%	22.21%	2.16%	3.10%	33.09%
PT	18.38%	23.03%	20.96%	1.66%	2.40%	33.54%
RO	25.01%	20.72%	19.90%	1.96%	2.21%	30.16%
SK	27.78%	21.98%	18.86%	1.79%	2.27%	27.29%
SL	18.65%	22.84%	22.27%	1.71%	2.56%	31.95%
SV	21.10%	21.33%	19.62%	1.87%	2.28%	33.78%
MEAN	18.87%	23.00%	21.43%	1.89%	2.47%	32.32%
SD	7.91%	2.40%	2.07%	0.45%	0.57%	3.85%
CV	0.419	0.104	0.097	0.239	0.231	0.119
RANGE	39%-8%	27%-16%	26%-18%	3%-1%	3%-1%	40%-26%

The first thing to be noted is that the miscellaneous category accounts for about one third of the total sessions for each language. This may seem a high proportion but at a closer look it turns out that the selected five categories account for more than half of the searches and have a much larger coverage than the dozen coming under the heading *Other.* One word of caution is necessary before discussing these results: Macro-category E (Mixed Strategy) was not singled out because including it into the picture in a meaningful way was too complicated. Since sessions longer than two strings are fewer than two-string sessions, no exact breakdown of the mixed strategies was deemed necessary at this stage. The automatic labeling was based on the first couple of strings of each session so that evaluation turned out to be consistent. In the case of a longer session, the rest of the session was generally not considered to lead to an uneven distribution of mixed strategies because the initial strategies employed are likely to be distributed across all categories.

The first result that emerges from Table 18 is that averages between two forms of reduction in category C (or expansion in category D) are quite close. However, there is a very marked gap between the total averages of reduction and expansion, showing that expansion, and in particular left expansion, occurs much more rarely than trimming. Left trim is the most frequent type of strategy employed, whereas left expansion is the least common.

A closer look at the results for individual languages reveals that BG mostly resorts to the resubmission strategy (A1) as opposed to e.g. FI or EL, which hardly use it. This trend is clearly reflected in the type/token chart (Figure 44) where longer high-low lines reflect the high proportion of strategies falling in A1. Repeated strings seem a direct consequence of a deliberate search strategy rather than an accidental match between different strings in a language subset. The only possible exceptions to this trend may come from PL and PT. Polish uses A1 relatively little but has a token/type delta above average, i.e. has more repeated searches than is justified by the deployment of A1. Conversely, Portuguese makes an average use of A1 but has a delta (well) below average, i.e. most resubmissions are intentional or resubmissions occur together with other strategies or are concentrated in the same longer search sessions. Table 12 in Section 5.6 shows that PT has the second highest value for average session length, suggesting that the third hypothesis is more likely.

Resubmission (A1) is quite common in all languages but has a much higher SD (and CV) than C and D. A greater variability should therefore be expected across languages, as confirmed by the range values. Range is possibly the simplest measure of dispersion and shows the difference between the lowest value and the highest value, in this case in percentage points (Rasinger 2008: 124). The range for A1 is almost 31 percentage points

whereas for the other categories it does not exceed 11 points. The two outliers in the case of A1 are BG (38.97%) and EL (7.91%). At a closer look, new languages seem to occupy the higher end and old languages tend to appear in the lower half of the rank. Bulgarian and Greek are also outliers in both types of reduction strategies (C1 and C2), only this time the other way round: EL is being the most active trimmer and BG the least active one. This means that BG translators tend to repeat the same query instead of trimming/expanding it, whereas EL translators tend to start off with a longer query and then reduce it. In the case of the expansion categories, the picture is slightly different. BG hardly uses left expansion — almost on a par with LT — whereas ES uses it the most. NL seems to favor right expansion (D2) while LT neglects it. In sum, Lithuanian translators favor expansion as a macro-strategy the least, whereas both ES and NL seem to use it more than other languages. NL is the language that resorts to "other" strategies in 40% of its searches. On the other hand, BG translators use strategies other than A1 very rarely. The "Other" group includes at least a dozen sub-types and accounts on average for over one third of the sessions. This means that the relative weight for each remaining category is very low with respect to the automatically identified ones.

It was previously suggested that mixed strategies (category F) would be distributed across the other strategy groups due to the automatic categorization. To substantiate this assumption, a manual categorization of the sessions of one sample language was carried out.

6.1.2.1 MANUAL CHECK ON FINNISH

The purpose of this manual evaluation was to have a better understanding of the reliability of the script for the automatic labeling of sessions. Finnish sessions were manually classified according to the developed taxonomy and results were compared to the ones obtained automatically with the script.

In this case, the analysis had to be more precise and the Mixed Strategy category had to be appropriately developed. Compared to the original taxonomy (see Section 6.1.1), a few additional categories were added that were not previously identified. Table 19 provides an overview of all used categories with a verbatim example from the logs as well as an abstract representation of their structure using letters of the alphabet.

Table 19. Overview of manually identified categories and sub-categories (with examples) and exemplification of the abstract structure of each session.

Cat. Label	Name[121]	Exemplification
A.	**Resubmission**	
A1	Repeated query	ABC = ABC
	induction of new teachers > induction of new teachers	
A2	Wildcards	AB* = AB*
	receive revenues > receiv* revenues	
B.	**Formal changes**	
B1	Casing	abc > ABC
	equivalent NOEC > equivalent Noec	
B2	Removal (/Addition) of punctuation	A-BC > ABC
	non discrimination of people with disabilities > non-discrimination of people with disabilities	
B3	Locale changes (US-UK)	ABZ > ABS
	International organization for migration > International organisation for migration	
C.	**Reduction**	
C1	Left trim	ABC > BC
	team-orientated staff management > staff management	
C2	Right trim	ABC > AB
	naval security operations > naval security	
C3	Middle trim	ABC > AC
	communication system of the fishing vessel > communication system of the vessel	
C4	Cross trim	ABC > B
	bring the worlds of education training and work closer together > worlds of education	
C5	Removal of plural (or genitive) 's'	ABCs > ABC
	weighing records > weighing record	
D.	**Expansion**	
D1	Left expansion	BC > ABC
	international convergence > The development of international convergence	
D2	Right expansion	AB > ABC
	adopt the procedures for the adoption > adopt the procedures for the adoption and revision	
D3	Middle expansion	AC > ABC
	guidelines for employment policies > guidelines for the employment policies[122]	
D4	Cross expansion	B > ABC
	Risk committee > a risk committee which shall be composed	
D5	Addition of plural (or genitive) 's'	ABC > ABCs
	sovereign debt instrument > sovereign debt instruments	
F.[123]	**Replacement**	
F1	Tense change (or third person 's')	ABCed > ABC
	determining additional elements > determine additional elements	
F2	Paraphrasing	ABC > AxyC

[121] The parenthesis indicates a possible, less common, alternative.
[122] From the point of view of the concordancer, these two strings are equivalent because the word 'the' is ignored in the search.
[123] Letter E should come first, but it was originally used for mixed strategies. To keep the logical sequence of the categories outlined earlier, category F has been brought forward in the table.

	reductions in catch limits > reduce catch limits	
F3	Synonym/Antonym	ABC > ZBC
	Decision to prolong the authorisation > Decision to continue the authorisation financial instability > financial stability	
F4	Typo Fix	zBC > ABC
	ventral asia strategy > central asia strategy	
F5	Word substitution	ABCD > ABCE
	greeting cards > business cards	
F6	Intersection (same as E1 but simultaneous)	AB > BC or BC>AB
	Interdepartmental consultation and coordination > organising and responding to interdepartmental consultations	
E.	**Mixed strategy** (strategies in any order)	
E1	Expansion + Trim	AB > ABC > BC
	ABCD>BCD>ACD>CD: economic and social market conditions > social market conditions > economic market conditions > market conditions	
E2	Mixed Trim	ABCD > ABC > BC (or ABCD > BCD > A)
	ABCD>ABC>BC: responsible for the submission of sales notes and takeover declarations shall submit > responsible for the submission of sales notes > submission of sales notes	
E3	Typo Fix + Trim (/Expansion)	zBC > ABC > AB
	simplified transit procedcure > simplified transit procedure > simplified transit procedure for rail	
E4	Trim (/Expansion) + Repeated query	ABC > AB > AB
	detecting divers or swimmers > divers or swimmers > swimmers > swimmers	
E5	Typo Fix + Repeated query	zBC > ABC > ABC
	shortlist of succesfull candidates > shortlist of succesful candidates > shortlist of successful candidates	
E6	Repeated query + Replacement	ABC > ABD > ABD
	ABC>ABC>ABD>ABC: reverse voting > reverse voting > reverse vote > reverse voting	
E7	Removal (/Addition) of plural 's' + Repeated query	ABCs > ABC > ABC
	ABCs>ABC>ABCs: troop allowances > troop allowance > troop allowances	
E8	Typo Fix + Plural 's'	aBCs > ABCs > ABC
	aBC>ABC>ABCs: risk free interest rate term structur > risk free interest rate term structure > risk free interest rate term structures	
E9	Trim (/Expansion) + Formal change	abcd > abc > ABC
	study on Decentralised energy production - current barriers > Decentralised energy production - current barriers > decentralised energy production - current barriers	
E10	Plural 's' + Replacement	ABCs > ABC > ABE
	blue collar workers > green collar workers > green collar worker	
E11	Trim (/Expansion) + Replacement	ABC > BC > BD
	{replacement as locale change}: whether to authorize or deny > whether to authorise or deny > authorise or deny	
E12	Plural 's' + Trim (/Expansion)	ABCDs > ABCD > ABC
	ABC>DBC>BCs: anchoring the expectation > anchore the expectation > anchoring expectations	
E13	Typo Fix + Replacement	aBC > ABC > ABD
	Single Market > internalt Market > internal Market	
G.	**False Positives**	(here due to 'with')
	progressed well with their conclusion > delegated acts in accordance with Article	

The mixed category is meant to cover the majority of sessions longer than two strings because this is where different strategies are more easily combined. However, there are cases where the same strategy (e.g. "Right Trim") is used throughout the same session and in these cases a single strategy label is applied. Given the very large number of possible combinations, most mixed categories do not distinguish between strategies of expansion and trim and in the case of sessions of three or more strings, the actual sequence is often not taken into account. This is exemplified in Table 20 where a few examples belonging to the same category (E11 – "Trim/Expansion + Replacement") are provided. The only sub-category to be kept apart from the main category "Replacement" is "Typo Fix" because it was felt that a typo fix would in any event negatively impact the outcome of the search, as the word would not be found in the database. Typos are one common cause of a failed search both in concordance searches (Macklovitch *et al.* 2000: 1204) and Web queries (Wang *et al.* 2003: 756-757). The most frequent strategies related to word Replacement to be found in the Mixed Strategy group are F5 – "Word Substitution" and F1 – "Tense Change".

Table 20. Examples of sessions labeled E11.

Search Session	**Strategies** (main category: E11)
EX #1	
launch a flight service launching a flight service flight service	> Tense Change (F1) > Left Trim (C1)
EX #2	
QuickTime uncompressed file uncompressed file compressed file	> Left Trim (C1) > Synonym/Antonym (F3)
EX #3	
that are at low levels stocks are at low levels stocks at low levels	> Replacement (F5) > Middle trim (C3)

The classification of Mixed Strategies is not meant to be fine-grained because too many specific categories would impair the understanding of the big picture that is the main purpose of this research project. After these preliminary considerations, results from the manual comparison and the automatic labeling for the Finnish subset, summarized in Table 21, can finally be discussed.

Table 21. Results from the manual categorization vs. results from the PHP script.

SCRIPT	Count	%	MANUAL	Count	%	Delta Count	Delta Ratios
A1	371	9.14%	**A1**	324	7.74%	47	1.40%
			A2	2	0.05%		
			B1	5	0.12%		
			B2	18	0.43%		
			B3	2	0.05%		
C1	1040	25.62%	**C1**	959	22.91%	81	2.71%
C2	936	23.05%	**C2**	890	21.26%	46	1.79%
			C3	177	4.23%		
			C4	38	0.91%		
			C5	169	4.04%		
D1	72	1.77%	**D1**	63	1.51%	9	0.27%
D2	100	2.46%	**D2**	86	2.05%	14	0.41%
			D3	22	0.53%		
			D4	3	0.07%		
			D5	67	1.60%		
			E1	168	4.01%		
			E10	9	0.22%		
			E11	64	1.53%		
			E12	72	1.72%		
			E13	4	0.10%		
			E2	132	3.15%		
			E3	24	0.57%		
			E4	121	2.89%		
			E5	13	0.31%		
			E6	13	0.31%		
			E7	24	0.57%		
			E8	5	0.12%		
			E9	4	0.10%		
			F1	44	1.05%		
			F2	26	0.62%		
			F3	25	0.60%		
			F4	144	3.44%		
			F5	75	1.79%		
			F6	93	2.22%		
Other	1541	37.96%	**N/A**	301	7.19%	1240	30.77%
Total	**4060**	**100%**		**4186**	**100%**	**126**	**3.01%**

First of all, the total number of sessions in the manually labeled group exceeds that of the automatically labeled groups by about 130 strings. Caution is necessary in the interpretation of both this delta and the "N/A" category emerging from the manual labeling. The label "N/A" comprised about 7% of the sessions and was used to indicate false positives (e.g. 'All data elements defined in the Annexes > name of the data element > Element or attribute name'; 'calling for a moratorium on the use of the death penalty > total abolition in all states which still practise the death penalty'). In other cases, there was a proper session but it was in the wrong source language (noise in the data) and here too, the label N/A was used. Finally, there were other instances where there was a single

string that was clearly not part of a session (e.g. 'limit itself to ≠ scope set by > scope set') and in this case the outsider was labeled as N/A. In sum, the label "N/A" was used either as a replacement for a non-session but also to identify outsider strings within a session. As a consequence, the number of labels in the manually categorized dataset turned out to be higher that the automatically generated one.

All automatically labeled categories have higher counts than their counterparts in the manual categorization. Individual delta values between the manual and automatic classifications range from less than half a percentage point to almost three points (i.e. about 80 strings), which can be considered a good result given that proportions in each category are very similar. The deltas can be explained by the greater diversification in the manual categorization and by the fact that the script was generally not able to distinguish between a two-string session and longer ones. Overall, the automatic labeling seems to provide reliable results for each category, covering about 60% of the total sessions[124].

As for the actual categories, consistent choices had to be made in a few special cases. When sessions longer than three strings occurred, usually more than two different macro-strategies were involved, which posed a problem for the labeling. If the same strategies occurred more than once, the most representative category was chosen (e.g. 'Development Fund for the Electronics and Information Industry > Electronics and Information Industry > Information Industry > electronics Industry' = E2,[125] i.e. 2 trims + 1 replacement). Alternatively, the chronological sequence was followed and the label was attributed on the basis of the first two strategies encountered (e.g. 'autonomous budget line > autonomous budget lines > autonomous budget lines > autonomous' = E7, i.e. addition of plural 's' + resubmission + trim). Another special case occurred when two strategies were applied simultaneously in a two-string session (e.g. 'start and finish positions > start position' = plural 's' + trim). In this case preference was given to the otherwise less frequent strategy, which in the given example would be C3 – "Middle Trim". Where two strategies in a string turned out to be equally common (one usually being F4 – Typo Fix), a Mixed Strategy was chosen (e.g. 'State ransmitting the message > **t**ransmitting the message' = E3)[126].

The "Other" category accounts for almost 40% of the strings in the categorization from the script. An additional purpose of the manual categorization was to identify which individual categories would need to be handled separately because of their representativeness. In the C macro-category, strategies C3 – "Middle Trim" (4.23%) and C5 – "Removal of Plural 's'" (4.04%) are the most frequent in absolute terms after C1 and C2. The F macro-category follows with F4 – "Typo Fix" (3.44%) and F6 – "Intersection" (2.22%). Due to the prominence of macro-category C, some Mixed Strategies where Expansion and Trim are involved also score quite high: E1 – "Expansion + Trim" (4.01%), E2 – "Mixed Trim" (3.15%) and E4 – "Trim + Repeated Query" (2.89%). Unfortunately, due to the Mixed Strategy category and the above-mentioned uncertainties in the labeling, reliable values for recall and precision could not be calculated.

This manual check on a randomly selected language served the purpose of verifying the reliability of the script for the automatic labeling of the strings. Ideally, the script should be able to distinguish longer sessions and additional trim categories such as C3 and C5. This is technically possible, as Huang and Efthimiadis' work has shown (see above). Other

---

[124] Manual labeling for the same 5 categories covered about 55% of the total sessions identified.
[125] The script labeled this session as C1, evidently only taking into account the first two strings.
[126] Such choices were only required for a handful of cases, which should contribute to dispelling the impression that the results of the analysis are biased in some way.

categories beyond these two could in principle be identified (e.g. B sub-categories, tense changes (F1), word substitution (F5) and typos (F4)) by building an automatic classifier using the proposed unsupervised algorithms (2009: 79ff.) and a finer-grained study would undoubtedly benefit from a more accurate classifier and more accurate precision/recall calculations. For the time being, however, the script will suffice to study the distribution of a selection of strategies.

6.1.2.2 REDUCTION VS. EXPANSION

Translators seem to prefer to start off by submitting a longer query and gradually trim away portions to increase their chances of finding a match (recall). This trend is confirmed by the contextual inquiry study carried out by Désilets *et al.* (2009), who reported that "[…] subjects seemed very adept at scanning a list of potential solutions, and rapidly sifting grain from chaff," particularly when the resource used was a corpus-based one, such as a list of Google hits or a bilingual concordancer, where bad or irrelevant solutions are likely to be mixed with some good ones. In a previous paper, Désilets *et al.* (2008a: 341) made a claim based on their contextual inquiry results:

> When translators consult a resource (e.g. Terminology Database, Translation Memory) to resolve a translation problem, they seem to care more about recall than precision. In other words, translators do not mind seeing a list of mostly poor suggestions, as long as it contains at least a few good ones. Translators are highly skilled at quickly scanning lists of potential solutions to a translation problem, and identifying which ones (if any) are most appropriate for their current situation.

Hutchins (2005: 13) had previously noted a similar trend when stating that "[t]ranslators find often that inexact examples are as important for them as exact repetitions" but often "too many irrelevant examples [are retrieved] or too many potentially useful examples are missed." Redundancies in the results are sometimes an added issue because they impede an efficient retrieval and display of the results. This is often the case with concordance searching on a large TM, where the amount of time needed to evaluate the results might even prove counterproductive for the translator (Benito 2009). On the other hand, an empty search (i.e. one with no results) could be seen as more detrimental than a result-rich query.

Interestingly, translators' inclination toward greater recall seems to run counter to Web search behavior, where users were found to increase the length of the query in the course of the session, thus narrowing the information need and increasing precision (Huang *et al.* 2003 in Jansen *et al.* 2009: 1360). In Web searching,

> the query length, measured in terms, is longer at the end of a session relative to the beginning. [Researchers] further noted that the query terms in the beginning of the sessions were more general than those at the end of the sessions. This suggests that these users go through a process of query reformulation to narrow their information need and that there may be a correlation between longer queries and more specific information expressions (Jansen *et al*. 2009: 1360).

6.2 TOOL SETTINGS

String length has been presented as one way of balancing recall and precision when submitting a search. In this section, tool settings will be presented as an additional means to fine-tune retrieval. Settings and filters will be systematically analyzed as was done with string length: in the overall dataset of 724,000 strings and in the two subsets of search sessions and spot searches.

To study tool settings, Euramis logs will be broken down into individual components, each containing an item of metadata. Every feature of the way in which translators use Euramis will be discussed across all languages by means of frequency counts. Only the "sentence" field where the text string is contained will not be taken into consideration. It is the only 'open' field where any value can be entered and it has already been discussed in quantitative terms in the previous section. Whenever possible and meaningful, individual filters will be combined and their joint distribution will be discussed where patterns could emerge to shed further light on search strategies and information needs.

6.2.1 AUTOMATIC METADATA

As described in Sub-section 3.2.3.2, a log from the Euramis concordancer contains the following metadata: date and time stamp, (user ID)[127], institution code, source language, target language, search mode, search interface, searched database, search method, execution time, number of results. If *advanced* search mode is selected, the following filters become available: requesting DG, year, document type, document number, directionality and maximum number of results. However, if the selected interface is Quest, the number of available filters is reduced to the ones in the simple interface.

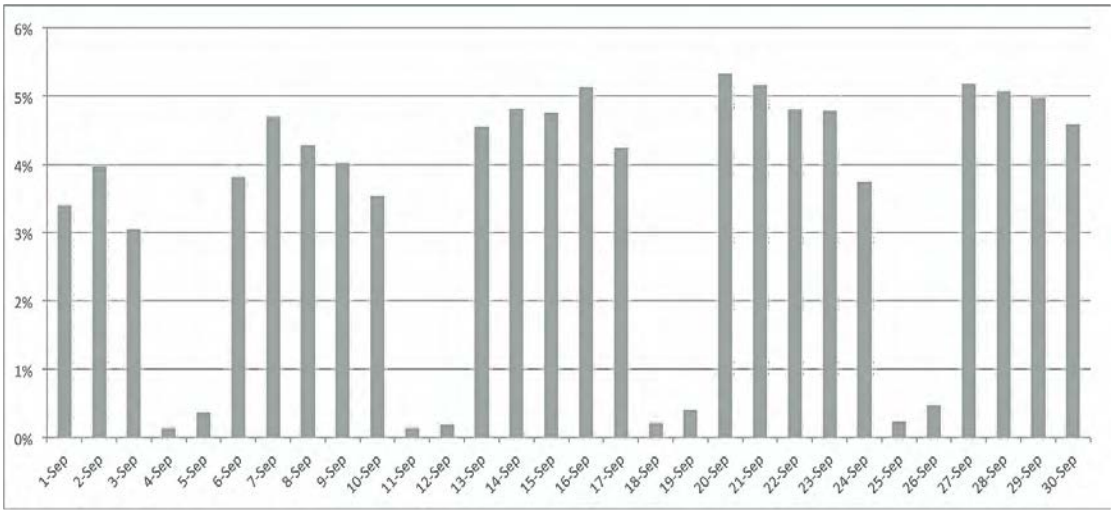
Some metadata items are automatically added by the system while others have to be manually selected by the user. The first part of the analysis will deal with user-independent settings, i.e. the metadata that the system automatically adds to the log. Date and time stamps as well as execution time and results provide a good overview of usage and performance trends. Source and target language selection together with institution code will not be dealt with here[128].

6.2.1.1 DATE AND TIME STAMP

Date and time logs are interesting pieces of information in studying user activity across the month and the time of day. Figure 47 shows, in percentage terms, how many requests were sent to Euramis each day of the month.

[127] As explained in Sub-section 3.2.3.1, this information had to be removed for privacy reasons.
[128] Source language is not relevant in the EN>ALL dataset. For a discussion about source and target languages, see Section 5.4, where institution code distribution is also dealt with.

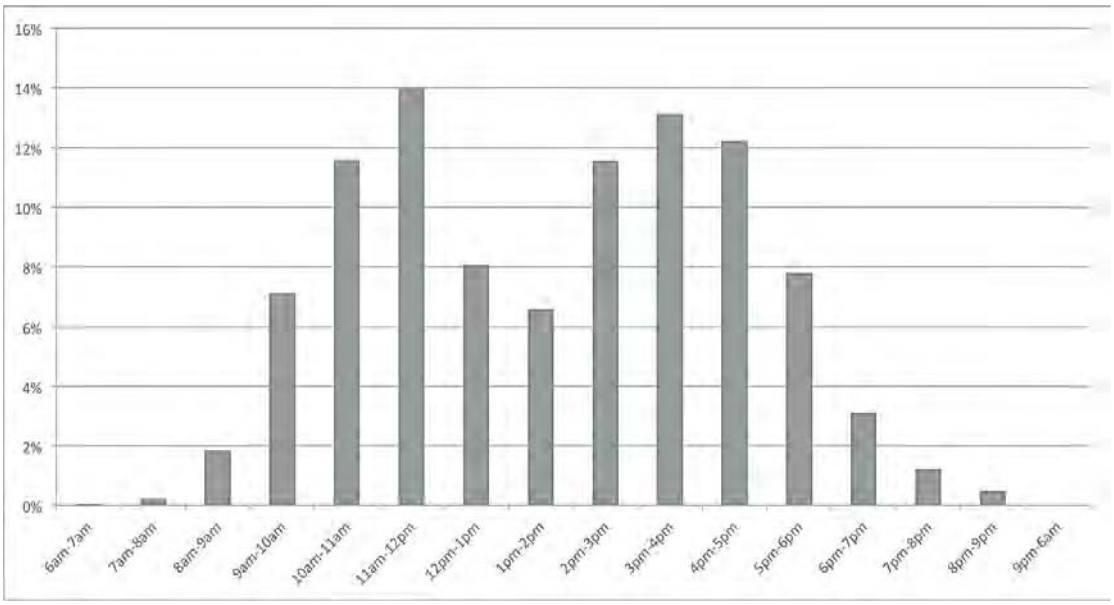
Figure 47. Distribution of requests each day of the selected month (September 2010).



The first thing to be noted is that some activity, however little, was logged for every day of the month, including weekends. As discussed in Section 5.4.1, lower activity is usually expected at the beginning of September and the peak is reached in the very last week, possibly matching an increase in the workload. The busiest day was clearly September 20th and it may be partly due to a presumably higher workload for the EP, in that September 20th was session week in Strasbourg. On the other hand, workload at the EC is expected to remain fairly constant over the month.

Figure 48 shows how activity in Euramis is distributed on an hourly basis over a single day, between 6am and 9pm. Activity clearly peaks just before lunchtime and rises again in mid-afternoon. This piece of information might be useful to try and monitor whether activity peaks cause longer response time.

Figure 48. Distribution of searches on an hourly basis.



6.2.1.2 EXECUTION TIME AND RESULTS

Execution time is probably less interesting for this study as it still relates to the performance of the system. If users have to wait too long for results to be displayed, they are likely to move on to the next best tool for their current need and should this happen

too often, users are likely to give up on using the tool. Statistics from the EN>ALL dataset show that over half of the results (54.10%)[129] were returned in less than one second[130]. Over one third of the results (36.62%) were retrieved and displayed in between one and three seconds. Instances where the system took longer than 4 or even 10 seconds were much rarer (7.42% and 1.84%, respectively). These results could be considered indicative of the "tolerance" of users towards system response time, which should not exceed a couple of seconds' waiting time. In the case of Euramis, execution time is closely linked to the number of results retrieved, the specific search submitted and also the size of the repository to be queried.

In the Euramis interface, the selected maximum number of results indicates the amount of results to be displayed after each search. The default number is set at 30 but the user can manually increase the amount, provided the advanced search interface is selected. Users can choose from a set of options (50, 100 or 200 results), but this feature does not seem too popular, in that the default value was changed in only 4.14% of the searches. In the event that the system finds at least one result beyond the set maximum, a button will be displayed in the results page where the user can click to see more results. Changing the number of results displayed can be useful when users aim for a very wide overview of contexts or translation suggestions or when they know there are many results and do not want them to be limited to the first 30 only. Counter-intuitively, the architecture of the system is such that a search involving a very large pool of results (e.g. common words and a common language pair) will be faster than a search into an unusual language pair with an unusual wording. In the former case, the maximum number of results is reached much faster due to the large pool of data. Execution times might be reinterpreted in the light of this feature, meaning that the vast majority of the searches involves a combination of elements (e.g. language pair and word frequency) for which the system has a large pool of data.

Success rate is measured by looking at how many searches produced at least one result (i.e. a potential solution to the translation problem) as opposed to failed searches that returned no results at all and therefore could not be of any use to the translator. A closer look into retrieval effectiveness shows that for almost one third of the searches (31.34%) the users did not get any results from the system, which means that slightly over two third of the searches (68.65%) were successful. The underlying assumption is that a search returning no results is useless by definition; in other words: "Looking up a chunk and retrieving no translation units is a waste of time" (Melby 2006) — whereas getting at least one result turns the search into a potentially useful one. A more precise break down of the retrieval effectiveness in Euramis is provided in Table 22.

[129] This value also includes the instances of the so called "time-out", a situation in which the system automatically stops the retrieval because it is taking too long. This instances are logged by the system as -1 and account for less than 1%.
[130] See Appendix C – Table C.9 for comprehensive statistics.

Table 22. Percentage distribution of failed and successful searches for each of the three subsets (724,000, session and spot) with a further breakdown for successful searches.

	Overall (724k)	Sessions	Spot
% zero	31.33%	50.07%	20.83%
% successful	68.66%	49.92%	79.12%
SD	1.64%	3.16%	1.77%
Success Result Range[131]			
1 to 29	27.62%	20.38%	31.75%
30	39.55%	28.20%	45.83%
> 30	1.48%	1.34%	1.57%

A more precise breakdown within the sessions subset (not reported in Table 22, see Appendix C; Table C.1) shows that only 29% of the first searches in a session returned at least one result whereas the percentage of success increased considerably (65%) when the remaining part of the session was examined (i.e. from the second to the last search). The increase in the number of results suggests that the change in search strategy (query refinement) was more effective and satisfied the current information need. The "spot" column in the table above seems to confirm this view in that there is only one unsuccessful search every five request[132]. Getting results at the first attempt automatically reduces the need of submitting a new search. The lower half of Table 22 (result range) reveals a more homogeneous picture. The most likely event is a search that returns 30 results (i.e. the default value). Thirty has been kept separated from the 1-29 range because results could in fact be 30 or more. In the latter case an option appears on the screen inviting the user to browse through additional results. Obtaining more than 30 results occurs more seldom because this option is closely linked to the instances where the maximum number of results was increased (i.e. 4.14%).

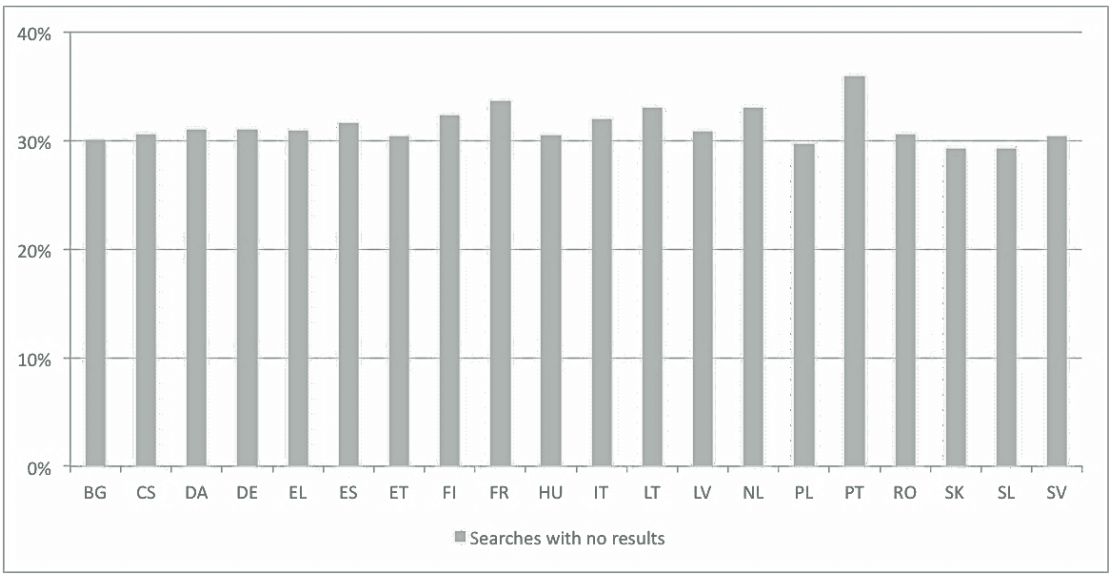
The number of failed searches varies distinctly between each subset, as can be inferred from SD values. When the whole dataset is considered, the least variability is found between the percentages of failed searches across languages. For the main dataset, the ratio between no results and successful searches is 1:2 but for sessions it is in fact 1:1. This may be one of the reasons, if not the main reason, why a spot search develops into a search session. The user is not satisfied with the results provided by the system and decides to change search strategy and submit a different query.

Even though the overall average of no results is 31.33%, some variability is expected across languages. A more accurate break down of unsuccessful searches per language is provided in Figure 49.

[131] Result range percentages were calculated by normalizing the counts on the total number of searches with results for each subset. The values should add up to the corresponding percentage indicated in the "% success" line and not 100%.

[132] Possibly, a number of unsuccessful searches would not be even noted by the user if Quest was used as the search interface because a different resource likely responded and returned some hits.

Figure 49. Distribution of unsuccessful searches (i.e. those producing zero results) across language pairs, normalized by the total number of strings per language subset.



The threshold for unsuccessful searches seems quite stable across the whole dataset, lying at about 30% in the vast majority of cases. The technical reason why results cannot be retrieved was outlined in Sub-section 3.2.3.2 but it is somewhat surprising how consistent the amount of unsuccessful searches is across languages. Interestingly, this result is in line with that of a previous study (Macklovitch *et al.* 2000: 1204), where approximately 39% of the submitted queries[133] returned no match. In particular, the percentage of no results increased from 65% for five-word queries to 78% for seven-word queries up to 100% for fourteen-word queries and the non-response rate was not expected to change considerably after the database size had increased (Macklovitch *et al.* 2008: 414). A search without results may have several causes, e.g. the matching string missing from the database or a problem in the retrieval of the searched-for string that can be subsumed under the label "sparse data problem". Johnson *et al.* (2006) exemplify this by showing that a vocabulary of 25,000 words can generate almost 625 million distinct bigrams and over 15 trillion trigrams, most of which will very likely not be found in the repository.

6.2.2 SEARCH SETTINGS

In the following subsections, the attention will shift from the information automatically added by the system to the available options for refining and filtering the searches which can be manually selected from the chosen interface, i.e. Quest, Euramis Simple or Euramis Advanced.

6.2.2.1 INTERFACE AND SEARCH MODE

The different ways to access each tool have been discussed in Section 3.2.3 and only the main points will be briefly reprised here. Quest has a much simpler Web interface than Euramis, where only source and target languages and a choice of two search methods are available but the search can also be launched directly from the text editor thanks to a

[133] The paper does not provide an exact amount of the total queries examined but it may be interpreted either as the total number of queries submitted between 1997 and 2000 or as the total queries submitted in one month.

customized Microsoft Word macro. If results from Euramis are displayed, the Euramis result page appears nested into the Quest result window and then the user can switch to any of the other resources available. Euramis can also be accessed as an independent tool via the intranet portal, a Word shortcut or by going directly to the relevant page in the Web browser. Once the user opens the Euramis interface, either simple or advanced, some settings can be customized.

Statistics about interface usage show that Quest and the Euramis Portal are used with a ratio of about 2:1 in the overall dataset (63.23% and 36.76%, respectively). The same ratio can be found in the case of spot searches (66.91% and 33.08%) but it tends to converge when search sessions are considered (56.97% and 43.02%)[134]. Search sessions seem to be more common for the Euramis Portal. However, both SD and CV are higher in this case, suggesting that there is a less homogeneous behavior on the part of the translators for the different languages. For example, BG and EL are the fondest users of the Portal whereas EL and NL are the top two users of the metasearch engine. The overall popularity of Quest may be tentatively explained by two factors. On the one hand, it can be directly accessed from the text editor, where the source text portion is selected by highlighting and not typing or copying/pasting; on the other hand, Quest hosts multiple resources, of which up to four can be queried at the same time. Pooled resources, intuitiveness and fewer options seem to be the winning features.

Whenever a search is submitted, the system recognizes the type of resource and the interface used (i.e. simple or advanced). An overview of the distribution of searches for each interface and string group is provided in Table 23.

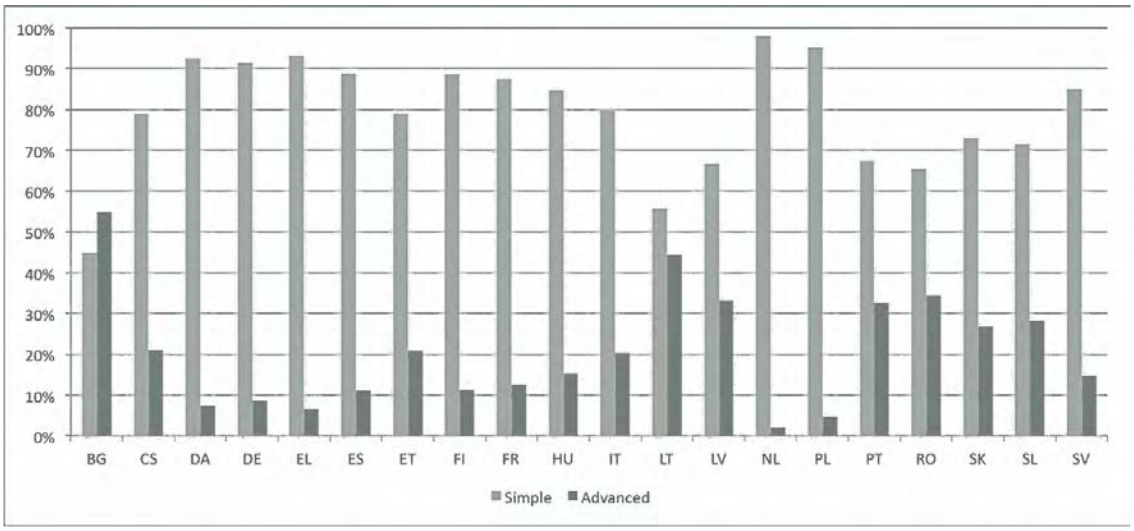
Table 23. Percentage distribution of search modes for each group of strings.

	Main (724k)	Sessions	Spot
Simple mode (S)	79.39%	74.77%	82.19%
Advanced mode (A)	20.60%	25.22%	17.80%
SD	14%	17%	12%
Quest (included in 'S')	63.23%	56.97%	66.91%

Overall, simple mode was used with a ratio of about 5:1 with respect to advanced mode. As anticipated in Chapter 3, advanced filters are not too common, possibly due to the longer response time in the case of refined searches. The percentage of advanced searches increases slightly in the case of sessions but falls again when spot searches are considered. This suggests a possible relation between search sessions and more complex queries involving different filters. However, overall averages tend to give only a partial view of the actual distribution of search modes so the distribution of simple and advanced interfaces was calculated for each language subset (Figure 50).

[134] See Appendix A for the complete tables.

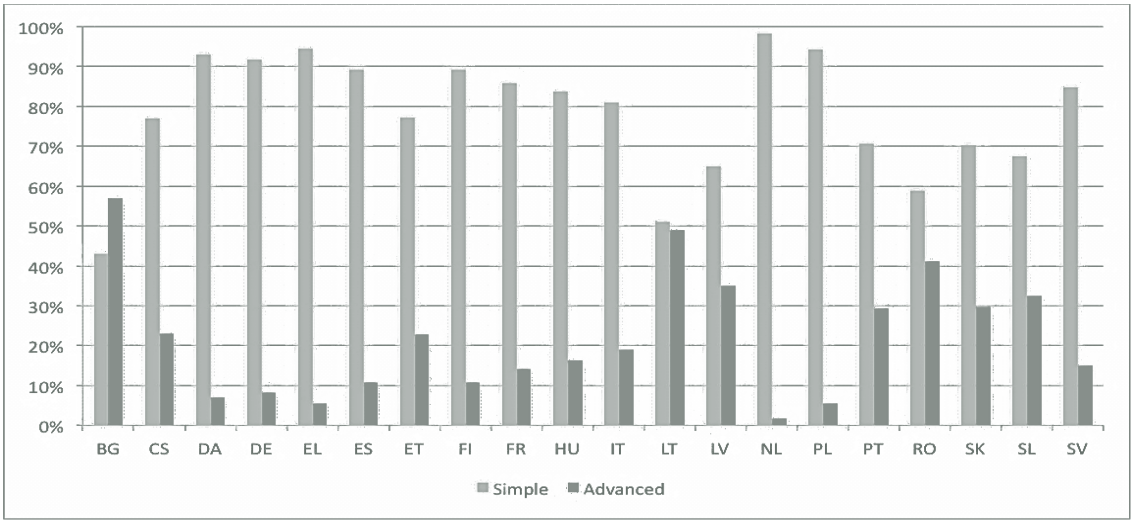
Figure 50. Distribution of search mode (Simple and Advanced) across the target languages (normalized by total strings per language subset).



As expected, advanced mode is used very little with a few notable exceptions: BG has more advanced than simple searches, LT has close percentages between the two and LV, RO and PT[135] have the highest percentages of advanced searches. Except for PT, they are all 'new' languages though they belong to different families.

Advanced mode implies settings selection, which in turn contribute to increase the precision of the search. Previous results have shown that higher precision (i.e. more limitations) can contribute to unsuccessful searches hence the link between advanced search mode and search effectiveness can be quickly and intuitively established. The cross-tabulation in Figure 51 aimed to check whether advanced mode could be related to an increase in the number of zero results, i.e. whether by selecting filters the number of unsuccessful searches would increase.

Figure 51. Distribution of unsuccessful searches (zero results) between simple and advanced mode for each target language, normalized by the total number of searches per target language.



[135] PT translators at DGT have an additional shortcut that launches queries in advanced mode directly form Word which could explain the higher amounts of advanced searches for this language.

This chart reproduces quite faithfully the situation that was outlined in the previous chart (Figure 50). Again, the only exception in the general trend is BG, where the zero results in advanced mode are higher than those in simple mode. In a few instances (e.g. LT and RO), the gap between the two columns is less pronounced, suggesting either that simple mode searches were more successful or advanced mode searches produced more zero results for them than other languages. A sample check showed that both hypotheses are possibly correct: there can be an increase in the number of zero results in advanced mode parallel to an increase of successful searches in simple mode. Other languages, on the other hand, show the opposite scenario: more zero results in simple mode and less zero results in advanced mode, with smaller delta values as a result. The sample check was conducted as follows: two languages with notable differences in the two charts were selected (i.e. SL and LT) and checked against one language where proportions are maintained across the two charts kept (FI). Two proportions were calculated in order to obtain the expected amount of zero results for both S and A mode, had proportions been perfectly maintained. For each language: [(tot mode S):(tot mode A)=(zero mode S):x] and [(tot mode S):(tot mode A)=x:(zero mode A)]. The results were compared to the actual values of x. Compared to the expected results for x in SL and LT, there was an increase in the number of zero results in mode A and a decrease in the number of zero results in mode S; for the control language (FI), the opposite trend was registered with smaller delta values in the beginning.

An additional perspective will be provided as for possible causes of unsuccessful searches, which cross-tabulates the number of failed searches with the submitting institution, to see whether there are any relations between the two (Figure 52).

Figure 52. Distribution of unsuccessful searches (zero results) across different submitting institutions, normalized by the total number of searches for each institution.

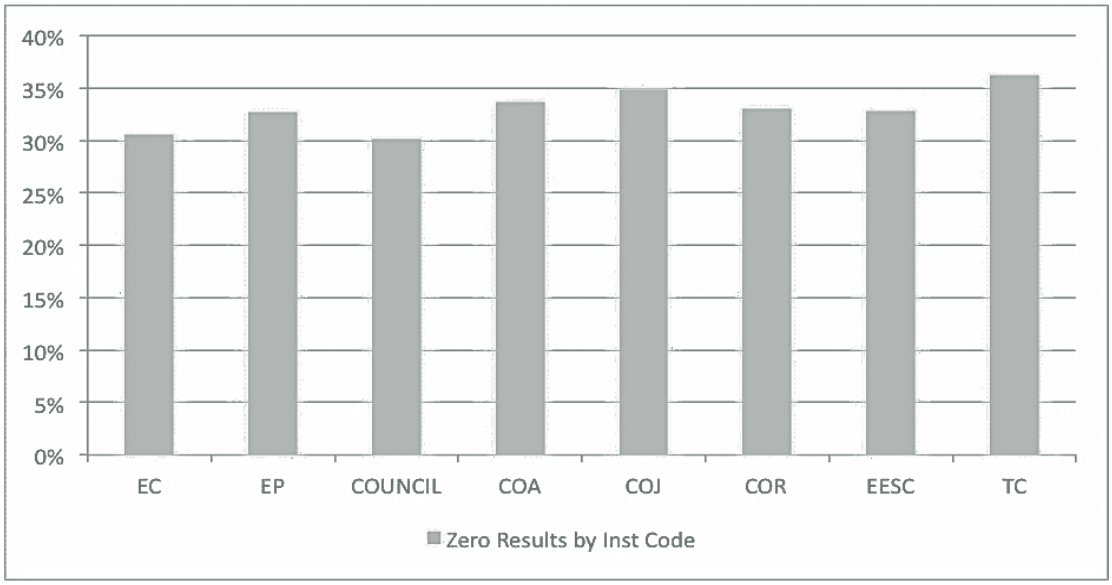


Figure 52 shows that there are no outliers among the institutions and, at a closer look, the percentages of unsuccessful searches are very close to the ones highlighted in the analysis per individual target language. The hypothesis of a baseline failure rate for IR systems of about 30% seems to hold.

6.2.2.2 SEARCH METHOD AND DIRECTIONALITY

Search method has a different location on the search interface and is the only setting offered in the main search page of Quest. With this setting, the system is told how

stringent the retrieval of matches should be, i.e. high (exact), medium (basic) or low (any word) (see Sub-section 3.2.3.2). The default value for Euramis is "basic", i.e. all words in the source (but a few stop words for English and French) have to be matched. In Euramis, the user can choose between two alternatives (basic/exact) in simple mode, which become three in advanced mode. Quest only gives a choice between 'exact string' or 'all words' (i.e. any word). The possible combinations between search mode and search method are summarized in Table 24 together with their distributions.

Table 24. Distribution of search method according to search mode for the main dataset (724,000).

	Basic (B)	Exact (E)	Any (A)
Simple (S)	74.09%	<1%	5.69%
Advanced (A)	17.50%	1.71%	1.00%

Basic search is the default value for both simple and advances search interfaces. Results clearly show that the search method filter is not particularly popular among users. In particular, exact search is a common strategy to get more targeted results in Web search (where it is usually started by adding quotes at both ends of the string) but concordance users seem to prefer noisier results from Basic search.

Similar to search method is the filter of directionality, which determines how extended the search can be, i.e. how many language combinations should be allowed. The default option is 'indirect', i.e. the system looks for matches in both directions, considering all results for one language irrespective of its status as source or target. Alternatively, the user can select either 'direct' or 'reverse' mode from the advanced interface to limit the search to one directionality only. A thorough explanation of the inner workings of directionality is provided in Blatt (1998: 97ff.), even though it refers to a much older version of the software. Unsurprisingly, the overwhelming majority of the searches are performed using the default indirect mode with some 2.5-3% of searches submitted using direct mode, whereas usage of the sole reverse mode is negligible (<1%).

6.2.3 DOCUMENT-BASED FILTERS

The remaining filters are all related to specific information (metadata) characterizing the segments to be retrieved. In this case, the filters do not affect the overall search process but rather make the extraction process more selective because they impose some conditions for the match. More specifically, the user can restrict the search to segments stored in a particular database or filter results by the translation requesting service and document type or number. Another filter refines the search by year(s). All these filters can be combined within the same search, but this analysis will only consider one filter at a time. For every filter, there is a default value that is automatically used by the system. Consequently, all fields in the logs are always filled with the default value unless a different parameter is specified by the user.

6.2.3.1 DATABASE FILTER

The database filter is the only filter that can also be selected in the simple interface and basically represents the available internal Translation Memories. TMs in Euramis are organized in several databases that can be domain specific (e.g. 'Budget') and/or dedicated to a specific institution (e.g. 'EP-Budget'). Some databases are interinstitutional whereas others are only accessible by a specific institution. As of 2010, there were 10

interinstitutional TMs, grouped according to the institution owning them. Table 25 provides an overview of shared resources among institutions and dedicated TMs with a brief description of their content as found in information material for the EC and EP.

Table 25. Overview of available TMs in Euramis. The first half of the table lists the interinstitutional memories, the bottom half lists TMs accessible to a single institution.

INTERINSTITUTIONAL/SHARED			
Ownership	**Name**	**Type**	**Description**
Commission			
	Legis-Juris	final	Legislation and jurisprudence
	Budget	final	Budget
	Nomenclatures	final	Combined Nomenclature
Parliament			
	EP-Committees	working	All Committee documents related to procedures and documents dealt with in the Plenary
	EP-Standard	working	n/a
	EP-Basic References	final	Final versions of basic EP reference documents: Conventions, Agreements, documents related to EP's Human Resources Policy and activities
	Legis-Process	final	Final documents involved in legislative procedures
Council			
	Council-Master	working	Council documents and related background information
EESC/CoR			
	EESC/COR	n/a	All EESC/COR administrative and consultative documents
Court of Auditors			
	ECA	final	Final ECA reports and reference documents on audit methodology and administrative procedures.
DEDICATED/INTERNAL			
Ownership	**Name**	**Type**	**Description**
Commission			
	ExternalRelations	n/a	RELEX family (last 5 years)
	Policies	n/a	Operational DGs (last 5 years)
	Services	n/a	Internal and general services (last 5 years)
	Archive	final	Old DGs and services (until 1999) and current DGs' old documents
	Web	n/a	Web translations
	NormativeMem	n/a	Standard documents
	OtherDoc	n/a	COM, SEC, other institutions, international organizations
Parliament			
	EP-Budget	n/a	

If no database is selected, the default value is "search all available databases" (i.e. "*") and this happens over 98% of the times. Despite the very small percentage of searches filtered by TM selection, an additional count was performed to see how often searches were refined by database in each search mode. Results[136] show that in simple mode the

[136] Results normalized by the total count for each mode. See Appendix C – Table C.6.

database filter was selected in only 1% of the searches and for advanced mode the percentage was slightly higher (mean value of 2.71%). Averages for individual languages, however, ranged from slightly above zero (e.g. SV, ET and DA) to an outlier value of almost 16% for German[137]. As previously noted for search mode, the analysis also considers sessions and spot searches. Percentage values are slightly higher for sessions (1.32% for simple and 3.08% for advanced) and lower for spot searches (0.83% and 2.40%, respectively).

Users can select multiple databases at once so it would be interesting to have a closer look at the most popular databases, despite the small number of overall searches. A full list of the available databases with a short description of their content is provided in Table 25. Percentages are in this case negligible, so the ranking will be based on absolute counts from the main dataset (EN>ALL) and only the most popular databases will be mentioned. Council-Master (shared), Policies (EC, internal) and Web (EC, internal) are the databases that appeared most frequently as individual selections. Multiple selections ranged from 2 to 18 databases in the same search but in this analysis only combinations exceeding 100 searches per TM were taken into account. The most popular combinations of databases (in ascending order according to the number of resources selected at once and frequency count in brackets) are:

1. Legis-Juris/EP-Committees/Council-Master/Legis-Process (387)
2. Legis-Juris/EP-Committees/EP-Standard/EP-Basic-References/Council-Master/EESC-COR/Legis-Process (134)
3. Legis-Juris/Budget/EP-Budget/EP-Committees/EP-Standard/EP-Basic-References/Council-Master/Legis-Process (286)
4. ExternRelations/Policies/Services/Web/NormativeMem/Legis-Juris/EP-Committees/Council-Master/Legis-Process (118)
5. Legis-Juris/Budget/Nomenclatures/EP-Budget/EP-Committees/EP-Standard/EP-Basic-References/Council-Master/EESC-COR/ECA/Legis-Process (345)

Some combinations (e.g. 1) only rely on shared Translation Memories while others combine shared resources with internal memories.

6.2.3.2 REQUESTING SERVICE AND DOCUMENT TYPE

Another filter enables users to refine the search according to the requesting service, i.e. who commissioned the translation (from a Directorate General to a parliamentary group or committee). Instances of advanced searches where the DG filter was selected amounted to 3.51%[138] for the whole dataset. CV is close to 1, suggesting some differences between languages. The range is similar to the one found in the database filter, going from values close to zero (e.g. SL, DA, EL) to over 18% in the case of Spanish. Search sessions registered a higher mean (5.41%) while the mean for spot searches was 4.36%. Multiple DGs can be selected, to the extent that in some cases several dozens DGs were used for the same search. Once again, the top three individual DGs have been established on the basis of absolute frequency counts. Web is the top selected requesting DG, followed by ESTAT (Eurostat) and ECFIN (Economic and Financial Affairs), suggesting that those queries belonged to specific domains.

For filtering according to document type and number, a specific code designating the type of documents or a more specific numeric reference for the document can be selected. Unlike the previous two cases (i.e. requesting service and database), *Doc. Type* and *Doc.*

[137] CV for advanced mode was close to 1.4 against values of hardly above 0.5 in the other cases.
[138] 4.75% if the mean is calculated from the individual means for each language.

Num. were not used by all languages. In fact, just a small group of languages used it at all and percentage values normalized by the total advanced searches did not reach 0.1%. The only languages using either filter on a few occasions were BG and HU. Given the negligible values, neither the list of selected documents nor percentage values for sessions and spots will be examined.

6.2.3.3 YEAR(S)

The *Year* filter filters segments so that they match the selected year. This information is found in the metadata attributed to the document, which justifies instances of "2011" in the logs even though the logs were collected in 2010. The *Year* filter was used on average in almost one fifth of the total searches for advanced mode (19.87%)[139] with the usual oscillations between sessions (21.03%) and spot searches (18.97%). The reason for the popularity of this filter may reside in the way results are displayed by the system. Retrieval works on a "first come, first served" basis, in that the concordancer searches the database following the chronological order of upload, hence "older" documents are the first to be searched. This implies that more often than not, top matches displayed are not necessarily the most recent ones, but simply the first ones to be found by the system. If users want to retrieve more recent documents, they have to manually refine their search.

A user can choose between single and multiple years. Multiple years generally range from 2 to a dozen with a couple of outliers where whole decades were selected. Similarly to string length, the most frequent type of multiple year selection covers two years (almost 12,000), followed by one (~10,000), three (~10,000) and four (3,000); very few instances of longer time spans were found. However, there were only 9 different combinations of multiple years when two years were chosen and only 6 combinations for three-year selections. This means that most of the searches using the *Year* filter are concentrated around a small number of very popular years as detailed in Table 26.

Table 26. Distribution of the most popular single and multiple selections of years.

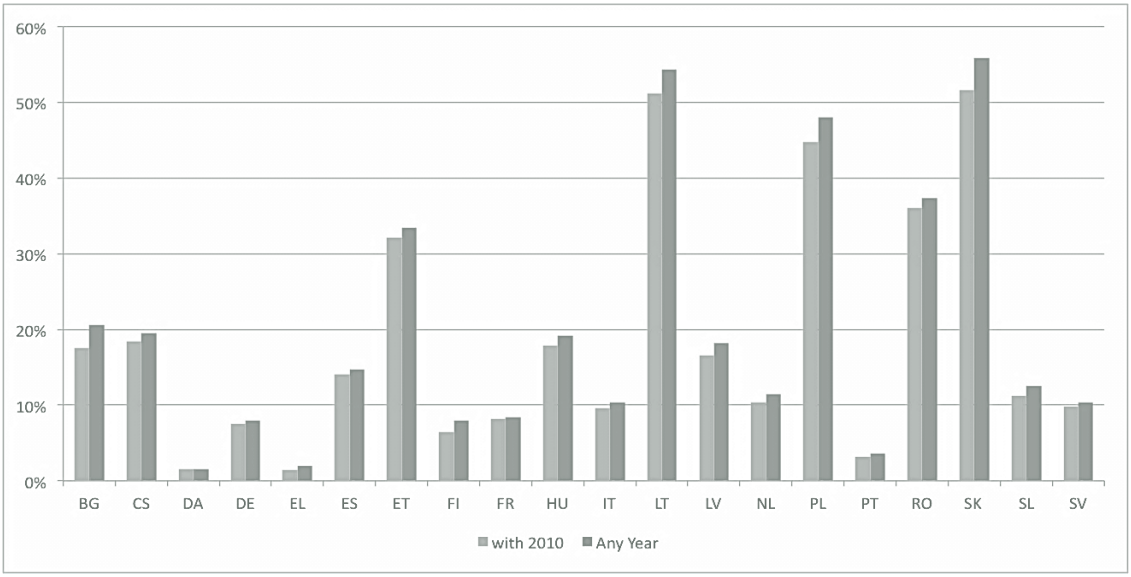
SELECTED YEAR(S)	% (on tot_ADV)	total count (724k)
2010, 2009	7.95%	11762
2010	5.56%	8232
2010, 2009, 2008	5.30%	7847
2011, 2010, 2009	1.36%	2009
2010, 2009, 2008, 2007	1.17%	1724
2009	1.09%	1619
2011, 2010, 2009, 2008	0.68%	1001

Selections with smaller counts than the ones provided in Table 26 did not seem relevant in terms of size. In all but one instance in Table 26, the current year (2010) and the previous year (2009) appear, whereas the upcoming year (2011) features twice. When multiple years were selected, they were usually consecutive years. These are good indicators that the main reason for using the *Year* filter was to get the most recent results displayed. With only data from 2010, there is not enough evidence to establish whether this was incidental (e.g. due to terminological changes introduced by the Lisbon treaty) or a trend that occurs systematically every year. The previously cited Contextual Inquiry study (Karamanis *et al.* 2010) substantiates the hypothesis of a common preference

[139] The average calculated directly from the total (724,000) is actually higher (25.01%).

among translators, who are reported to pay attention first of all to the date of creation and the author of a given TM segment. Based on the present results, the most common filtering strategy covers a 1-3 year span in anti-chronological order. However, differences can be expected among individual languages and so a breakdown of the *Year* filter across all language pairs is shown in Figure 53 in a two-bar chart; one bar includes all combinations where the current year (2010) was included, the other provides an overview of the overall distribution of the Year filter.

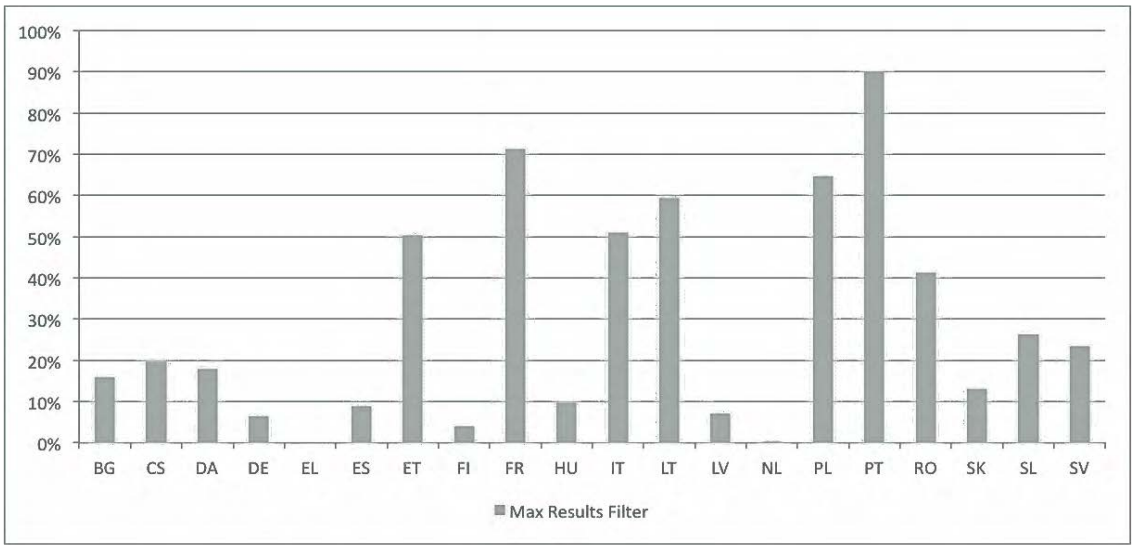
Figure 53. Distribution of the Year *filter compared with the distribution of searches where "2010" was selected. Results were normalized by the total number of advanced searches per language and show that 2010 was selected in the vast majority of the searches.*



There are no noteworthy differences between the two bars for each language, though LT, PL and SK, seem to particularly favor the possibility of filtering results by year, while DA, PT and EL hardly ever chose it.

It is interesting to compare each right bar (i.e. any year) with the corresponding bar in Figure 50 showing the overall distribution of simple and advanced search. PT had a ratio of advanced searches above average (above 30%) but for this filter almost no activity is found. Conversely, PL translators used the advanced filter in less than 5% of the cases in the earlier chart but they use the *Year* filter in almost every other advanced search. This means that PT translators are more interested in other types of filters than the chronological one, whereas filtering by year is very important for PL translators. This last statement can be quickly checked by looking at another filter that is used relatively often, i.e. the maximum number of results. The same distribution calculated for the *Year* filter is reproduced for this filter but a different scenario is obtained, as shown in Figure 54.

Figure 54. Distribution of the Max Results filter across target languages, normalized by the total number of advanced searches per language.



By looking at the chart, it becomes evident what filter PT translators tend to use: 90% of the advanced searches by PT translators use the Max Results filter. A similar preference for this kind of filter can be found for FR and PL. For PL the total of the two filters (Year and Maximum number of results) actually exceeds 100%. This can be explained by the fact that translators can select multiple filters during the same search; for PL it means that there are instances where both filters are selected at the same time. During informal interviews with translators from various language units, one PT translator pointed out that he had set up Euramis in such a way that it would automatically use advanced mode by default so as to obtain the largest number of results allowed. Because of the limited use of such advanced features, this result may quite convincingly be an example of idiosyncratic behavior of a small group of users.

6.2.4 FILTERS OVERVIEW

In this chapter, automatically stored metadata and available filters have been studied separately to highlight some of the main trends in user behavior and identify the most popular filters. The main findings are summarized in Table 27 but percentages should not be taken too strictly because, as previously discussed, in some cases (e.g. Database and DG) there is great variability across languages and users can also select multiple filters within the same search.

Table 27. Distribution of non-default advanced filters for each subset of searches. Percentages were calculated as an overall average: first, averages for each language were obtained and then numbers were normalized on the total number of advanced searches (except for the first line).

Filter Type (not default)	Total (724k)	Session	Spot
Mode Advanced (not Simple)[140]	20.60%	25.22%	17.80%
Database (not *)[141]	2.71%	3.08%	2.40%
Method (not Basic)	7.40%	8.37%	6.76%
DG (not *)	4.75%	5.41%	4.36%
Year (not *)	19.87%	21.03%	18.9%
DocType (not *)	0.08%	0.09%	0.08%
DocNum (not *)	0.09%	0.06%	0.11%
Direction (not Indirect)	0.89%	2.74%	3.02%
MaxResults (not 30)	4.22%	5.08%	3.70%

Advanced search mode was used in only one search out of five and therefore any filter (but the *Year* filter) is bound to have low representativeness in absolute counts. This result is not too surprising given that advanced searches are known to take longer. The chronological filter turned out to be the most popular filter for selecting the current or the most recent years. This result indicates that filtering results by year is possibly a more relevant filter for translators than e.g. database (TM) selection, possibly due to the way results are displayed (see Sub-section 3.2.3.2). Moreover, either the *Year* filter does not particularly affect response time or translators consider it so important that they are willing to wait.

Filters for Year, Method and MaxResults, all of them used by virtually all languages, can also be selected in Web searching. In a Google search, for example, the basic search method can be changed into an exact search by adding quotes to the string; the *Year* filter is quite common and it can be added directly from the sidebar to the result page; finally, MaxResults can be adjusted from the setting page, where users are offered the options of 10, 20, 30, 40, 50 and 100 results per page (as of October 2012).

Advanced filters can be seen as a good example of the differences that can emerge between overall trends and idiosyncratic behaviors. In multiple occasions, high values of SD and CV were found which could sometimes be complemented and verified with field observations. Because of the unbalanced nature of many results, advanced filters will not be discussed further and will be only referred to collectively, unless otherwise specified. However, these findings are relevant because they help to make informed choices in the selection of subsets for future stages of the analysis, as they provide a comprehensive picture of most variables.

On the other hand, this could be seen as a first result in the analysis of general trends in the use of CAT tools and concordancers. Irrespective of the language pair, translators seem to resort to search strategies in the same proportions, which in turn closely resemble Web search behavior, with the notable difference of a tendency to reduce string length when queries are reformulated. At this stage, only the search strategy component

[140] Normalized by the total count of 724,000 strings.

[141] Databases could also be searched in Simple mode but in the chart only results for advanced mode are displayed. Average percentage values in the case of simple mode search are: 1.00%, 1.32% and 0.83%, respectively.

has been dealt with from the perspective of interaction with the tool and setting selection, after having looked at the quantitative component of string length. The nature of the query (seen as a Problem Unit) will be addressed in the following chapter.

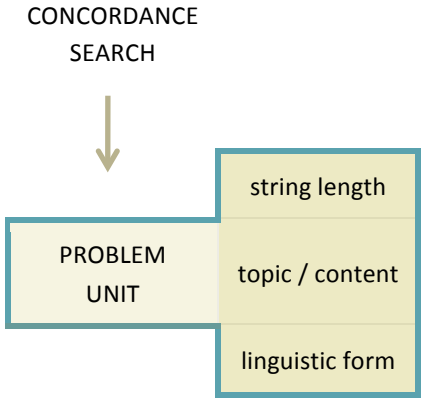
6.3 KEY CONCEPTS

- Distribution of string length for the whole dataset is in line with results from previous studies on the bilingual concordancer *TransSearch* and Web queries. The vast majority of strings is between 2 and 3 words long.
- There are some differences between searches seen as types of problems (types) and instances of problems (tokens).
- Average results for search sessions tend to be consistently higher than for spot searches whereas results for a language subset are generally halfway.
- The most frequent query refinement (search) strategy is reduction, i.e. trimming of the left- or right-most part of the string; resubmission is also frequent.
- String reduction favors recall and is in countertendency with Web searching, where string expansion (i.e. precision) is preferred.
- Consistently, almost a third of the searches in the overall dataset are unsuccessful; the same has emerged in other studies on concordancing tools and Web searching.
- Users tend to maintain default settings and perform simple and basic searches without much filtering.
- Advanced mode is known to make retrieval slower and is used in only 20% of the searches. It often corresponds to the choice of the *Year* filter.

CHAPTER 7:
ANALYSIS OF THE 'PROBLEM UNIT' COMPONENT

In Chapter 6, the strategic component of a concordance search was analyzed in its two subcomponents: string length and search settings. As anticipated in Secion 4.6, concordance searches can be further detailed when the focus shifts to the text string, which has been labeled Problem Unit. This long and articulated chapter will be structured in three main sections, each dealing with one of the previously identified subcomponents, i.e. string length, string content and the linguistic form, as summarized in Figure 55.

Figure 55. Breakdown of a Problem Unit into its various levels of analysis.



In the first part (Section 7.1), string length will be taken into account once more. This time, however, it will not be seen as a component of a search strategy but rather as an intrinsic property of the text string and used to group searches to identify possible approaches to concordance searching. The second part (Section 7.2) will attempt a semantic analysis of the queries so as to achieve a semi-automatic clustering of the strings according to semantic domains. For this study, exsiting content-based approaches to query clustering and categorization in the field of Web serarch logs analysis will be taken into account and adapted to the specific context of translation-related searches. Finally, the linguistic form of the strings will be addressed in Section 7.3. This is possibly the part of the whole study where no systematic quantitative analysis across languages was possible for reasons that will be detailed in due course. Initially, search strings will be analyzed from a syntactic perspective and part-of-speech tagging will be performed to have at least some quantifiable data. Existing problem categories in empirical studies in tranlsation will then be discussed to identify those that can be successfully implemented and operationalized. This small review will lead the way to a more cognitive-oriented discussion of the search strings with a particular focus on the tip-of-the-tongue state. The last part of the chapter will go back to the concept of concordance searches as manifestations of (implicit and/or explicit) information needs and propose a different classification of translation problems into scenarios by taking an Information Retrieval perspective.

7.1 STRING LENGTH

In Section 6.1, string length was described as one way to modulate recall and precision and filter results. However, string length can also be seen as a fundamental element of a search string: a string can range in length from a single word to a whole paragraph. A continuum could be an appropriate visual representation of the possible variety of string lengths (Figure 56). The leftmost end of the continuum would represent single-word queries; these are the hardest to interpret particularly in the spot searches group (i.e. in isolation) due to polysemy and the lack of context. Both factors are likely to negatively impact retrieval precision and produce false positives and irrelevant results. In addition, if the retrieved segments are long, the user may spend more time to identify the corresponding translation for the searched item. This has been partly addressed using word-alignment in transpotting and other highlighting systems both on the source and target side (see Section 3.2). Still, the concordancer does not seem the best solution for this type of short lexical items. As opposed to multi-word units, single words are often best found in a (online) bilingual dictionary, a specialized glossary or a term bank, which may provide a more targeted solution to the problem.

At the other end of the spectrum, full sentences and strings up to a paragraph are to be found. In this case, the query contains too much information and recall will inevitably suffer because precision needs to be too high. Intuitively, for analytical purposes the longer the string, the less clear the problematic item becomes up to the point where one may wonder whether there was a problem in the first place. When the string becomes too long, the system automatically performs the search based on a maximum number of characters (230 in the case of Euramis) and transpotting would no longer help. This type of search is better compared to a standard sentence match from a TM rather than a search for a specific sub-segment. There can be several reasons why a user would decide to search for a long chunk of text, the most likely being document retrieval[142]. Translators working with a TM system usually have a local memory that is automatically generated (as is the case at the EC) or that users populate semi-automatically before translating by using specific interfaces (e.g. Twist and Shout! at the EP). While working, translators may encounter a sentence in the text that is not found in the local TM but that they recall having translated or found elsewhere. Euramis has a function that enables users to download the document from the results page by clicking on the corresponding sentence pair (see Sub-section 3.2.3.2). Long searches may therefore not be necessarily related to a sub-segmental translation problem. As a consequence, in such cases the concordancer cannot be said to serve the purpose of helping translators satisfy a specific information need (i.e. solve a translation problem), which is why long search segments have eventually been discarded from the dataset used for the present study[143].

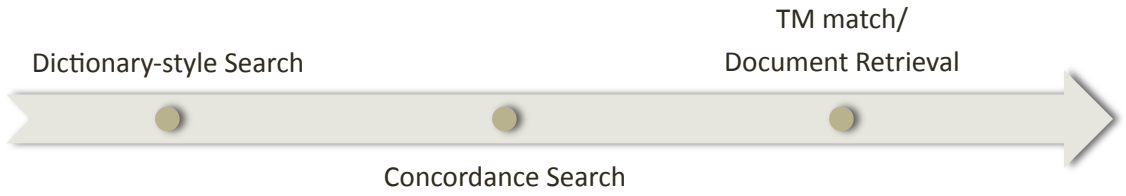
The central part of the continuum (i.e. multi-word units) corresponds to standard concordance searches; the architecture of the system is such that it also well suited for searches of this kind. The three segments that emerge from the continuum are taken to

[142] In these cases, the translator would either search for a long chunk of text or choose some distinctive keywords that would help identify the text. Retrieved documents can then be used as an external reference or can be aligned and imported as a TMX into the workbench. For Macklovitch *et al.* (2000: 1204), length was a variable that correlated with the likelihood of an empty (i.e. unsuccessful) query.

[143] Document retrieval can also be performed using distinctive keywords. Unless the translator specifically states his/her intent, this scenario cannot be distinguished from a standard multi-word concordance search and can therefore not be considered further in the discussion of document retrieval.

represent three main approaches to searching Euramis according to the length of the searched string. They can range from single-word searches to long stretches of texts possibly for document retrieval while the central part corresponds to the concordance-proper search range, i.e. where the concordancer is likely to yield the best results. Figure 56 provides a graphical representation of the main possible approaches to concordance searching that will be later operationalized to study the distribution of each approach in the dataset.

Figure 56. Continuum representing the most likely search approach as search string length increases.



After having identified the central area of the continuum as the most relevant for this study, the question arises as to how long a string should be to fall into the 'concordance proper' part of the continuum. Given the above discussion, minimum length should be two words. Maximum length is more complex to determine because there is no set criterion to determine at what length a translation problem 'dissolves' in a longer string. A well-defined and quantifiable definition of translation problem would be of help but such definition does not seem to exist. Alternative solutions need to be found to determine an operational cut-off length in the continuum.

7.1.1 CUT-OFF LENGTH

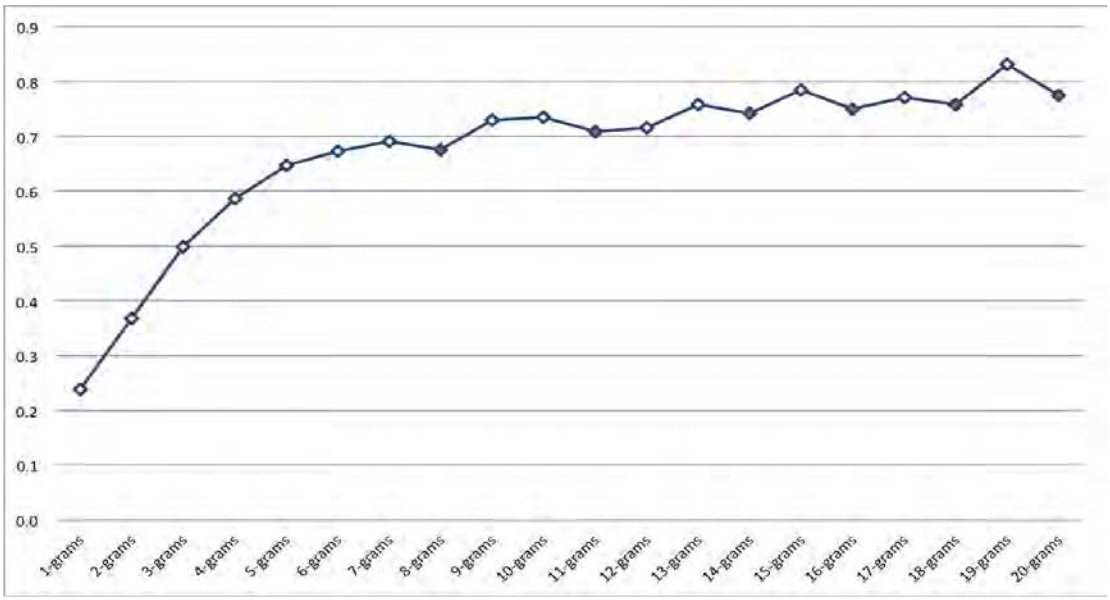
Most of the analyses related presented in this section were carried out in the preliminary phases of the study. As a consequence, they are mostly based on an earlier version of the dataset. On that account, the relevant dataset will be specified at each stage. Despite the resizing of the dataset from one analytical stage to the other, language distributions were shown to remain fairly constant between different versions of the dataset (see Section 5.5).

The analysis started with an attempt at identifying a cut-off length that could be justified by observing the dataset instead of selecting an arbitrary value. The measure chosen for this analysis was the type/token ratio. The type/token ratio is a known measure in corpus studies and "refers to the relationship between the total number of running words in a corpus and the number of different words used" (Olohan 2004: 80). The string is here taken as unit, such that types represent different queries and tokens all instances of queries found in the dataset. The type/token ratio is therefore an indicator of query variation within the same subset: the higher the ratio, the more variation there is. Type/token ratio can relate to concordancer usage in that an increase in the ratio would signify more variability. The segments identified in the continuum (Figure 56) are expected to produce different values of type/token ratios. For example, in Dictionary-like searches recurring searches can be hypothesized if many translators have to resort to the concordancer to look up a specific item and type/token ratio will be lower (i.e. closer to zero) because types will be much fewer than tokens. Conversely, document retrieval is likely to depend more on an individual need rather than a recurrent search pattern because the longer the string, the less likely it will match an existing one. The document

retrieval approach can therefore be identified by higher type/token ratios, i.e. closer to 1. Based on these assumptions, concordance searches are also expected to fall within a more or less specific range of values in type/token ratio.

Based on the 742,033 dataset, the queries were treated as *n-grams* and grouped according to their length, ranging from 1 to 20 words. "N-gram" is a label that is commonly used in Computational Linguistics to refer to a string of words (or characters) of *n* length, hence 1-grams are single-words, 2-grams contain two words and so forth. The average length of an English sentence seems to range between 15 and 20 words according to existing literature[144]. A 20-gram was considered a sentence-like segment – a string long enough to fall into the right-most part of the abovementioned continuum[145]. For each n-gram group, type/token ratio was calculated and results were transferred into a chart (Figure 57). A marked delta between two data points would signify a change in variability, i.e. in one group the same strings are searched over and over again whereas the other group would be characterized by fewer repeated strings. The cut-off length is expected to fall between two data points where a trend change occurs, more specifically after a drop in the type/token ratio value. No *a priori* delta values were hypothesized but it was nonetheless expected that the cut-off would occur somewhere between 5 and 12 words, i.e. a plausible range for substring matches below sentence level.

Figure 57. Distribution of type/token ratio for each n-gram group from 1- to 20 (742,000 dataset).



As expected, there is a steep rise in the type/token ratio for the shorter n-grams, with 1-grams (i.e. strings similar to dictionary lookups) having the highest value. This is due to a higher probability of finding a single word repeated verbatim than a longer string. Delta values for strings shorter than 7 words will not be taken into account further because no drop in type/token ratio occurs. In some cases, the delta is negative because the ratio

[144] Among others: Matthews (1961: 535); O'Neill (2009: 15).

[145] Here are a few examples of 20-grams: 'any fees imposed for the procedures on imported products should be no higher than the actual cost of the service'; 'the commission invites each arm of the budgetary authority to inform the other arm and the commission on its intentions'; 'if no comments are received by that deadline the texts of the replies will be deemed to have been adopted'.

decreased slightly, e.g. between 7- and 8-grams[146]. The largest delta values in absolute terms in the chosen range (i.e. 7- to 12-grams) are found between 8-/9-grams[147] and 12-/13-grams[148]. Based on the delta values of the type/token ratio, the possible upper cut-off length could be one between 8-/9-grams and 11-/12-grams but no value seems to justify the choice of one over the other.

An alternative (or integration) to the type/token ratio comes from another concept used in research in linguistics and corpus studies may provide some help in defining a possible cut-off length. In his studies on collocation[149], Sinclair used the concept of *span*, defined as "the amount of text within which collocation between items is said to occur" (Krishnamurthy 2004: 10). The span was calculated by testing significance for a node word[150] with surrounding words that resulted in a collocation window of −4 and +4 words from the node (2004: 42). Later, the optimal span was recalculated to include 5 words to the left and 5 words to the right of the node (2004: xix). If the window is symmetrical, the total span can therefore include up to 9 (8+1 node) or 11 (10+1 node) words, but clearly "the wider the span, the lower the significance in general" (2004: xxvii). In the case of problem units, they can include one specific troublesome word[151] or a particular combination of words that proves problematic as a whole. To some extent, the text string of a search log resembles a collocational window, though the word "collocation" referring to a problem unit is here purposely avoided (see Sub-section 7.3.2.2). The problem unit, be it a single word or a group of words, can be seen as a node around which the string expands (or shrinks). Therefore, it may seem legitimate to use the collocational window as a benchmark to find an appropriate cut-off length for the searches. For the sake of simplicity, only symmetrical windows will considered, though examples of asymmetrical collocational windows might be just as valid:

> "[…] the span does not have to be equally wide to the left and to the right. For some purposes, for instance if you are looking for subjects of certain verbs, it can be better to just look at the five words to the left: – 5" (Lindquist 2009: 73).

A third justification for the identification of the cut-off length takes into account a cognitive perspective. A search string can be seen to represent a piece of the source text that the translator was focusing on at a given moment (see Section 4.2) and that was therefore occupying the translator's working memory. As noted by Dragsted (2004: 36),

> [w]orking memory is traditionally defined as a readily accessible repository for temporary storage of information that is being consciously processed in any range of cognitive tasks […].

Researchers seem to agree on the limited capacity of the working memory, even though no consensus can be found on the (exact) number of items the working memory can

---

[146] Examples of 7- and 8-grams are: 'setting up an eu rapid response capability'; 'a strategy for smart sustainable and inclusive growth'.
[147] Examples of 9-grams are: 'un convention on the rights of persons with disabilities'; 'the commission should be empowered to adopt delegated acts'.
[148] Examples of 13-grams are: 'internal rules on the implementation of the general budget of the european union'; 'it shall not affect the validity of the delegated acts already in force'.
[149] Collocations are defined as "the co-occurrence of two items in a text within a specified environment" (Krishnamurthy 2004: 10).
[150] *Node* is the term chosen by Sinclair (1991: 115) to refer to the word being studied, whereas *collocate* refers to "any word that occurs in the specified environment of the node", i.e. the span.
[151] This scenario seems confirmed by Dragsted (2004: 57): "the presence of a problematic item in the source text will reduce the number of items in a translation unit, possibly to only one word."

handle simultaneously, which can range from five to nine but is most frequently set at seven[152]. Results from an experiment on segmentation (Dragsted 2004: 123ff.) have shown that the typical translation unit size for professional translators was of about 4 words on average, but could include "exceptionally long" segments of more than 10 words (2004: 126). Longer segments seemed more frequently processed by professional translators than students. It is interesting to note that the cut-off length for "exceptionally long" segments was set at 10 words. Once again, indicators seem to point to a string of about 10 words as the maximum length for the textual units that the translator's attention focuses on.

An additional check was carried out by looking at examples of strings for each n-gram subset. One of the basic quantitative checks in corpus-based studies uses frequency lists, i.e. a list of all the words in the corpus with the respective total number of occurrences. For the purpose of this analysis, a standard word list was not considered very helpful, so frequency lists were constructed on the basis of multiple occurrences of the same string, as previously done by Macklovitch *et al.* (2008: 415). They obtained frequency counts for each length-based class, reporting the top twenty searches for n-gram group ranging from 2 to 4 words. Table 28 compares the top 10 searches for 2- and 3-grams for TransSearch and Euramis, respectively.

Table 28. Comparison of top 10 most searched bi-grams and tri-grams in TransSearch (Macklovitch et al. *2008: 415) and Euramis.*

2-grams				3-grams			
Freq	**TransSearch**	**Freq**	**Euramis**	**Freq**	**TransSearch**	**Freq**	**Euramis**
1195	as such	199	europe 2020	1131	as a result	205	single market act
1046	over time	172	economic governance	994	in terms of	180	europe 2020 strategy
743	consistent with	171	european semester	913	at this time	151	memorandum of understanding
708	in turn	159	impact assessment	887	in conjunction with	132	terms of reference
680	hard work	152	digital agenda	817	in support of	132	level playing field
669	along with	149	flagship initiative	777	with respect to	124	millennium development goals
655	subject to	135	smart regulation	772	as part of	122	credit default swaps
649	based on	128	innovation union	722	in line with	107	lifelong learning programme
609	as required	127	resource efficiency	716	in keeping with	94	credit rating agencies
587	as appropriate	124	legal certainty	707	make a difference	92	enterprise europe network

The first thing to be noted is the considerable gap in the frequency counts between TransSearch and Euramis. If normalization is carried out the picture changes quite remarkably. For example, the most searched-for query in TransSearch has a frequency count of 1195 vs. 199 for Euramis. The dataset in the TransSearch study amounted to over 7.2 million queries, while the Euramis dataset included slightly over 742,000 searches. Expressed in percentages, the two frequencies therefore become 0.017% and 0.027% of the total strings for each subset. However small, Euramis string turns out to be proportionately more frequent than that of TransSearch. The other observation regards the type of searches. TransSearch strings correspond almost exclusively to prepositional

[152] For a more detailed discussion about working memory capacity, see Dragsted (2004: 42ff.).

groups while Euramis searches include lexical items, chiefly compound nouns. A possible explanation for this discrepancy is that Euramis tends to be used self-referentially. EU translators use it to translate EU-related material by accessing previously translated EU texts. TransSearch, on the other hand, is accessed by users who do not necessarily work for the Canadian government or translate parliamentary debates and court rulings. Greater variability should therefore be expected in the lexical queries submitted to TransSearch that do not make it to the top ten. In Euramis, a string like 'as amended' ranks 98th and has a frequency count of 72 (close to 0.01%) while 'in terms of' is 294th and amounts to 47 hits, i.e. 0.006% (vs. 0.014% in TransSearch). Clearly, percentage values from frequency counts are too small to be really informative but they are still useful to put raw frequencies in perspective, if needed.

7.1.1.1 STRING CATEGORIES

The frequency counts suggested that there are some recurring types of searches but no clear or exhaustive categories could be obtained from the results. Moreover, short strings such as bigrams and trigrams, despite being most frequent, are too short to be useful as candidate cut-off lengths. After looking at the top and bottom 50 ranked strings in the whole dataset and for each language subset, some potential recurring categories were highlighted such as such acronyms, formulae, collocations, compounds, prepositional phrases and Eurojargon, but no categorization was deemed so satisfactory as to be systematically implemented. Strings were then grouped according to their length and the top 30 searches for each n-gram subset (from 5 to 15 words) were considered so as to see whether the top searches could be ascribed to a particular category. The top 10 most frequent 8- and 11-grams in Euramis are listed in Table 29.

Table 29. Top 10 most frequent 8- and 11-grams with absolute frequency counts (742,000 dataset).

Freq	8-grams (Top 10)	Freq	11-grams (Top 10)
121	treaty on the functioning of the european union	54	high representative of the union for foreign affairs and security policy
52	acting in accordance with the ordinary legislative procedure	35	after transmission of the draft legislative act to the national parliaments
39	charter of fundamental rights of the european union	33	european convention for the protection of human rights and fundamental freedoms
28	unlocking the potential of cultural and creative industries	29	guidelines for the examination of state aid to fisheries and aquaculture
27	high representative for foreign affairs and security policy	21	convention on the elimination of all forms of discrimination against women
26	committee on civil liberties justice and home affairs	20	international convention on the elimination of all forms of racial discrimination
26	council of bars and law societies of europe	20	advisory group on the food chain and animal and plant health
22	towards adequate sustainable and safe european pension systems	13	establishing the organisation and functioning of the european external action service
21	european network of civil aviation safety investigation authorities	12	aquaculture inland fishing processing and marketing of fishery and aquaculture products
21	action plan on enhancing the security of explosives	11	non acceptance of the appropriate measures for the fisheries insurance scheme

Two main types of strings seem to emerge from Table 29: strings that are exclusively nominal (e.g. 'charter of fundamental rights of the european union', 'non acceptance of the

appropriate measures for the fisheries insurance scheme') and strings that starts with or include a verbal form, although of the type that is closest to a nominal form (e.g. 'establishing the organisation and functioning of the european external action service', 'acting in accordance with the ordinary legislative procedure'). At a closer look, the nominal strings seem to refer to people, organizations or documents and could collectively referred to as *Named Entities.*

7.1.1.2 NAMED ENTITIES

Named Entities (NE) are a very well known category in IR and Natural Language Processing (NLP) because they can be quite challenging to identify and extract, particularly in a multilingual context. A named entity is "a phrase that uniquely refers to that object by its proper name, acronym, nickname or abbreviation" (Strassel 2006: 1) but the category can also include dates, times, monetary amounts and percentages[153] (Sekine & Nobata 2004: 1977; Nadeau & Sekine 2007: 5). One of the reasons why Named Entities are challenging is the variability in the spelling and forms in which they (especially proper names) appear within the same or between two or more languages, which is problematic when searching databases, the Internet and other repositories. This was one of the reasons behind the creation of a multilingual resource such as *JRC-Names*[154] by the European Commission (Steinberger *et al.* 2011: 104). For the purpose of this analysis, the NE category will be sub-divided into four main types of entities on the basis of the examples found in the data. The first sub-category will be labeled "Classic" and includes instances that most closely resemble the standard interpretation of Named Entities such as those listed in Strassel (2006: 1), i.e. person, title/role, organization, location. The remaining sub-categories can be considered *ad hoc* categories of Named Entities that include references to official documents, abstract entities and events. The manual categorization was carried out of the basis of some general "rules" and trigger words found in the string (Table 30). The concept of trigger words can be compared to that of *topic signature* employed in text classification and applied to named entity recognition by Fleischmann & Hovy (2002: 3), i.e. a list of terms that signal the membership of a named entity to a specific category.

Table 30. List of trigger words for the identification of Named Entities, grouped by sub-category.

Classic	Document	Abstract Concept	Event
agency	agreement	action plan/ plan of action	conference
association	budget	framework	day (+ prep)
authority	charter	guidelines	week (+ prep)
body	code of conduct	instrument	
centre	communication	network	
commissioner	convention	network	
committee	covenant	panel	
confederation	directive	policy	
directorate general	dossier	principles	

[153] Dates, times and the other numerical values listed actually evoke the concept of "placeable" in a TM system.

[154] JRC-Names is a free multilingual collection of named entities covering over 200,000 person and organization names in up to 20 different languages.

fund	green paper	programme
(advisory) group	guide	scheme
(high-level) group		
(expert) group		
institute	handbook	strategy
organization	protocol	system
representative	regulation	
society	report	
task force	treaty	

The category *Classic* includes references to persons' titles (e.g. 'Commissioner') and organizations (e.g. 'Agency', 'Institute')[155]. The sub-category *Document* includes several types of (official) written documentation, such as 'treaty', 'charter' and 'directive'. *Abstract Concept* contains relatively loose and sometime blurry concepts that can either refer to an existing entity or be intended in a prospective and general sense but nonetheless have their official denomination (e.g. 'strategy', 'instrument', 'policy'). Finally, the sub-category *Events* contains festivities or titles of conferences.

After populating each category, special attention was given to the overall structure of the string. Trigger words alone were not enough to isolate a string representing a named entity and a few additional parameters had to be born in mind. For a string to be labeled as NE, it had to meet some specific criteria. They are listed below together with examples of strings that do not comply with the stated condition.

For it to be labeled a Named Entity, a string had to:

- contain no verbal forms (e.g. 'memorandum of understanding ~~signed at~~ the ninth meeting of the conference of the parties cop9') or predicative expressions (e.g. 'financial regulation ~~applicable to~~ the general budget of the european communities…')
- be clearly identifiable (e.g. 'report ~~from the commission to the european parliament and the council~~', '~~deep and comprehensive~~ free trade agreement')
- contain no indefinite article (e.g. '~~a~~ strategy for smart sustainable and inclusive growth')
- contain no additional prepositional groups (e.g. '~~as a consequence of the entry into force of~~ the lisbon treaty', 'agreement ~~in the form of an exchange of letters~~')
- have the trigger word as the nominal head (e.g. '~~europol heads of~~ high tech crime units task force').

7.1.1.3 OTHER STRING TYPES

The other type of string encountered in the dataset contained one or more verb form(s). The label *Verb/–ing form* will be used whenever the string has a predominantly verbal nature (e.g. 'should therefore be amended accordingly') or when the first word of the string is an "-ing" form that reminds of citations contained in documents (e.g. resolutions), such as 'having regard to the treaty on the functioning of the european union'[156]. Should they actually be full sentence citations[157], they would fall under the

[155] As previously stated, abbreviations are included in the category and are also present in the dataset, but will not be taken into account in that they are usually found in the 1-gram category.
[156] This string clearly contains a NE, but because of the verb there is no certainty as to whether or not the named entity was the main focus of the translator's attention.

category of TM matches. Unfortunately, the available information was not sufficient to determine whether or not a segment was an actual citation, a portion of a longer sentence or an item in an action list (e.g. 'establishing the organisation and functioning of the european external action service'). Because of this uncertainty, such strings will be given the same label as other verbal forms.

An additional category will be used to refer to self-contained strings without reference to any particular domain (e.g. 'subject to its possible conclusion at a later date'). These strings will be labeled LGP (Language for General Purposes; see Section 7.3.2 for additional details on the LGP labeling).

7.1.1.4 N-GRAMS AND CATEGORY DISTRIBUTION

On the basis of the outlined criteria, a manual labeling of the first 30 strings of each n-gram subset was carried out after looking at frequency counts (Table 31). A clear diminishing trend in frequency counts is to be found, with shorter strings being quite popular and longer strings being hardly sought in the overall dataset. In the case of 5-grams the lowest frequency was higher than some top frequencies for other categories but in the case of 15-grams, the lowest frequency added up to only 4, and starting from 9-grams the frequency was always below 10. Representativeness in strings past rank 30 would be too low and therefore they were excluded.

Table 31 provides an overview of the distribution of each category and sub-category. 'Documents' and 'Classic' are the labels used most frequently and show two contrasting trends. *Classic* NEs are highest between 5-grams and 9-grams but are hardly to be found in longer strings. Vice versa, the *Documents* count tends to increase the longer the string becomes. The remaining three categories seem quite evenly distributed. Finally, the number of strings in the *Unlabeled* category is rather low below 9-grams, then it increases and peaks within the 11-gram subset.

Table 31. Overview of the distribution of the main categories for Named Entity recognition for each sub-set and range values for frequency counts for the 742,033 dataset.

(742,033)	5-gram	6-gram	7-gram	8-gram	9-gram	10-gram	11-gram	12-gram	13-gram	14-gram	15-gram	Total
Classic	7	14	7	10	6	4	3	1	0	2	1	55
Docs	3	3	5	6	4	5	3	8	8	3	9	57
Abstract	7	2	6	3	4	4	2	3	5	1	2	39
Event	0	2	2	0	0	0	0	1	1	0	0	6
LGP	2	1	0	1	1	1	1	3	1	9	3	23
Verb/-ING	1	0	5	4	3	2	2	4	2	4	6	33
Unlabeled	10	8	5	6	12	14	19	10	13	11	9	117
Total	30	30	30	30	30	30	30	30	30	30	30	
Max Freq	100	53	51	121	51	21	54	32	31	16	13	
Min Freq	26	18	14	13	9	8	8	7	5	5	4	
Delta Freq	74	35	37	108	42	13	46	25	26	11	9	

This brief analysis once again highlighted the 8-/9-gram and the 10-/11-gram subsets as dividing points in a general trend. Interestingly enough, these are the same cut-off lengths

[157] *Citations* in EU documents are the sentences "stating the legal basis for the act and listing the procedural steps" and generally begin with "having regard to" (DGT 2012: 60).

previously identified when type/token ratio was calculated and collocational window and working memory capacity were discussed.

This result can be taken as a good indicator as to what an adequate cut-off length may be. The choice seems to be between the two ranges identified by Sinclair (in Krishnamurthy 2004), i.e. the ±4 or the ±5 windows, which translates into 8-/9-grams and 10-/11-grams respectively, depending as to whether the node word is included. 9-grams and 11-grams have similar range and delta values whereas 8-grams and 10-grams have exceptionally high and low delta values, respectively. The most searched-for string in 9-grams was 'standing committee on the food chain and animal health', which is one of the standing committees in the area of food safety[158] whereas 'high representative of the union for foreign affairs and security policy', i.e. the official title of Commissioner Ashton[159], was the most searched-for 11-gram. Both are examples of a non-TM type of search, because in this case both strings represent a single entity – it just so happens that the name is a rather long one. The string 'high representative' is ranked 24th in the 2-gram frequency list, indicating that the title of Commissioner Ashton occurs many times as a search string. On the basis of all of the above, the final decision was to choose 11-grams as the arbitrary cut-off length. Any string above 11 words will be taken as an example of a search following a TM match approach where no particular translation problem can be identified, or at least no problem distinct from the string in its entirety. Indeed the most frequent search for the 12-grams group is a citation, namely 'having regard to the opinion of the european economic and social committee'. After establishing the cut-off length at 11 words to signal a change in the approach to the concordancer, distributions can be calculated for all languages in the final dataset of 724,000 strings as well as for search sessions and spot searches. A customized Python script was developed[160] that would divide the strings for each language subset into three groups: single words, 2- to 11-word string and queries longer than 11 words. The aim was to examine how often each approach (Dictionary-style Search, Concordance Search, TM-Match, see Section 7.1) was used. Results are summarized in Table 32.

Table 32. Distribution of the three categories of n-grams representing three different approaches to concordance searches for each of the three analyzed datasets.

(Token Count)	**1-grams**	**2/11-grams**	**> 11-grams**
***Overall* (724k)**			
Mean	13.82%	83.38%	2.80%
SD	1.66%	1.85%	0.65%
CV	0.120	0.022	0.230
Search Sessions			
Mean	8.01%	88.06%	3.93%
SD	1.12%	1.70%	0.99%
CV	0.140	0.019	0.253
Spot Searches			
Mean	14.74%	82.67%	2.58%
SD	1.94%	2.10%	0.57%
CV	0.131	0.025	0.222

[158] http://ec.europa.eu/food/fs/rc/index_en.html [last accessed: October 2012].
[159] http://ec.europa.eu/commission_2010-2014/index_en.htm [last accessed: October 2012].
[160] The author is indebted to Sabine Hunsiker, Linguistic Solutions Architect at Euroscript Deutschland, for her help in writing the script.

In all three datasets, the concordance approach (2/11-grams) unsurprisingly dominates. More interesting considerations can be made if the other categories are considered. Single-words were looked up almost once every five searches in the overall dataset whereas strings longer than 11 words were relatively rare. In the search session subset, 1-grams decreased by almost six percentage points compared to the overall figure whereas the other two categories increased slightly. Spot searches, on the other hand, followed their usual trend and decreased in the case of concordance and TM searches with respect to the main dataset, increasing slightly only in the case of single-word searches. These results suggest that search sessions mostly deal with multi-word units and longer strings, in agreement with the main theoretical purpose of a concordance search. Vice versa, spot searches tend towards the short end. A finer-grained breakdown of the central group of n-grams would be necessary to make more detailed observations about the tendency to use shorter or longer strings. It could easily be obtained by changing some parameters in the Python script but this was beyond the scope of this specific analysis.

As for averages across languages, spot searches have two of the highest SD values, indicating greater variability within the subset. Without going too much into detail (comprehensive statistics can be found in Appendix B), Greek and Spanish are the most active users of 1-grams, particularly in spot searches, whereas Latvian is the language with most single-word searches in the session group. At the other end of the spectrum, Bulgarian tends to use very long queries in all three instances and Portuguese follows at close distance. Finally, French is unmistakably the most active user in all cases within the concordancer proper range, followed by Dutch. Once again, there does not seem to be a clear trend to justify any specific language clustering.

This analysis has shown how problem units can eventually be of very different sizes and that there is no clear pattern as to what size is the most common. In Section 2.9, problem units were presented as a special kind of translation unit[161]. No clear relationship can be established between the two in terms of size (e.g. problem units are generally shorter than translation units) with the available data but clearly they are both very dynamic concepts that cannot be framed into a fixed length.

A good example comes from a more-in-depth analysis of the occurrences of specific strings. The title of Commissioner Ashton was previously mentioned as the most common search for the 11-gram subset and a smaller portion of that string ("high representative") ranked quite high in the frequency list of bi-grams. In a search of this kind, the most obvious information need is that of finding the target language equivalent for the title of Commissioner Ashton. From a translation perspective, this could be regarded as a straightforward information need, because for proper nouns and titles there tends to be a 1:1 correspondence between source and target, provided that a translation exists in both languages. A range of different string lengths can be used to search for a NE: from entering the proper noun only, in the expectation that the corresponding title will appear somewhere in the aligned segments, to a (much) longer query, possibly quoting the exact wording of the title found in the source text.

If all occurrences of "high representative" are searched within the dataset and then filtered to contain only those strings with no additional items to the named entity, a range

[161] Readers should be reminded that this study only deals with translation problems that can be quantified and that there are higher-levels types of problems (e.g. cohesion problems) that cannot be resolved with available computerized translation support.

of realizations of this NE is found. Table 33 lists all occurrences together with their frequency in the main dataset (724,000) and length expressed in words. There are a total of 59 different versions of this NE for a total frequency of 353 searches[162] (83 is the frequency of the most frequent single string in the list). The whole list containing just the proper noun and/or the title can be ascribed to the same information need whose frequency equals the sum of all occurrences (i.e. 353). This amounts to about 1.5 times more searches than the top searched string in the main frequency list of search strings, i.e. 'youth on the move' (freq.= 221).

Table 33. Possible realizations of one single named entity, i.e. the title of Commissioner Ashton, calculated on the main dataset of 724,000 strings (all searches lowercased).

Freq	#wds	String	Freq	#wds	String
5	1	ashton	3	7	high representative for foreign and security policy
2	2	baroness ashton	27	8	high representative for foreign affairs and security policy
2	2	catherine ashton	1	8	high representative for foreign relations and security policy
83	2	high representative	4	8	high representative of the union for foreign affairs
1	3	catherine ashton high	1	9	vice president of the european commission and high representative
1	3	chief catherine ashton	3	9	high representative for the common foreign and security policy
5	3	eu high representative	1	9	high representative of the european union for foreign affairs
2	3	high representative ashton	6	9	eu high representative for foreign affairs and security policy
1	3	ashton high representative	1	10	high representative for foreign affairs and security policy vice president
2	3	high representative vp	1	11	catherine ashton eu high representative for foreign affairs and security policy
1	3	high representative cfsp	6	11	european union high representative for the common foreign and security policy
21	3	high representative vice-president (vice president/vicepresident)	53	11	high representative of the union for foreign affairs and security policy
1	4	high representative catherine asthon	3	11	high representative of the eu for foreign affairs and security policy
6	4	high representative vice-president ashton	2	11	eu high representative for foreign affairs and security policy commission vice-president
1	4	eu high representative vice-president	1	12	eu high representative of the union for foreign affairs and security policy
1	4	vice-president-high representative ashton	3	12	the high representative of the union for foreign affairs and security policy
3	4	high representative for the cfsp	17	12	high representative of the european union for foreign affairs and security policy
1	4	vice-president-high representative	1	12	commission high representative of the union for foreign affairs and security policy
4	4	high representative and vice-president	7	12	vice-president of the commission high representative of the union for foreign affairs
2	5	eu high representative vice-president ashton	3	13	eu high representative of the european union for foreign affairs and security policy

[162] No distinction is made here between search sessions and spot searches, even though multiple strings could belong to the same search session, i.e. they emerge from the same information need.

1	5	catherine ashton the high representative	1	13	the high representative for foreign affairs and security policy vice-president of the commission
2	5	high representative for foreign affairs	1	14	the high representative of the union for foreign affairs and security policy and vice-president
1	5	high representative for foreign relations	1	15	high representative of the union for foreign affairs and security policy vice-president of the commission
1	5	eu high representative catherine ashton	3	15	vice president of the european commission and high representative for foreign affairs and security policy
3	5	high representative and vice-president ashton	13	15	vice-president of the commission high representative of the union for foreign affairs and security policy
6	6	high representative of the european union	2	16	eu vice president of the european commission and high representative for foreign affairs and security policy
2	6	high representative vp of the commission	3	16	vice-president of the commission and high representative of the union for foreign affairs and security policy
1	6	vice-president of the commission high representative	3	16	high representative of the union for foreign affairs and security policy and vice-president of the commission
9	6	high representative vice-president of the commission	3	17	eu high representative of the union for foreign affairs and security policy and vice-president of the commission
8	6	high representative of the union (for)			
		TOTAL	**353**		**Average Freq. = 5.98; SD Freq. = 13; CV Freq. = 2.173**

The average length of a string in this selection is just below 8 words but the actual values range from 1 word ('ashton') to 17 words ('eu high representative of the union for foreign affairs and security policy and vice-president of the commission'). Within the chosen cut-off range, 62 searches would have been missed out because they were longer than the set threshold of 11 words. Theoretically, all these strings can be easily recognized as different realizations of the same information need. However, automatic recognition and clustering of strings is in fact much more challenging due to a well-known phenomenon in Natural Language Processing: the *sparse data problem*. NLP uses many statistical methods based on collections of training data. Probability estimation in e.g. IR and machine translation however, cannot be precise because, irrespective of the size of the training corpus, relatively common events may not be estimated reliably. In the words of Katz (1987: 400):

> Sparseness of data is an inherent property of any real text, and it is a problem that one always encounters while collecting frequency statistics on words and word sequences (m-grams) from a text of finite size. This means that even for a very large data collection, the maximum likelihood estimation method does not allow us to adequately estimate probabilities of rare but nevertheless possible word sequences – many sequences occur only once ("singletons")[163]; many more do not occur at all. Inadequacy of the maximum likelihood estimator and the necessity to estimate the probabilities of m-grams which did not occur in the text constitute the essence of the problem.

In the case of concordance searches, this means that the same information need, i.e. finding the target language version of a named entity, has been expressed in 59 different ways. Because of data sparseness, co-occurrences of words within a named entity cannot be predicted in such a way that they can be successfully identified automatically.

[163] In corpus-based studies, *word forms* occurring only once in the corpus are called "hapax legomena". Given the previous adaptations of the concepts of token and types to string level, a hapax (or "singleton") could be considered a string that appears only once in a frequency list.

Frequency lists are without doubt a good starting point for studying the most recurring types of problems but they are not enough to account for the actual frequency of a problem category or a specific problem.

7.1.1.5 CORE STRINGS

It was previously pointed out that the longer the string, the less evident the translation problem becomes, and the lower its frequency. As a consequence, short strings will likely be high up in a frequency list and could be seen as a *core (problem) string*, i.e. a string where the translator's attention is clearly focused on and one that is likely to appear in combination with other words in longer queries. In the case of Named Entities, however, core strings may simply represent the most effective retrieval element for that named entity because they provide a clear reference to that entity. This is very likely the case of 'high representative', in that the full string has 10+ words.

Because a core string can be expected to occur in longer strings, the concept of collocational window could be a useful starting point for a quantitative analysis. A customized Python script[164] was developed to quantify the collocational window of one or multiple core string(s). First, candidate core strings are chosen and these are then identified within the dataset. Candidate core strings can be chosen based on their ranking in a frequency list or can be manually entered as a list in a separate file that will be used as input for the script. The script looks for the matching strings in the dataset and counts the number of words to the left and to the right of the core string. The top 100 strings were selected and manually analyzed using the tentative categories outlined earlier (see Sub-section 7.1.1.1). The 30 top ranked strings are summarized in Table 34

Table 34. Top 30 most searched strings in the main dataset (724,000) with absolute frequency counts.

Freq	Search String	Freq	Search String
221	youth on the move	132	terms of reference
200	single market act	131	esma
198	europe 2020	130	level playing field
195	european external action service	127	resource efficiency
194	tfeu	126	digital agenda for europe
177	europe 2020 strategy	125	innovation union
171	economic governance	123	cip
170	european semester	122	millennium development goals
155	erdf	121	legal certainty
155	impact assessment	121	capacity building
151	digital agenda	118	credit default swaps
146	memorandum of understanding	117	stability and growth pact
136	flagship initiative	117	treaty on the functioning of the european union
135	smart regulation	115	european regional development fund
135	eeas	114	eafrd

Twenty-three per cent of the strings of the top 100 are labeled Named Entities (10% Abstract, 7% Classic, 6% Documents) with an additional 13% of acronyms and abbreviations that could technically be included into the NE count. LGP strings accounted

[164] The author is indebted to Philip John Gorinski, graduate student at Universität des Saarlandes, for his help in writing the script.

for 5% while the rest remained unlabeled. Average length is 2.3 words, ranging from 1 to 8 words. The most frequent are bi-grams (38%) followed by 1-grams (27%) and tri-grams (20%). With the exception of single-word acronyms, all Named Entities are to be found in strings of three words or more. A sample analysis of the collocational window was carried out using two strings: one acronym ('erdf', rank 9) and one 4-gram ('european regional development fund', rank 29). Table 35 summarizes the results.

Table 35. Distribution of collocational windows related frequencies for two sample strings extracted from the top 100 most frequent strings in the overall database (724,000).

#9 erdf				#29 european regional development fund			
Left	**String**	**Right**	**Freq.**	**Left**	**String**	**Right**	**Freq.**
0	**1**	**0**	**155**	**0**	**4**	**0**	**126**
0	1	1	20	2	4	0	9
4	**1**	**0**	**15**	**0**	**4**	**1**	**6**
0	1	2	9	0	4	3	3
0	1	4	9	3	4	0	3
3	1	0	7	29	4	8	2
10	1	4	6	0	4	2	2
1	1	1	6	1	4	0	2
2	1	0	5	0	4	6	2
3	1	1	4	27	4	5	1
4	1	1	4	4	4	0	1
5	1	0	3	0	4	7	1
4	1	5	2	29	4	5	1
2	1	4	2	29	4	3	1
33	1	7	2	0	4	4	1
4	1	2	2	29	4	7	1
1	1	0	2	7	4	8	1
33	1	6	1	4	4	2	1
34	1	22	1	29	4	6	1
0	1	3	1	19	4	11	1
31	1	4	1	24	4	29	1
0	1	5	1	2	4	5	1
3	1	5	1	27	4	29	1
6	1	28	1	25	4	12	1
9	1	1	1	2	4	6	1
3	1	2	1	32	4	7	1
36	1	6	1	8	4	11	1
10	1	1	1	2	4	11	1
33	1	2	1	30	4	10	1
33	1	5	1				
10	1	12	1				
6	1	5	1				
33	1	4	1				
2	1	1	1				
18	1	5	1				
Total freq.			**271**	**Total freq.**			**175**

For each string four columns are provided: *string* contains the length in number of words of the core string (here 1 and 4, respectively) and is surrounded by the columns with the width of the collocational window to the right and left; finally, the fourth column provides the weighted frequency of each combination. Unsurprisingly, the acronym totals more combinations than the longer string and has a higher total frequency (271 vs. 175). The absolute frequency for each core string alone can be found in the first row, where the collocational window is [0, *string*, 0], and represents 57% and 72% of the total occurrences, respectively. A few window sizes seem relatively popular [4,1,0] and [2,4,0] whereas for the remaining window sizes, frequency drops quickly down to one. In particular, there are a few instances where the number of words at either side of the core string exceeds 30. These are very likely queries where one very long sentence was searched-for and the relevance of the isolated string becomes negligible. If all three columns are added together for each search type, the 11-word cut-off is exceeded in many occasions.

A final look at the strings reveals that they are actually related to one another, in that 'erdf' is nothing but the acronym for 'european regional development fund'. They refer to the same named entity and can be considered two different representations of the same information need, provided the referent for the acronym is known. This can be seen reflected quite well in the respective windows. On the one hand, there is a single-word preceded by four words [4,1,0] repeated 15 times; on the other hand a string of four words followed by a single word [0,4,1] 6 times. This is the case where the spelled out form is accompanied by the acronym and according to the overall frequency list this occurs 4 times total. Numbers are unfortunately not particularly high but they should suffice to raise awareness about the challenges of quantifying the exact frequency of an information need.

There are many methodological issues to be raised here, such as finding the best way to handle this extreme variability: should different instances of a problem with the same named entity ('erdf' vs. 'european regional development fund') be kept separate or added? In the latter case, the NE would account for almost 450 searches (overlaps included and without cut-off length) on the basis of sheer frequency counts. The case of 'high representative' is slightly different because the vast majority of the longer searches already contained the bi-gram. When the strings with the named entity "title" were examined earlier, total frequency was 353. If the Python script is run using the core string 'high representative', it totals 556 with windows exceeding by far 25 words. How could the 353 instances of the same named entity be singled out from the 556 found by the automatic script?

These remain open questions in the present research project because the answer depends eventually on the understanding of the "core string" and the specific interest in exactly quantifying each instance of a problem. When translation tools are developed, overall trends are more relevant, while the present analysis only served the purpose of raising awareness on specific aspects to consider when providing support to translators. Nonetheless, being able to systematically analyze the size of collocational windows for a specific string may prove useful to better quantify segmentation patterns and, possibly, cognitive loads. Named Entities in the broad sense can be said to represent one frequent category of problems for translators, as suggested by the high frequency with which acronyms and NEs are looked up. Introducing the concept of Named Entities means that we are moving a step forward from pure quantitative analysis and starting to consider the

semantic content of a string. The next section attempts to carry out a quantitative study of the distribution of strings in terms of subject domains.

7.2 STRING CONTENT

This part of the chapter will deal with the semantic content of the strings. More specifically, the Problem Units will be labeled according to their content in an attempt to group together strings that may relate to a given (specialized) domain and those for which no specific domain can be identified.

A previous study with translation professionals (Désilets *et al.* 2009) has found that the problems encountered by translators can be divided into two main categories according to their degree of specialization. Researchers identified Language for Special Purposes (LSP) problems and Language for General Purposes (LGP) problems, which were found to be distributed in two groups of about the same size. This result will be taken as a baseline reference for this part of the study. A preliminary study on this topic (Valli 2011) was carried out which already confirmed the original results but at that time a different dataset[165] was used. To make the analysis consistent, the study needs to be replicated with the final 724,000 dataset. The data volume from the Contextual Inquiry conducted by Désilets *et al.* (2009) was of a manageable size and researchers could perform their categorization of the whole dataset by hand. Unfortunately, the present data volume does not allow for a systematic and unbiased manual categorization and, for this reason, methodologies will be surveyed to try and find a suitable approach to classify strings automatically. Pu *et al*. (2002: 619) state the methodological problem in computational terms:

> The problem is to develop an automatic categorization method that is effective in classifying each term t in T into one or more appropriate categories in C that indicate the subject domain(s) of t's search interests.

First of all, the scope of LSP and LGP needs to be specified because translation environments in Désilets *et al.* (2009) and the present study are different. LSP strings are expected to mainly relate to topics concerning the EU whereas LGP strings are expected to be content-neutral[166]. Approaches to automatic labeling can be drawn directly from the field of Web search log analysis. Many studies aim at grouping (i.e. clustering) the logs in order to establish the areas and/or domains where users are most active. A number of different methodologies have been proposed over the past decade, some of which have been evaluated as potential methodologies for the present analysis. Clustering techniques include co-occurrence of terms (Ross & Wolfram 2000), classification in topical categories (Jansen & Booth 2010) and autocategorization based on feature terms, i.e. using seed terms previously categorized by hand into a predefined taxonomy (Pu *et al.* 2002). Generally speaking, they all require either major manual work or additional automatized operations that could not be carried out here. The lack of context is a well-known limitation in this data type and originates from the nature of a search query, which is written in a particular way — often keyword-like. For the purposes of this study, the classification will be carried out on a string-by-string basis and by only considering the textual information each string contains: if a string can be ascribed to an LSP domain, it is

[165] For the preliminary study, only about 510,000 searches were used because searches with no results had been removed.

[166] Examples of LGP strings in the dataset are: 'greeting card'; 'determined accordingly'; 'come across'; 'timely'; 'must be proportionate'; 'at the invitation of'; 'gave its consent'; 'will continue to give'; 'provides input'; 'is aware that'; 'applicable option'; 'against this background.'

labeled accordingly; and all remaining strings will be considered instances of LGP. The data volume requires labels to be automatically generated if the need for data sampling is to be reduced as much as possible.

Sampling is a common strategy in data analysis, particularly when little or no automatization is possible, but it can result in a partial picture of the data. Early translation process studies using TAPs (Krings 1986a: 181) stated that strong interpersonal variation ("individuelle Differenzierungen") was to be observed and that every subject behaves virtually in a different way, particularly when it comes to the use of external resources (1986a: 179):

> Die Wahl zwischen [...] Strategien ist sicherlich keine individuelle Konstante, sondern hängt offensichtlich vom Umfang der subjektiv beim ersten Lesen erfahrenen Schwierigkeiten ab.

In the case of Euramis, no detailed information is available on individual users or their information needs, so further sampling of the data may reduce the representativeness of the sample. In addition, looking at the whole dataset may highlight overall trends in the searches that might not emerge from a smaller sample[167]. Because data is well structured, it is computationally easy to process. The main problem lies in finding a suitable and structured classification that could yield results without much manual intervention, e.g. a "reference" corpus. In Web log analysis, some methodological approaches rely on a manual classification of strings and an *ad hoc* taxonomy, at least in the initial stage. The topical categories and taxonomies identified and used in previous studies are presented in Table 36.

Table 36. Examples of found topical categorizations of Web queries (categories are often non-mutually exclusive).

Ross & Wolfram 2000	**Pu *et al*. 2002 Macrocategories [Subcategories]**	**Spink *et al.* 2002**	**Jansen & Booth 2010**
Animals	Arts & Humanities	Commerce, Travel, Employment or Economy	Auto
Communication	[Arts]	Computers or Internet	Business
Computing	Business & Finance	Education or Humanities	Computing
Education	[Banks, Electronic Industry, Business Information, Personal Finance]	Entertainment or Recreation	Entertainment
Free Object	Computers & Networks	Government	Games
Games	[Telecom Industry, Software, Hardware, Network Services, Search Engines]	Health or Sciences	Health
Gaming, Lottery	Entertainment	Non-English or Unknown	Holiday
Government, Law	[Stars, Popular Music, Entertainment News]	People, Places or Things	Home
Graphic Arts	Recreation & Chat	Performing or Fine Arts	Misspellings
Groups	[Animation, Chat, Sports]	Sex and Pornography	News
Health, Medicine	Games	Society, Culture, Ethnicity or Religion	Organization
History	[Computer Games, Game Codes]		Other
Jobs, Business	Science & Technology		Places

[167] Looking at trends rather than individual preferences can also be useful from the perspective of tool developers, who need to take into account what the majority of users prefer and do.

Multimedia	[Bibliographic Info., Science]		Porn
Music	Shopping		Research
News	[Food & Restaurants, Festivals, Mobile Products]		Shopping
Organizations	Media & News		Sports
Persons	[News]		Travel
Pictures	Politics & Society		URL
Places	[Local Culture]		
Publication	Adult		
Reference	[Sex Photos, Sex Info]		
Science	Travel		
Sexuality	[Travel Abroad, Travel Info]		
Sports			
Stories			
Trade			
Travel			
TV, Films			
Web Or Network			

Generally speaking, no additional details are provided as to what criteria were used for labeling queries and no examples of queries for each category are given. Quite obviously, a number of the listed categories are not relevant to the present study (e.g. holidays, games, shopping, sexuality, TV & films) and other important categories are missing (e.g. European Union). None of these categorizations can be successfully applied 'as is' to the present study. Viable alternatives had to be found without resorting to manual evaluators, which would have complicated the classification and created additional issues with inter-rater agreement scores, for example. The aim was to find a compromise taxonomy that could be fit for immediate use and account for the vast majority of the domains wherein the EU operates, while keeping human intervention to a minimum.

To effectively target domains of EU activity, the Europa.eu[168] portal was initially used to try and find a suitable categorization. The first option was to use the domains of the EU Factsheets from the European Parliament's website[169] but they turned out to be too generic to be effective. The second option was to resort to the Summaries of EU Legislation found on the Commission's website[170]. The content is organized under 32 headings, which seemed reasonable for a finer-grained classification of LSP domains. The practical problem was to determine how each string could be automatically assigned to one category. One option was to crawl the relevant Web pages and extract keywords for each domain but this would have been too onerous given the limited time and resources available. The solution was eventually found thaks to EuroVoc[171].

[168] http://europa.eu/index_en.htm [last accessed: December 2012].
[169]http://www.europarl.europa.eu/aboutparliament/en/0044c3dd41/EU-fact-sheets.html [last accessed: October 2012].
[170] http://europa.eu/legislation_summaries/index_en.htm [last accessed: October 2012].
[171] http://eurovoc.europa.eu/drupal/ [last accessed: October 2012].

7.2.1 EUROVOC[172]

EuroVoc[173] is a multilingual thesaurus first developed in the 1980s to serve as a means to index documentation within the EU institutions as well as in other national bodies such as national parliamentary libraries. Originally developed by the European Parliament and the Commission's Publication Office in cooperation with national organizations, EuroVoc has been managed by the Publication Office since 2008. EuroVoc is now also used by the Council of Europe, public administrations and, for translation purposes, some private companies. For the most part, the entries in the thesaurus are short strings of up to four words. Originally selected from international glossaries, the terms in the thesaurus have been regularly augmented over the years using proposals from national parliamentary libraries, the EU and individual users, among others. The proposed terms should not reflect a country-specific reality or be too specific but they should be as widely applicable as possible. For the first multilingual version of EuroVoc, translations of terms were provided by EC translators but at that time there were no specific guidelines for the task. With the new content management system adopted for EuroVoc, translators are required to provide a translation for the preferred terms, i.e. terms used as descriptors to index content and documents, the definition and possibly the history note of the entry. Currently, Eurovoc is available not only in 22 official EU languages plus Croatian and Serbian, but also in Basque, Catalan and Russian (EC 2012). Before presenting EuroVoc in further detail, an overview of the main headings of the above-mentioned resources (i.e. EuroVoc, the EU Factsheets and the Summaries of EU Legislation) will be compared in Table 37.

Table 37. Fields and headings of the three candidate external taxonomies considered for the study.

EU FACTSHEETS	SUMMARIES of EU LEGISLATION	EUROVOC (domain code)
How the EU works	Agriculture	Agri-Foodstuffs (60)
Citizens' Europe	Audiovisual and media	Agriculture, Forestry & Fisheries (56)
The Internal Market	Budget	Business & Competition (40)
Common Policies	Competition	Economics (16)
Economic and Monetary Union	Consumers	Education & Communications (32)
EU's External Relations	Culture	Employment & Working Conditions (44)
	Customs	Energy (66)
	Development	Environment (52)
	Economic and monetary affairs	European Communities (10)
	Education, training, youth	Finance (24)
	Employment and social affairs	Geography (72)
	Energy	Industry (68)
	Enlargement	International Organisations (76)
	Enterprise	International Relations (08)
	Environment	Law (12)

[172] Reproduced and adapted from the original language editions of the *Eurovoc Thesaurus (Edition 4.3)* © European Communities, 2008. Responsibility for the reproduction and adaptation lies entirely with Paola Valli.

[173] The information about the history and structure of EuroVoc (version 4.3) is partially obtained from http://eurovoc.europa.eu/drupal/?q=abouteurovoc&cl=en [last accessed: October 2012]. On 18th December 2012, release 4.4 of EuroVoc became available which contains an updated list of descriptors according to the Lisbon Treaty and about 100 new and/or updated terms.

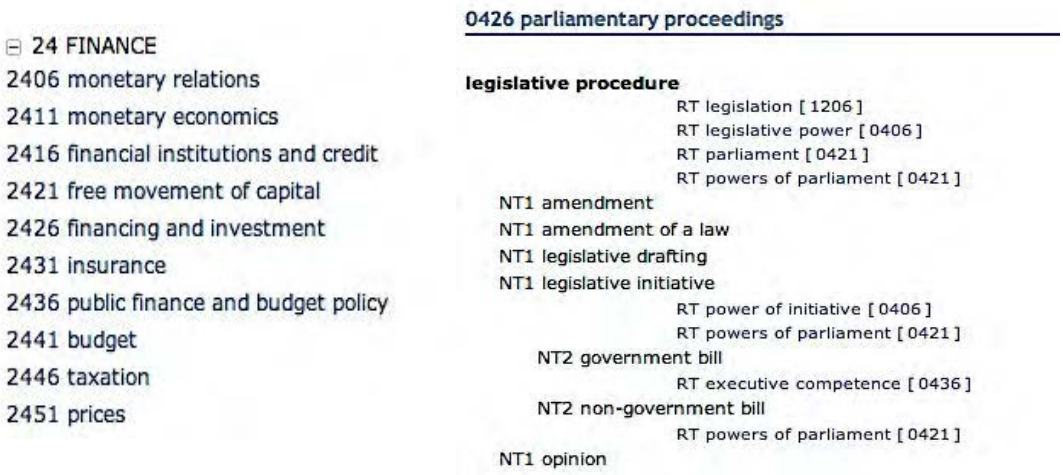
	External relations	Politics (04)
	External trade	Production, Technology & Research (64)
	Fight against fraud	Science (36)
	Food safety	Social Questions (28)
	Foreign and security policy	Trade (20)
	Humanitarian aid	Transport (48)
	Human rights	
	Information society	
	Institutional affairs	
	Internal market	
	Justice, freedom and security	
	Maritime affairs and fisheries	
	Public health	
	Regional policy	
	Research and innovation	
	Taxation	
	Transport	

Compared to the Factsheets and the Summary of EU legislation, EuroVoc seems to strike a good balance in terms of domain coverage, granularity and number of domains. In addition, it does not require keyword extraction or other technical operations because the descriptors are readily available. Poliquen *et al.* (2003) point out that there is a difference between performing a *keyword extraction* and a *keyword assignment*. The former identifies keywords as they appear in the texts, whereas the latter selects the appropriate keyword from an external controlled vocabulary used as reference (generally a thesaurus). Because the controlled vocabulary lists descriptor terms, it may well be that the descriptors are not found as such in the text. They are meant to be somewhat abstract and artificial so as to provide a more exhaustive coverage of a given field. This is also the reason why EuroVoc is considered a *conceptual thesaurus,* as opposed to a *natural language thesaurus*, where descriptors tend to be concrete nouns (Poliquen *et al.* 2003). A more detailed description of the structure of EuroVoc will be provided in the next sub-section.

7.2.1.1 STRUCTURE OF EUROVOC

EuroVoc descriptors (v. 4.3) are organized following a hierarchical structure. First, there are fields identified by a two-digit code (e.g. 24-Finance), each field containing a minimum of three micro-thesauri in turn identified by a four-digit code, i.e. the two-digit field code plus two-digit descriptor identifier (e.g. 2441-Budget) (Figure 58). There are 21 fields altogether, for a total of 127 micro-thesauri containing over 6,700 descriptor terms (EC 2012).

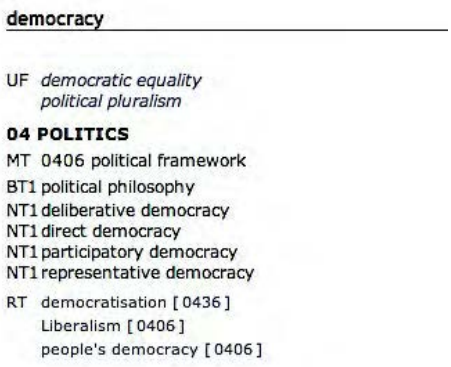
Figure 58. Hierarchical structure of fields and micro-thesauri (left) and breakdown of the hierarchical structure at micro-thesaurus level and below (right).



Within each micro-thesaurus (e.g. 0426 – Parliamentary Proceedings), there are the so-called top terms (e.g. Legislative Procedure), i.e. the first descriptors below the micro-thesaurus level. According to the chosen perspective (top-down or bottom-up), EuroVoc provides a list of multiple levels of narrower terms (NT) or broader terms (BT). For example, 'NT1 –Amendment' means that the term is one level below its corresponding top term; 'NT2 – Government Bill' means that there are two levels between the current descriptor and its corresponding top term, and so forth. There are also non-descriptor terms for each level within the micro-thesaurus that point to the corresponding descriptor term in the micro-thesaurus to facilitate the indexing. These non-descriptors are linked to a specific term in the hierarchy but belong to a different micro-thesaurus and their relationship to the current term is indicated as RT (related term), i.e. there is an associative relationship between the terms belonging to different micro-thesauri.

In the terminological list (Figure 59), USE or UF (used for) means that there is an equivalent relationship between the terms, e.g. democracy and political pluralism. For indexing purposes, however, only 'democracy' will be used, as it is the preferred term. Preferred terms are always used as descriptors, non-preferred terms are not and will only serve as pointers to a preferred term.

Figure 59. Example of a terminological list in EuroVoc.



The whole list of preferred and non-preferred terms provides a comprehensive account of validated terms encompassing the areas where the EU is actively working as well as national points of view, with an emphasis on parliamentary activities, all of which can be

classified as examples of LSP. Because of the way EuroVoc has been maintained and has expanded over time, the fields are assumed to be sufficiently populated and balanced, making EuroVoc a reliable source for clustering. According to the EuroVoc webpage[174], however, the multilingual thesaurus has some limitations:

> EuroVoc has been designed to meet the needs of systems of general documentation on the activities of the European Union; it is not suitable for indexing and searching for specialised documents; EuroVoc cannot claim to cover the various national situations at a sufficiently detailed level; however, efforts are being made to take account of the needs of users outside the EU institutions.

In fact, the nominal form and average length of the descriptors make EuroVoc a particularly well suited basis for a reference corpus. There have already been some attempts to automatically annotate multilingual text collections with EuroVoc in the early 2000 (Pouliquen *et al.* 2003). Less than one third of the documents in the training corpus used in the study (587 out of the almost 60,000 texts of 8 different types) explicitly contained the descriptors which had been manually assigned to the texts (5.6 descriptors per text on average). A descriptor could be automatically assigned only if the string occurred explicitly in the text but then the automatically assigned descriptor was not chosen by the human evaluator in 9 out of 10 times. These figures can appear discouraging at first but they refer to a study serving different purposes and using different data from the present analysis (i.e. whole texts as opposed to the text strings considered here).

7.2.2 METHODOLOGY

The aim of this analysis is to (automatically) compare Euramis strings to EuroVoc descriptors in order to establish how much they overlap. Eurovoc descriptors will be matched against the search strings in Euramis and whenever a match is found, the corresponding two-digit field descriptor is to be appended to the search log. Eventually, search strings will be filtered by this code. The various steps of this content analysis are detailed in the following sub-sections.

7.2.2.1 PRE-PROCESSING OF THE DESCRIPTOR LIST

Some minor pre-processing of the descriptor list was necessary in order to turn the over 15,000 descriptors into a usable reference corpus for the analysis. The only piece of information to be used in the analysis was the field code. Based on the hierarchical structure of the descriptors, each preferred and non-preferred term was given the corresponding two-digit field code.

The EuroVoc webpage[175] provides information about some of the grouping conventions used, which are useful to know for anticipating future weaknesses in the analysis:

> The grouping of descriptors into fields is to a certain extent arbitrary. One of EuroVoc's distinctive features is the limitation of polyhierarchy. Descriptors which could fit into two or more subject fields are thus generally assigned only to the field which seems the most natural for users, in order to facilitate the management of the thesaurus and limit its volume.

---

[174] http://eurovoc.europa.eu/drupal/?q=node/304&cl=en [last accessed: November 2012].
[175] http://eurovoc.europa.eu/drupal/?q=node/305&cl=en;
http://eurovoc.europa.eu/drupal/?q=node/316&cl=en [last accessed: November 2012]. Emphasis in the original.

> In all the language versions, the preferred terms and non preferred terms are generally in the **singular**, and the **plural** is used when the singular does not correspond to normal usage.
>
> **Abbreviations** have been avoided as far as possible in order to make the thesaurus easier to understand and use. Only the acronyms of well-known international organisations have been taken as preferred terms. Specific associations have been modelled to represent the relationships between a term (**full name**) and its **acronym** or **short name**.

These apparently neutral conventions may have a great impact on the domain analysis to be carried out. For example, without lemmatizing the search strings and removing e.g. the plural form, a simple match between descriptors and strings is likely to fail. Ideally, a fuzzy match approach should be adopted to make up for morphological and syntactic variations. At the same time, the choice of avoiding acronyms and abbreviations as much as possible implies that most acronyms contained in the search strings will not be matched, even though they should in fact belong to the LSP group of problems.

An alternative solution was devised so as to try and make up for the missing acronyms. A Eurojargon dictionary (Davies 2004) was deemed fit for purpose and contained, among others, explanations for thousands of acronyms and abbreviations for European projects, schemes and agencies. This seemed appropriate at first because it would isolate as an independent category exactly those words that are recognized as Eurojargon. When looking at lists of Eurojargon terms, some expressions are re-semantizations from other domains or are related to typical EU word-forming procedures (e.g. 'directive', 'regulation' – see Cosmai 2007; Goffin 1994). In the EU portal, there is only a short glossary about Eurojargon[176] containing very specific terms. Those examples are indeed very useful for laymen and common citizens who are not familiar with EU terminology and procedures. However, most of these items are quite unlikely to be looked up by staff translators on their own because they are very common words for someone working internally (e.g. 'regulation' and 'directive'). By the same token, words such as 'Eurocrat' or 'Fortress Europe' are expressions which are unlikely to appear on a daily basis in EU documents to be translated. This means that those Eurojargon words, if at all present in the queries, will not be helpful for a macro-categorization.

The only drawback was that the most recent available edition of the Eurojargon dictionary dated back to 2004, i.e. six years before the current dataset was collected. While this will not impact on most established acronyms, it could be a problem for the most recent ones, which are also the most likely to cause problems to translators. Unfortunately, there was no ready solution that would not involve some kind of information extraction task, so it was decided not to update the list but for a few items. Noise in the results both in terms of false positives and false negatives was to be expected in both subgroups, i.e. LGP strings labeled as LSP and actual LSP strings not labeled as such. To verify this, some preliminary tests were conducted.

7.2.2.2 PRELIMINARY TESTS

After importing all descriptors and dictionary entries in a single file[177], some duplicates were removed from the dictionary entries but no additional editing was performed because the first trial run would use the original list to test the methodology and define

[176] http://europa.eu/abc/eurojargon/index_en.htm [last accessed September 2012].
[177] The entries from the Eurojargon dictionary were given the field code '00EJ' so that they could be easily told apart from the official EuroVoc field codes.

editing criteria, if needed. A Perl script was written[178] that would take every descriptor and match it against a 20,000 sample of Euramis search strings. A provisional output was generated to check how the matching performed. From a quick check of the results, matching seemed to work, in some cases producing reasonably high volumes of matches and there was an encouragingly low amount of false positives, often due to a domain descriptor that did not seem to fit very well the domain in which it appeared (e.g. the descriptor 'economic theory' in the domain 'science').

A further check on recall and precision was carried out using a random sample of 100 strings. Using a random number generator[179], 100 random ID numbers were obtained between 1 and 19,918, i.e. the number of different IDs for the EN-IT sample used in the test. The strings corresponding to the generated IDs were extracted and imported into an excel file. At this point, the strings were manually evaluated and assigned to either the LGP group or one (or more) Eurovoc domain(s), i.e. the LSP group. This resulted in 44 LGP strings and 56 LSP strings[180]. The same string sample was then run against the test version of the script and the *unedited* descriptor list. 53 strings were not labeled (LGP = 53%) against 47% of LSP. Percentage values for the two trials are quite close but a closer look at the results was still needed because matching LSP strings from the two samples had to be compared to ensure that they were indeed the same.

It turned out that 38 automatically labeled LSP strings matched the manual categorization, which means that there was 68% recall and 81% precision[181]. A small domain count was made and it turned out that Finance (24) was the most represented domain, followed by International Relations (08) and European Communities (10). The mismatched strings were used to further evaluate the strings that could/should be edited to improve the matching. Some of the identified false positives and negatives could be grouped under general categories:

- Plural 's': Some descriptors in the reference corpus may appear in both forms but in all likelihood this is not the case for most of the other words. A point was made to ensure that the most relevant keywords were included in both forms;
- Polysemous or homonymous descriptors: Some of these could be mistaken for verbal forms or grammatical words, sometimes due to casing (e.g. 'who' vs. 'WHO', 'lead' vs. '(to) lead', 'will' vs. '(to) will');
- Hyphenated words: these sometimes seemed to be recognized independently of the hyphen, while some other times they generated no match;
- Duplicate descriptors: excessive domain coverage in the 00EJ field causes overlaps with other domains due, among others, to duplicate descriptors (the domain accounts alone for roughly a third of the total descriptors). Double, dubious and overlapping entries were amended accordingly.

False negatives looked like being mainly caused by missing descriptors. Furthermore, the manual analysis highlighted the presence of strings in languages other than English,

---

[178] The author is indebted to Daniel Hardt, associate professor at Copenhagen Business School, for his help in writing the script.

[179] The Random Integer Generator function that was used that can be found at http://www.random.org/ [last accessed: March 2011].

[180] The manual classification was not always straightforward as it had to be born in mind that the matching was based on formal resemblance between string and descriptors and not by semantic association, i.e. deliberately adding contextual information to the strings.

[181] The number of retrieved Relevant items (38 LSP) as a proportion of all relevant items (56 manually categorized). As stated in Section 4.4, precision is measured by dividing the number of retrieved relevant items (38) by the number of retrieved items (47 LSP strings).

which was indicative of some noise in the corpus. Long search strings are one of the overall problem areas, due to the imbalance between search content and matching descriptors. This suggests that automatic string labeling is far from being watertight but at least any "mistakes" would be consistent throughout the analysis. Furthermore, long strings raise the question as to whether the previously identified cut-off length of 11 words should have been employed for this analysis. Eventually, it was decided not to limit the analysis to a specific length range but to allow multiple matches within the same string, instead. In this way, fields could be more evenly represented while statistical calculations were made slightly more complex.

7.2.2.3 EDITING OF THE DESCRIPTORS

The preliminary analysis showed that there was room for improving the matching scores by slightly amending the reference corpus. First of all, the most populated fields (i.e. 00EJ, 72 and 76) were examined to remove double descriptor entries introduced after adding the dictionary. Subsequently, false negatives were examined and some missing descriptors were added to the list. This was done heuristically by looking for single keywords in the strings that could be used effectively. Where possible, internal variability was exploited in that both singular and plural forms were added and each was assigned to a different domain to make up for the issue of homographs. In other cases, descriptors that returned false positives (particularly in the LSP group, e.g. "WHO" vs. "who") were removed from the dictionary. This generally involved eliminating homographs with common words, particularly auxiliary verbs (e.g. 'will' used as 'testament' in the list).

The editing was obviously carried out on the basis of the small sample available and was aimed at testing whether recall and precision could be improved rather than altering the original descriptor list. The added descriptors received a special code in order for them to be easily recognizable so that – if necessary – they could be easily removed. After editing the descriptors, the same randomly generated list of strings was run against the edited descriptor list. The test produced 57 matches for LSP strings (just as in the manual case) and, of these, 51 matched the manually identified ones. Compared to the previous results, precision slightly increased whereas recall improved considerably. The relatively small size of the sample allowed for effective improvements by simply editing a few strings.

A new random sample of 100 strings was therefore generated and extracted. This time the script was run directly on the edited version of the descriptor list. According to the manual categorization, there should be a 50-50 ratio between LSP and LGP strings. The script produced 53 matches for LSP and 47 for LGP, suggesting that the edited version of the descriptors produced more matches than the manual categorization. For LSP strings 80% recall was found and precision amounted to 75%, meaning that recall improved but precision was reduced compared to the first test. This can be explained by the fact that the edited list contained some one-word keywords whose purpose was to increase coverage (= recall) but this obviously negatively affected precision. Results were nonetheless satisfactory and the slightly edited version of the descriptor list was chosen for the main analysis. Repeated analyses were conduced at different stages of the study before the final version of the dataset was developed. At each stage, additional checks were made on the Eurojargon dictionary to clean potential noise sources (mostly acronyms that were, in fact, homographs of English words, e.g. home, heart, scale, impact, step, rule, core, scope). In the final version of the descriptor list, false positives and false negatives were still present but they were expected to even each other out, at least for the whole dataset. However, this meant that some domains would possibly be underrepresented due to missing descriptors. This is the case of e.g. the science field (36), which was negatively affected by the lack of descriptors to match e.g. the names of

chemical substances found in the searches. On the other hand, other fields containing very common descriptors such as European Communities (10) and Eurojargon (00) were likely to be overrepresented. No adjustment measure was found for this problem other than re-writing the descriptor list, which was neither possible nor reasonable. Results should therefore be taken with caution, particularly when discussing the least populated fields and domains.

7.2.3 DOMAIN ANALYSIS TEST-PHASE

After fine-tuning the reference list and finalizing the Perl script, the analysis on the 724,000 dataset was finally carried out. The script compared each descriptor against each search string and appended the field codes of the matching descriptors to the log. The script was run on each language and for each level of analysis, namely language subset, search sessions and spot searches[182]. Normalized percentage distributions for each group are summarized in Table 38, showing the amount of LGP and LSP strings found as well as the amount of LSP strings where multiple matches were appended.

Table 38. Distribution of LGP and LSP strings for each language pair at each of the three levels normalized by the total number of searches for each language.

Lang Pair	Main			Sessions			Spot		
	LGP	**LSP**	**MULTI**	**LGP**	**LSP**	**MULTI**	**LGP**	**LSP**	**MULTI**
BG	41.88%	58.12%	37.90%	40.32%	59.68%	39.20%	43.05%	56.95%	36.70%
CS	39.03%	60.97%	37.85%	37.18%	62.82%	40.02%	39.94%	60.06%	36.58%
DA	42.79%	57.21%	34.79%	41.42%	58.58%	36.72%	43.40%	56.60%	33.91%
DE	41.79%	58.21%	34.47%	38.99%	61.01%	37.14%	43.02%	56.98%	33.21%
EL	45.08%	54.92%	34.85%	42.24%	57.76%	37.25%	46.49%	53.51%	33.51%
ES	45.65%	54.35%	33.27%	45.15%	54.85%	33.85%	45.90%	54.10%	32.84%
ET	42.93%	57.07%	34.86%	41.54%	58.46%	36.33%	43.64%	56.36%	33.99%
FI	42.13%	57.87%	33.02%	40.67%	59.33%	33.11%	42.77%	57.23%	32.96%
FR	46.56%	53.44%	33.20%	45.09%	54.91%	34.21%	47.05%	52.95%	32.86%
HU	39.48%	60.52%	37.05%	37.08%	62.92%	39.22%	40.81%	59.19%	35.64%
IT	42.14%	57.86%	34.27%	40.16%	59.84%	36.24%	43.27%	56.73%	33.08%
LT	41.45%	58.55%	35.17%	40.68%	59.32%	36.50%	41.87%	58.13%	34.10%
LV	42.32%	57.68%	36.93%	40.83%	59.17%	37.99%	43.18%	56.82%	36.17%
NL	42.93%	57.07%	35.78%	42.91%	57.09%	37.93%	42.73%	57.27%	34.53%
PL	40.72%	59.28%	36.55%	39.29%	60.71%	37.49%	41.27%	58.73%	35.97%
PT	41.50%	58.50%	38.44%	39.55%	60.45%	40.81%	42.73%	57.27%	36.77%
RO	40.06%	59.94%	39.07%	39.25%	60.75%	41.22%	40.47%	59.53%	37.72%
SK	40.17%	59.83%	37.87%	37.88%	62.12%	39.51%	41.57%	58.43%	36.53%
SL	42.61%	57.39%	36.98%	40.26%	59.74%	40.65%	43.85%	56.15%	34.92%
SV	41.48%	58.52%	36.07%	38.81%	61.19%	37.39%	42.94%	57.06%	35.21%
MEAN	42.13%	57.87%	35.92%	40.47%	59.54%	37.64%	43.00%	57.00%	34.86%
SD	1.93%	1.93%	1.83%	2.21%	2.21%	2.28%	1.86%	1.86%	1.56%
CV	0.046	0.033	0.051	0.055	0.037	0.061	0.043	0.033	0.045

[182] The analysis proved to be very time consuming if run on an average laptop. For instance, the first run of the script on the EN>ALL dataset was launched at 2am and terminated at 9.20pm whereas the average processing time per one language subset was of about one hour. This explains why so much time was devoted to preliminary analyses and multiple checks.

The usual variability in sessions and spot searches compared to the main dataset is confirmed, with an overall ratio of about 40-60 between LGP and LSP. This suggests that there is no huge difference between LSP and LGP problems, as previously noted by Désilets *et al*. (2009). LSP problems seem to occur more frequently than LGP ones, particularly in the case of search sessions. As for individual language pairs, FR tends to consistently show the highest proportion of LGP strings and CS the lowest. This might be due to the considerably large size of the French subset, which may not be well covered by the EuroVoc descriptors. EL and ES also tend to have higher percentages of LGP strings whereas HU, RO and SK request more LSP strings. Despite the fact that SD and CV have rather low values, i.e. there is not much variability within the different language pairs, a trend seems to nonetheless emerge such that newer languages require more support for LSP problems whereas older languages look for general language strings.

As for the distribution of multiple matches, it was expected that strings with the highest average lengths (PT and BG) would also have the highest amount of matches and vice versa for the shortest ones (ES, DA, SK). This was partly verified, in that PT and BG were at the higher end whereas ES and DA at the tail end. However, no target language systematically ranked at the very top or bottom. Instead, RO consistently had the highest amount of multiple matches, suggesting more "content dense" searches, whereas FI often had the lowest amount of multiple matches while SK, despite using short strings, was found in the top positions, with queries that were possibly even more "content dense". With regard to multiple matches, the mean percentage values are in line with those for sessions and spot searches, suggesting that there is no increase in multiple matches for search sessions, as one might have expected. The percentage values for each language are an indicator of the likelihood that a concordance search is of the LGP or LSP kind. There are, however, some EuroVoc fields or descriptors that are more likely to be found in the strings (e.g. descriptors in the European Communities field) and that may increase match percentages. For example, one corpus-based study comparing national and European parliamentary texts found that the most frequent words in documents from the EU parliamentary debates were indeed 'European', 'Parliament', 'the Council' and 'the Commission' (Danielsson 2003).

The next step in the analysis was the average distribution of each domain across the 20 language pairs, which came before the analysis of the distribution for each language individually. Distributions for each language pair were analyzed in the preliminary studies on the reference descriptors but results were inconclusive because there were too many variables to be considered simultaneously. Grouping by language family was also attempted (Valli 2011) but proved not ideal due to the imbalances in language distribution (Figure 62) whereas the age criterion did not produce fine-grained results (Figure 63). Average domain distribution was therefore calculated for each level of analysis, as summarized in Figure 60, while Table 39 lists the domains considered in the analysis with descriptor codes in ascending order.

Figure 60. Average distribution per domain calculated for each of the three level of analysis (Main, Session and Spot), normalized by the total number of LSP strings (724,000).

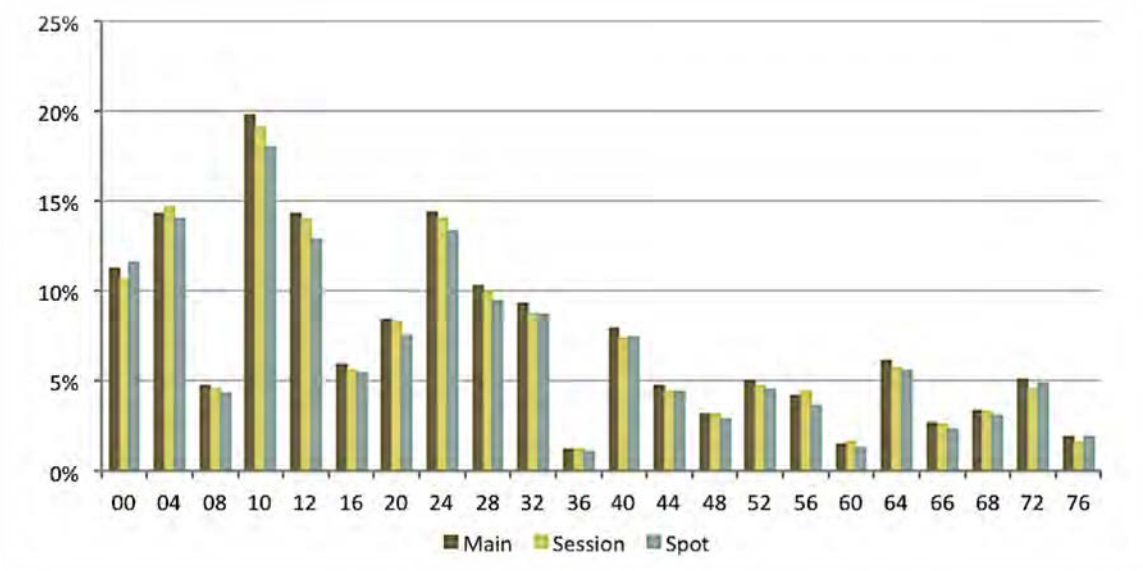


Table 39. List of domains and codes employed in the descriptors list (in ascending descriptor order).

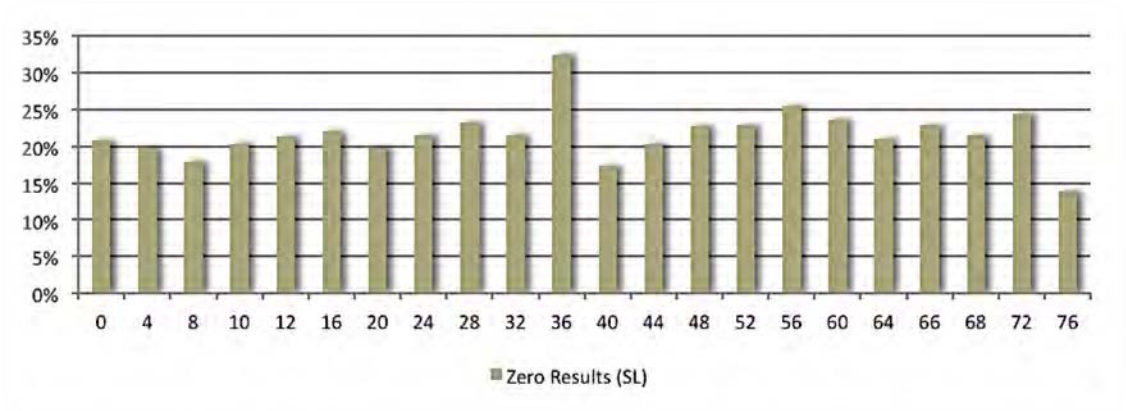
DOMAIN	LABEL
00	EUROJARGON *(added)*
04	POLITICS
08	INTERNATIONAL RELATIONS
10	EUROPEAN COMMUNITIES
12	LAW
16	ECONOMICS
20	TRADE
24	FINANCE
28	SOCIAL QUESTIONS
32	EDUCATION AND COMMUNICATIONS
36	SCIENCE
40	BUSINESS AND COMPETITION
44	EMPLOYMENT AND WORKING CONDITIONS
48	TRANSPORT
52	ENVIRONMENT
56	AGRICULTURE, FORESTRY AND FISHERIES
60	AGRI-FOODSTUFFS
64	PRODUCTION, TECHNOLOGY AND RESEARCH
66	ENERGY
68	INDUSTRY
72	GEOGRAPHY
76	INTERNATIONAL ORGANISATIONS

As can be seen in Appendix D (Tables D.2, D.3 and D.4), the highest SD is found for the three most populated fields, i.e. 10 (European Communities), 12 (Law), 24 (Finance) and indeed, a more visible gap can be noted between the 'session' and 'spot' sub-groups. This suggests that the most populated domains are more common in search sessions, possibly because the same items re-appear within the sessions. Field 36 (Science) is by far the least populated but in this case a potential bias should be pointed out, in that the variability in e.g. chemical substances or other technical terms could not be properly accounted for in the descriptors. Overall, the proportions for each field seem to be

maintained across the three levels and can be used to estimate the likelihood that an LSP search belonged to a given domain.

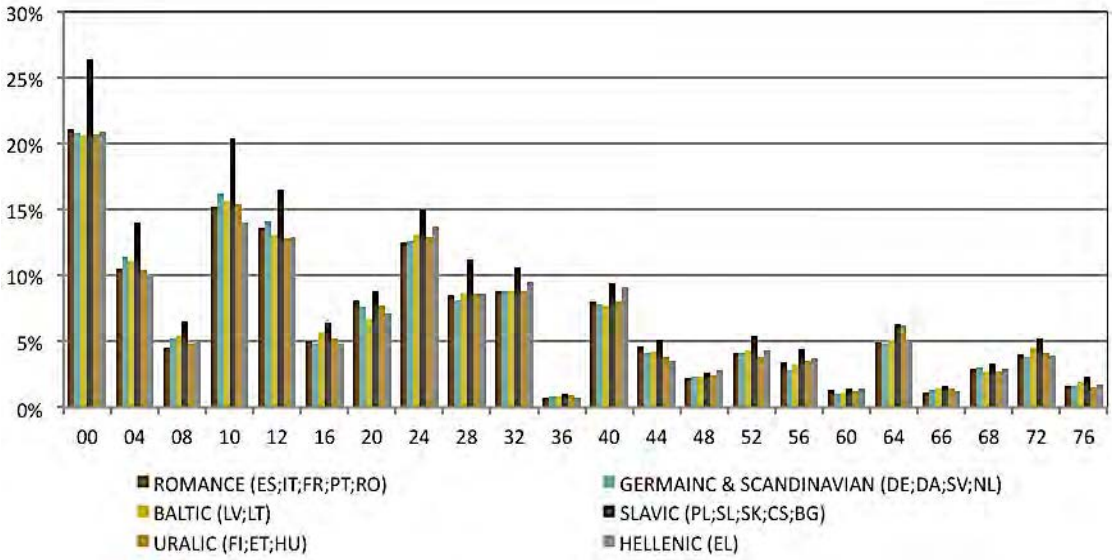
A quick check (with SL taken as sample) into the distribution of zero results per domain, indeed suggests that domain 36 is the most challenging for users because it clearly has the highest proportion of unsuccessful searches (Figure 61).

Figure 61. Distribution of domains and zero results (SL sample taken from the 724,000 set).



As previously mentioned, earlier analyses looked at the distributions according to the language family. Because a different dataset was used for the analysis at that time (about 510,000 strings), results in Figure 62 cannot be directly compared with the ones discussed above; they only serve as an indication of trends.

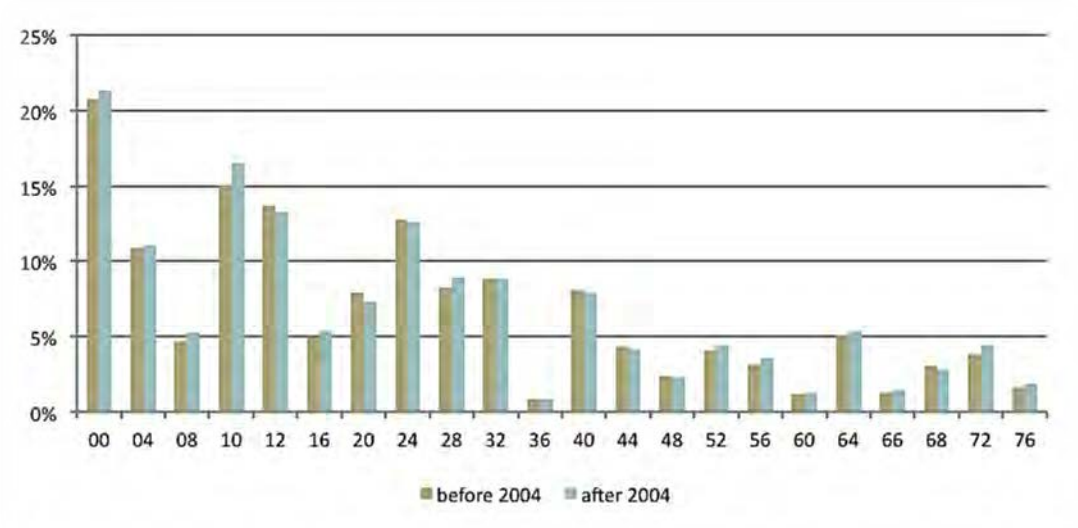
Figure 62. Distribution of domains according to language families. The dataset used amounted to some 510,000 queries and was normalized using the total amount of LSP strings.



Slavic languages clearly stand out in most domains, whereas the two other major families (Romance and Germanic) are often on a par. This observation, however, cannot be generalized to all languages within a single family because, as previously noted (see Section 5.5.2), there can be forms of internal compensation that do not emerge from this representation. A further breakdown into individual languages is technically possible but it would make data interpretation and comparison extremely hard because each of the 20 domains would have to contain 20 columns instead of six.

From a chronological perspective, in the EU context Slavic languages are all newer languages as opposed to most Germanic and Romance languages. This suggests that a different clustering approach could be attempted as shown in Figure 63.

Figure 63. Distribution of domains according to the language age clustering (510,000 normalized by total LSP strings).



Unfortunately, differences are not as marked as one might have expected from the above considerations. A few domains (in particular n. 10) show a more marked gap between the two groups but generally speaking all languages seem to behave consistently or rather the clustering by age produces too general results.

The reasons behind the popularity of a given domain may be due to several intervening factors but the main reason seems linked to the actual volumes of translation for each domain, the most popular being "politics, law and economics" (Wagner *et al.* 2002: 44). In order to have a more accurate estimate of the incidence of the searches in each language, frequencies for each domain should in principle be normalized against the actual number of documents translated into each language whose content relates to any of the EuroVoc fields. There are some general domains that obviously are more or less always present and others that are not, and results could just be a consequence of some domains occurring more often than others.

The fact that Slavic languages are most active within each domain suggests that they prefer using the concordancer to solve LSP problems. Alternatively, it may be a direct consequence of a poorer performance of the pre-translation phase in the workflow in terms of exact matches, leaving translators with more text to translate. However, given the TM sizes provided in Table 10 (Chapter 5), this latter hypothesis seems less likely, in that TM sizes are comparable across most languages.

In the final part of this section, results from this study on domain distribution will be combined with some other items from the tool settings to check whether domain distribution could be affected by search strategy components. For this analysis, the 510,000 dataset will still be used (Valli 2011). The joint frequency distribution between the domains of the descriptors and the institution field was not informative as far as the main trends were concerned, because EC, EP and Council consistently covered the vast majority of the searches. A closer look into the percentages (all below 5%) for the remaining institutions provides further insight into search behavior. For example, the Court of Auditors peaks in the domains relating to economics and finance (16, 24) and

agriculture (56, 60) whereas the Court of Justice only peaks at domain 12 (law). For other institutions the distribution was less clear but this may still be an indication that the number of searches for a domain could be linked to the volume of translated documents. Another joint frequency distribution was generated between domains and search mode, as shown in Figure 64.

Figure 64. Joint frequency distribution of search mode and descriptors domains, normalized by the total strings in each domain.

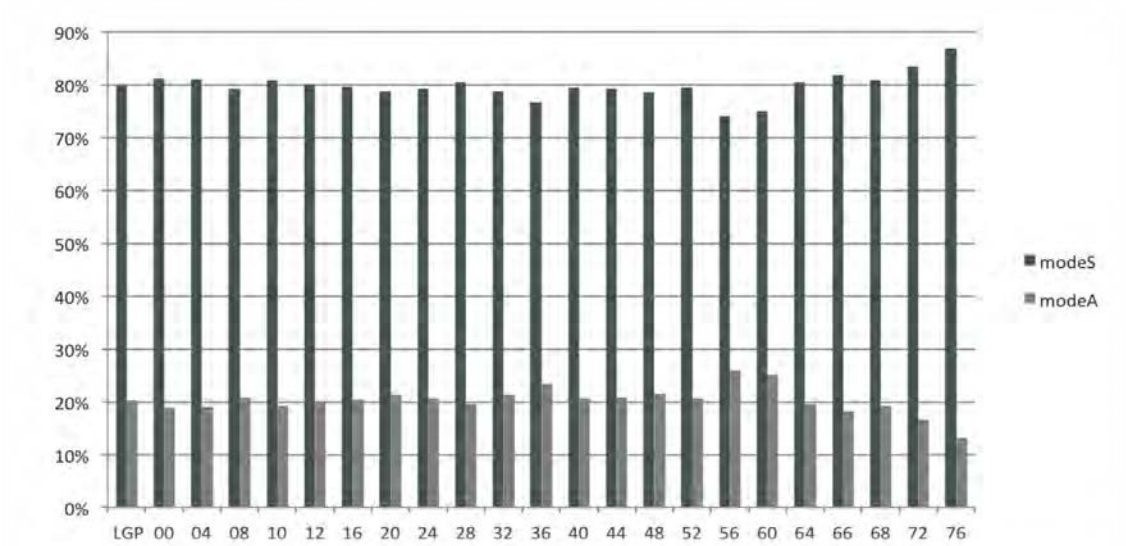


Figure 64 clearly shows that both LGP and LSP domains are mostly found in conjunction with simple search mode. Domains 56 and 60 (both relating to agriculture) seem to the have the highest proportions of advanced filters, about 25%. Domains 76 (international organizations) and 72 (geography) are below the average of about 20%, meaning that advanced filters are rarely used in these cases.

This section focused on the analysis of string content in order to automatically process the strings from a semantic perspective and build on the findings of previous studies conducted on a smaller scale. After reviewing a number of approaches in the field of Web log analysis, an *ad hoc* methodology was devised to automatically label the strings based on their content and assign them to a macro-category (LSP or LGP). The labels (i.e. domain codes) were attributed according to the similarities between the strings and a collection of descriptors distributed across about 20 domains for a finer-grained classification of the LSP strings. Any string that did not receive a label was classified as LGP. Overall the distribution of searches into the macro-categories LSP and LGP turned out to be not too far off the benchmark with a consistent ratio of about 3:2 across the three levels of analysis. Using an earlier version of the dataset, domain distribution was calculated using the two suggested grouping criteria, namely language families and language 'age'. Irrespective of the chosen data representation, the same four domains can be identified as the most populated (Eurojargon, European Communities, Law and Finance). The same result was obtained when the distribution was corrected for multiple matches in Appendix D (see Table D.5). Additional distributions were then calculated to see whether a relation could be hypothesized between domains and number of failed searches (Figure 61) or search mode (Figure 64). With only one or two outliers, results of the cross-tabulations show a rather uniform distribution across domains.

This tentative study showed that in principle, strings can be automatically categorized into domains, though the specific results obtained may not be too informative (e.g. the

most populated domains are rather unsurprising). The outcome of the analysis relies heavily on the descriptor list and its organization into domains, which in this case was not addressed from a methodological perspective in that the classification was readily available.

After examining the strings from a quantitative and a content-related perspective, the string analysis will consider the linguistic form of the strings, i.e. the third level of analysis for the Problem Unit component.

7.3 LINGUISTIC FORM OF CONCORDANCE SEARCH STRINGS

In the previous sections and chapters, search strings were found to resemble Web queries under many aspects and some existing methodologies were borrowed for their analysis. Similarly, existing studies will be examined for relevant methodological approaches for the study of the linguistic form of the searches. An automatic and systematic categorization of strings as described earlier was not feasible in the case of the linguistic analysis for a number of reasons that will be detailed in due course. The analysis and discussion will therefore be conducted qualitatively using several examples but no comprehensive statistics covering all language pairs will be provided. For most analyses, the strings have been lowercased to ensure better matching and reduce the number of variables. All examples of searches provided in the following sections will also be lowercased for consistency.

7.3.1 LINGUISTIC CATEGORIES OF WEB QUERIES

A preliminary step in the analysis was the scrutiny of the few, available linguistic-oriented taxonomies of Web queries. The overwhelming majority of Web queries are known to be noun phrases, usually with both nominal head and modifier, yet sometimes "it is not clear to what lexical category a term belongs" (Jansen *et al.* 2000a), particularly if the analysis is carried out at term level. In their study, Jansen *et al.* (2000a) performed a lexical analysis of the first 511 queries in their dataset by studying lexical patterns. However, their investigation dates back to over a decade ago, when the Internet was not as widely used as it is today, and it should best be considered of an exploratory nature. Web query language turned out not to be comparable to common English usage, particularly at the level of syntax. Overall, no grammatical consistency was found in the search logs and eventually five syntactic categories were identified: (i) adjective and noun phrase, (ii) grammatically correct, (iii) verbal phrase, (iv) random category and (v) miscellaneous. The first group represented the attributive construction and it was the most populated. The second group included all queries that took the form of a WH-question whereas the category "verbal phrase" included non-complete English sentences where at least a verb or a "verbal" (i.e. *-ing* form) appeared. The last two categories do not bear very informative labels and were respectively defined as "a series of words of varying lexical categories and which defied syntactical categorization" and "any query pattern represented less than 10 times" which included URLs, email addresses and proper names (Jansen *et al.* 2000a,b). In spite of the fact that virtually all these five categories could be populated with examples from the Euramis concordance logs, the proposed classification was unsuitable for the planned analysis. However, the conclusion that "[…] this particular strategy [Adjective and Noun Phrases] either works best or is the default for many human users when they are not sure what syntax applies" (Jansen *et al.* 2000b: 174) is worth keeping in mind.

A more recent study (Barr *et al.* 2008) used part-of-speech (POS) tagging to parse Web search logs. It built on previous findings that highlighted potential POS ambiguities in queries and aimed at resolving them by developing a trained tagger that considered inter-annotator agreement scores. Researchers sampled queries from the *Yahoo!* search engine and their results confirmed that the majority of queries (over 70%) were noun phrases. The most common tag in their tag set of 19 unique classes was "proper noun" which amounted to 40% of all query terms (2008: 1022-3). However, "[t]he sparse textual information in search queries presents difficulties beyond standard corpora, not only for part-of-speech tagging software, but also for human labelers" (Barr *et al.* 2008: 1022). Inter-rater agreement was particularly compromised by proper nouns. A string such as 'stillwater chamber of commerce[183]' could be considered either as one single proper noun or could be split so that only the first word-token would count as a proper noun. Such labeling issues are particularly relevant in the light of the earlier discussion about the different realizations of the string 'high representative' (see Sub-section 7.1.1.4). POS-tagging showed that Web queries present peculiarities of usage with respect to published texts, which affects the distribution and variety of POS-tags. One of the instances in the Euramis searches was the inconsistent or infrequent capitalization of nouns, otherwise used as a reliable indicator for tagging texts. The same issue was found in Web queries, where capitalization was used inconsistently: 83.2% out of a sample of about 290,000 queries were all-lowercase and 16.8% contained some capitalization, of which 3.9% were all-caps (Barr *et al.* 2008). Irrespective of casing, some degree of syntactical structure emerged from the tagging, and 7 categories were identified: (i) noun phrase (ca. 70%), (ii) URL, (iii) word salad, (iv) other query, (v) unknown, (vi) verb phrase, (vii) question (2008: 1027). Once again, the main problem seems to lie in the choice of sometimes very blurry labels (e.g. "word salad", "other query") that make these studies difficult to replicate, especially on different sets of data. On the other hand, linguistic categories such as "verb phrase" and "noun phrase" seem to provide a more solid basis for further considerations about the nature of the strings, as opposed to other commonly used labels that will be discussed later on in the chapter.

7.3.1.1 POS-TAGGING

In order to verify whether some of the claims about Web queries could be applicable to concordance searches, the dataset was POS-tagged. The first issue was finding a POS-tagger[184] that would only parse the "*sentence*" field while preserving the rest of the string (i.e. its metadata). Some suggestions on available taggers were found in the literature (e.g. Barr *et al.* 2008: 1024) but these taggers involved a lot of manual work and training was always involved for which this project lacked adequate resources. A solution was found in the IMS Open Corpus Workbench (CWB)[185], which provides a collection of tools for querying and managing large text corpora using linguistic annotation. Different subsets were selected for testing purposes but each had first to be pre-processed in order to work with the corpus query processor (CQP), the central component that performs the queries.

---

[183] Example provided by the authors (2008: 1023). Strings were all lowercased in the pre-processing of the data.

[184] Some common examples are the Stanford POS tagger or the TreeTagger; a more comprehensive list can be found at http://www-nlp.stanford.edu/links/statnlp.html#Taggers [last accessed: October 2012].

[185] CWB was originally developed at IMS Stuttgart and in 2007 it was released as open-source software; available at http://cwb.sourceforge.net/ [last accessed: October 2012].

Some customized Perl scripts[186] were developed that would first isolate the "sentence" field by turning all the remaining fields into XML attributes (such as the language combination), which could be selected for querying purposes. With the second script, the strings in the file were annotated (i.e. POS-tagged) using the TreeTagger, because tokenization is necessary for the corpus to be indexed into CWB (Evert & Hardie 2011). There are two versions of the TreeTagger tag set, the older with 36 tags, the newer with 58, the increase mostly due to the distinction between auxiliaries and main verbs in all most frequent tenses[187]. Once the corpus was readily annotated, it was compiled and queried with CQP. Before statistics could be collected, the tagging had to be checked to ensure it was of acceptable quality.

One hundred random strings were extracted from the Estonian subset and tagged by the system. The results were compared with a human tagged version of the same subset. Out of 290 tags (the total number of tokens in the sample strings), about 20 tags differed between the automatic and manual tagging and about a dozen were considered ambiguous. Ambiguities were not marked as errors because without context there was not way to tell which was more likely. For one-word strings, in case of doubts the first entry in a dictionary was taken as the most likely tag in the manual analysis. These results seemed promising enough to move on to the next stage, where CQP was used to study POS distribution. However, the logic behind the tagging was not always clear or predictable, particularly for gerunds, past participles and homographs in one-word strings. To verify the distribution of POS tags, only the subset of n-grams ranging from 2 to 11 words was considered and extracted from the provisional 740,000 dataset covering 20 target languages. A small sample of the results of the tagging are shown in Table 40.

Table 40. Example of POS-tagged strings. Found tags (to be found after the '/'): CC (coordinate conj.); DT (determiner); IN (preposition/subord. conj,); JJ (adjective); MD (modal); NN (noun, singular or mass); NNS (noun plural); NP (proper noun, singular); NPS (proper noun, plural); POS (possessive ending); VB (verb, base form); VBG (verb, gerund or present participle).

	<string>[]*[pos=".*"][]*</string>;
1	<Ubiquitous/JJ high-speed/JJ connectivity/NN>
6	<high-speed/JJ connectivity/NN>
10	<Web-based/JJ learning/NN>
14	<mobile/JJ communications/NNS>
18	<radio/NN bandwidth/NN>
22	<location/NN signals/NNS>
26	<Open/NP Access/NP Navigation/NP>
31	<trunk-level/NN transmission/NN>
35	<utility/NN companies/NNS>
39	<virtuous/JJ cycle/NN>
43	<will/MD continue/VB to/TO give/VB>
49	<spending/NN plans/NNS>
53	<Transitional/NP Federal/NP Institutions/NPS>
58	<Security/NP and/CC Stabilisation/NP plan/NN (/(NSSP/NP>
66	<Security/NP and/CC Stabilisation/NP plan/NN>
72	<Security/NP and/CC Stabilisation/NP>

[186] The author is indebted to Dr. Hannah Kermes, Research Assistant at Universität des Saarlandes, for her help in writing and adapting the scripts.
[187] The 36 tag set only distinguishes tenses/verb forms for the sole generic POS 'verb'.

77	<troop/NN allowances/NNS>
81	<European/NP Committee/NP of/IN Social/NP Rights/NPS>
88	<European/NP Convention/NP on/IN the/DT Exercise/NN of/IN Children/NP 's/POS Rights/NNS>
99	<Convention/NP on/IN Contact/NP concerning/VBG Children/NNS>
106	<European/NP Convention/NP on/IN the/DT Adoption/NN of/IN Children/NNS>
115	<about/IN future/NN>

Overall, tagging seems satisfactory but there are some obvious issues with the label PN (Proper Nouns). The system clearly relies on casing to identify these instances. However, casing is not always reliable due to the different modalities of input in the concordancer and the fact that the default concordance search is case-insensitive. Lines 88 and 99 display some inconsistent tagging where 'Children' is first interpreted as proper noun and then as plural noun. A more striking example can be found in the one-gram 'who'. On its own, it most likely refers to the acronym WHO, i.e. World Health Organization, rather than the pronoun 'who'. In the dataset, it appears both upper- and lower-cased but the system consistently tags it as a pronoun. Another thorny category is the gerund (VBG) that can be verb, noun or adjective and the past participle can also be used as adjective. These are the main reasons why the 1-gram corpus was not used in this study and still the tagging and resulting statistics might not be completely reliable and therefore should be handled with caution.

First of all, the distribution of each part-of-speech was examined. By using the "group" command in CQP, overall frequency counts for each POS were obtained (Table 41).

Table 41. Aggregate distribution of the main categories of POS-tags for ca. 605,000 strings.

POS-tag	Count
tot NN|NP(S)	568,857
tot JJ	210,642
tot VB*	173,280
tot DT	104,855
tot CC	50,071
tot RB*	31,393
tot MD	10,218

Unsurprisingly, the most populated category by far was that of nouns (NN) and proper nouns (PN), both in the singular and plural forms. The second most populated category (with less than half the occurrences of the nominal category) was adjectives (JJ), possibly because adjectives usually appear in conjunction with nouns. The category of verbs (VB; base and other verb forms) ranked third, but modal verbs (MD) should perhaps be added to the VB group. Conjunctions and adverbs were rarer. Adjectives came in a higher number of occurrences with respect to the aggregated volume of verb forms, including past forms, gerunds and conjugated verbs. These figures should only be taken as indicative of the respective volumes for POS tags as the count suffers from some systematic problems which will be detailed below.

These results are in line with another study (Johnson *et al.* 2006) aimed at better targeting Web searching, which calculated the most frequent POS templates for strings. Aside from the top (grammatical) bigrams and trigrams that convey little meaning, the extraction of high-utility phrases (or content phrases) showed the following most

frequent POS templates: A+N, N+N, A+A+N, A+N+N, N+A+N, N+N+N, and N+Pronoun+N (2006: 3). Almost all potential content phrases in that study could be covered by a set of 40 bigrams and 61 trigrams, excluding phrases containing stop-words. Similarly, Barr *et al.* (2008: 1028) highlighted the telegraphic nature of queries, which were much more likely to have an NP structure of type Adj+N rather than Det+Adj+N. In order to study the most common combinations of POS-tags in the Euramis corpus, a frequency count of POS-patterns was performed, as summarized in Table 42 together with some examples for the top 20 combinations.

Table 42. Most frequent POS patterns in the n-grams string corpus. Only the top 20 POS combinations out of some 60,000 found are shown.

Rank	Count	POS-pattern	Example strings
1	51921	NN/NN	radio bandwidth, gender identity, interoperability constituents, silver economy
2	49456	JJ/NN	criminal record, virtuous cycle, indicative list, total cost, reasonable access
3	28999	NN/NNS	core values, entity vehicles, business start-ups, quality checks, contract agents
4	26911	JJ/NNS	basic metals, mobile communications, corrective measures, natural events
5	15902	NP/NP	EU Roadmap, Van Rompuy, Southern Corridor, Crown Corporation, HR CFSP
6	11238	NP/NP/NP	European Youth Forum, Working Time Directive, GATT Kennedy Round
7	10073	JJ/NN/NN	direct heating process, international investment strategy, public health capacity
8	7194	NP/NN	Nokia case, Gulf region, Hague programme, H1N1 influenza, BEREC office
9	6807	NN/IN/NN	freedom of establishment, case by case, diffusion with water, autonomy of decision
10	6493	JJ/NN/NNS	secondary steel plants, annual lease payments, external aid instruments
11	6224	NN/NN/NN	service provider contract, business registration number, research funding program
12	4792	NP/NNS[188]	GHG savings, IACS controls, Israel colonies, IGC states, EU funds, US authorities
13	4497	NP/NP/NP/NP	European Atomic Energy Community, Regional Innovation Performance Index
14	4423	VB[189]/NN	take effect, obtain feedback, recognise asylum, bring relief, ensure coherence
15	4100	VBG/NN	rolling plan, continuing education, ensuing risk, losing momentum, creating value
16	4054	JJ/JJ/NN	foreign direct investment, solid scientific basis, pulpy whole fruit
17	3995	NN/IN[190]/NNS	growth for jobs, number of observations, body of laws, mapping of services
18	3817	NN/NN	tender notice, evidence base, emergency planning, wood oil, asset freeze
19	3250	VBG/NNS	supporting data, implementing laws, vending ingredients, exporting countries
20	3196	VBN/NNS	delegated acts, regulated articles, estimated numbers, reasoned opinions

Results clearly show that the vast majority of these strings are nominal, including singular and plural nouns as well as many proper nouns. The examples provided are meant to illustrate different realizations of each pattern, but in some cases a remarkable amount of false positives was found, as can be anticipated by looking at some examples. In the case of VBG and VBN, there are instances where the word is probably used as an adjective and it is not entirely clear whether the form should be labeled as an adjective or as a VBG. For example, 'integrated hub' is labeled JJ+NN, just like 'limited company'; 'ranking system' is also tagged JJ+NN, just like 'outstanding balance'. However limited, such instances raise questions as to how the tagging should be interpreted. Another problematic case is that of homographs that can be used both as nouns and as verbs. The pattern VB+NN may have

[188] This is a tricky pattern due to lots of false positives (capitalization).
[189] Another tricky pattern due to lots of false positives (homographs).
[190] Both here and in pattern n.9 the impression was that the vast majority of IN were instances of the prepositions 'of', but in the examples greater diversification was sought.

suffered from this ambiguity and many strings may be in fact be false positives, even without additional context (e.g. 'cover page', 'stem cell', 'double interest', 'welcome dinner'). The most affected categories in terms of false positives were possibly the ones distinguishing between proper and common nouns, i.e. NP and NN. Casing was once again the culprit. The system associates a capitalized word to a proper noun, which is useful to identify acronyms or Named Entities (in the broad sense) (e.g. 'European Maritime, Safety Agency', 'External Action Service'), but in fact inconsistent capitalization negatively affects this rule (e.g. 'Employment Strategy', 'Freelance Management', 'Dear President'). Moreover, a capitalized word may simply indicate the beginning of a sentence (e.g. 'Old age', 'Verification service', 'Bacon crisp', 'Weak practice'). By considering all noun forms as one aggregated category these issues could be eliminated, but the usability of such large category remains questionable.

One study aimed at distinguishing Named Entities (i.e. proper nouns) from other types of text strings (Vincze *et al.* 2011: 291) has found that humans are better at identifying Named Entities (NE) as opposed to other categories of multi-word units (MWU). In the study, manual categorization of a small corpus of Wikipedia articles showed higher aggregated counts for NE than MWU, suggesting that the incidence of NE is not negligible. When terminology is considered, the most represented POS is the 'noun' with a proportion between 84% and 98%. The great imbalance between nouns and other parts of speech is due to the tendency of most languages to use (complex) nouns over other parts of speech to label concepts (L'Homme 2005: 1119): users tend to refer to nominal strings as terms and such high numbers of nouns in the searches somehow justify the labeling of the concordancer as a terminological tool.

This brief overview showed that it is indeed possible to study strings using POS-tags. However, several shortcomings (e.g. the unknown ratio of false positives) have also been highlighted that make finer-grained results unsuitable for a systematic study based on frequencies. Results were often found to be in line with previous findings in Web search log analysis and Natural Language Processing. As expected, POS tagging was not particularly efficient because of the potential ambiguities both in form and meaning for each searched word (e.g. acronyms that resemble lexical words where the only difference is formatting, which is generally disregarded by the system).

7.3.1.2 VARIATION ACROSS STRINGS IN SEARCH SESSIONS

The aforementioned linguistic categories in Web searching were labeled "adjective and noun phrase" or "verbal phrase", i.e. the categorization was carried out at the syntactic level. Following the definitions in Bussmann (1996), a Noun Phrase (NP) is understood as made up of a head and one or more attributes whereas an Adjective Phrase (AdjP) has an adjective as head that can be modified by an adverb of degree or a complement (1996: 20); a Verb Phrase (VP) consists of a verb, its (obligatory) complements and its (optional) adjuncts. Prepositional Phrases (PP) have a preposition and a NP that serves as its object. An additional level was included to account for longer chunks of text introduced by a coordinate conjunction in the 2/11-grams dataset. They can take the form of a coordinate phrase or a coordinate clause, where the clause is seen as structurally independent (Bussmann 1996: 716).

There are two main approaches to investigate strings in terms of their syntactic constituents. The first only considers individual strings, e.g. the strings found in the spot search subsets; the other approach looks at the dynamic group of strings, i.e. search sessions. The underlying difference between the two approaches lies in the granularity of the analysis. The former assigns the whole sting to a category, whereas the latter looks at

the changes that occurred from one search to the next and categorizes the strings on the basis of variation in successive searches, i.e. the delta portion of the string.

The categorization for the static analysis was based on a judgment of the syntactic nature of the string as a whole and included strings from the spot subset as well as strings from search session A1 (i.e. resubmission), where the search is repeated multiple times without changes and can be seen as a single search. The strings that belong to this group have static boundaries and the problem unit (as the focus of attention) can only be inferred from individual search instances of various lengths. Table 43 provides a rough classification of the main types of identified problematic items in the subset of n-grams comprising 2 to 11 words.

Table 43. Classification of queries in spot searches group and session category A1 with verbatim examples from the logs.

PHRASE	**Examples**
NOUN PHRASE	apartment houses, cooperative societies, auxiliary staff
ADJECTIVE PHRASE	unforeseen financial, royalty-free non-exclusive, large corporate
VERB PHRASE	play their part in the labour market, address concerns, push forward, to fully implement
PREPOSITIONAL PHRASE	on the account of, in parallel with, after submission to, among other, out of law
ADVERBIAL PHRASE	just before, specially on, early enough
CLAUSE	
INDEPENDENT	how to get paid for goods and services; The objective of the inspection was to ensure that; These findings are being followed up
DEPENDENT	since the application should be dismissed; After which they would be deemed approved

Table 43 only provides a few examples for each category taken verbatim from the logs but there are also instances where there is a combination of phrases (e.g. 'at regular intervals after each session', 'enter into force with immediate effect', 'take account of the implications of the Lisbon Treaty'). Given the results of the analysis of the lexical categories, the vast majority of strings are noun phrases and to a lesser extent verb phrases. At the same time, this classification is not particularly informative because it does not offer a considerably different picture from the previous POS-tagging analysis. Moreover, the focus of attention in terms of string size does not change in the case of static searches and no additional information about problem units can be obtained.

The dynamic analysis of searches takes into account search sessions. This study is more complex than the previous one because more variables are involved. The question arises as to if and how the translator's focus of attention changes with respect to the problem unit when the temporal dimension of a search session is considered. Given the adopted definition of search session (see Section 5.6), the information need is not going to change within a search episode but the focus of attention can be directed to different textual elements. The first string in a session can be taken to represent the working source-text segmentation made by the translator. While working, the translator has first to cognitively process the ST and it is only at a second stage that s/he submits the query. During the session, the string changes in size, implying that the core problem unit was either framed in the very first string or left as a shortened string at the end of the session. The first string does not always necessarily match the problem unit. The assumption is that the first search operation is carried out according to the Principle of Least Effort (Azzopardi 2009: 557): the user tends to submit the search that s/he feels best suited for

his/her current information need. Should the search be unfruitful, a different search strategy is resorted to and the user will most likely trim the original string, according to the results reported in Sub-section 6.1.2.2.

In analyzing search sessions, a key distinction should be drawn between the expansion and reduction categories. Strings in the reduction category (C) are progressively shortened. Probably, the translator first entered the portion of text which s/he was focusing on but then decided to modify the search string in order to increase recall, e.g. after the system returned no results (failed search). When trimming a string within a session, the information need (core problem) is assumed not to change so the trimming has to be relevant and effective with respect to the original problem (information need). In other words, the translator must have removed "superfluous" non-core parts from the string to center the search on the core problematic item instead. In the expansion category, on the other hand, the translator seems to immediately focus on the core item. In this case, the search may result in high recall but low precision. Therefore, the search needs to be refined by expanding the (relevant) context of the string. This search strategy seems to complement the previous one, which means that in the case of an expansion strategy the core problem unit might found at the beginning.

These changes, however, do not necessarily affect the syntactic category of the whole string, despite the fact that a change in size has occurred (e.g. 'authentication dialogue box': NP > 'dialogue box': NP; 'rural development policy': NP > 'rural development': NP). In some other cases, the syntactic category changes but the shifts seem more casual than predictable and cannot be accounted for in any systematic way (e.g. 'set any legal framework for' VP > 'any legal framework for' NP >> expected: 'set any framework' VP; 'in connection with the EU 2020 strategy' PP > 'EU 2020 strategy' NP >> expected: 'in connection with': PP). As a consequence, the string categorization that was applied in the static part of the analysis and to category A1 is not helpful if used with categories C (reduction) and D (expansion). The reduction and expansion categories involve a more or less marked change in string length because of the loss or addition of characters or phrases. The identification of the problem unit becomes less straightforward and creates an objective problem in framing how the search string evolves in time.

A way out of the problem could be to focus on the changes that took place during the search session and use these as the basis for classifying the strings. This means looking at the portion of the string that was added or removed (i.e. the "delta string") rather than focusing on the initial and final string. This approach is more helpful from a syntactic point of view because in a session with two queries (i.e. the vast majority), there will only be one delta string. In addition, the delta string can be related to the wider context of the longer string in the session in order to solve potential ambiguities. From the perspective of the problem unit, classifying the possibly "superfluous" element also means identifying the "problematic" one. A number of categories for the delta strings were found (NP, VP, PP) which could be further broken down taking into account the strategy employed (expansion or reduction)[191] and the affected syntactic component, e.g. head or modifier.

Examples of queries for each search strategy and delta category are provided in Tables 44, 45 and 46. Not all cells could be filled with examples because in some instances there was a syntactical and/or logical incompatibility with the direction of the trim (or expansion) and the natural position of the lexical item (for example, an attributive noun added to the right of its head). In some other cases, there were simply no examples in the dataset (for instance, an adjunct to be added to the left of the string in category D1).

[191] The analysis was carried out for the automatically labeled categories, i.e. C1, C2, D1 and D2.

Table 44. Verbatim examples of additions and deletion of elements in noun phrases for the trim dynamic category (C1, C2).

TRIM	CATEGORY C1 / LEFT TRIM		CATEGORY C2 / RIGHT TRIM	
NOUN PHRASE (Head + Attributes)				
of HEAD				
	meat imported for human consumption	imported for human consumption	operational implementation period	operational implementation
	accreditation for decentralised implementation	decentralised implementation	zero emission standards	zero emission
of ATTR.				
Attr. Adj.	detailed relevant developments	relevant developments	Minimum amount present	Minimum amount
	technological champion	champion		
Attr. Noun	uranium enrichment facility	enrichment facility	…	
Genitive Attr.	Commission's roadmap for a low-carbon economy	roadmap for a low-carbon economy	course of the life of the contract	course of the life
Prep. Attr.	together with the Annexes and the declaration made unilaterally by the European Union attached to the Final Act	attached to the Final Act	charging systems for electric car batteries	charging systems
Adv. Attr.	hereinafter the authorised agents	the authorised agents	rate foreseen therein	rate foreseen
	formerly isolated areas	isolated areas		
Infinitive Group	to appoint ad personam	ad personam	Parties to the UNFCCC to encourage the widest participation in the UNFCCC process	Parties to the UNFCCC
(non-) Restrictive Clause	that was designed to increase	designed to increase	policies that target Roma women, who are victims of twofold discrimination	policies that target Roma women,
Apposition	Land Carinthia	Carinthia	National Grid Initiatives (NGI)	National Grid Initiatives

Table 45. Verbatim examples of additions and deletion of elements in noun phrases for the expansion dynamic category (D1, D2).

EXPANSION	CATEGORY D1 / LEFT EXPANSION		CATEGORY D2 / RIGHT EXPANSION	
NOUN PHRASE (Head + Attributes)				
of HEAD				
	Common European Asylum System	creation of a Common European Asylum System	further implemented	further implemented inquiry
	Fiscal Studies	Institute for Fiscal Studies	primary market	primary market operations
of ATTR.				

Attr. Adj.	regulators	nursing regulators	…	
Attr. Noun	DIAMAP	FP7 DIAMAP	…	
Genitive Attr.	land management	watershed's land management	Court of Justice	Court of Justice of the european union
Prep. Attr.	…		quarterly report	quarterly report on the euro area
Adv. Attr.	capable	financially capable	…	
	interconnected	technically interconnected		
Infinitive Group	…		procurement procedures	procurement procedures to be implemented by
(non-) Restrictive Clause	tackled	risks that have to be tackled	documentary evidence	documentary evidence which has to accompany consignments of animal by-products for the purposes of traceability
Apposition	ashton	vice-president ashton	High Representative	High Representative ashton

Table 46. Delta strings for VP and Prepositional Phrases.

TRIM/ EXPANS.	**CATEGORY C1 / LEFT TRIM**		**CATEGORY C2 / RIGHT TRIM**		**CATEGORY D1 / LEFT EXPANSION**		**CATEGORY D2 / RIGHT EXPANSION**	
VERB PHRASE (Verb + Complements + Adjuncts)								
of VERB/AUX	take the preliminary view	preliminary view	Shared values are to be expressed	Shared values	earlier diagnosis	encourage earlier diagnosis	My services	My services are available
	will be assisted by EU funds	assisted by EU funds			be considered to represent	shall be considered to represent	Some of the Member States	Some of the Member States have published
of COMPL.	Balance to be recovered	recovered	will highlight everything	will highlight	undertaken	actions to be undertaken	ability to assume	ability to assume the obligations
of ADJUNCTs	sufficiently state the reasons for its decision	state the reasons for its decision	the activities organised in every EU country	the activities organised	…		to be implemented	to be implemented immediately
			allocate costs efficiently	allocate costs				
PREPOSITIONAL PHRASE (Preposition + Noun Phrase)								
of PREP. GROUP	In the absence of solid evidence	solid evidence	court order issued in relation to	court order issued	EU regulations	in line with EU regulations	…	
of PREP.	at the closure stage of	closure stage of	By way of derogation to	By way of derogation	case law	under the case law	of the whole allocation	of the whole allocation for
of NP	…		sound strategy for risk management	sound strategy for	…		affected by	affected by cost increases

The examples in the tables show how varied and to some extent unpredictable search sessions are in terms of core problem units. There are instances where only one or two words are missing or added (e.g. 'operational implementation ~~period'~~) and others where as much as 70% of the string is removed or added (e.g. 'tackled > risks that have to be tackled').

All the most common parts of speech, i.e. nouns and proper nouns, adjectives, verbs (both main and auxiliary), prepositions and adverbs (in descending order) seem involved in these changes. Barr *et al.* (2008) have studied Web query reformulation and applied POS tagging to see what element of a query was changed within the same search session. They used a set of automatically tagged queries and calculated change probabilities for each part of speech. Proper nouns were found to be the most represented tag in the corpus but adjectives were the lexical category most likely to be reformulated. The word class with the highest reformulation probability was however numbers (e.g. change of year, model number). A change in number was understood to modify the core search meaning (i.e. it reflects a different search need), similarly to nouns and proper nouns. Reformulation of adjectives, on the other hand, implied a refinement strategy without altering the search intent. According to Barr *et al.* (2008), modification of the remaining parts of speech occurred quite rarely, suggesting that they were not affecting document retrieval as much. However, their study focused on modifications affecting a single word in a search string rather than additions or deletions. In this sense, examples would be better compared with macro-category F (replacement) in the present study. According to the manual categorization performed on the Finnish subset (see Sub-section 6.1.2.1), replacements affected mostly nouns and adjectives (sub-category F5 – word replacement) and, to a lesser extent, verbs in the case of tense change (sub-category F1). Typo-fixes (sub-category F4) were the main cause of word replacements but were not necessarily related to a specific word-type, although nouns were the most likely to be affected by typos due to their higher frequency in the dataset.

So far, only small modifications at phrase level have been considered but there were also instances where even greater units were involved in the expansion or trimming, i.e. clauses or coordinate phrases for which examples are provided in Table 47.

Table 47. Examples of addition and deletion of coordinate phrases or clauses for each dynamic category.

COORDINATE PHRASE/CLAUSE							
CATEGORY C1 **LEFT TRIM**		**CATEGORY C2** **RIGHT TRIM**		**CATEGORY D1** **LEFT EXPANSION**		**CATEGORY D2** **RIGHT EXPANSION**	
in bulk or wholesale packages	wholesale packages	original design and periodic reform	original design	economic profitability	financial and economic profitability	farmer associations	farmer associations and groups
		under this Directive or under the Regulation	under this Directive	cultural industries	creativity and cultural industries	conditions of detention	conditions of detention as well as guarantees

This small analysis highlighted a number of changes that can affect phrases and their components. The most articulated group was found to be the Noun Phrase, particularly at the attribute level, even though it also was the sub-group with most empty cells, especially for the expansion category. Generally speaking, strings belonging to sessions

where the end part was modified (right trim/expansion) seemed to change more markedly in terms of length than strings in the left trim/expansion categories. Trimming or adding whole semantic or syntactic units (e.g. attributes, adjuncts) at the end of a string rather than at the beginning seems cognitively easier (or more relevant) and is probably a consequence of the syntax of English.

Ambiguity in the strings due to the lack of context can becomes a problem when trying to label a search from a linguistic perspective. Consider for example, strings such as 'targets as a reference' or 'government expenditure targets'. In isolation, the first 'targets' could be considered a verb and the string would become a VP, whereas the second will initially be read as a noun (NP) but at a closer look the verb option cannot be excluded. If the whole string is searched for in the Web, the original sentence is found, i.e. 'The assessment of effective action will benefit from taking compliance with general government expenditure targets as a reference, in conjunction with the implementation of planned specific revenue measures', from which it becomes clear that the word 'targets' has to be interpreted as a noun. Unfortunately, this additional check cannot be performed for the vast majority of the strings, particularly in the case of spot searches.

The linguistic analysis of syntactic categories, at both lexical and phrasal level has been useful to highlight once again the similarities with Web queries. However, it has not provided much new information nor suggested how to approach existing ambiguities. The analysis has to move beyond the syntactic level and look at the data from another perspective.

7.3.2 PROBLEM CATEGORIES IN EMPIRICAL STUDIES OF TRANSLATION

Empirical studies have occasionally used categories for translation problems either in experiments or in translation teaching (see Chapter 2). In this section, a reprise of the main translation problem categories in the literature will be briefly discussed and their usability for the present analysis[192] will be assessed.

The first categories to be identified back in the mid-1980s were reception problems (*Rezeptionsprobleme*), production problems (*Wiedergabeprobleme*) and combined reception-production problems (Krings 1986a: 144-152), in the attempt to link the problematic element to the L2, the L1 or both. A few years later, Lörscher (1991a: 201-217) defined translation problems as manifesting themselves in the source-language text and grouped them into three categories. His first and eventually most populated category (ca. 70%) was called lexical problems, i.e. "single lexemes of the SL text for which the subject has no corresponding TL lexemes available" (1991a: 202), e.g. 'Eigenschaften und Merkmale'. The second category (syntactic problems, ca. 8%) was concerned with the syntactic arrangement of the lexemes (e.g. 'wann welches Deutsch mit wem geschprochen wird') while the third group (lexico-syntactic problems, ca. 22%) either included both levels or was used when no distinction between the two could be made (1991a: 203), e.g. 'einen kommunikationsorientierten Unterricht' or 'the poetry of the race'. The main problem with both tripartite categorizations is the presence of the third "hybrid" group, not helpful for a systematic analysis. A further source-language based categorization was proposed by Campbell (1999) who identified problematic (i.e. "difficult") source text chunks on the basis of evidence in the translated version. He came to the conclusion that

[192] In this brief review no reference will be made to aspects such as the specific experimental conditions and the types of subjects etc. despite the fact that they are likely to have played a role in the choice of the categories.

difficulty was related to word class (verbs and adjectives scoring high, nouns being evenly distributed and function words found at the lower end), distributed meaning (e.g. in collocations or idioms), complex noun phrases (i.e. concatenated nouns), abstractness, frequency and familiarity. These categories embraced a wide range of perspectives, from parts-of-speech to semantics. They went beyond the level of "linguistic form" and introduced a subjective component with the category "frequency and familiarity". A two-tier classification was used by Désilets *et al.* (2009) in a Contextual Inquiry study with professional translators. Thanks to direct observations and interviews, they first identified the two categories of LSP and LGP problems to study the degree of specialization of problems. Secondly, they grouped problems according to their nature into eight categories (Table 48) and defined problems as "difficulties encountered during the process of translation, such as terminology, phraseology and named entities". Unfortunately, no precise methodology was provided to explain the choice of labels and the criteria used to perform the classification. For example, 'Junior High School' was capitalized as if it was a proper noun, was labeled an LGP problem on the specialization axis but then fell into the category "Term" when its nature was considered (Axis 2).

Table 48. Classification of translation problems according to Désilets et al. *(2009).*

Specialization (Axis #1)	**Examples provided**
LGP	Junior High School / outcome / value for the money / Jordan / City Hall / Go Huskies!
LSP	adjunct professor / cover the allegations / MCH program
Nature (Axis #2)	**Examples provided**
Term	adjunct professor / Junior High School / currency (Finance) / letter carrier depot
Highly polysemic vocabulary	determine / outcome / step / grave
Phrases (collocations, idiomatic expressions, phraseology)	value for the money / cover the allegation / wooden-hearted / on short notice
Named entities	Jordan / Pearson International Airport / MCH program / Kelowna accord
Typography	*capitalization, hyphenation*
Cultural realities	Go Huskies!
Official translation of a sentence	*official translation of a quote from a legal judgment*
Misc. General Language	inconsistency / disempower / initially / young and growing population

In this taxonomy, semantics is still mixed with syntax and lexical categories, without clearly identifiable boundaries. For example, there are some phraseological units that might be considered terminology (e.g. 'value for the money') and others that may be simply labeled LGP problems, such as 'young and growing population'. A previous simplified categorization (Désilets, Farley *et al.* 2008) provided two examples for each of the chosen five categories (Table 49) but here too levels are still mixed. In this case, LGP and terminology seem to include only single-words, whereas phraseology covers longer units.

Table 49. Simplified categorization of types of translation problems as found in Désilets, Farley et al. *(2008).*

Type of Translation Problem	Examples provided
Terminology	subsidiary, fuel-oil
Phraseology	on short notice, for more than a decade
General Language, Polysemic Words	grave, fiery, step
Cultural/Country-specific Realities	Go Huskies! Liberal Indian Affairs critic
Named Entities	Sun (company), Xinjian Uighur autonomous region

A different take on problematic items can be found in Angelone (2010), where problem solving is associated with uncertainty, operationalized as an interruption of the translation flow. In his analysis, he examines the textual level, the behavioral level and the translation locus but the present discussion will only touch upon the textual level. The textual level is aimed at locating the problem-solving behavior "at a particular location of difficulty in the text" (2010: 28) and takes into account the following levels: lexis, term, collocation, phrasal, syntax, sentential and macrolevel. However, no further definitions or examples are provided as to what is precisely understood under each label. Adopting the same categories for Euramis becomes quite challenging and some of the listed categories would have to be left out given the specific nature of the dataset in this study.

A recent doctoral dissertation deals specifically with the topic of "how translation trainees use the Web as an external resource to satisfy their information needs within the context of domain-specific translation" (Enríquez Raído 2011: 145-6), which bears evident resemblance to the present study but has a wider scope because it investigates *any* type of information need for which the Web can be used. Enríquez Raído's data show that the most frequent information need concerns general lexical items (2011: 414) but, at a closer look, a good deal of the identified items could in fact be ascribed to terminological and/or thematic needs (p. 415). Although no systematic labeling of the problems encountered by the subjects is performed, a number of categories are mentioned in the discussion of results. Information needs (both common and individual) are associated with nouns, adjectives, verbs, adverbs, multi-word expressions, collocations, acronyms (p. 348) and proper names (p. 422); participants' feedback added the categories "unknown words", "allosemic words" (i.e. words that are used in an unusual sense), "polysemous words" and "false friends" (p. 348). Finally, one additional category includes unreported general lexical problems "for which [students] were mainly looking for confirmation (or 'reassurance') of already existing tentative solutions" (p. 486). Here, too, different levels are combined, from lexical to semantic categories and from lexico-syntax to familiarity, but no scope is provided for the given categories.

Eventually, none of these classifications can be directly transposed in a formal language to be used with a script. The main shortcoming in most of the covered classifications is the use of intuitive categories that are loosely defined and allowed to overlap or intertwine. Before attempting a further categorization of strings from a linguistic perspective, a closer look into the commonly chosen labels is necessary to better understand the scope of each category. A known problem when using taxonomies and classifications for units of language is the lack of frequency data, particularly for phrases (Stubbs 2002: 215). This is due to both a deeply rooted lexicographic tradition which tends to consider only selected examples and the variability of the units in question. The smallest unit to be analyzed is the word, generally nouns, and the categories employed are meant to discuss polysemy or parts of speech but additional categories need to be found for strings longer than one word. Quigley (2005: 29-31) identifies a cline that ranges from a single lexeme (e.g.

'baseload') to longer structures that include several lexemes but cannot be considered unit-like (e.g. 'Unregulated electronic information processing industry'); some of the most frequent labels are discussed in the following sub-sections.

7.3.2.1 COMPOUNDS

A compound is defined as lexical unit consisting of a set invariable combination of two or more words which has a single referent or designatum (Quigley 2005: 30) but in fact there are many types of compounds without there being clear-cut boundaries for their identification. Three main types can be nonetheless identified:

1. Close compounds, i.e. words written solid without any space or hyphen;
2. Hyphenated compounds, such as 'fuel-oil' in Table 49;
3. Open compounds, i.e. words separated by a space, which is usually the form taken by new compounds.

Different types of compound have uneven frequency distributions, which makes compound analysis quite challenging. According to a recent study (Maguire *et al.* 2010: 67), some compounds come with a lexicalized definition but many compounds are unique or occur only once in a dataset, as is the case for almost 70% of the 400,000 types found in the BNC (Lapata & Lascarides 2003 in Maguire *et al.* 2010: 67). This is a concrete problem in the field of Computational Linguistics, where the best performing methods to interpret compounds focus on 2- and 3-word NN compounds. These systems rely either on domain-specific hand-coded semantic taxonomies or statistical models built on large collections of unlabeled data, but the large number of rare and infrequent NN compounds makes probability estimation unreliable (Girju 2008: 189).

Another problem derives from the potential polysemy of many compounds, which are by necessity context-dependent. Their meaning cannot be derived without knowing the context because the same structure can have equally probable interpretations just as different structures may have hardly any semantic differences (Berg 2006: 212-4). The study conducted by Berg (2006) focused mainly on four-word compounds both in English and German. He considered nominal (NN) compound strings such as 'air traffic control', 'child language acquisition research group', 'meat safety assurance scheme' and 'fork lift truck driver' which are considered *unmarked*, i.e. their constituent structure reflects the semantic structure. In the specific case of English and German the compounds comply with the "right-hand head rule", i.e. they have a modifier-head structure, which implies a left branching of the compound in the case of recursivity. Left-branching compounds are perceived as the unmarked structure, therefore they are more easily understood than right-branching compounds (2006: 224). This theoretical framework is in line with findings from psycholinguistic studies according to which people exploit statistical regularities when interpreting novel compounds (Maguire *et al.* 2010: 51). Previous studies (e.g. Gagné & Shoben 1997 cit. in Maguire *et al.* 2010) have identified a set of 16 possible relations between noun and modifier, and findings from experiments have shown that difficulties arise when there is an unfitting relationship between the noun and the modifier(s). For example, 'plastic' is usually associated with the relation <made of> so the combination 'plastic bag' is easier to understand than the combination 'plastic crisis' where there is a relation of type <about>. A similar example found in the dataset for the present study could be 'carbon emission' or 'carbon leakage' compared to 'carbon tax' or 'carbon footprint'. In other words, the problem originates from the semantic relation between constituents, which is often implicit. The concept of the unexpected relationship within compounds was also developed by Girju (2008). To correctly interpret a NN compound, one requires information ranging from word knowledge to lexico-syntactic

and discourse information and much depends also on the syntactic and semantic directionality of the compound (2008: 186). The standard case involves two lexical nouns with a particular syntactic directionality (usually N_1 is the syntactic modifier and N_2 is the syntactic head) encoding a semantic relation with two semantic arguments (Arg_1 and Arg_2) with the same semantic directionality. One example is 'beer glass' where N_1 modifies N_2 and Arg_1 expresses a PURPOSE relation with respect to Arg_2. Compound interpretation is undermined whenever there is a mismatch between the syntactic and the semantic directionalities, i.e. when N_1 does not match Arg_1. In addition, Maguire *et al.* (2010: 55) point out that the analysis is challenged when only the most frequent combinations are considered because there are thousands of possible combinations for a given noun and restricting the scope to the top frequencies may result in an unrepresentative sample.

These studies are generally carried out with native speakers and in a monolingual environment. In the case of translation, the scenario is even more complex because contrastive studies would be needed to better evaluate particular structures. One hypothesis is that family proximity between the two languages involved in translation makes the transfer of some semantic structures unproblematic whereas the same semantic structure could represent an issue for a speaker of a more distant language family. In other words, some problems may arise from interference between languages and the lack of a statistical basis to interpret compounds. In the particular context of the EU there can be instances of artificial interference that may add to the difficulties of interpreting and translating compounds. Some examples of this kind of interference were provided by a number of EU translators in their replies to a questionnaire (DGT 2010a: 71ff.):

> The German equivalent of the term *Lisbon process* was translated under the influence of the English drafting language as *Lissabon-Prozess* using and thereby distorting the term *Prozess*, originally used in German language to mean a 'judicial trial' or 'natural processes' (2010a: 92).
>
> Autonomous concepts of EU law expressed by autonomous term (*conformity assessment, type approval*) are often calqued from the original language into the other official languages (2010a: 73).
>
> In some legal acts, the term 'reception' was erroneously translated into Portuguese as *receoção* as a result of reciprocals translinguistic lexical attraction with the French term *réception.* [...] The expression *retrait des navires* was also translated under reciprocal translinguistic attraction by *retirada de navios,* however, the correct term would have been *abate de navios* (2010a:92)[193].

Compound interpretation can be ascribed to two main approaches, i.e. statistic-based and schema-based theories. According to Maguire *et al.* (2010: 66), "[...] people can activate conceptual knowledge selectively by exploiting regular patterns that exist in compounding, thus avoiding the consideration of irrelevant information", i.e. the activation of a full conceptual schema. Their results (2010: 57) show a significant association between semantic content and combinatory use, i.e. similar concepts combine in similar ways. In other words, the more similar concepts are, the more likely they will combine with the same nouns but the combination probability varies according to whether the noun is used as a head or modifier (2010: 63). Moreover, productivity is closely linked to interpretation, in that a productive pattern is strongly associated with a particular semantic interpretation of the compound, which explains why some very

[193] All emphases in the examples cited were found in the original.

infrequent structures in a language are perceived as difficult. For example, some EU-specific expressions (such as the Polish EU-term for 'certificate of conformity', which is different from and broader than its national equivalent, and the Finnish term for 'female bovines' which was specifically created at the EU level and even attracted criticism among Finnish speakers; DGT 2010a: 75, 79) are perceived as difficult by translators even though they have an internal unmarked structure. In sum, "statistics based on the interaction of semantic categories are more informative than statistics based on the relation preference of individual words" (Maguire *et al.* 2010: 65).

This brief theoretical overview on compounding has highlighted how entangled the syntactic and the semantic levels are in groups of two or more nouns. They may be considered two sides of the same coin and this dual nature is partly found in the two main approaches to the study of collocations: the statistical approach (e.g. that of Firth, Sinclair and Smadja) and the semantic-lexicographic approach (associated with the names of Hausmann, Benson and Mel'čuk). For others (Partington 1998: 15 in Mollin 2009: 177), there are not two but three different approaches: (i) textual (Sinclair 1991), (ii) statistical (Halliday 1985) — the standard approach in contemporary corpus linguistics — and (iii) associative (such as Hoey's theory of "lexical priming", 2005).

7.3.2.2 COLLOCATIONS

That of "collocation" is a particularly challenging label given that admittedly "[t]he term collocation does not cover the same range of linguistic phenomena for all linguists [...] and unfortunately, not all researchers spell out their definition of the phenomenon" (Mollin 2009: 176). The concept of collocation was first suggested by Firth (1957) as a new category of meaning that looked at lexical co-occurrences in a text from a quantitative perspective and from there collocation theory developed[194].

In her definition, Quigley (2005: 28) resorts to the idea of a continuum and describes collocation as "a bond of varying strength, from loose to tightly cohesive, between two or more words". This can range from a transitional combination to a compound, but it is ultimately impossible to know the exact strength of this word group (2005: 29). A collocation is generally understood to mean "the occurrence of two or more words within a short space of each other in a text" (Sinclair 1991: 170); more specifically, collocations are "unidades fraseológicas fijadas sólo en la norma, es decir, sintagmas completamente libres a los que el uso les ha conferido cierto grado de restricción combinatoria" (Corpas Pastor 2003: 135). The word combinations that could be labeled collocations are not entirely free, in that there is a greater (but not fixed) likelihood than chance for the words to occur together[195].

In the framework of the statistical approach, this greater-than-chance likelihood is measured with some standard statistical measures such as the *p*-value, the *t*-score and probability measurements. Stubbs (1995) used raw frequencies, Mutual Information (MI) and the *t*-score to study collocations and calculated an I-value for lexical collocates and a T-value that identified lexical and grammatical collocates. The statistical value of verifying a null-hypothesis that assumes *no* correlation between two phenomena in the language was challenged by Kilgariff (2005) who maintained that language is *never* random, therefore the null hypothesis will never be true because "almost any combination is more frequent than expected by chance in real language" (Mollin 2009: 192). Statistically significant deviations are generally sought in real corpora compared to hypothetical

[194] See Krishnamurthy (2001) for an essential historical overview.
[195] Also known as "certain mutual expectancy" (Jackson 1988: 96 in Quigley 2005: 28).

corpora but textual data are never proper random samples and "this calls into question whether such statistics can reasonably be used on language data" (Stubbs 1995: 29). Significance in statistical analysis is not always reached due to the high natural variability of language and because statistical measures are often transformed to fit known value sets. For this reason, raw frequencies and joint co-occurrences should always be taken into consideration in the discussion of statistical results. Absolute frequencies together with a range of association measures such as MI, *z*-score and log-likelihood models were also employed by Mollin (2009: 177), who adopted a corpus-based statistical approach to collocations. Her analysis highlighted the possible biases of association measures. For example, some measures favor very infrequent terms with respect to grammatical collocations, which are in turn overestimated by other measures.

The non-randomness of word combinations appears partly substantiated by Quigley (2005: 28), who states that there is an intimate relationship between collocations and verb complementation (i.e. valency) in that many Noun-Verb collocations require a specific distribution of semantic roles and collocations should not necessarily be considered as binary units. Association measures are not suitable to account for the complexity of linguistic phenomena because they are only able to consider two variables at a time, as already established earlier (Dias *et al.* 1999). In addition, there is no consensus in the literature as to the cut-off point below which instances of node and collocate should be discarded (Dayrell 2007: 383), though some studies use a frequency of 4 instances.

This is where the psycholinguistic associative approach becomes very relevant, in particular the concept of lexical priming (see Hoey 2005) according to which "words are 'primed' for use through experience" (Mollin 2009: 178). Results show that the strength of the collocation in the corpus (in statistical terms) does not correlate with the strength of association (Mollin 2009: 185) and, what is more, there seems to be no systematic relationship between frequency of co-occurrence and the frequency of association found in experiments. In addition, Hunston (2002: 20 in McGee 2009: 81) affirms that humans have a weak intuition with respect to collocations, frequency, semantic prosody and phraseology, and combinations are often difficult to evaluate without (statistical) information from corpus data. This is confirmed by Stubbs (2002: 219-220), who states that native speakers cannot spontaneously and systematically retrieve the most frequent two-word collocations (node and top collocate) but will recognize them "in retrospect – as entirely banal".

Additional limitations in word combinations are also discussed in Beitzel *et al.* (2007: 13), who address the concept of "selectional preference", according to which words tend to prefer syntactic arguments that belong to particular semantic classes. L'Homme and Bertrand (2000: 497) use a different label to refer to word combinations involving a term (a 'keyword', for example a noun) and a co-occurrent (e.g. a verb, an adjective or another noun). They call them *Specialized Lexical Combinations*, i.e. concept-bound word combinations to be found in specialized domains which differ from *Collocations*, i.e. word groups used in LGP. A concept-based labeling distinguishes between *lexical collocations* where the co-occurrent combines with a single terminological unit (1:1) and *conceptual collocations* where there is a 1:n relationship between the co-occurrent and other terminological units (Heid 1994: 239 in L'Homme & Bertrand 2000: 498). According to their findings, conceptual collocations are highly productive in specialized languages with one specific term being used more frequently over the other semantically-related terms.

McGee (2009: 97) moves a step further when he maintains that longer sequences in human lexicons appear not to be collocations but language chains whose lexical

components seem less available to memory searches according to the responses elicited from the subjects in the course of an experiment. For example, the dyad 'possible exception' is usually embedded in a larger chain such as 'with the possible exception of'. The noun collocate 'exception' of the adjective 'possible' seems to be less accessible to memory searches than unit-like complete collocations such as 'good idea' which often occurs without the determiner. In other words, "most frequent noun collocates typically combine with the adjective in a larger chain of language, and they are 'incomplete' as 'bare' collocates" (McGee 2009: 98).

The challenge posed by collocations has also been addressed in a practical sense by tool developers who have put together systems for collocation analysis and retrieval. One example is COLLOCATION | ANALYZER (Kimmes & Koopman 2010) a tool to retrieve, verify and compare collocations developed with translators in mind. Another resource is the Sketch Engine[196] a multilingual Web-based program whose core functions are to show possible concordances and provide information about the grammatical and collocational behavior of words.

7.3.2.3 LANGUAGE CHAINS AND PHRASEOLOGY

The notion of linear sequences of uninterrupted word-forms has received attention from several scholars who have labeled them in different ways (e.g. *clusters*, *chains*, *recurrent word-combinations, lexical bundles*) as there is no standard term for such strings. Irrespective of the chosen label, the understanding of such units is usually linked to quantitative aspects (i.e. frequency), which brings them intuitively close to the problem category "phraseology". In its wide sense,

> la fraseologia [...] engloba todas aquellas combinaciones formadas por al menos dos palabras y cuyo límite superior se sitúa en la oración compuesta, caracterizadas por una alta frequencia de aparición en la lengua y de coaparición de sus elementos integrantes, así como la institucionalización, la estabilidad, la idiomaticidad y la variación que dichas unidades presentan en diverso grado (Corpas Pastor 2003: 135).

For Stubbs (2002: 238), there are different aspects of English phraseology to be captured which correspond to different underlying concepts. The concept of collocation refers to the habitual co-occurrence of two content words within a small span, whereas the concept of colligation identifies frequent co-selections of a content word and an associated grammatical frame (that he also calls "chain"). He defines a chain as a linear sequence of two or more uninterrupted word-forms, which occur more than once in a text corpus, but also acknowledges that several terms are used to refer to these word groups, such as dyads/tryads, clusters, recurrent word combinations, statistical phrases, lexical bundles and n-grams (Stubbs 2002: 230). In the same fashion, there are no standard terms for the abstract grammatical sequences underlying these strings, variously called canonical form, construction, extended lexical unit, frame, pattern and template.

By finding frequently occurring chains in different texts, evidence can be provided about units of language use[197]. The statement "[p]eople speak in set phrases — rather than in separate words" (Mel'čuk 1998: 23) seems to effectively summarize this last point. Set phrases (or phrasemes) are one of the biggest challenges for theoretical linguistics and lexicographers alike. Mel'čuk explicitly acknowledges the non-uniformity in the use of the

---

[196] http://www.sketchengine.co.uk/ (last accessed: December 2012).

[197] Units of language *use* should not be confused with language units because often times the chains are not complete syntactic or semantic units (Stubbs 2002: 230).

label "collocation"[198], which he sees as the most prominent subclass of set phrases. As such, collocations need to be defined by their distinctive features with respect to set phrases that are not collocations. The main property of a phraseme is non-compositionality, i.e. they are lexical units that need to be treated as a whole. Interestingly, when seen as lexical units, phrasemes outnumber words 10:1 in any language (Mel'čuk 1998: 24). However, different languages may have different ways to encode and interpret the relationship between nouns in nominal phrases and compounds. According to Corpas Pastor (2003: 194),

> [...] la fraseología constituye uno de los aspectos más problemáticos en traducción. Además de la dificultad de detectar tales unidades en el texto de origen, resulta complicado determinar las estrategias de traducción adecuadas en cada momento. Salvo los casos raros de equivalencia total (por ejemplo, en los europeísmos), la mayoría de las veces se trata de equivalencia nula, aparente, o, en el mejor de los casos, parcial.

Girju (2008: 186) conducted a contrastive analysis between English and Romance languages (ES, IT, FR, PT, RO) with a special focus on Romanian, focusing primarily on compositional noun phrases and more specifically nominal phrases (NPN) and noun compounds (NN). While English tended to use both NN (right-headed) and NPN (left-headed) structures, Romance languages preferred the NPN structure with very few occurrences of NN compounds[199], which were usually restricted to few semantic categories or a specific text genre. For example, European Union texts (from the Europarl corpus[200]) have instances of NN compounds such as 'legge quadro' or 'Stato membro' in Italian, which precisely map the NN structure in English. If a NN English compound becomes a nominal phrase (NPN) in Romance languages, the preposition is the element that should encode the semantic relationship and as such it should cover the same semantic range as intended in the English compound. However, the NN>NPN transformation is not necessarily straightforward and easy to match semantically. For example, in English 'tea cup' and 'sailor suit' only encode the semantic relation PURPOSE but the same words in a NPN structure ('cup of tea' and 'suit of the sailor') encode CONTENT-CONTAINER and MEASURE and POSSESSION, respectively (Girju 2008: 187). In her experiment, annotations of compounds from two corpora resulted in instances (5-8% of relevant occurrences) where an example could belong to multiple semantic categories in the same context (2008: 196).

In particular, Girju (2008: 186) noted that there are very few studies dealing specifically with the function of prepositions in natural language processing (NLP) applications, despite the fact that prepositions are probably the most polysemous category in the language and a notoriously problematic one (Dragsted 2004: 109). The cross-linguistic dimension explored in her study has hardly been handled in the works on automatic interpretation of nominal phrases and compounds and there have not been any investigations on the *role* of prepositions in automatic NP interpretation between e.g. English and Romance languages (Girju 2008: 191). Just as with tokens in a text corpus, there is agreement that a limited number of semantic relations have a high frequency of occurrences but there is no clear picture as to their number and level of abstraction. Girju's experiment highlighted consistent prepositional choices across Romance languages in the encoding of a NN English compound into a NPN structure, e.g. for the

---

[198] "[...] there is, as far as I know, no universally accepted formal definition of collocations nor a proposal for their uniform and systematic treatment" (Mel'čuk 1998: 23).
[199] With the possible exception of Romanian, which uses a genitive marked NN compound.
[200] http://www.statmt.org/europarl/ [last accessed: December 2012].

semantic relations PURPOSE and LOCATION. In particular, the preposition "de/di" (i.e. "of") was frequently used but in fact it can be considered semantically unspecified, thus covering the vast majority of the remaining semantic relations. Her experiment also highlighted that each Romance language gives a contribution to the interpretation of an English instance; combining information from different languages may thus help interpret a structure that is semantically ambiguous, i.e. where multiple NPN combinations are possible due to various noun senses (2008: 212-216). This seems to apply very well to Euramis, where the user has the possibility of selecting multiple target languages (see Sub-section 3.2.3.2) to double check the proposed solutions against solutions in other languages and increase the chances of getting results in case of no results for one language. In this case, the chosen language combinations will also depend on the language proximity between target languages so that a user may select a language that he does not actually know or speak just because it is close enough to the one s/he is interested in. The study of multiple language combinations has not been carried out in the present study and searches where multiple target languages were selected have been initially discarded but they may nonetheless provide interesting insights into which combinations users perceive as useful and relevant.

7.3.2.4 FORMULAIC SEQUENCES

Formulaic sequences possibly fall in the realm of set phrases; from a cross-linguistic perspective, they are likely to extend beyond the level of phraseology up to the category of routine translations of longer stretches of text. Just as Quigley (2005) identified a continuum in the case of compounds, Wray and Perkins (2000: 1) describe a cline for *formulaic sequences*. They use this neutral label to describe a phenomenon that encompasses strings of various lengths that appear to be stored and retrieved from memory as self-contained chunks. More specifically, a formulaic sequence is

> a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (Wray & Perkins 2000: 1).

At one end of the spectrum, there are tightly idiomatic and immutable strings which are semantically opaque and syntactically irregular (e.g. 'by and large'); at the other end there are transparent and flexible structures with slots for open class items (e.g. 'NP be-TENSE sorry to keep-TENSE you waiting'). These structures can be categorized according to the extent of their fixedness from non-idiomatic to idiomatic, like one of the problem categories listed in Campbell (1999). The authors propose the following exemplification: 'under the table' (free combination), 'under attack' (restricted collocation), 'under the microscope' (figurative idiom), 'under the weather' (pure idiom) (Wray & Perkins 2000: 6). "Formulaicity" manifests itself in text strings where the relation of each item to the rest is relatively fixed and where there is limited scope to substitute one item by another of the same category. Once again, the authors underline "the looseness of the terminology, which makes it extremely difficult to be sure when like is being compared with like" (Wray & Perkins 2000: 3) and, as a case in point, they list over 40 terms found in the literature to refer to one or more types or subtypes of formulaic language.

7.3.2.5 TERMS

Term and *terminology* are possibly among the notions that are most frequently associated with translation problems as terminology is known to require a high level of accuracy and is often a fundamental component of a high quality translation. However, the concept of "term" has to be further investigated to better understand the actual scope within which the label can be applied. One of the main issues is the ambiguous use of the word *term* in the specialized literature: it sometimes refers to the combination of form and meaning in specialized vocabulary; in other cases it is a formal label for a specialized concept; in other cases still, it is used in a more generic sense without referring to any specific theoretical approach (L'Homme 2005: 1112-3). Most of the classifications encountered so far do not explicitly distinguish between specialized and general context, with the possible exception of the Specialized Lexical Combinations in L'Homme and Bertrand (2000). Terminology is one of the distinctive features of LSP texts and its constituents are terminological units, understood as units of knowledge and communication. Given the results obtained in the analysis of the search strings against EuroVoc (Section 7.2), one may argue that there is a considerable amount of terminology in the strings.

In the integrated theory proposed by Cabré Castellví (2003: 181ff.), i.e. the Communicative Theory of Terminology, terminological units are examined in their linguistic and semantic components as indivisible combinations of form and content[201]. They are seen as lexical units with or without a syntactic structure and may coincide with units belonging to general discourse but acquiring a discreet meaning within a subject field (2003: 184). The reason why word combinations in the analyzed strings cannot be systematically labeled as 'terminology' is that

> [i]n a theory of natural language the terminological units are not perceived as separate from the words which constitute a speaker's lexical space but as special meanings of the lexical units at a speaker's command (Cabré Castellví 2003: 189).

In practical terms, there are no distinguishing features of terminological units with respect to lexical units in phonological, morphological, syntactic terms but only in semantics and pragmatics. In other words,

> a lexical unit is by itself neither terminological nor general but [...] it is general by default and acquires a special or terminological meaning when this is activated by the pragmatic characteristics of the discourse (Cabré Castellví 2003: 189-190).

Terminological units are better referred to as *units of special meaning* because any lexical unit has the potential of being a terminological unit, but the condition of "terminological unit" does not exist prior to its usage in a specific communicative context (Cabré Castellví 2003: 190).

> La particularité du terme, par rapport aux autres unités lexicales d'une langue, est d'avoir un sens spécialisé, c'est-à-dire un sens qui peut être mis en rapport avec un domaine de spécialité. [...] La définition de « terme », contrairement à celle qui est donnée pour d'autres unités linguistiques, est donc relative. Elle dépend de la délimitation qu'on a faite d'un domaine spécialisé et les objectifs visés par une description terminologique (L'Homme 2005: 1125).

[201] The other theoretical models in terminology are the General Theory of Terminology (GTT), Socioterminology, Textual Terminology, the sociocognitive approach and Computational Terminology. They are reviewed e.g. in L'Homme 2005 and Cabré Castellví 2003.

This is possibly the most suitable approach to discuss terminology in the context of the present corpus of search strings in that there is no way to establish what the original communicative context was despite the fact that the EU itself may be considered a specialized context *a priori*. In this sense, it may be hypothesized that one type of information need derives from the specificity of the context rather than from the combination of words: their meaning and translation as terminological units have to be checked. A string such as 'organic unit' might be labeled as a terminological problem, i.e. one for which an appropriate translation exists for a specialized domain, as the string 'organic unit' can refer to e.g. a military context, biology or an organizational chart. Conversely, a string such as 'expiry review' is more likely to represent a syntactic problem. Even though it tends to appear in an unambiguous context, the semantic relation between the constituents may not be entirely transparent for all languages. An Italian translator, for instance, may be faced with two possible standard renditions of strings of the kind N+N, such as 'water distribution': one using a N+P+N structure (e.g. "distribuzione dell'acqua"); the other with a right-expansion (N+A; "distribuzione idrica"). The translator may already have in mind options for both structures because they are virtually possible and grammatically acceptable, but s/he would need to verify whether both solutions are actually used. If so, s/he would have to make sure that the domains of the translations match the source text domain and/or verify which solution is most frequent or most recent. In the case of 'expiry review', neither syntactic rendition may work 'as is'. One official translation found in EU documents is "riesame in previsione della scadenza", which means that the preposition slot in the N+P+N structure was rendered with an even more explicit P+N+P structure (i.e. "in view of"), the final string being N+P+N+P+N. Not all syntactic transformations from the initial NN structure are equally straightforward in spite of the fact that the source text element is clearly contextualized and understandable.

A close relationship between the problem (*what*) and the rationale behind the search (*why*) emerges when the query logs are examined from a linguistic perspective. Saying that Euramis is used for terminological searches implies that the main usage of the tool is for finding domain-specific translations for a given item, but about 40% of the searches have been labeled as LGP problems (see Section 7.3.2), which are possibly less related to terminology than LSP strings. However, a blurry list of concepts and labels that are not well-defined and often overlapping, like the ones discussed in the preivous sub-sections, is not ideal to carry out a systematic analysis of the searches and discuss the units found in the corpus. Choosing one label over another will produce a partial or even distorted view of the information need behind the search, particularly in cases where no additional information about the context is available. Therefore, none of the categories discussed so far can be effectively used in the analysis and a viable alternative for labeling the strings has to be found.

7.3.2.6 MULTI-WORD UNITS

One of the main problems encountered in the studies reviewed here is the multitude of labels and approaches to the study of closely related linguistic phenomena, suggesting a blurry dividing line between categories. Another issue is the frequent use of metaphors such as "cline" or "continuum" to account for phenomena of variation. Such metaphors, however, do not lend themselves to being encoded into a script or a formal language. This is the main reason why none of the previously discussed categories can be effectively used 'as is'.

Another common label for word combinations is *Multi Word Units* (MWUs) or Multiword Lexical Units, used in lexicography, corpus-based studies, Natural Language Processing

and information extraction, and sometimes used as alternative to *n-grams* in more computation-oriented fields. Lexicographers (Quigley 2005: 27) have also used the same label to include compounds, phrasal and prepositional verbs and possibly also abbreviations, acronyms and proper nouns. MWUs are defined in lexicography as a string of two or more words acting as a single lexeme and are used as superordinate term, whereas collocations (of varying degrees of strength), translational combinations[202] and compounds are subordinate species (Quigley 2005:28). MWUs still remain a thorny issue for lexicographers because there is no accepted approach as to how they should be entered in dictionaries. A similar concept is that of *Multiword Expressions* (MWEs), i.e. "lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasy" (Vincze *et al.* 2011: 289). Subtypes include compounds, verb-particle constructions (i.e. phrasal verbs), idioms, light-verb constructions (i.e. a semantically weak verb that accompanies a semantically full noun).

In the field of Natural Language Processing, Monti *et al.* (2011: 11) define a MWU as a group of two or more words (or terms) in a language lexicon that generally conveys a single meaning. They also agree that it is difficult to precisely delimit the concept and propose concurrent labels such as 'multi-word expression', 'fixed expression', 'idiom', 'compound word' or 'collocation' to show how varied their scope can be. Because "a linguistic specification of multiword lexical units seems to be a never ending task" (Dias *et al.* 1999: 12), a statistical specification of multi-word units is proposed, according to which they are seen as "a group of words that occur together more often than expected by chance" (Dias *et al.* 1999: 11), just like collocations, and partly overlap with compound nouns, compound verbs, adverbial locutions, prepositional locutions and frozen forms. From this definition, three types of MWUs are identified, based on their structure: (i) contiguous, (ii) non-contiguous with gaps to be filled by interchangeable words and (iii) free sequence of words.

MWUs can be extracted using their syntactic regularities on the basis of POS-tagging, shallow morphosyntactic information combined with statistics or using purely statistical and language independent approaches. MWUs may appear relatively easy to identify but are in fact a thorny issue for many applications in Natural Language Processing because there is extreme variability in the number of words, syntactic categories and relations as well as the flexibility of the expression. Different processing solutions can be adopted based on the degree of variability of the unit (i.e. the part of the continuum where the MWU is placed), provided that a semantic unity is found. For example, compound words with almost no variability and a specific grammatical function need to be lemmatized, whereas MWUs with high degree of variability need to be handled with rules (Monti *et al.* 2011: 12). A concrete problem occurs when there are nested expressions, i.e. MWEs containing another MWE (e.g. *'carbon monoxide* leak') (Vincze *et al.* 2011: 292), which may be solved with hierarchical annotation.

This brief overview of multi-word units in the literature has highlighted existing differences in the way the label is used and understood by researchers. Despite the inevitable differences, some common features can be found:

1. The definition of MWU presupposes that a word combination is perceived as a (lexical or semantic) entity.
2. The concept is used in different fields: lexicography, phraseology, information extraction and machine translation.

[202] "[L]ess variable than collocations but not yet as unchanging as compounds" (Quigley 2005: 29).

3. There seems to be some agreement about the superordinate nature of the label "multi-word units" with respect to labels such as compounds, collocations, idioms and prepositional locutions.

In sum, the concept of Multi-Word Unit can be said to be interdisciplinary, superordinate and produce recognizable units. These three features seem to meet basic requirements for describing the strings in the dataset. The choice of the label is particularly important because this analysis aims to be descriptive and as comprehensive as possible. However, a given label may produce some undesired associations and consequently unsubstantiated assumptions may arise, as previously exemplified. The concept of Multi-Word Unit does not presuppose anything but the fact that the user perceived the string as a unit (or part of a larger unit) when the search was launched.

The only exception in this respect may be represented by Named Entities because they are generally not listed as a subtype of MWUs like collocations, idiom and phraseology, though theoretically they fit the given definitions. Multi-word Named Entities are proper nouns, traditionally belonging to the categories of person, organization and location names; they "can be composed of any words or even characters and their meaning cannot be traced back to their parts" (Vincze *et al.* 2011: 289) but they can also be a component of in a nested expression, e.g. *'FBI* special agent' (2011: 292). Person names are often more challenging than other names (e.g. location names) in that they involve a deeper semantic analysis of the surrounding text and have greater variability (Fleischmann & Hovy 2002: 1). A clear boundary should however be established to determine the extent to which the label NE is applicable to some MWUs, given the examples of possible NEs discussed earlier (see Sub-section 7.1.1.2). This involves finding an operational definition for NEs to possibly include references to elements such as documents, projects and funds.

To sum up, Multi-Word Unit will be the label applied to the strings ranging from 2 to 11 words in length, with the tentative subcategory of Named Entities. Following Dias *et al.* (1999: 13) the concept of MWUs can be operationalized as "specific contiguous or non-contiguous *n*-grams in a window of ten words" (5:5). The MWUs in the Euramis corpus were not extracted automatically; they were deliberately selected by the user as a meaningful word combination wherein an information need lies. In this sense, they can be seen as manifestations of the way humans segment and possibly parse a text they have to translate[203].

Different statistical methods to extract MWUs from text corpora (i.e. debates of the European Parliament) in four languages (Portuguese, English, Italian and French) were compared and evaluated in a study by Dias *et al.* (1999). Irrespective of the purely statistical considerations, all tested models tended to extract more MWUs in Italian and the smallest amount in French. Precision rates seemed to confirm this trend, i.e. best for Italian, worst for French. Researches further distinguished between contiguous and non-contiguous MWUs (1999: 18). The former were further subdivided in (i) noun phrases, (ii) verbal lexical units, (iii) prepositional/conjunctive/adverbial locutions and (iv) prepositional/relative/coordination structures. The latter also were grouped in (i) noun phrases, (ii) verbal phrases, (iii) syntactical structures and (iv) templates for "long idiomatic domain dependent phrases" (1999: 19). For each language, over 70% of the extracted MWUs were noun phrases, in line with a previous finding according to which

[203] This is an important specification as it has been proven that reading for translation involves different patterns of eye movements (i.e. visual attention) than e.g. reading for monolingual text comprehension (Jakobsen & Jensen 2008) and differences in pause patterns have also been found in writing tasks (writing for translation vs. monolingual text production) (Immonen 2006).

more than 70% of technical terms are MWUs (Justeson 1993 cited in Dias *et al.* 1999: 19). Interestingly, the researchers found that the distribution of MWUs across languages was remarkably similar. More specifically FR, EN and IT mostly had noun phrases and verbal units, whereas PT had more locutions than verbal units, possibly due to the domain in question (i.e. legal). In addition, over 60% of extracted MWUs had a frequency of only 2 in the selected corpus, irrespective of the language, suggesting that the percentage of hapax could also be quite high. In the present study, the percentage of hapax legomena calculated as string types amounted to 46% for 1-grams; it quickly rose to 68% for 3-grams and stabilized around 80% from 6-grams onwards (see Appendix D – Table D.1). Results cannot however be directly compared across studies because all strings in the Euramis dataset are in English and the target language is just a parameter. However, striking similarities among languages were also found at various levels as discussed in Chapter 6 and the study by Dias *et al.* suggests that this may not be accidental. Provided that similar additional results can be collected, generalizations of some linguistic phenomena will be possible on the basis of consistent empirical results.

7.3.3 THE BILINGUAL MENTAL LEXICON

A searched-for segment implies that the translator's attention is focused on that particular text portion, which can now be referred to as a Multi-Word Unit. In the previous section, some suggestions emerged as to why some types of Multi-Word Units may be problematic for a translator. Cognitively, motivations included mismatches between the semantic and syntactic structure of the string or an unexpected relationship between its components (see 7.3.2.1). Language interference is also expected to play a role, particularly in the case of an EU translator, who may be asked to work on a text drafted in his/her L2, L3 or even L4. Ultimately, the concept of translation problem seems to boil down to a question of accessibility to a given MWU in the translator's mental lexicon. This would account for the great variety of searches that sometimes are linked to individual traits.

The mental lexicon can be described as "[...] a repository of all the words in a certain language an individual knows" (Stamenov *et al.* 2010: 325). Broadly speaking, a translator's brain may be compared to a bilingual brain where a bilingual lexicon is found.

> The underlying assumption is that the acquisition of more than one language requires at the very least a change to or expansion of the existing lexicon, if not the formation of language-specific components, and this is likely to manifest in some way at the physiological level (Meuter 2009: 1-2).

Neurolinguistic studies on bilinguals "fairly consistently point to proficiency as the determining factor in efficient language use" (Meuter 2009: 18). When discussing translation problems, differences were found not only between novice and professional translators but also in experiments where specialized translators were given a text type from a different domain or a general purpose text, e.g. a newspaper article (O'Brien 2009: 262). Specialized terminology for an expert in the field is generally not problematic (i.e. routine lexicon) but it may become so if his/her area of specialization differs (i.e. non-routine lexicon).

The notion of problem tends to become a highly subjective concept, increasingly linked to acquired knowledge as well as the accessibility of this knowledge. In the case of translation, there are at least two sets of (linguistic) notions that need to be accessed and coordinated, i.e. information about source and target. Eventually, a problem can be thought to occur whenever the translator fails to retrieve the relevant information about

the source (or the target) language and/or "coordinate" the retrieved pieces of monolingual information. Drawing a parallel with IR, at any given point in time, the translator possesses a Knowledge State (KS) which is a collection of Knowledge Items (KI)[204] that are used in the task at hand. At some point during the translation task, the translator cannot access a necessary KI and a problem (i.e. an information need) arises. The KS can change in time, i.e. a transition into a new KS occurs, for two different reasons: (i) by internal inference which takes place without a perception system (*inferential transition*) or (ii) by receiving information, i.e. the subject perceives something from the outside world leading to a new KS (*non-inferential transition*) (Mizzaro 1996: 235-236). A fairly direct comparison can be made between the two reasons behind a change of KS and the two main types of support available to a translator: internal and external support (see Section 4.1). The choice of either support is not necessarily mutually exclusive but there may be instances where external support backs a weak internal support, as may be the case with the categories of interaction between internal and external support in the classification by the PACTE group (2005: 615-616). This can happen when the translator theoretically knows both source and target items but for some reason is unable to retrieve the translation. In this case, internal support is insufficient and external support needs to be employed to eventually solve the problem. A related phenomenon is known in neurolinguistic studies with monolingual and bilingual subjects as tip-of-the-tongue (TOT) state.

7.3.3.1 THE TOT STATE

The tip-of-the-tongue (TOT) state is known as an occasional and temporary retrieval failure when the desired item is momentarily inaccessible to the subject but a feeling of imminent retrieval is nonetheless experienced[205].

> The speaker is certain that he knows a momentarily unavailable word (the target), feels close to recalling it and frequently has access to partial target attributes and/or related words (associates) during word search (Ecke 2009: 185).

As can be noted by the reference to "the speaker" in the quote, this refers to oral experiments generally conducted with monolingual and bilingual subjects[206] who are in some cases affected by aphasia or other brain lesions. The main focus of these studies is on verbal communications and retrieval is generally manifested through speaking. In this sense, parallels might be more easily drawn with interpreters while the question arises as to the relevance of these findings for written translation. According to Diamond and Shreve (2010: 291),

> the bilingual brain of a translator or interpreter might be expected to exhibit differences relative to that of a typical bilingual using both languages primarily for routine speech production and comprehension. Over time, with repeated practice in the performance of specific cross-language tasks, the nature and operation of the lexico-semantic system, and perhaps other cognitive structures, of a language mediation professional could be expected to be altered in response not only to

[204] Mizzaro (1996: 234) also lists possible alternatives for the concepts of KSs and KIs such as logical theories, semantic nets, sets of beliefs, situations, recursive models, minds and ideas.
[205] This experience is sometimes contrasted with the Feeling-of-Knowing (FOK) judgment, in which the subject feels that he will eventually be able to recall the item (Schwartz 2008: 9). Clearly, both judgments have many overlapping traits.
[206] In his review of existing studies on bilingual lexical retrieval, Ecke (2009: 186) actually reports that there is relatively little research on TOTs in bilinguals.

> developing language proficiency, but to cross-language task performance proficiency as well.

For interpreters, the emphasis is on switching between speech production and listening comprehension whereas for translators it lies between reading in one language and writing in another (Diamond & Shreve 2010: 291). According to Pyers *et al.* (2009: 324), one of the causes of a TOT is a phonological blocker, in that the bilingual subject has both SL/TL phonological representations always active and naturally thinks of the word in the other language. Irrespective of all phonological considerations, a written manifestation of a successful retrieval, i.e. the typing of the translation, could still be considered comparable to verbalizing the retrieved item. The moment when the translator looks up a string in a concordancer instead of typing the TL string, very likely signifies some degree of retrieval difficulty.

In practical terms, speakers experience a TOT when they fail to retrieve a word or name they are sure they know. Despite the TOT state, speakers may still be able to provide some information about the target word form (Kroll & De Groot 2005: 393). A brief overview of studies on TOT states highlighted some recurrent findings that can be relevant for the present analysis. First of all, evidence shows that bilinguals experience more TOTs than monolinguals (Ecke 2009: 185) and this suggests that the mechanism underlying TOTs is sensitive to the existence of more than one lexicon and/or phonological systems (Pyers *et al.* 2009: 323). Since translators can be said to have a bilingual brain and to work inter-linguistically, they can be expected to experience more TOTs than people using just one language in their work.

Second, Schwartz (2008: 11) reports that there seems to be a link between TOT state and working memory, in that they share the neurocognitive monitoring processes. Items involved in TOTs may also affect working memory. Working memory has previously been discussed in relation to translation problems and cognitive load (see Chapter 2), hence the interplay between TOT and working memory is worth investigating. In language production, there appear to be separate access stages for meaning and form because people are often able to retrieve alternative words that are either meaning- or form-related to the target word (Pyers *et al.* 2009: 323). In the case of TOT experiments, form should be rather read as "phonological encoding" because when this level fails completely or in part, semantic and syntactic information has generally been specified already and this is seen as evidence for a two-stage encoding process (Ecke 2009: 185). Pyers *et al.* (2009: 324) identify as the second cause of TOTs a "semantic blocker" because meaning-related alternate words are often involved in a TOT. In this view, TOTs are supposed to take place at an early stage of retrieval where the lexical representation is chosen. This probably happens in bilinguals, where translation equivalents could function as blockers because their meanings are almost entirely overlapping. With respect to the form level, the majority of TOTs seem to occur via a semantic blocking, i.e. at the earlier retrieval locus. This may be particularly relevant to explain the documented difficulties in translating e.g. terminology.

Third, experiments have shown that the type of target item to be retrieved can have an impact on the likelihood of experiencing a TOT state. In particular, proper nouns are known to be challenging for retrieval and are the most commonly reported type of TOT target (Kroll & De Groot 2005: 393-4). Word frequency is also taken to affect TOT states. In this particular case, however there are differences between TOTs experienced by monolinguals and bilinguals. Ecke (2009: 203) and Kroll and De Groot (2005: 393) report that monolinguals generally experience TOTs with (very) low-frequency words. However, bilinguals are found to also experience TOTs with words of relatively high-frequency

ranges as a consequence of reduced use of individual word forms (in both L_1 and L_2) compared to monolinguals. In particular, bilinguals are reportedly disadvantaged when retrieving nouns, verbs and adjectives but not so much when retrieving cognate word and proper names[207]. Frequency thus represents the third type of potential blocker after phonology and semantics (Pyers *et al.* 2009: 324). In this case, researchers do not assume a direct competition for selection between languages but rather attribute the blocking effect to a weak connection between meaning and form or rather incomplete activation of the target lexical representation (2009: 328). As bilinguals divide their language production between two languages (possibly even more, in the case of translators), they use each language less frequently than monolinguals. This is also known as the *weaker links account,* according to which increased TOT rates are due to relatively weak connections in the bilingual lexical system without considering effects of cross-language coactivation (Gollan & Acenas 2004: 264).

So far, the elements pertaining to the blocking hypothesis have been discussed. This hypothesis assumes that TOTs are caused by interfering competitors (usually more frequently or recently used) that are activated prior to the target item and subsequently block/inhibit its retrieval (Ecke 2009: 186). A different approach proposes the incomplete activation hypothesis, according to which there is an incomplete activation (retrieval) of the target (from a phonological perspective). In other words,

> the target word's meaning and syntactic attributes have been specified, but the activation of the target's phonology fails, due to weakened connections between the nodes of the two representational levels. [...] [M]eaning-form connections become weak with infrequent or nonrecent target word use and aging (Ecke 2009: 186).

Should it be possible to replace the phonological level with the target's form in a written representation, then this second hypothesis could be also considered applicable to a translation problem. Irrespective of the chosen hypothesis, findings seem to converge on the conclusion that switching between languages involves processing costs, both for translators and interpreters (Diamond & Shreve 2010: 308). If TOT states are considered applicable to translation (and interpreting), the experienced failures in the retrieval of the target expression may be considered a type of translation problem. If the translator feels s/he can retrieve the target item, the information need may be reduced to finding a simple cue to trigger the retrieval (e.g. the first syllable) and a translation support such as the concordancer can easily provide this type of help. Cues of this kind have been shown to assist recall in studies with bilinguals on TOTs and include the target's initial letter (or sound), the letters (or sounds) of its first syllable and sound-related words. In particular, words sharing the first syllable albeit differing in syntactic class have been proven particularly effective (Ecke 2009: 202).

Compared to a complete failure to produce the target word, a TOT should not be considered as a "serious" type of translation problem because it has a temporary nature and fragmentary information is nonetheless available to the translator. Directionality can however play a role in a TOT state as far as target retrieval is concerned. As explained by Ecke (2009: 202),

> [s]imilarities and differences in lexical retrieval have been suggested for stable lexis, usually of dominant languages or L1s, and unstable lexis, usually of non-

[207] The presence of proper names as a non disadvantageous retrieval target might be explained either by the fact that they are to be understood as proper names not requiring a translation or they are equally problematic for monolinguals and bilinguals to retrieve.

> dominant languages or L2s: TOTs with words of less stable L2s generate more sound-related associates and fewer meaning-related associates compared to TOTs of more stable L1s, suggesting that their resolution is more form-driven and less meaning-driven.

In this section, TOT states have been introduced as a minor translation problem, an information need which is related to some retrieval block from the bilingual mental lexicon that can however be easily solved with different forms of cueing. The concept of cueing has been addressed in the literature as *prompting* when discussing a very similar phenomenon to a TOT state.

> [The idea of prompting] is based on the following intuition. Even if we know what a word in a source language means, from time to time its translation equivalent in target language may not be available when we need it. Sometimes we can translate a word immediately; sometimes, however, we have to search in our memory for a longer time (for a whole set of different reasons) in order to find the correct translation equivalent. During this time we "look around" and, suddenly, we have an "aha" experience: the word we looked for has popped up. Prompting serves the function of alleviating the search for an appropriate word in target language under the assumption that translators already know the word in the source language and its meaning, but have the problem of finding a word in target language that matches its meaning (Stamenov 2009: 240-1).

In conclusion, language switching is a cognitively demanding activity that is an integral component of a translation task. However, translation very likely involves a number of other systems that are not necessarily language-specific. When discussing translation tasks and its component processes, one should take into account that translation is "a complex higher-order and problem-solving activity rather than a primarily linguistic one" (Diamond & Shreve 2010: 309).

7.3.4 IMPLICIT VS. EXPLICIT INFORMATION NEEDS

Translation as a problem-solving activity is by all means not a new concept in translation studies (see Section 2.6) but if the problem-solving part is replaced by the notion of information retrieval, a new light can be shed on the topic. As previously noted (see Section 4.4), information retrieval presupposes an initial information need that has to be satisfied. Problem solving involves taking action to obtain the desired result, which in this case means retrieving the missing information. TAP experiments in translation have shown that automated translation processes exist and are performed without conscious control by the (experienced) translators (Bernardini 2001: 249-250). Automaticity of processing can be interrupted or altered by non-routine task conditions that are likely to be noted and verbalized. This distinction may be brought a step further by adding the technological component of translation support. Process automatization could be revisited in the light of technological advances. The review of available forms of translation support in Chapter 3 showed how cognitively cheap current forms of external support can be.

During informal interviews with staff translators and tool developers at the EC, it was noted that there was some discrepancy between what people said they were doing with the concordancer (and were expected to do) and what they actually did. Officially, the tool was known to aid with terminological searches and document retrieval but when the actual search logs were examined together with the translator, s/he reported additional rationales, such as verifying a translation or getting help in choosing among alternative translations they had in mind or disambiguating a syntactic construction. There seem to

be instances where the translator only requires the concordancer for a quick check of a translation solution — a check that takes place almost automatically. When asked about his/her activity, the translator is more likely to report about major problems or noteworthy issues that were cognitively demanding, similarly to what happens with traditional verbal protocols.

In this sense, it seems justifiable to distinguish between *implicit* information needs and *explicit* information needs. The former are the ones that would hardly be labeled as problems but which also caused an interruption in the workflow, however minor, whereas the latter are the ones the translator is likely to report on because for some reason s/he has become aware of the problem (or the knowledge gap). Implicit information needs are automatized and represent and un-marked search activity. They may be tentatively considered as LGP problems or double checks where the translator does not perceive the item as unknown or problematic. On the other hand, explicit information needs are non-automatized and represent a marked type of search, i.e. an item that the translators would most likely label as (relatively) problematic. Researchers have generally focused on the explicit type of information need to develop problem categories, possibly because there were no systematic ways to log translators' activities comprehensively or because "minor" problems were not deemed sufficiently informative. Implicit needs, however, are just as important[208] and could be tapped to obtain information to fine tune and customize support tools to the actual needs of a translator. One possible example of an implicit information need could be Sinclair's "collocational frameworks" (or "grammatical frameworks"), i.e. patterning of grammatical words with its intermediate word or "collocate" (Renouf & Sinclair 1991: 129). Let us consider, for example, a given grammatical structure such as "/IN []* /IN" as extracted from the Euramis subset of 2- to 11-grams, where "IN" stands for any preposition[209] and "[]*" stands for any number of words, i.e. a structure corresponding to a query that will return a string of any length that begins and ends with a preposition. The most frequent combinations of this type are summarized in Table 50.

Table 50. Most frequent combinations for the "/IN [] /IN" pattern. Dataset: 605,000 strings between 2 and 11 words. Total number of strings matching the pattern: 2086). [DT=determiner, NN(S)=noun, VBN=past participle, IN=preposition].*

Pattern	Frequency Count	Examples
IN DT NN IN	544	for a term of, on the fringe of, in the face of, in the light of
IN NN IN	473	in relation with, in application of, by virtue of, in parallel with, in breach of
IN VBN IN	173	as regulated in, as adapted by, as stated at, as defined in
IN NNS IN	84	in terms of, by means of

Most of these patterns would probably fall in the LGP group of strings and would not be mentioned by the translator when asked about the problematic items found in the text. This is of course just a hypothesis on the perception of such strings by the translators. A fully-fledged experiment would be needed to verify that these searches can be used as examples of implicit information needs.

---

[208] In translation process studies, too, the equal importance of problems and non-problems was recognized by some scholars, e.g. Jääskeläinen (1993: 101), Jakobsen (1999: 15) and Enriquez Raído (2011: 49)

[209] To be precise, the POS-tag IN includes prepositions and subordinate conjunctions but excludes the preposition "TO".

The distinction between implicit and explicit information needs can be mapped onto different translation problems. An information need was previously (Section 4.4) defined as "the gap between people's current information and information sufficiency threshold[210]" (Lu & Yuan 2011: 134) or "the amount of information people feel they further need to adequately handle a given task [here: translation]" (Lu & Yuan 2011: 135). To satisfy the need, people would try to apply the principle of least effort when looking for the solution. In addition to this principle, information retrieval also resorts to the concept of *sufficiency,* which states that "people always strike a balance between minimizing their efforts and maximizing their decision confidence" (Lu & Yuan 2011: 135). The relative importance of one factor will take precedence over another and influence decision-making (e.g. quality and accessibility of the source), meaning that a given information need can be perceived as more pressing than another and the principle of least effort will be adjusted accordingly.

Sufficiency eventually depends on the information need, just as different types of information needs correspond to different effort levels. Lu and Yuan (2011: 135ff.) distinguish three levels of information needs: (i) high, (ii) medium and (iii) low. A high information need means that the information seeker has a high level of uncertainty and equivocation, limited knowledge about the subject and low confidence in judging the results. Therefore, precedence will be given to quality of the source over accessibility. In the case of the concordancer, this could mean checking the metadata or filtering by database or year, but it may also correspond to a search session, where more time is devoted to finding an answer or making crosschecks. Vice versa, a low information need implies that people "are very focused on the information-seeking process, meaning that they know exactly what they are looking for" (2011: 142). According to Lu and Yuan, only high-quality sources can serve the need of the information seeker but in the case of translation this may be debatable, as shown in the Contextual Inquiries with professional translators. If translators know exactly what information they are after, they should also be able to quickly evaluate results. In this case, accessibility might be equally (if not more) important as quality. A very generic information source could be easily tolerated as long as it is readily available, because people already have the means to judge the quality of information and are very clear about their search aim. The information seeker is focused on the information-seeking process and may already possess considerable knowledge about the topic. In the case of a concordance search, a low information need may simply mean that the translator already has a target language version available and simply wishes to double-check it, differently from source selection in "traditional" information seeking where both accessibility and quality were found to be critical in the case of low information needs (2011: 142). As a consequence, the concordance search will be more targeted, very likely in the form of a spot search. Unsuccessful searches are probably not tolerated and the translator may well give up the search after the first failed attempt because no additional time will be allotted for a low information need. This is where a low-quality quickly accessible source may still help, as opposed to a high-quality one that produces no results. Finally, medium information needs may be considered as one that strikes a balance between accessibility and quality of the source. In the case of a medium information need, a "good-enough" answer would suffice and the "best" one is not considered necessary (2011: 142).

---

[210] The information sufficiency threshold is understood as the level of (adequate) information individuals need in order to make decisions and which would curb further need for information for that task.

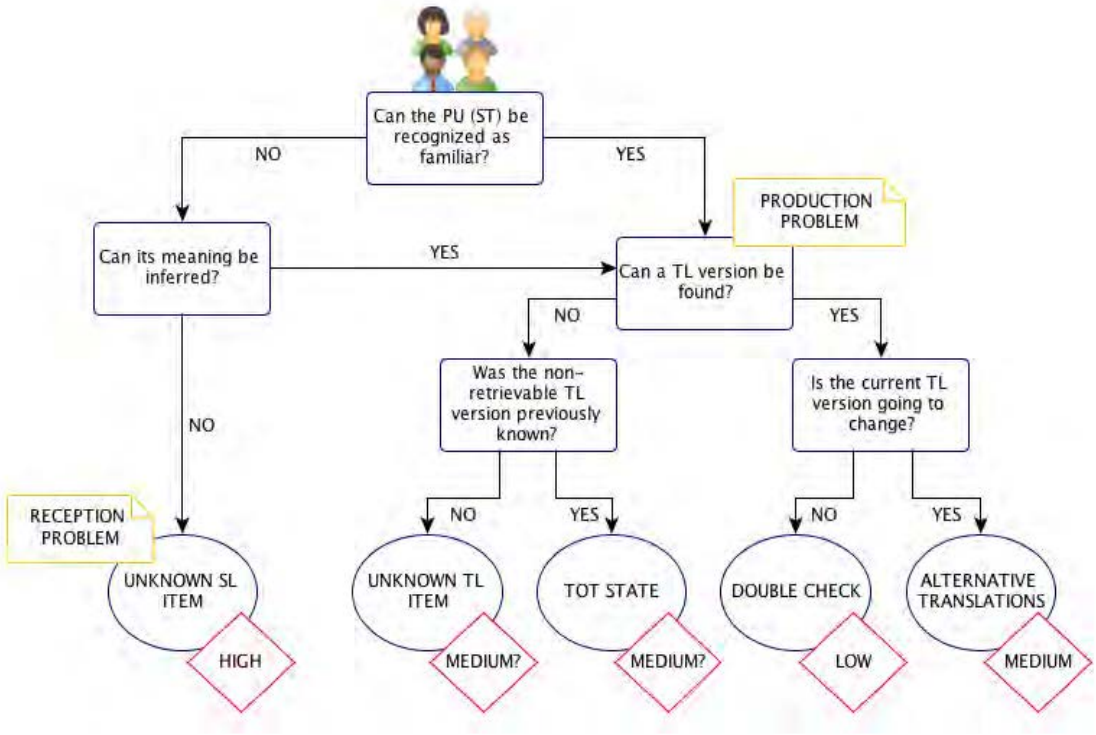
With this categorization in mind, the distinction between implicit and explicit information needs can be enriched with hypotheses about the type of information need involved. It would seem that in the case of explicit information needs, translators would remember and verbalize about the problem, and the information need would range from high to possibly medium, i.e. the cases where translators are still dependent on the source of information. On the other hand, implicit information needs can be though of ranging from medium to low, in that the translator may not even perceive them as problems because confidence is sufficiently high. Search behavior might be less goal-dependent than one may assume and more conditioned by the status of the information seeker with respect to the current information need. Eventually, it may all boil down to a trade-off between quality and accessibility of information, where accessibility could be also interpreted in terms of finding the appropriate balance between precision and recall. Lu and Yuan (2011: 142) conclude that "[…] individuals skillfully adjust their information-seeking strategies according to their information needs." Once again, concordance-searching behavior resembles in many respects Web searching and information retrieval but while Web searching can have a variety of purposes (see Section 4.5), concordance searching has always as one ultimate goal, i.e. the retrieval of a target language version for an SL element.

7.3.5 TRANSLATION PROBLEM SCENARIOS

The extensive discussion of the linguistic nature of translation problems in the previous sections ended by presenting problematic instances as subjective information needs. Information needs in turn can range from high to medium and low, according to a number of subject-dependent variables (e.g. expertise or domain knowledge). Using traditional problem categories for classification purposes is not very effective, because of labeling uncertainties and the difficulty in accommodating the newly introduced concept of implicit information needs. Implicit information needs have been associated with medium and low levels. The translator may not be aware of the translation problem because it is not perceived as such, particularly in the case of a double check. In a double check, the translator already has a target version available, which means that s/he was able to successfully retrieve a target from their bilingual lexicon.

A new perspective on translation problems as information needs can be seen taking shape, which centers on the cognitive resources of the translator. All concepts discussed in the previous sections can be put together in a diagram which offers a number of scenarios from which translation problems can be studied, all of them including both conscious and unconscious conditions (Figure 65).

Figure 65. Diagram summarizing the main possible scenarios behind a translation-related information need. The yellow boxes refer to the traditional dichotomy in TPR between reception and production problems whereas the red diamonds specify the hypothesized level of information need.



The actual translation problem categories are found in the circles at the end of the flow chart, together with the corresponding suggested level of information need. The preceding nodes encountered on a given path lead to a problem type and may be used to better target translation support according to the required level of information need.

Whenever a translator focuses on an text portion to be translated the brain also focuses on the relevant "Area of Interest" in the source text and determines whether the item is known or unknown ("Can the Problem Unit from the source text be recognized as familiar?"), where *known* item means that the SL text can be found in the mental lexicon or that it can at least be matched against some existing item. This becomes clear if a comment such as "*I've never heard this word before*" (PACTE 2011b: 334) is considered. The diagram however deliberately avoids the word "known" in favor of "familiar" because there can be instances where the item is in fact not known beforehand but it may be linked and/or matched against an existing one. The adjective "familiar" was chosen in connection to the phenomenon of *family resemblance* discussed by Wittgenstein (1953 in Rosch & Mervis 1975: 574-5).

> A family resemblance relationship consists of a set of items of the form AB, BC, CD, DE. That is, each item has at least one, and probably several, elements in common with one or more other items, but no, or few, elements are common to all items.

One example of familiarity may be cognate words, defined as pairs of words in the same or different languages that are similar in form and meaning and whose origin is shared (true cognates) or different (false cognates) (Stamenov 2009: 219-220). A more operational definition of cognates is found in Simard *et al.* (1993: 7):

> Cognates are pairs of words in different languages which, usually due to a common etymology, share phonological or orthographic properties as well as semantic properties, so that they are often employed as mutual translations.

The "familiar" word in the source language might not have been in the mental lexicon but if a cognate word was, chances are that the source language item will be perceived as known because it formally resembles the word in the lexicon.

Information retrieval from the lexicon is understood in terms of *lexical access,* which involves "matching a seen or heard word form with the associated word meaning stored in the lexicon" (Stamenov *et al.* 2010: 325). Lexical access occurs in two stages: lexical activation and lexical selection. In the phase of lexical activation, the stimulus triggers candidate words (with their corresponding meanings) according to the form of the stimulus word (phonological or visual) (2010: 325). Form seems to take slight precedence over meaning and the familiarity check might therefore be first performed at form level. Meaning then follows and the translator thus attempts to retrieve or deduce the meaning of the unit ("Can its meaning be inferred?"). If this step is not successful, neither form nor meaning can be retrieved and the item is therefore perceived as unknown. At this point, the translator will most likely turn to external support for understanding the SL element, which represents a reception problem.

> *Reception problems* occur when a subject has difficulties in receiving a source-language text segment, i.e. in the intake of information from an SL lexeme or combination of lexemes, and in the subsequent sense constitution (Lörscher 1991a: 95; emphasis in the original).

In this case, the translator relies exclusively on the external resources available to understand and translate the problematic item. Due to the high level of uncertainty and dependency on the external resource, this type of problem can be given the status of high information need. A possible example of the Problem Type "Unknown SL" from authentic interviews is the following (Karamanis *et al.* 2010):

> "*in most cases if there is a difficult term someone researches it and it goes to the TM, after the review it stays in the TM and this is the final decision about it, if I am a new translator and I come across this term I trust the TM*".

An even more explicit one can be found in PACTE (2011b: 334): "*I didn't know how to say it in X*". However, there can also be instances where the translator does not recognize the item at the formal level but is nonetheless able to guess its meaning, using e.g. the context and some intuition. In this case, the scenario converges into a different node that belongs to a different path.

Starting again from the top node in the diagram, this alternative path will be now explored. This time, the translator is able to recognize both form and meaning of the given item[211] and the question arises as to whether s/he is able to produce a target language version ("Can a TL version be found?"). This is taken to coincide with the phase where the translator supposedly retrieves the corresponding target version from the bilingual lexicon. The actual cognitive process is beyond the scope of this discussion and no distinction will be made between form and meaning. Suffice it to say that any further branching from this node falls in the scope of production problems because all subsequent nodes are related to the TL rendition of a source language item[212].

---

[211] In this case both levels are included in the concept of familiarity, somehow suggesting that there can be meaning without form but no form without meaning. In the case of the converging path, obviously, form is not recognized.

[212] This stage could be linked to the "retrieval strategies" and further sub-strategies (Krings 1986b: 270) that are linked to a recall problem. They refer to "a learner's conscious attempt to recall a known lexical item".

> *Production problems* occur when a subject has difficulties in finding a target-language text segment which s/he considers equivalent to the corresponding source-language text segment (Lörscher 1991a: 95; emphasis in the original).

At this point, the TL item is either retrievable or non-retrievable. If it is non-retrievable, a further check is necessary to be able to distinguish between two different information needs, i.e. (i) find a (new) translation for the SL concept and (ii) find a way to recover the non-retrievable TL from (long-term) memory ("Was the non-retrievable TL version previously known?"). If the translator feels s/he knows the target language version in question, the information need is reduced to retrieving the TL item from the bilingual mental lexicon. This scenario closely resembles the TOT state described earlier: the translator knows the item but simply cannot retrieve the target form (i.e. verbalize it in the TL)[213]. This can apply with virtually any type of string, from LGP to LSP, and may be affected by a number of factors such as frequency of use, tiredness, expertise and language combination.

However, there can also be instances where the translator understood the source text but was not able to retrieve or find the target language version because it never was in the mental lexicon. This is the case of an unknown TL item, possibly frequent when dealing with (technical) terminology or specialized phraseology. Solutions are likely obtained using forms of external support. Differently from reception problems, the translator now evaluates the source language side of a concordance search (much) faster and only has to spot the corresponding TL item in the translated version. Both scenarios have been tentatively labeled as medium level information needs. These instances were particularly hard to frame it into a single information need level because depending on a number of internal and external factors (e.g. time pressure, domain knowledge), accessibility and quality of the resource can easily take precedence one over the other.

Finally, going up one level, if the translator *can* find a TL version, it means s/he is able to retrieve or produce a translation for the SL item. However, this first translation proposal may not be the same as the one eventually used in the translation. This is why an additional node is necessary ("Is the current TL version going to change?"). A 'NO' answer means that the translator is quite sure about his/her choice and only needs the concordance search to double check the TL version, be it for consistency reasons, compliance with specific requirements or other reasons. In this case, the information need is low because the translator has a clear picture of what s/he is after and could virtually do without the external resource. Here is one example for the problem type "Double Check" extracted from a Contextual Inquiry (Karamanis *et al.* 2011: 40):

> *"In most cases the translator is not really stuck as in they don't have a clue about what a term means. I can easily find what 'stacking' means e.g. with a dictionary or online, so it would be more helpful for me to know what he [team leader] thinks or what my team agrees with, rather than starting a debate with a freelancer whom I have never worked with."*

If however the current TL version *is* going to change, it means that the translator is not completely satisfied with the initial translation (e.g. because of register, frequency of use, domain or style) and is hoping to find an alternative to the current TL version. A few

---

[213] A similar situation is reported in Krings when discussing "retrieval strategies" of language learners. This was the case when subjects had troubles recalling a lexical item that they had already learnt but "as soon as the term reappeared in their minds they considered the problem solved" (1986b: 270).

examples for problem type "Alternative Translations" can be obtained from the transcripts of retrospective protocols and Contextual Inquiry (Karamanis *et al.* 2010):

> *"I searched for some of the words even if I knew their meaning (...) I thought the dictionary definition might not work. So I translated "A Giant Stirs" as "Um gigante acorda" [A giant awakes] to give it a sense of rising, and not of moving"* (Buchweitz & Alves 2006: 258).
>
> *"so again 'arrangement' can have several meanings, I know they are talking about the pictures, again I do concordance [searches Concordance], in this case they are all the same but if this was not the same, if here it said something different from this, I would trust this user more [...]"*
>
> *"you see this, we did it ourselves, maybe the TM is inconsistent sometimes so in the case of 'device' there are a few ways to translate it, we decided to go for this among ourselves"*
>
> Similarly to the "stacking options" example [...], the problem was not that the translator needed help to find out what "device" meant but that he had to choose between the various ways in which this term had been translated in the past.

In this case, external resources become more relevant than in a simple double check, in particular the accessibility of the source, which is why this information need was labeled as "medium". These examples prove a point already brought forward by e.g. Jääskeläinen (1993: 106) and Sirén and Hakkarainen (2002: 77): problems do not necessarily involve serious or difficult processing to be solved and strategic behavior is related both to problems and to unproblematic decisions. In a real situation, translators are not asking themselves all these questions nor are they considering scenarios other than their current information need. The resource used to satisfy a given information need may be adopted as the prototypical resource for a given type of information need. The following can be seen as an example of Complex Problem Type "Unknown SL/TL" followed by a "Double Check" using mixed resources, i.e. online dictionary and Web search engine (Karamanis *et al.* 2010):

> the translator was challenged by the phrase 'dual throttle'. He searched an online dictionary which included 'forum discussions'. After looking at the meanings of the dictionary entries for 'throttle', he looked at the links in the forum:
>
> *"let's see 'full throttle' [clicks on 'full- throttle' link in the forum discussion] and someone says, this is another window, 'blabla martin Scorsese is filming in full throttle', I don't like this, I go back, 'full throttle' is about cinema mine is technical, [pause] this is by someone else [clicks on 'air-throttle valve' link] [reads in Spanish] so it's the valve of the accelerator, OK, so next we go to Google to check if it means what I assume that it means [searches Google], let's see how many, 'full throttle' [pause] this is proz, [pause] 'half throttle', 'half acceleration' so it's definitely acceleration so no further question."*

At a closer look, these categories of information needs have probably not turned into categories that can be directly operationalized. In fact, they have taken a user-centered perspective that can help to better understand the degree of support sought by different groups of translators. New forms of support could emerge from the understanding that e.g. professional translators mainly need to double-check a translation proposal whereas translation trainees require more support for unknown TL items. Findings from past studies support the user-centered approach that helps to identify general trends on the one hand and allows some room to account for individual preferences on the other. For

example, the fact that different subjects selected different TM matches[214] as being most helpful to solve a translation difficulty (Fifer 2007: 101 in Bowker & Barlow 2008: 17) may indicate that there were different information needs underlying a given translation difficulty. Different information needs as well as problem types might explain why

> [the quality of the solutions proposed] can be highly subjective, and translators often do not agree among themselves on what constitutes a quality solution to a given translation problem (Désilets, Farley *et al.* 2008: 8).

The different scenarios are presented once again in a more structured way in Table 51, where the most likely path across the flowchart is specified.

Table 51. Summary of problem scenarios with their most likely flowchart path. Items in brackets are theoretically possible but less likely to occur in practice.

Problem Level	Problem Type (Scenario)	Retrieval of SL Form	Retrieval of SL Meaning	Retrieval of TL Form	TL in Mental Lexicon	Change in TL Version	Info Need Level	Type of Info Need	Automatiz.
Reception	Unknown SL item	NO	NO	NO	NO	/	HIGH	Explicit	NO
Production	Unknown TL item	YES/NO	YES	NO	NO	/	MEDIUM?	Explicit	NO
Production	TOT State	YES (NO?)	YES	NO	YES	/	MEDIUM?	Implicit	YES
Production	Double Check	YES (NO?)	YES	YES	YES	NO	LOW	Implicit	YES
Production	Alternative Translations	YES (NO?)	YES	YES	YES	YES	MEDIUM	Explicit?	YES/NO

This change in perspective was deemed useful to try and account for the large number of variables that characterize a translation task and that affect the way in which translation problems and information needs are perceived and rationalized. One difficulty with general classifications of translation problems is that they tend to match similar searches (on the surface level) with a specific resource, e.g. a terminology database is generally linked to "terminology problems" (likely a high to medium information need) but in fact it can very well be used for a simple double-check (low information need). By doing so, they implicitly attribute the same importance to all problems of the same kind, which in fact is not necessarily the case, as the transcripts from the other studies have shown. This overview is only meant to be an initial theoretical model built 'bottom-up'; the model will need to be checked in a more systematic way against real scenarios. This user-centered approach may prove useful to effectively interact with the users when eliciting data rather than using traditional categories and concepts such as the labels described earlier in this chapter. Taking concordance searches, fixations or pauses as a starting point, the translator could be asked whether or not a given item was recognized as familiar and according to the answer provided, the next question on the path can be asked, as suggested in Figure 65, which eventually would lead to a problem category and a level of information need. A nominal string may thus be linked to a high information need if the source element is not recognized or understood by the subject but it can just as likely become a low information need in case of a double-check. This approach can help to further differentiate between novice and professional translators (the former more likely to experience high information needs, the latter medium to low) or simply gain further insight in the translation behavior and different translation styles.

[214] The 2nd ranked match was judged to be more useful than the highest ranked match in almost 40% of the instances (Fifer 2007: 103 in Bowker & Barlow 2008: 11).

7.4 KEY CONCEPTS

- Translation support forms can be related to string length on a continuum that goes from dictionary-style search to concordancing to document retrieval; as string length increases, translators are more likely to move from one to the other.
- The length of strings likely to be searched for mainly through concordancing was established to be between 2 and 11 words; this interval was arrived at considering type/token ratio, the collocational span, working memory and segmentation in translation.
- Nominal strings are by far the most frequent type of search; the sub-type "Named Entities" deserves special attention.
- The traditional concept of named entity should probably be adapted to account for a whole range of topical signatures for the Named Entities found in the corpus.
- Search strings are also affected by the "sparse data problem" which makes classification and clustering quite challenging.
- Short and frequent searches were considered as instances of core problem strings and examined in the context of longer strings and with a view to further clustering some searches triggered by the same information need.
- Content analysis attempted to distribute strings between the LSP (~60%) and LGP (~40%) problem groups, based on a list of descriptors.
- A finer-grained domain distribution highlighted European Communities, Law and Finance as the most popular EuroVoc fields.
- Various clustering approaches were applied to search strings from which it emerged that Slavic languages are the most active searchers in most domains.
- Reference to linguistic analysis of Web queries was found not to be useful for labeling concordance searches; POS tagging as a labeling and classifying method was tested.
- POS tagging confirmed findings of previous studies that nouns are by far the most represented lexical category in search strings and that the most frequent searches are for noun phrases. However, POS tagging could not be systematically employed and only provided approximate results.
- Instead of classifying isolated strings, search sessions were considered and the "delta string" between the first and the last query was classified from a syntactic perspective.
- Traditional categories of translation problems were reviewed and the existing literature for categories such as compounds, collocation, phraseology and terminology was further analyzed to check whether any label could be successfully applied to the present analysis.
- Eventually, "Multi-Word Unit" was chosen as general label for referring to the strings.
- Translation problems were then related to the structure and functioning of the bilingual mental lexicon.
- The Tip-of-the-Tongue state was described and presented as a possible cause for an information need.
- While translating, the translator faces both implicit and explicit information needs. However, translators are expected to consciously retain only the latter type and consequently the former have likely been under-examined in previous studies.
- Search behavior, just like translating, is made up of automatized and non-automatized patterns in response to implicit and explicit information needs.
- Translation problems can be classified according to the level of information need perceived by the user: high, medium and low.
- Five different scenarios of information need were suggested which account for the majority of translation problems experienced by translators.

CHAPTER 8: CONCLUSIONS

The main purpose of this study was to try and answer the three main research questions detailed in Section 1.1, namely:

How do translation problems vary across different language pairs? [RQ1]

What are the types of searches translators submit into a concordancer? [RQ2]

How do translators interact with a concordancer as a translation aid? [RQ3]

The concordancer was chosen as the object for the study because it is very often used by professional translators in their working practice and can be seen as "a source of external support since it has to be looked up by the translator and works differently from the standard solutions provided by the TM" (Alves & Liparini Campos 2009b: 201). One additional purpose was to present the concordancer as a valuable source of information on the translation process and suggest that greater attention should be devoted by the research community to this long-established tool. In particular, the concordancer might be a useful and unobtrusive way to elicit additional User Activity Data for triangulation purposes. By triangulating search logs with pauses and fixations, researchers may have a more objective basis to identify segmentation patterns and problem units by measuring pause length immediately before a search. Search logs could provide an empirical basis to identify the most appropriate cut-off length for pauses and possibly determine whether their length should vary according to other parameters such as expertise or text difficulty.

Before addressing the main research questions, concordance searches were presented as manifestations of translation problems and included in the theoretical framework used for process research as an additional data type for triangulation. To access the concordancer and submit a search, the translator has first to interrupt the ongoing task and leave the translating environment. This interruption in TT production can be explained in terms of an information need that the translator is not able to satisfy just by using his/her internal support, i.e. cognitive resources. A concordance search contains a source text item that triggered the use of external support and, just like other types of problem indicators, can be used to try and isolate Problem Units (PUs).

The relationship between PUs and other kinds of units in translation was explored at the end of the literature review in Chapter 2 where Problem Units (and Translation Problems) were presented as a special kind of Attention Units (or Translation Units), which in turn could be thought of as manifestations of underlying Cognitive Units. This hierarchy can be easily applied to instances where the "unit" can be clearly established, as is the case with the segmentation spontaneously produced by consulting various forms of translation support. For other types of problems that cannot be broken down as easily into units, such as cohesion problems, hardly any non-relational form of external support would come in useful. Theoretical considerations made at sentence and sub-sentence level cannot be generalized to *all* possible instances of problems, even though they can be expected to cover the vast majority of problems as "[m]ost instances of external support involve web searches or dictionary look-ups to find translation alternatives for specific terms" (Alves & Liparini Campos 2009b: 203). This sub-segment dimension (as opposed to problems at higher levels) seems justified by Levelt's model of language processing (1989, in Campbell 1999: 37), according to which lexis takes precedence in language

processing over syntax so that difficulties arise if no lemma exists in the mental lexicon or if a retrieval failure occurs.

8.1 [RQ1] THE LANGUAGE PAIR

The language pair was initially chosen as the main independent variable and was used whenever possible in the analysis. The first research question assumed that variability existed across language pairs, particularly at the level of language families. This possibility was discussed in greater detail in Section 5.5.2, where the distribution of languages across families was studied. The 20 target languages in the dataset are distributed rather unevenly across language families, ranging from a maximum of five languages to a minimum of one language per group. Such uneven distribution prevented direct comparisons of the results across language families. In addition, taking into account their relative age as official EU language, some language families fall entirely either in the group of the "new" languages or that of "old" languages one which turned out to be the only level at which differences could be systematically highlighted. Because of this overlap, it was not possible to clearly establish whether the differences were due to the language family categorization or rather to the relative age of the languages, which was not a linguistic parameter. For this reason, the initial hypothesis assuming differences across language families could not be verified and consequently the research question about differences across languages is still pending.

In the course of the study (see Section 5.6), three new levels of analysis were identified (language sub-set, search sessions and spot searches), which were expected to increase the chances of finding differences among languages. Irrespective of the chosen level of analysis, findings were rather balanced across languages and interestingly, most languages tended to behave consistently at each level so that small differences were generally maintained throughout the whole dataset. The quantitative analyses in Chapters 5 and 6, in particular, highlighted Bulgarian as the possibly sole outlier and confirmed the chronological criterion for clustering as the only criterion that provided some kind of consistent differences across languages.

As for the analysis of the Problem Unit in Chapter 7, methodological limitations prevented a systematic and quantitative analysis of the problems, which means that no comprehensive data for each language could be obtained, as opposed to the analysis in Chapter 6. The initial frequency lists that were used as resources to perform a quantitative study did not prove effective in the end. The main problem in this respect was the difficulty in finding suitable operational definitions for the categories needed to perform any linguistic analysis.

In sum, the answer to the first research question is that much fewer differences than expected emerged from the comparison of the language pairs, at least in terms of search strategies. Based on the quantitative findings it could be hypothesized that not many differences across languages should be expected from a linguistic perspective, though this statement could not be verified empirically. Overall, global factors seem to prevail over language-specific elements. These partial results seem nonetheless to justify generalizations from existing empirical studies covering a limited number of language pairs and in particular those reviewed in Chapter 2. For example, as regards text difficulty, "the source text can be an independent source of translation difficulty and [...] a substantial proportion of the items can be equally difficult to translate into typologically different languages" (Campbell 1999: 33).

8.2 [RQ2] THE PROBLEM UNIT

The second research question was first addressed indirectly in Chapter 2, where the concept of Problem Unit was discussed from a theoretical viewpoint and related to the concept of unit of translation.

The Problem Unit, i.e. the actual searched-for string, was considered one of the two main components of a concordance search and was further broken down into string length, string content and linguistic form to account for the three main perspectives from which a text string could be analyzed. The Problem Unit was then reduced to include strings of up to 11 words in an attempt to distinguish between concordance searches that could contain a problem of some kind and longer searches that resembled a Translation Memory search approach. In this perspective, three main approaches for using the concordancer were found, which correspond to three different "types of searches": dictionary-style search in the case of single-word searches, concordance proper for the middle range (2-11 words) and TM approach or document retrieval in the case of much longer strings.

Concordance searches were categorized based on whether or not the searched-for string could be linked to a specific topic. This classification was obtained by performing a semi-automatic categorization, which produced a content-oriented distribution of searches between LGP and LSP and a finer-grained classification of strings into more specific LSP domains. Within the LSP group, more precise results emerged about the most frequent, likely and/or problematic domains, i.e. European Communities, politics, law and finance. Results were not necessarily straightforward because higher frequencies could intuitively mean that strings belonging to some domains (e.g. law and finance) were perceived as more difficult (problematic) while, on the other hand, high frequency counts in the domain "European Communities" could rather be explained by the greater likelihood for EU-related items to appear in a string than represent a translation problem.

Some domains seemed to contain a high number of Named Entities (NE), which was a quite common category both within the range of "proper" concordancing and in the single-word subgroup, where they featured as acronyms. Named Entities were analyzed as a separate category because of their peculiarity from a translation perspective. However, the definition of NE was partially reworked to broaden its scope and include a range of "non standard" NE according to the general understanding of the concept in the field of Natural Language Processing. In quantitative terms, Named Entities — in their spelled-out form as well as acronyms — were quite high up in the frequency lists, which can be explained in a number of ways. First of all, Named Entities can be seen as an intrinsically problematic element that translators feel the need to look up very frequently, in line to some findings of neurolinguistic studies where proper nouns were found to be challenging for retrieval (see Sub-section 7.3.3.1). Should that be the case, target text variation as an indicator of problems and/or difficulties as suggested by Campbell (1999, 2000) and Dragsted (2012) should be partially revised because NE often translate with virtually 1:1 correspondence between source and target text. Another explanation could however be that the concordancer is considered a particularly well-suited resource for this type of search (at EU level). In any case, what these results certainly suggest is that there is room for some *automatized* form of support for the translation of Named Entities, given that the concordancer is accessed *manually* and NEs can be considered as a special kind of search string.

The most challenging part of the analysis, however, turned out to be the categorization of strings from a linguistic perspective. In order to process such large dataset, a clear-cut

operational categorization had to be found. However, existing taxonomies and categorizations were not suited to the task because they are generally not explicitly defined and could not be replicated. Definitions of frequently recurring concepts in existing studies were examined but after a short review of the literature, the conclusion was reached that the only operational category to be used was the "Multi-Word Unit", which however is not particularly informative in terms of the initial understanding of the type of search. During this stage, no quantitative analysis of the strings could be performed systematically. More sophisticated computational approaches would probably provide some solutions to this challenge which seems rooted in both a theoretical problem of definitions and agreement among scholars and the peculiar nature of language as an object of study through computational methods.

Nonetheless, some smaller-scale analyses have established that the most frequent type of search involves nouns and noun phrases and that it would be possible to isolate a core search string and analyze its distribution in the dataset so as to obtain aggregated frequency counts from any level of analysis. Frequency counts turned out to be a tempting and easily available resource. However, they can also be dangerous on a large-scale analysis because comparisons can be easily skewed and ultimately "the corpus frequency of words may not necessarily correlate with the accessibility of items in the mental lexicon" (Campbell 1999: 56) nor with the concept of "familiarity". Frequency counts do not necessarily relate to difficulty nor can the same source-text item be always ascribed to the same problem category, suggesting that there should be another way of looking at concordance searches.

A concordance search as a Problem Unit can be seen as a special kind of Translation Unit (see Section 2.9). In this perspective, all concordance searches would become Problem Units because the search has been triggered by a knowledge gap of some kind. However, there cannot be a 1:1 correspondence between a Problem Unit and the underlying "problem" because of the dynamic dimension of the searches (i.e. the search sessions), where multiple searches within a search episode are in fact triggered by the same information need. In this sense, Problem Units within a search session would rather represent different realizations of one underlying problem. Overall, they tend to be nominal, are rather balanced between LGP and LSP strings and a sub-category of Named Entities can be quite easily singled out. Ultimately, the Problem Unit can be useful for operationalizing the manifestation of a "problem" by applying clear boundaries to a source text string but it can be hardly used as an absolute (static) concept.

8.3 [RQ3] THE SEARCH STRATEGY

The third research question was aimed at studying users' behavior in terms of interactions with the concordancer at the level of both search sessions and spot searches. User behavior is here understood in terms of a search strategy for translation, searching or problem-solving purposes.

The study of the search strategy represented the quantitative part of analysis, in that search logs were examined by systematically looking at the average string length, the distribution of metadata and any other additional search filter used for each of the 20 languages. Additional differences between languages were sometimes highlighted by combining different parameters, for example looking at patterns of (un)successful searches (see Sub-section 6.2.1.2). Overall, results were once again quite close across languages, suggesting that global search strategies prevailed over language-specific strategies (e.g. those that emerged for Bulgarian). Knowing that global strategies seem to

take precedence over language-specific ones suggests that looking at general search trends is enough to obtain data about the search habits of translators and that language customization is not necessarily a priority.

Search strings were quite short, generally between two and three words per query, similarly to what was found for Web searching. Simlar to Web searching, users were not particularly interested in using advanced settings and filters and when they did, they opted for the Year filter to rank results according to a chronological perspective. Results for tool settings are relevant for tool developers because they show what pieces of information translators perceive as most relevant and useful, thereby providing developers with first-hand information about the features that translators would like to see implemented in a tool.

Translators clearly favored a pooled form of translation support (Quest) over the standalone concordance tool (Euramis only), suggesting that a combination of different types of translation resources is their preferred approach to problem-solving. Consequently, there does not seem to be great interest in adjusting the search strategy to better interact with a specific resource. Rather, a larger pool of translation resources increases the chances of obtaining at least one useful and usable solution, which is the main intent behind a search as already pointed out by Sharoff *et al.* (2006: 745). This type of interaction is a clear sign of the prominence of recall over precision in concordance searching.

In terms of the overall approach to searching, translators were found to operate more frequently at the string level than with the search settings. When it comes to search sessions and query refinements, they are more likely to shorten (or expand) the string rather than change search settings to adjust recall and precision. In particular, strings in a session were found to be generally longer at first and then progressively shortened either from the left or right end. All operations identified at string level were classified in a taxonomy of reformulation strategies. In addition to the reduction strategy, other five main search strategies were identified, namely resubmission, formal changes, expansion, replacement and mixed strategy. Each was further broken down into sub-categories, though the vast majority was hardly represented in the dataset beside some specific sub-categories of expansion, reduction and resubmission.

Based on these results, a relationship between Translation Units and Problem Units could be hypothesized in terms of length. The underlying assumption is that PUs will hardly ever be longer than TUs and most likely have the same size. Statistics about search sessions highlighted that the main trend for the strings is to become shorter rather than longer in the course of a search session. If the search session is considered as a refinement of the query to make it more targeted for the actual problem, then Translation Units can be identified with the initial query whereas the Problem Unit can be identified with the shorter text string (vice versa in the case of expansion). If however the session was a result of a first unsuccessful search, then it becomes harder to establish that the PU was contained in a larger TU because an unsuccessful search necessarily involves a change in the strategy.

8.4 INFORMATION NEEDS

The analysis of the Problem Unit in Chapter 7 showed that a different perspective had to be taken in order to attempt some kind of classification of translation problems. At a closer look, evidence from existing studies revealed that problems "do not necessarily refer to something serious or difficult" (Sirén & Hakkarainen 2002: 77). Evidence from

TAPs and interviews highlighted that often translators are not struggling with a source text expression but only seek confirmation and use translation support to take "unproblematic decisions" (Jääskeläinen 1993: 106). When talking about the translation process, translators themselves are likely to gloss over unproblematic decisions, provided they are aware of them in the first place. Therefore, researchers have traditionally focused on problematic instances rather than considering the overall picture. If a concordance search is indeed a problem indicator, then *all* concordance searches should be a manifestation of some kind of "problem". Translation problems were presented as a manifestation of knowledge gaps associated with a corresponding information need. By resolving the related information need, the translation problem is also automatically addressed. This is the main reason why the label "translation problem" has been eventually complemented in this study by "information need", which can better account for those instances that translators themselves would not call "problematic" in the first place because the information need is likely to be "low" (as opposed to "high" or "medium"). Conversely, translators possibly seem more aware of "problematic" items because, by addressing a high information need, they eventually increase their knowledge base with the solution to that problem in terms of e.g. adding a new concept or a new translation to their knowledge base ("What did I learn with this translation?"). Lower information needs may be linked to elements already present in the translator's cognitive system that tend to be processed automatically without the translator necessarily becoming aware of it or explicitly verbalize about it:

> [...] the interview data of this study shows that some of the participants [...] omitted to report the information needs that they considered unproblematic. The more general nature of the unreported needs in fact seems to support the assumption that the participants tended not to report on searches that involved unproblematic processing and that primarily aimed at confirming preexisting solutions (Enríquez Raído 2011: 486).

Findings have shown that problems encountered by translators do not differ considerably, at least on the surface level: "all participants had a similar number of more or less problematic items (generally associated with technical terms) in both tasks" (Enríquez Raído 2011: 486) but nonetheless the underlying level of information need may still differ, which would make similar searches become in fact different information needs.

Prior work in Information Retrieval and the notion of information need were used to develop a diagram (Section 7.3.5) that begins with the identification of an element that cannot be processed as "automatically" as others during a translation task. Through a number of possible paths with binary choices (yes/no) at each node, five main categories of information needs have been proposed according to the type of underlying issue (Unknown SL item; Unknown TL item; Tip-of-the-Tongue State; Double Check; Alternative Translation). Each category has been assigned to a level of information need which affects the quality and accessibility of the source perceived by the user. The five types of information needs could be broadly grouped into two macro-categories of problems in the traditional sense: reception problems, if no lemma exists in the translator's mental lexicon, or production problems, if it is a matter of retrieval failure. Krings' third mixed category (comprehension and production problems) could potentially also exist because in the case of an unknown source language item, chances are that the translator would not know the target language version either. However, this instance would probably represent a sequence of information needs ("first understand the SL and then worry about the TL") rather than two concurrent ones. Krings' third category had already been rejected by Jääskeläinen (1987: 36) because

> a problem's nature changes so considerably when a comprehension problem turns out to be a production problem as well, that there are really two separate problems connected with the same item in the source text, rather than one combined problem.

The diagram is very much subject-oriented because it was felt that the intra-subjective perspective on difficulties was a very important element when dealing with information needs, as shown by experiments where great variability was generally found among participants because "not all translators and interpreters find the same items difficult" (Séguinot 2000: 145). Similarly, Enríquez Raído (2011: 486) noted that

> [i]n contrast to my original position, in which I envisaged information seeking as motivated by the need to solve a (translation) problem [...], the results of this study appear to support Case's statement that information seeking is sometimes motivated by 'a desire to simply have more or less of some quality; more information; stimulation, or assurance; or less uncertainty, boredom overload, or anxiety' (2008: 88).

With these categories of information needs, classifying searches becomes a dynamic task (the same search can fall into different categories) as opposed to the more "static" (albeit sometimes blurry) traditional classifications of translation problems at sentence and sub-sentence level.

8.5 FUTURE AVENUES OF RESEARCH

If deemed appropriate, the diagram arrived at in Chapter 7 could be easily transposed into questions for retrospective interviews or written questionnaires so that the participants themselves can point out the type of information need, and this information can then be used to verify quantitative experimental data. Furthermore, the questions can be asked to virtually anyone, from language learners to translators with various levels of expertise, and it would be interesting to test the level of awareness that translators have about the medium to low types of information needs. In this way, some answers could be provided to some pending questions, for example,

> [w]ill it be possible to figure out whether a translation pause is due to unknown terminology [...], or for instance, [...] whether it is due to a more complicated understanding (and translating) problems? (Carl 2009a)

The usefulness of a translation tool is closely linked to its ability to effectively respond to users' needs. For concordancing tools, this may be more challenging given the broad spectrum of searches that are submitted to the system. Unfortunately, information needs cannot be easily told apart by simply looking at the searched-for item and, as previously suggested, different displays of results may be necessary for the same searched string as different information needs can underlie the same query (e.g. chronological display or frequency-based ranking). Display criteria can also include similarity of context and identity in the TM metadata in addition to the quality and accessibility of the resource. The challenge is to define behavioral indicators to identify the correct level of information need and develop criteria to assign a specific display of results to a given information need (e.g. a high information need would possibly prioritize context whereas frequency would work better for a medium information need). The way results are displayed has a direct impact on the usefulness and effectiveness of the search. The tool may even find the required information but if it is not able to display it in an adequate way for the user, the search might still prove unsuccessful. The search rationale can vary greatly from search to search and from user to user and the architecture of each tool has different features and

capabilities. Data type, data storage and retrieval mechanisms are also critical elements in the performance of the tool. Results from this study on professional translators (albeit from a quite specialized domain) could be taken as a baseline for future studies that analyze search logs from other (commerical) standalone concordancers to study translators' behavior in different domains (e.g. localization) as well as the interaction between non-professional translators (i.e. general Internet users) and the concordancer so as to better provide better support to the different user groups.

A customized tool that is able to guess the information need and the appropriate way to display results would require a trial phase where different solutions (results displays) are offered to the users who can manually select the one they find most useful. Once enough data have been collected, tool architecture can be appropriately fine-tuned. With a considerable amount of field studies, information needs could even be related to a specific type of resource (e.g. a bilingual dictionary used for a high information need due to an unknown SL item) so that each resource would be optimized for the most frequent type of search performed. Given the increasing tendency to combine resources in the same tool, such differentiation may be unrealistic but it could still be interesting to study which criteria should be considered and followed when presenting translators with search results.

Standalone concordancing tools are not yet systematically used (particularly outside the EU) and it would be hard to collect enough data. Using an unfamiliar concordancing tool in experiments may also invalidate the results. A possibly easy way to systematically collect data on concordance searches would be to add a logging feature in a traditional TM system that keeps track of concordance searches even in off-line working mode, as has already been attempted with OmegaT.

On a methodological note, before engaging in large-scale studies using concordancers, a few experiments should be carried out to verify the usability of search logs in the field of translation process research and, in particular, to determine how they can best be used in conjunction with key-logging and eye-tracking.

REFERENCES

Abdi H. (2010) Coefficient of Variation. In *Encyclopedia of Research Design*, edited by N. Salkind. Thousand Oaks, CA: Sage, 169-171.

Agar M. (1991) The Biculture in Bilingual. *Language in Society* 20:167-181.

Alcina A. (2008) Translation Technologies. Scope, Tools and Resources. *Target* 20 (1):79-102.

Alves F. (1997) A Formação de Tradutores a Partir de uma Abordagem Cognitiva: Reflexões de um Projeto de Ensino. *TradTerm* 4 (2):19-40.

Alves F., ed. (2003) *Triangulating Translation: Perspective in Process Oriented Research.* Amsterdam/Philadelphia: John Benjamins.

Alves F. (2006) Unidades de Tradução. O que São e Como Operá-las. In *Traduzir com Autonomia: Estratégias para o Tradutor em Formação*, edited by F. Alves, C. Magalhães and A. Pagano. São Paulo: Contexto, 29-38.

Alves F. & Couto Vale D. (2009) Probing the Unit of Translation in Time: Aspects of the Design and Development of a Web Application for Storing, Annotating, and Querying Translation Process Data. *Across Languages and Cultures* 10 (2):251-273.

Alves F. & Gonçalves J. L. V. R. (2003) A Relevance Theory Approach to the Investigation of Inferential Processes in Translation. In *Triangulating Translation*, edited by F. Alves. Amsterdam/Philadelphia: John Benjamins, 3-24.

Alves F. & Liparini Campos T. (2009a) Chains of Cognitive Implication in Orientation and Revision During the Translation Process. *Current Issues in Language Studies* 1:75-95.

Alves F. & Liparini Campos T. (2009b) Translation Technology in Time: Investigating the Impact of Translation Memory Systems and Time Pressure on Types of Internal and External Support. In *Behind the Mind. Methods, Models and Results in Translation Process Research*, edited by S. Göpferich, A. L. Jakobsen and I. M. Mees. Copenhagen: Samfundslitteratur, 191-218.

Alves F., Pagano A., Neumann S., et al. (2010) Translation Units and Grammatical Shifts: Towards an Integration of Product- and Process-Based Translation Research. In *Translation and Cognition*, edited by G. M. Shreve and E. Angelone. Amsterdam/Philadelphia: John Benjamins, 109-142.

Angelone E. (2010) Uncertainty, Uncertainty Management and Metacognitive Problem Solving in the Translation Task. In *Translation and Cognition*, edited by G. M. Shreve and E. Angelone. Amsterdam/Philadelphia: John Benjamins, 17-40.

Angelone E. & Shreve G. M. (2011) Uncertainty Management, Metacognitive Bundling in Problem Solving, and Translation Quality. In *Cognitive Explorations of Translation*, edited by S. O' Brien. London/New York: Continuum, 108-130.

Anthony L. (2004) AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. *Proceedings of IWLeL 2002 - An Interactive Workshop on Language e-Learning*, 7-13.

Arampatzis A. & Jaap K. (2008) A Study of Query Length. *Proceedings of SIGIR'08.*

Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 811-812.

Austermühl F. (2001) *Electronic Tools for Translators*. Manchester: St. Jerome Publishing.

Azzopardi L. (2009) Query Side Evaluation. *Proceedings of SIGIR'09. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. New York: ACM, 556-563.

Baeza-Yates R., Calderón-Benavides L. & Gonzáles-Caro C. (2006) The Intention Behind Web Queries. *Proceedings of SPIRE 2006. Proceedings of the 13th international conference on String Processing and Information Retrieval*, Glasgow. Berlin/Heidelberg: Springer, 98-109.

Barkhudarov L. (1993) The Problem of the Unit of Translation. In *Translation as a Social Action. Russian and Bulgarian Perspectives*, edited by P. Zlateva. London/New York: Routledge, 39-46.

Barlow M. (1996) Analysing Parallel Texts with Paraconc. *Proceedings of Joint International Conference ALLC-ACH*, University of Bergen, Norway, 25-29 June, 25-27. Available from http://gandalf.aksis.uib.no/allc/barlow.pdf

Barlow M. (2002) ParaConc: Concordance Software for Multilingual Parallel Corpora. *Proceedings of 3rd LREC Conference*, Las Palmas, Spain, 27-28 May, 20-24. Available from http://www.athel.com/paraweb.pdf

Barlow M. (2004) Parallel Concordancing and Translation. *Proceedings of ASLIB Translating and the Computer 26*, London, 18-19 November. Available from http://www.mt-archive.info/Aslib-2004-Barlow.pdf

Barr C., Rosie J. & Regelson M. (2008) The Linguistic Structure of English Web-Search Queries. *Proceedings of EMNLP '08 - Conference on Empirical Methods in Natural Language Processing*, ACL, 1021-1030. Available from http://www.cs.cmu.edu/afs/cs/user/rosie/www/papers/emnlp2008.pos.pdf

Bassnett-McGuire S. (1991) *Translation Studies*. Revised Edition ed. London/New York: Routledge.

Batista B. (2007) *O Impacto dos Sistemas de Memória e Tradução nos Processos de Revisão de Tradutores Profissionais Brasileiros*. Federal University of Minas Gerais, Belo Horizonte, Brazil. Unpublished MA thesis.

Beitzel S. M., Jensen E. C., Chowdhury A., et al. (2007) Temporal Analysis of a Very Large Topically Categorized Web Query Log. *Journal of the American Society for Information Science and Technology* 58 (2):166-178.

Bell R. T. (1991) *Translation and Translating: Theory and Practice*. London: Longman.

Benito D. (2009) Future Trends in Translation Memory. *Revista Tradumàtica - Traducció i Tecnologies de la informació i la Comunicació* 7. Available from http://ddd.uab.cat/pub/tradumatica/15787559n7a7.pdf

Bennet P. (1994) The Translation Unit in Human and Machine. *Babel* 40 (1):12-20.

Berg T. (2006) The Internal Structure of Four-Noun Compounds in English and German. *Corpus Linguistics and Linguistic Theory* 2 (2):197-231.

Bernardini S. (2001) Think-Aloud Protocols in Translation Research. Achievements, Limits, Future Prospects. *Target* 13 (2):241-263.

Blatt A. (1998a) Euramis Alignment and Translation Memory Technology. *Terminologie et*

Traduction 1998 (1):74-114.

Blatt A. (1998b) Euramis: Added Value by Integration. *Terminologie et Traduction* 1998 (1):59-73.

Borlund P. (2003) The Concept of Relevance in Ir. *Journal of the American Society for Information Science and Technology* 54 (10):913-925.

Bourdaillet J., Huet S., Gotti F., et al. (2009) Enhancing the Bilingual Concordancer TransSearch with Word-Level Alignment. *Proceedings of 22nd Canadian Conference on Artificial Intelligence 2009*, Kelowna, Canada, 25-27 May, 27-38. Available from http://www.iro.umontreal.ca/~felipe/bib2webV0.81/cv/papers/paper-cai-2009.pdf

Bourdaillet J., Huet S., Langlais P., et al. (2010) TransSearch: From a Bilingual Concordancer to a Translation Finder. *Machine Translation,* 24 (3-4):1-36.

Bowker L. (2002) *Computer-Aided Translation Technology. A Practical Introduction*. Ottawa: University of Ottawa Press.

Bowker L. & Barlow M. (2004) Bilingual Concordancers and Translation Memories: A Comparative Evaluation. *2nd International Workshop on Language Resources for Translation Work, Research and Training*, Stroudsburg: ACL, 70-83.

Bowker L. & Barlow M. (2008) A Comparative Evaluation of Bilingual Concordancers and Translation Memory Systems. In *Topics in Language Resources for Translation and Localisation*, edited by E. Y. Rodrigo. Amsterdam/Philadelphia: John Benjamins, 1-22.

Bowker L. & Pearson J. (2002) *Working with Specialized Language: A Practical Guide to Using Corpora*. New York: Routledge.

Broder A. (2002) A Taxonomy of Web Search. *SIGIR Forum* 36 (2):3-10.

Buchweitz A. & Alves F. (2006) Cognitive Adaptation in Translation: An Interface between Language Direction, Time, and Recursiveness in Target Text Production. *Letras de Hoje* 41 (2):241-272.

Bussmann H. (1996) *Routledge Dictionary of Language and Linguistics*. London/New York: Routledge.

Cabré Castellví M. T. (2003) Theories of Terminology. Their Description, Prescription and Explanation. *Terminology* 9 (2):163-199.

Callison-Burch C., Bannard C. & Schroeder J. (2004) Searchable Translation Memories. Paper read at Translating and the Computer (26), London, UK, 18-19 November. Available from http://www.cs.jhu.edu/~ccb/publications/searchable-translation-memories.pdf

Campbell S. (1999) A Cognitive Approach to Source Text Difficulty in Translation. *Target* 11 (1):33-63.

Campbell S. (2000) Choice Network Analysis in Translation Research. In *Intercultural Faultlines. Research Models in Translation Studies: Textual and Cognitive Aspects*, edited by M. Olohan. Manchester: St. Jerome, 29-42.

Canfora G. & Cerulo L. (2004) A Taxonomy of Information Retrieval Models and Tools. *Journal of Computing and Information Technology - CIT* 12 (3):175-194.

Carl M. (2009a) Grounding Translation Tools in Translator's Activity Data. Paper read at MT Summit XII, Ottawa, Canada, 26-30 August. Available from http://www.mt-

archive.info/MTS-2009-Carl.pdf

Carl M. (2009b) Triangulating Product and Process Data: Quantifying Alignment Units with Keystroke Data. In *Methodology, Technology and Innovation in Translation Process Research*: Copenhagen Studies in Language, 225-246.

Carl M. & Jakobsen A. L. (2009) Towards Statistical Modelling of Translators' Activity Data. *International Journal of Speech Technology* 12 (4):125-138.

Carl M., Jakobsen A. L. & Jensen K. T. H. (2008a) Studying Human Translation Behaviour with User-Activity Data. *Proceedings of NLPCS 2008*, Barcelona, Spain, 1-5 June, INSTICC Press, 114-123. Available from http://openarchive.cbs.dk/bitstream/handle/10398/8044/UAD-3.pdf?sequence=1

Carl M., Jakobsen A. L. & Jensen K. T. H. (2008b) Modelling Human Translator Behaviour with User-Activity Data. *Proceedings of 12th EAMT Conference*, Hamburg, Germany, 22-23 September, 21-26. Available from http://www.mt-archive.info/EAMT-2008-Carl.pdf

Case D. O. (2008) *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. 2nd ed. Bingley: Emerald.

Catford J. C. (1965) *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. London: Oxford University Press.

Chesterman A. (1997) *Memes of Translation: The Spread of Ideas in Translation Theory*. Amsterdam/Philadelphia: John Benjamins.

Chesterman A. (2011) Process Models and Their Assumptions. Paper read at Text-Process-Text Conference, Stockholm, Sweden, 17-19 November.

Chesterman A. & Wagner E. (2002) *Can Theory Help Translators? A Dialogue between the Ivory Tower and the Wordface*. Manchester: St. Jerome.

Chuang S.-L. & Chien L.-F. (2002) Towards Automatic Generation of Query Taxonomy: A Hierarchical Query Clustering Approach. *Proceedings of ICDM. IEEE International Conference on Data Mining*, Maebashi City, Japan, 9-12 December. Washington, DC: IEEE Computer Society, 75-82.

Cole C. (2011) A Theory of Information Need for Information Retrieval That Connects Information to Knowledge. *Journal of the American Society for Information Science and Technology* 62 (7):1216-1231.

Colominas C. (2008) Towards Chunk-Based Translation Memories. *Babel* 54 (4):343-354.

Corbin J. & Strauss A. (1990) Grounded Theory Research: Procedures, Canons and Evaluative Criteria. *Zeitschrift für Soziologie* 19 (6):418-427.

Corpas Pastor G. (2003) *Diez Años de Investigación en Fraseología: Análisis Sintáctico-Semánticos, Contrastivos y Traductológicos*. Frankfurt am Main/Madrid: Vervuert/Iberoamericana.

Cosmai D. (2007) *Tradurre Per L'unione Europea*. 2nd ed. Milano: Hoepli.

Danielsson P. (2003) Units of Meaning in Translation - How to Make Real Use of Corpus Evidence. Paper read at ASLIB Translating and the Computer 25, London, UK, 20-21 November. Available from http://mt-archive.info/Aslib-2003-Danielsson.pdf

Davies E. (2004) *Eurojargon: A Dictionary of European Union Acronyms, Abbreviations and Terminology*. 7th ed. Manchester: EIA.

Dayrell C. (2007) A Quantitative Approach to Compare Collocational Patterns in Translated and Non-Translated Texts. *International Journal of Corpus Linguistics* 12 (3):375-414.

Dechert H. W. & Sandrock U. (1986) Thinking-Aloud Protocols: The Decomposition of Language Processing. In *Experimental Approaches to Second Language Learning*, edited by V. Cook. Oxford/New York: Pergamon, 111-126.

Désilets A., Brunette L., Melançon C., et al. (2008) Reliable Innovation: A Tecchie's Travels in the Land of Translators. Paper read at 8th AMTA Conference, Hawaii, 21-25 October. Available from http://www.amtaweb.org/papers/4.11_Desiletsetal2008.pdf

Désilets A., Farley B., Stojanovic M., et al. (2008) Webitext: Building Large Heterogeneous Translation Memories from Parallel Web Content. Paper read at Translating and the Computer (30), London, UK, 23-28 November. Available from http://mt-archive.info/Aslib-2008-Desilets.pdf

Désilets A., Melançon C., Patenaude G., et al. (2009) How Translators Use Tools and Resources to Resolve Translation Problems: An Ethnographic Study. *Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, August 26-30, 2009. Available from http://www.mt-archive.info/MTS-2009-Desilets-2.pdf.

DGT (2009a) *Translation Tools and Workflow*. Luxembourg: Publication Office of the EU.

DGT (2009b) *La Traduction à la Commission: 1958-2010*. Vol. 2, *Études sur la Traduction et le Multilinguisme*. Luxembourg: Publication Office of the EU.

DGT (2009c) *Translating for a Multilingual Community*. Luxembourg: Publication Office of the EU.

DGT (2010a) *Lawmaking in the EU Multilingual Environment*, *Studies on Translation and Multilingualism*. Luxembourg: Publication Office of the EU.

DGT (2010b) Translating for the English Department of DGT. Presentation held at Heriot Watt University, Edinburgh, 10 February. Slides available from http://www.strath.ac.uk/media/ps/careers/occupationalinfo/HW_presentation_final[1].ppt (accessed September 2012).

DGT (2010c) Translation at the EU Institutions & Euramis. Paper read at Information Meeting on Euramis, Luxembourg, 12 November. Slides available from http://ec.europa.eu/dgs/translation/workwithus/calls/open/cattools/cattools_documentation_en.htm (accessed November 2010).

DGT (2012) English Style Guide. A Handbook for Authors and Translators in the European Commission. Available from http://ec.europa.eu/translation/english/guidelines/documents/styleguide_english_dgt_en.pdf (accessed October 2012).

Diamond B. J. & Shreve G. M. (2010) Neural and Physiological Correlates of Translation and Interpreting in the Bilingual Brain: Recent Perspectives. *Translation and Cognition,* edited by G. M. Shreve and E. Angelone. Amsterdam/Philadelphia: John Benjamins, 289-321.

Dias G., Guilloré S. & Pereira Lopes J. G. (1999) Multilingual Aspects of Multiword Lexical Units. Paper read at *Workshop Language Technologies Multilingual Aspects*, Ljubljana, Slovenia, 8-11 July, 11-21. Available from http://www.di.ubi.pt/~ddg/publications/sle1999.pdf

Dollerup C. (2001) The World's Largest Translation Institution. Language Work at the European Commission (Part 2). *Language International* 6 (13):31-40.

Dragsted B. (2004) *Segmentation in Translation and Translation Memory Systems*. Copenhagen Business School, Copenhagen. PhD dissertation.

Dragsted B. (2005) Segmentation in Translation. Differences across Levels of Expertise and Difficulty. *Target* 17 (1):49-70.

Dragsted B. (2010) Coordination of Reading and Writing Processes in Translation. In *Translation and Cognition*, edited by G. M. Shreve and E. Angelone. Amsterdam/Philadelphia: John Benjamins, 41-62.

Dragsted B. (2012) Indicators of Difficulty in Translation - Correlating Product and Process Data. *Across Languages and Cultures* 13 (1):81-98.

Dragsted B. & Hansen I. G. (2008) Comprehension and Production in Translation: A Pilot Study on Segmentation and the Coordination of Reading and Writing Processes. In *Looking at Eyes. Eye-Tracking Studies of Reading and Translation Processing*, edited by S. Göpferich, A. L. Jakobsen and I. M. Mees. Copenhagen: Samfundslitteratur, 9-29.

Drugan J. (2004) Multilingual Document Management and Workflow in the European Institutions. *Proceedings of Aslib Conference: Translating and the Computer 26*, London, 18-19 November. Available from http://www.mt-archive.info/Aslib-2004-Drugan.pdf

EC (2008) Speaking for Europe. Languages in the European Union. Luxembourg: Publication Office of the EU.

EC (2010) *Translation at the European Commission - a History*. Luxembourg: Publication Office of the EU.

EC (2012) JEX-JRC Eurovoc Indexer. In *JRC Language Technology Resources*. Available from http://ipsc.jrc.ec.europa.eu/?id=60 (accessed November 2012).

Ecke P. (2009) The Tip-of-the-Tongue Phenomenon as a Window on (Bilingual) Lexical Retrieval. In *The Bilingual Mental Lexicon: Interdisciplinary Approaches*, edited by A. Pavlenko. Clevedon: Multilingual Matters, 185-208.

Ehrensberger-Dow M. & Künzli A. (2010) Methods of Accessing Metalinguistic Awareness: A Question of Quality? In *New Approaches in Translation Process Research*, edited by S. Göpferich, F. Alves and I. M. Mees. Copenhagen: Samfundslitteratur, 113-132.

Eisele A. & Lavecchia C. (2011) Using Statistical Machine Translation for Computer-Aided Translation at the European Commission. *Proceedings of JEC '11 "Bringing MT to the User: Research Meets Translation"*, Luxembourg, 14 October, 3-12. Available from http://mt-archive.info/JEC-2011-Eisele.pdf

Enríquez Raído V. (2011) *Investigating the Web Search Behaviors of Translation Students: An Exploratory and Multiple-Case Study*. Universitat Ramon Llull, Barcelona, Spain. PhD dissertation.

Evert S. & Hardie A. (2011) Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of Corpus Linguistics 2011 Conference*, Birmingham, UK, 20-22 July. Available from http://www.stefan-evert.de/PUB/EvertHardie2011.pdf

Færch C. (1984) Strategies in Production and Reception - Some Empirical Evidence. In

Interlanguage, edited by A. Davies, C. Criper and A. P. R. Howatt. Edinburgh: Edinburgh University Press, 49-70.

Federico M., Cattelan A. & Trombetti M. (2012) Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. *Proceedings of AMTA 2012*, San Diego, CA, 28 October-1 November. Available from http://www.mt-archive.info/AMTA-2012-Federico.pdf.

Fifer M. (2007) *The Fuzzy Factor: An Empirical Investigation of Fuzzy Matching in the Context of Translation Memory Systems*. University of Ottawa, Ottawa, Canada. MA thesis.

Firth J. R. (1957) Modes of Meaning. In *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press, 190-215.

Fleischman M. & Hovy E. (2002) Fine Grained Classification of Named Entities. *Proceedings of 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 24 August - 1 September, 1-7. Available from http://acl.ldc.upenn.edu/coling2002/proceedings/data/area-15/co-338.pdf

Fraser J. (1996) The Translator Investigates. Learning from Translation Process Analysis. *The Translator* 2 (1):65-79.

Gagné C. L. & Shoben E. J. (1997) Influence of Thematic Relations on the Comprehension of Modifier-Noun Combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition* 23:71-87.

Garcia I. (2009) Research on Translation Tools. In *Translation Research Projects 2*, edited by A. Pym and A. Perekrestenko. Tarragona: Intercultural Studies Group, 27-33.

Garcia I. (2012) Machines, Translations and Memories: Language Transfer in the Web Browser. *Perspectives: Studies in Translatology* 20 (4):451-461.

Gavioli L. (1999) Corpora and the Concordancer in Learning ESP: An Experiment in a Course for Interpreters and Translators. In *Transiti Linguistici E Culturali*, edited by G. Azzaro and M. Ulrych. Trieste: Editrice Università Trieste, 331-343.

Gerloff P. (1986) Second Language Learners' Reports on the Interpretative Process: Talk-Aloud Protocols of Translation. In *Interlingual and Intercultural Communication*, edited by J. House and S. Blum-Kulka. Tübingen: Narr, 243-262.

Gerloff P. (1987) Identifying the Unit of Analysis in Translation: Some Uses of Think-Aloud Protocol Data. In *Introspection in Second Language Research*, edited by C. Færch and G. Kasper. Clevedon/Philadelphia: Multilingual Matters, 135-158.

Gerloff P. (1988) *From French to English: A Look at the Translation Process in Students, Bilinguals, and Professional Translators*. Harvard University, Cambridge, MA. Unpublished PhD Dissertation.

Gil-Bardají A. (2010) La Résolution de Problèmes en Traduction : Quelques Pistes. *Meta* 60 (2):275-286.

Girju R. (2008) The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study. *Computational Linguistics* 35 (2):185-228.

Goffin R. (1994) L'eurolecte : Oui, Jargon Communautaire : Non. *Meta* 39 (4):636-642.

Gollan T. H. & Acenas L.-A. R. (2004) What is a Tot? Cognate and Translation Effects on Tip-of-the-Tongue States in Spanish-English and Tagalog-English Bilinguals.

Journal of Experimental Psychology: Learning, Memory, and Cognition 30 (1):246-269.

Gonzáles Davies M. & Scott-Tennent C. (2005) A Problem-Solving and Student-Centred Approach to the Translation of Cultural References. *Meta* 50 (1):160-179.

Göpferich S. (2009) Towards a Model of Translation Competence and Its Acquisition: The Longitudinal Study TransComp. In *Behind the Mind: Methods, Models and Results in Translation Process Research*, edited by S. Göpferich, A. L. Jakobsen and I. M. Mees. Copenhagen: Samfundslitteratur, 11-38.

Göpferich S. (2010) The Translation of Instructive Texts from a Cognitive Perspective: Novices and Professionals Compared. In *New Approaches in Translation Process Research*, edited by S. Göpferich, F. Alves and I. M. Mees. Copenhagen: Samfundslitteratur, 5-56.

Göpferich S. & Jääskeläinen R. (2009) Process Research into the Development of Translation Competence: Where Are We, and Where Do We Need to Go? *Across Languages and Cultures* 10 (2):169-191.

Gough J. (2012) Tools and Resources for Translation Professionals - Summary of Questionnaire Results. Paper read at *ASLIB Translating and the Computer 34*, London, UK, 29-30 November.

Grimes C., Tang D. & Russell D. M. (2007) Query Logs Alone are not enough. Paper read at *WWW 2007 – Workshop on Query Logs Analysis: Social and Technological Challenges*, Banff, Canada, 8-12 May. Available from http://www2007.org/workshops/paper_51.pdf

Gross A. (1998) Multiconcord - Ein Paralleles Konkordanzprogramm für den Fremdsprachenunerricht. *Tell & Call* 10.98 (4):36-42.

Halliday M. A. K. (1985) *An Introduction to Functional Grammar*. London: Arnold.

Hands F. (2012) The Translation Process. Paper read at *ASLIB Translating and the Computer 34*, London, UK, 29-30 November. Slides available from http://www.mt-archive.info/Aslib-2012-Hands-ppt.pdf

Heid U. (1994) On the Way Words Work Together - Topics in Lexical Combinatorics. *Proceedings of 6th EURALEX International Congress*, Amsterdam, The Netherlands, 30 August - 3 September, Vrije Universiteit, 226-257. Available from http://www.euralex.org/elx_proceedings/Euralex1994/27_Euralex_Ulrich[...].pdf

Herskovic J. R., Tanaka L. Y., Hersh W. H., et al. (2007) A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *Journal of the American Medical Informatics Association* 14 (2):212-220.

Hewson L. & Martin J. (1991) *Redefining Translation. The Variational Approach*. London/New York: Routledge.

Hoey M. (2005) *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Holmes J. S. (2004) The Name and Nature of Translation Studies. In *The Translation Studies Reader*, edited by L. Venuti. New York: Routledge, 180-192. Original edition, 1972.

Huang J. & Efthimiadis E. N. (2009) Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs. *Proceedings of CIKM 2009. 18th ACM Conference on*

Information and Knowledge Management, Hong Kong, 2-6 November. New York: ACM, 77-86. Available from http://jeffhuang.com/sigir09_submit.pdf

Huet S., Bourdaillet J. & Langlais P. (2009a) TS3: An Improved Version of the Bilingual Concordancer TransSearch. *Proceedings of 13th EAMT*, Barcelona, Spain, 13-15 May, 20-27. Available from http://www.iro.umontreal.ca/~felipe/bib2webV0.81/cv/papers/paper-eamt-2009.pdf

Huet S., Bourdaillet J. & Langlais P. (2009b) Intégration de l'alignement de mots dans le concordancier bilingue TransSearch. *Proceedings of TALN 2009*, 16è Conférence sur le Traitement Automatique des Langues Naturelles, Senlis, France, 24-26 June. Available from http://www.iro.umontreal.ca/~felipe/bib2webV0.81/cv/papers/paper-taln-2009b.pdf

Hunston S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hurtado Albir A. (2001) *Traduccíon y Traductología: Introduccíon a la Traductología.* Madrid: Cátedra.

Hutchins J. (2005) Current Commercial Machine Translation Systems and Computer-Based Translation Tools: System Types and Their Uses. *International Journal of Translation* 17 (1-2):5-38.

Hvelplund K. T. (2011) *Allocation of Cognitive Resources in Translation. An Eye-Tracking and Key-Logging Study*. Copenhagen Business School, Copenhagen. PhD dissertation.

Immonen S. (2006) Translation as a Writing Process. *Target* 18 (2):313-335.

Immonen S. (2011) Unravelling the Processing of Translation Units. *Across Languages and Cultures* 12 (2):235-257.

Immonen S. & Mäkisalo J. (2010) Pauses Reflecting the Processing of Syntactic Units in Monolingual Text Production and Translation. *Hermes - Journal of Language and Communication Studies* 44:45-61.

Jääskeläinen R. (1987) *What Happens in a Translation Process: Think-Aloud Protocols of Translation*. School of Translation Studies, University of Joensuu, Savonlinna. Unpublished MA thesis.

Jääskeläinen R. (1990) *Features of Successful Translation Process: A Think-Aloud Protocol Study*. School of Translation Studies, University of Joensuu, Savonlinna. PhD dissertation.

Jääskeläinen R. (1993) Investigating Translation Strategies. In *Recent Trends in Empirical Translation Research*, edited by S. Tirkkonen-Condit and J. Laffling. Joensuu: University of Joensuu, 99-120.

Jääskeläinen R. (1999) *Tapping the Process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating*. Joensuu: Joensuu yliopisto.

Jääskeläinen R. (2002) Think-Aloud Protocols Studies into Translation. An Annotated Bibliography. *Target* 14 (1):107-136.

Jääskeläinen R. (2010) Are All Professional Experts? Definitions of Expertise and Reinterpretation of Research Evidence in Process Studies. In *Translation and Cognition*, edited by G. M. Shreve and E. Angelone. Amsterdam/Philadelphia: John

Benjamins, 213-227.

Jääskeläinen R. (2011) Back to Basics: Designing a Study to Determine the Validity and Reliability of Verbal Report Data on Translation Processes. In *Cognitive Explorations of Translation*, edited by S. O' Brien. London/New York: Continuum, 15-29.

Jääskeläinen R. & Tirkkonen-Condit S. (1991) Automatised Processes in Professional Vs. Non-Professional Translation: A Think-Aloud Protocol Study. In *Empirical Research in Translation and Intercultural Studies: Selected Papers of the Transif Seminar, Savolinna 1988*, edited by S. Tirkkonen-Condit. Tübingen: Gunter Narr, 89-109.

Jackson H. (1988) *Words and Their Meaning*. London/New York: Longman.

Jakobsen A. L. (2002) Translation Drafting by Professional Translators and by Translation Students. In *Empirical Translation Studies*, edited by G. Hansen. Copenhagen: Samfundslitteratur, 191-204.

Jakobsen A. L. (2003) Effects of Think Aloud on Translation Speed, Revision, and Segmentation. In *Triangulating Translation*, edited by F. Alves. Amsterdam/Philadelphia: John Benjamins, 69-95.

Jakobsen A. L. (2005) Investigating Expert Translators' Processing Knowledge. In *Knowledge Systems and Translation*, edited by H. V. Dam, J. Engberg and H. Gerzymisch-Arbogast. Berlin: de Gruyter, 173-189.

Jakobsen A. L. & Jensen K. T. H. (2008) Eye Movement Behaviour across Four Different Types of Reading Task. In *Looking at Eyes. Eye-Tracking Studies of Reading and Translation Processing*, edited by S. Göpferich, A. L. Jakobsen and I. M. Mees. Copenhagen: Samfundslitteratur, 103-124.

Jansen B. J. (2006) Search Log Analysis: What It is, What's Been Done, How to Do It. *Library & Information Science Research* 28:407-432.

Jansen B. J. & Booth D. L. (2010) Classifying Web Queries by Topic and User Intent. *Proceedings of CHI EA 2010. CHI '10 Extended Abstracts on Human Factors in Computing Systems*, Atlanta, GA, 14-15 April, New York: ACM, 4285-4290. Available from http://faculty.ist.psu.edu/jjansen/academic/jansen_user_intent.pdf

Jansen B. J., Booth D. L. & Spink A. (2008) Determining the Informational, Navigational and Transactional Intent of Web Queries. *Information Processing and Management* 44:1251-1266.

Jansen B. J., Booth D. L. & Spink A. (2009) Patterns of Query Reformulation During Web Searching. *Journal of the American Society for Information Science and Technology* 60 (7):1358-1371.

Jansen B. J., Spink A. & Pfaff M. A. (2000a) Web Query Structure: Implications for IR System Design. *Proceedings of SCI'2000 – 4th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, FL, 23-26 July. Available from http://faculty.ist.psu.edu/jjansen/academic/pubs/sci2000/sci2000.pdf

Jansen B. J., Spink A. & Pfaff M. A. (2000b) Linguistic Aspects of Web Queries. *Proceedings of 63rd Annual Meeting of the American Society of Information Science (ASIS) v.37*, Chicago, IL, 13-16 November, 169-176. Available from http://faculty.ist.psu.edu/jjansen/academic/pubs/asis2000/asis2000.html

(accessed November 2010).

Jansen B. J., Spink A. & Saracevic T. (2000) Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management* 36 (2):207-227.

Jarvella R. J., Jensen A., Halskov Jensen E., et al. (2002) Towards Characterizing Translator Expertise, Knowledge and Know-How: Some Findings Using Taps and Experimental Methods. In *Translation Studies. Perspectives on an Emerging Discipline*, edited by A. Riccardi. Cambridge: Cambridge University Press, 172-197.

Jensen A. (1999) Time Pressure in Translation. In *Probing the Process in Translation: Methods and Results*, edited by G. Hansen. Copenhagen: Samfundslitteratur, 103-120.

Jensen A. & Jakobsen A. L. (2000) Translating under Time Pressure: An Empirical Investigation of Problem-Solving Activity and Translation Strategies by Non-Professional and Professional Translators. In *Translation in Context. Selected Contributions from the EST Congress, Granada 1998*, edited by A. Chesterman, N. G. San Salvador and Y. Gambier. Amsterdam/Philadelphia: John Benjamins, 105-116.

Jensen K. T. H. (2009) Indicators of Text Complexity. In *Behind the Mind: Methods, Models and Results in Translation Process Research*, edited by S. Göpferich, A. L. Jakobsen and I. M. Mees. Copenhagen: Samfundslitteratur, 61-80.

Johnson D., Malhotra V. & Vamplew P. (2006) More Effective Web Search Using Bigrams and Trigrams. *Webology* 3(4). Available from http://www.webology.ir/2006/v3n4/a35.html (accessed November 2010).

Just M. A. & Carpenter P. A. (1980) A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review* 87 (4):329-354.

Justeson J. (1993) Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *IBM Research Report* (RC 18906 (82591)).

Karamanis N., Luz S. & Doherty G. (2010) Translation Practice in the Workplace and Machine Translation. *Proceedings of EAMT 2010*, St. Raphael, France, 27-28 May. Available from https://scss.tcd.ie/Saturnino.Luz/publications/KaramanisLuzDoherty10tmt.pdf

Karamanis N., Luz S. & Doherty G. (2011) Translation Practice in the Workplace: Contextual Analysis and Implications for Machine Translation. *Machine Translation* 2011 (25):35-52.

Katz S. M. (1987) Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35 (3):400-401.

Kilgarriff A. (2005) Language Is Never, Ever, Ever, Random. *Corpus Linguistics and Linguistic Theory* 1 (2):263-276.

Kimmes A. & Koopman H. (2010) Collocation|Analyzer – an Electronic Tool for Collocation Retrieval and Verification. *T21N*. Available from www.t21n.com/homepage/articles/T21N-2010-11-Kimmes,Koopman.pdf

Kipling R. (1902) The Elephant's Child. In *Just So Stories*. London: Macmillan & co., 63-84.

Kiraly D. C. (1995) *Pathways to Translation. Pedagogy and Process*. Kent: Kent State University Press.

Kohen P. & Haddow B. (2009) Interactive Assistance to Human Translators Using Statistical Machine Translation Methods. *Proceedings of Machine Translation Summit XII*, Ottawa, Canada, 26-30 August. Available from http://www.mt-archive.info/MTS-2009-Koehn-2.pdf

Königs F. G. (1987) Was beim Übersetzen Passiert. *Die Neueren Sprachen* 86 (2):162-185.

Koskinen K. (2008) *Translating Institutions: An Ethnographic Study of EU Translation*. Manchester/Kinderhook: St. Jerome.

Kovačič I. (2000) Thinking-Aloud Protocols - Interview - Text Analysis. In *Tapping and Mapping the Process of Translation and Interpreting*, edited by S. Tirkkonen-Condit and R. Jääskeläinen. Amsterdam/Philadelphia: John Benjamins, 97-109.

Kowalska A. (2010) La Direction Générale de la Traduction de la Commission Européenne. Paper read at Campus d'été - Métiers des Langues et de la Traduction, Poitiers, 28 June-3 July.

Krings H.-P. (1986a) *Was in den Köpfen von Übersetzern Vorgeht*. Tübingen: Gunter Narr.

Krings H.-P. (1986b) Translation Problems and Translation Strategies of Advanced German Learners of French (L2). In *Interlingual and Intercultural Communication*, edited by J. House and S. Blum-Kulka. Tübingen: Narr, 263-276.

Krings H.-P. (1987) The Use of Introspective Data in Translation. In *Introspection in Second Language Research*, edited by C. Færch and G. Kasper. Clevedon/Philadelphia: Multilingual Matters, 159-176.

Krings H.-P. (2001) *Repairing Texts. Empirical Investigations of Machine Translation Post-Editing Processes*. Kent/London: Kent State University Press.

Krings H.-P. (2005) Wege ins Labyrinth - Fragestellung und Methoden der Übersetzungsprozessforschung im Überblick. *Meta* 50 (2):342-358.

Krishnamurthy R. (2001) Corpus, Collocation and Lexical Sets. *Proceedings of HUSSE (Hungarian Society for the Study of English) Thematic Conference*, University of Debrecen (HU), 21-22 June. Available from http://eprints.aston.ac.uk/5658/1/2001-HUSSE-corpus-collocation-lexical-sets.pdf

Krishnamurthy R., ed. (2004) *English Collocation Studies: The OTIS Report by John Sinclair, Susan Jones and Robert Daley*. London/New York: Continuum.

Kroll J. F. & De Groot A. M. B., eds. (2005) *Handbook of Bilingualism. Psycholinguistic Approaches*. Oxford: Oxford University Press.

Künzli A. (2001) Experts *Versus* Novices : L'utilisation de sources d'information pendant le processus de traduction. *Meta* 46 (3):507-523.

L'Homme M.-C. (2005) Sur la notion de «terme». *Meta* 50 (4):1112-1132.

L'Homme M.-C. & Bertrand C. (2000) Specialized Lexical Combinations: Should They Be Described as Collocations or in Terms of Selectional Restriction? Paper read at 9th Euralex International Congress, Stuttgart, Germany, 8-12 August, 497-506. Available from http://www.ling.umontreal.ca/lhomme/docs/berlhom.pdf

Lachat Leal C. (2003) *Estrategias y problemas de traducción.* Departamento de traducción e interpretación, Universidad de Granada, Granada. Tesis doctoral.

Lagoudaki E. (2006) *Translation Memories Survey 2006: Users' Perceptions around TM Use*. London: Imperial College London.

Lapata M. & Lascarides A. (2003) Detecting Novel Compounds: The Role of Distri-Butional Evidence. *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 12-17 April, 235-242.

Lauffer S. (2002) The Translation Process: An Analysis of Observational Methodology. *Cadernos de Tradução* 2 (10):59-74.

Laviosa S. (2002) *Corpus-Based Translation Studies*. Amsterdam/New York: Rodopi.

Leick J.-M. (1998) Euramis - The Ultimate Multilingual Blackbox? *Terminologie et Traduction* 1998 (1):52-58.

Levelt W. J. M. (1989) *Speaking from Intention to Articulation*. Cambridge, MA: MIT Press.

Lindquist H. (2009) *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Livbjerg I. & Mees I. M. (2002) Problem-Solving at Different Points in the Translation Process: Quantitative and Qualitative Data. In *Empirical Translation Studies: Process and Products*, edited by G. Hansen. Copenhagen: Samfundslitteratur, 147-190.

Livbjerg I. & Mees I. M. (2003) Patterns of Dictionary Use in non-Domain-Specific Translation. In *Triangulating Translation: Perspectives in Process Oriented Research*, edited by F. Alves. Amsterdam/Philadelphia: John Benjamins, 123-136.

Lorenzo M. P. (1999) La seguridad del traductor profesional en la traducción a una lengua extranjera. In *Probing the Process in Translation: Methods and Results*, edited by G. Hansen. Copenhagen: Samfundslitteratur, 121-134.

Lörscher W. (1986) Linguistic Aspects of Translation Processes: Towards an Analysis of Translation Performance. In *Interlingual and Intercultural Communication*, edited by J. House and S. Blum-Kulka. Tübingen: Gunter Narr, 277-292.

Lörscher W. (1991a) *Translation Performance, Translation Process and Translation Strategies. A Psycholinguistic Investigation*. Tübingen: Gunter Narr.

Lörscher W. (1991b) Thinking-Aloud as a Method for Collecting Data on Translation Process. In *Empirical Research in Translation and Intercultural Studies*, edited by S. Tirkkonen-Condit. Tübingen: Gunter Narr, 67-77.

Lörscher W. (1996) A Psycholinguistic Analysis of Translation Processes. *Meta* 41 (1):26-32.

Lörscher W. (2005) The Translation Process: Methods and Problems of its Investigation. *Meta* 50 (2):597-608.

Lossner K. (2012) Translation Tool Concordances Compared. *Translation Tribulations*. Available from http://www.translationtribulations.com/2012/01/recent-experience-when-tutoring-new.html (accessed October 2012).

Lou X. (1999) Linguistic Contribution to the Development of Translation Studies in China. *Meta* 44 (1):101-109

Lu L. & Yuan Y. C. (2011) Shall I Google it or Ask the Competent Villain Down the Hall? The Moderating Role of Information Need in Information Source Selection. *Journal of the American Society for Information Science and Technology* 62 (1):133-145.

Machado I. (2007) *Processos de Orientação Inicial e em Tempo Real e sua Interface com Sistemas de Memória de Tradução*. Federal University of Minas Gerais, Belo

Horizonte, Brazil. Unpublished MA thesis.

Macklovitch E., Lapalme G. & Gotti F. (2008) *TransSearch*: What are Translators Looking for? *Proceedings of 8th AMTA conference*, Hawaii, USA, 21-25 October, 412-419. Available from http://rali.iro.umontreal.ca/rali/sites/default/files/publis/amta08-TransSearch-final.pdf

Macklovitch E. & Russell G. (2000) What's Been Forgotten in Translation Memory. *Proceedings of 4th AMTA conference*, Mexico City, Mexico 10-14 October. London, UK: Springer Verlag, 137-146.

Macklovitch E., Simard M. & Langlais P. (2000) *TransSearch*: A Free Translation Memory on the World Wide Web. *Proceedings of 2nd LREC conference*, Athens, Greece, 30 May - 2 June, 1201-1208. Available from http://ccl.pku.edu.cn/doubtfire/nlp/Machine_Translation/Translation_Memory/TransSearch_A_Free_Translation_Memory_on_the_WWW.pdf

Maguire P., Wisniewski E. J. & Sorms G. (2010) A Corpus Study of Semantic Patterns in Compounding. *Corpus Linguistics and Linguistic Theory* 6 (1):49-73.

Malmkjær K. (2006) Translation Units. In *The Encyclopedia of Languages and Linguistics*, edited by K. Brown. Amsterdam: Elsevier, 92-93.

Martín-Mor A. (2011) *La Interferència Lingüística en Entorns de Traducció Assistida per Ordinador*. Recerca Empíricoexperimental, Departament de Traducció i d'Interpretació, Universitat Autònoma de Barcelona. PhD dissertation.

Matthews G. H. (1961) Analysis by Synthesis of Sentences of Natural Language. *Proceedings of International Conference on Machine Translation of Languages and Applied Languages Analysis*, Teddington, UK, 5-6 September, 532-540. Available from http://www.mt-archive.info/NPL-1961-Matthews.pdf

McGee I. (2009) Adjective-Noun Collocations in Elicited and Corpus Data: Similarities, Differences, and the Whys and Wherefores. *Corpus Linguistics and Linguistic Theory* 5 (1):79-103.

Mees I. M. (2009) Arnt Lykke Jakobsen: Portrait of an Innovator. In *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen*, edited by I. M. Mees, F. Alves and S. Göpferich. Copenhagen: Samfundslitteratur, 9-36.

Mel'čuk I. A. (1998) Collocations and Lexical Functions. In *Phraseology. Theory, Analysis and Applications*, edited by A. P. Cowie. Oxford: Clarendon Press, 23-53.

Melby A. K. (2006) MT+TM+QA: The Future is ours. *Revista Tradumàtica - Traducció i Tecnologies de la informació i la Comunicació* 4. Available from http://www.fti.uab.es/tradumatica/revista/num4/articles/04/04.pdf

Meuter R. (2009) Neurolinguistic Contributions to Understanding the Bilingual Mental Lexicon. In *The Bilingual Mental Lexicon: Interdisciplinary Approaches*, edited by A. Pavlenko. Clevedon: Multilingual Matters, 1-25.

Mizzaro S. (1996) A Cognitive Analysis of Information Retrieval. In *Proceedings of COLIS2 - Information Science: Integration in Perspective*, edited by P. Ingwersen and N. O. Pors, Copenhagen 13-16 October. Denmark: The Royal School of Librarianship, 233-250. Available from http://sole.dimi.uniud.it/~stefano.mizzaro/research/papers/colis.pdf

Mollin S. (2009) Combining Corpus Linguistics and Psychological Data on Word Co-Occurrences: Corpus Collocates Versus Word Associations. *Corpus Linguistics and Linguistic Theory* 5 (2):175-200.

Mondhal M. (1995) Lexical Search Strategies: A Study of Translation Process in a Brief Text. *Multilingua* 14 (2):183-204.

Mondhal M. & Jensen K. A. (1996) Lexical Search Strategies in Translation. *Meta* 41 (1):97-113.

Monti J., Barreiro A., Elia A., et al. (2011) *Taking on* New Challenges in Multi-Word Unit Processing for Machine Translation. Paper read at 2nd International Workshop on Free/Open-Source Rule-Based Machine Translation, Barcelona, Spain, 20-21 January, 11-19. Available from http://openaccess.uoc.edu/webapps/o2/handle/10609/5646

Muñoz Martín R. (2009a) Expertise and Environment in Translation. *Mutatis Mutandis* 2 (1):24-37.

Muñoz Martín R. (2009b) The Way We Were: Subject Profiling in Translation Process Research. In *Methodology, Technology and Innovation in Translation Process Research. A Tribute to Arnt Lykke Jakobsen*, edited by I. M. Mees, F. Alves and S. Göpferich. Copenhagen: Samfundslitteratur, 87-108.

Nadeau D. & Sekine S. (2007) A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes* 30 (1):3-26.

Newell A. & Simon H. A. (1972) *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newmark P. (1988) A Textbook of Translation, London: Prentice Hall.

Nida E. (1964) *Towards a Science of Translating*. Leiden: E.J. Brill.

Nida E. (1969) *The Theory and Practice of Translation*. Leiden: United Bible Society.

Nie N. H., ed. (1975) *SPSS: Statistical Package for the Social Sciences*. 2nd ed, New York: McGraw Hill.

Nord C. (1991) *Text Analysis in Translation*. Amsterdam/Atlanta: Rodopi.

Nord C. (2011) *Funktionsgerechtigkeit und Loyalität: Theorie, Methode und Didaktik des Funktionalen Übersetzens*. Berlin: Frank & Timme.

O'Brien S. (2008) Processing Fuzzy Matching in Translation Memory Tools: An Eye-Tracking Analysis. In *Looking at Eyes. Eye-Tracking Studies of Reading and Translation Processing*, edited by S. Göpferich, A. L. Jakobsen and I. M. Mees. Copenhagen: Samfundslitteratur, 79-102.

O'Brien S. (2009) Eye Tracking in Translation-Process Research: Methodological Challenges and Solutions. In *Methodology, Technology and Innovation in Translation Process Research*, edited by I. M. Mees, F. Alves and S. Göpferich. Copenhagen: Samfundslitteratur, 251-266.

O'Brien S. (2012) Translation as a Human-Computer Interaction. *Translation Spaces* 1 (1):101-122.

O'Brien S., O'Hagan M. & Flanagan M. (2010) Keeping an Eye on the UI Design of Translation Memory: How Do Translators Use the "Concordance" Feature?. In *Proceedings of ECCE'10 - 28th Annual European Conference on Cognitive Ergonomics*, Delft, The Netherlands, 25-27 August. New York: ACM, 187-190.

O'Neill A. (2009) Sentiment Mining for Natural Language Documents. *COMP3006 PROJECT REPORT*. Available from http://users.cecs.anu.edu.au/~ssanner/Papers/Alex_Report.pdf (accessed October 2012).

Oakes M. P. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Okidoo. (2012) *Where Does TradooIT Come From?* Available from http://en.wiki.tradooit.com/home/dou-vient-tradooit (accessed November 2012).

Olohan M. (2004) *Introducing Corpora in Translation Studies*. London/New York: Routledge.

Ondelli S. & Viale M. (2010) Evidenze quantitative sull'italiano tradotto in un corpus giornalistico. *Proceedings of JADT 2010 - 10th International Conference on Statistical Analysis of Textual Data*, Rome, 9-11 June, Milano: LED, 573-584. Available from http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-0573-0584_013-Ondelli.pdf

PACTE (2005) Investigating Translation Competence: Conceptual and Methodological Issues. *Meta* 50 (2):609-619.

PACTE (2009) Results of the Validation of the PACTE Translation Competence Model: Acceptability and Decision Making. *Across Languages and Cultures* 10 (2):207-230.

PACTE (2011a) Results of the Validation of the PACTE Translation Competence Model: Translation Project and Dynamic Translation Index. In *Cognitive Explorations of Translation*, edited by S. O'Brien. London/New York: Continuum, 30-56.

PACTE (2011b) Results of the Validation of the PACTE Translation Competence Model: Translation Problems and Translation Competence. In *Methods and Strategies of Process Research: Integrative Approaches in Translation Studies*, edited by C. Alvstad, A. Hild and E. Tiselius. Amsterdam/Philadelphia: John Benjamins, 317-343.

Palumbo G. (2008) *'Translating Science': An Empirical Investigation of Grammatical Metaphor as a Source of Difficulty for a Group of Translation Trainees in English-Italian Translation*. Department of Languages and Translation Studies, University of Surrey. PhD dissertation.

Palumbo G. (2009) *Key Terms in Translation Studies*. London/New York: Continuum.

Partington A. (1998) *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. Amsterdam/Philadelphia: John Benjamins.

Pavlović N. (2007) *Directionality in Collaborative Translation Process. A Study of Novice Translators*, Universitat Rovira i Virgili, University of Zagreb. PhD dissertation.

Pavlović N. & Jensen K. T. H. (2009) Eye Tracking Translation Directionality. In *Translation Research Projects 2*, edited by A. Pym and A. Perekrestenko. Tarragona: Intercultural Studies Group, 93-110.

Peressini S. (2010) Una panoramica delle principali tipologie di testi tradotti alla DGT. Paper read at PhD Seminar, Trieste, 7 May.

Pouliquen B., Steinberger R. & Ignat C. (2003) Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. Paper read at Ontologies and

Information Extraction - EUROLAN'2003, Bucharest, Romania, 28 July-8 August. Available from http://arxiv.org/pdf/cs/0609059.pdf

Poulis A. (2009) Language Technology in the European Parliament's Directorate General for Translation. Facts, Problems and Visions. Paper read at FLaReNet Forum - The European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe, Vienna, Austria, 12-13 February. Slides available from http://www.flarenet.eu/sites/default/files/Poulis_Presentation.pdf

Prahl B. & Petzolt S. (1997) Translation Problems and Translation Strategies Involved in Human and Machine Translation: Empirical Studies. In *Machine Translation and Translation Theory*, edited by C. Hauenschild and S. Heizmann. Berlin/New York: M. de Gruyter, 123-144.

Presas Corbella M. L. (1996) *Problemas de Traducció i Competència Traductora. Bases per a una Pedagogia de la Traducció*. Departament de traducció i d'interpretació, Universitat Autònoma de Barcelona, Bellaterra, Spain. Tesi doctoral.

Pu H.-T., Chuang S.-L. & Yang C. (2002) Subject Categorization of Query Terms for Exploring Web Users' Search Interests. *Journal of the American Society for Information Science and Technology* 53 (8):617-630.

Pyers J. E., Gollan T. H. & Emmorey K. (2009) Bimodal Bilinguals Reveal the Source of Tip-of-the-Tongue States. *Cognition* 112 (2009):323-329.

Pym A. (2003) Redefining Translation Competence in an Electronic Age. In Defence of a Minimalist Approach. *Meta* 48 (4):481-497.

Quah C. K. (2006) *Translation and Technology*. New York: Palgrave MacMillan.

Quigley K. (2005) Keeping Company in the City: Compounds in the Lexicon of the New Zealand Treasury. *New Zealand English Journal* 19:26-35.

Rasinger S. M. (2008) *Quantitative Research in Linguistics. An Introduction*. London/New York: Continuum.

Reiss K. (1974) Zur Bestimmung des Schwierigkeitsgrades von Übersetzungen. *Mitteilungsblatt für Dolmetscher und Übersetzer* 20 (3):1-6.

Renouf A. & Sinclair J. M. (1991) Collocational Frameworks in English. In *English Corpus Linguistics*, edited by K. Aijmer and B. Altenberg. London: Longman, 128-143.

Reppen R. (2001) Review of MonoConc Pro and Wordsmith Tools. *Language Learning & Technology* 5 (3):32-36.

Ronowicz E., Hehir J., Kaimi T., et al. (2005) Translator's Frequent Lexis Store and Dictionary Use as Factors in SLT Comprehension and Translation Speed - a Comparative Study of Professional, Paraprofessional and Novice Translators. *Meta* 50 (2):580-596.

Rosch E. & Mervis C. B. (1975) Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology* 7:573-605.

Rose D. E. & Levinson D. (2004) Understanding User Goals in Web Search. In *Proceedings of the 13th international conference on World Wide Web* – WWW '04, New York, USA, 17-22 May. New York: ACM, 13-19.

Ross N. C. & Wolfram D. (2000) End User Searching on the Internet: An Analysis of Term Pair Topics Submitted to the Excite Search Engine. *Journal of the American Society for Information Science and Technology* 51 (10):949-958.

Rusu A. (2009) Euramis Briefing for Trainees. Paper read at Training Seminars for EP Trainees, European Parliament, Luxembourg, 3 November.

Sager J. C. (1993) *Language Engineering and Translation. Consequences of Automation*. Amsterdam/Philadelphia: John Benjamins.

Sager J. C. (2001) Terminology - Theory. In *Routledge Encyclopedia of Translation Studies*, edited by M. Baker. London/New York: Routledge, 258-262.

Santorini B. (1991) *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. Available from http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.pdf (accessed March 2011).

Scarpa F. (1999) Corpus Evidence of the Translation of Genre-Specific Structures. *Textus* 12 (2):315-332.

Scarpa F. (2000) Using a Multilingual Parallel Concordancer in Advanced Translator Training into L1. *Rivista internazionale di tecnica della traduzione / International Journal of Translation* 5:63-73.

Scarpa F. (2006) Corpus-Based Quality Assessment of Specialist Translation: A Study Using Parallel and Comparable Corpora in English and Italian. In *Insights into Specialized Translation*, edited by M. Gotti and S. Saracevic. Bern: Peter Lang, 155-172.

Schiaffino R. (2011) Don't Search from the Wrong Side: A Reminder for SDL 2009 Users. *About Translation*. Available from http://www.aboutranslation.com (accessed October 2012).

Schou L., Dragsted B. & Carl M. (2009) Ten Years of Translog. In *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen*, edited by I. M. Mees, F. Alves and S. Göpferich. Copenhagen: Samfundslitteratur, 37-48.

Schwartz B. L. (2008) Working Memory Load Differentially Affects Tip-of-the-Tongue States and Feeling-of-Knowing Judgments. *Memory & Cognition* 36 (1):9-19.

Séguinot C. (1989) The Translation Process: An Experimental Study. In *The Translation Process*, edited by C. Séguinot. Toronto: H.G. Publications, York University, 21-53.

Séguinot C. (1991) A Study of Student Translation Strategies. In *Empirical Research in Translation and Intercultural Studies*, edited by S. Tirkkonen-Condit. Tübingen: Narr, 79-88.

Séguinot C. (2000) Management Issues in the Translation Process. In *Tapping and Mapping the Process of Translation and Interpreting*, edited by S. Tirkkonen-Condit and R. Jääskeläinen. Amsterdam/Philadelphia: John Benjamins, 143-148.

Sekine S. & Nobata C. (2004) Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Proceedings of LREC 2004 - 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 26-28 May, Paris: ELDA, 1977-1980. Available from http://webmail.cl.uni-heidelberg.de/~sourjiko/NER_Literatur/NEC_extended_Sekine.pdf

Sharoff S. (2006) Translation as Problem Solving: Uses of Comparable Corpora. *Proceedings of International Workshop LR4Trans-III (LREC 2006)*, Genoa, Italy, 22-28 May. Available from http://corpus.leeds.ac.uk/serge/publications/lrec2006-lr4trans.pdf.

Sharoff S., Babych B. & Hartley A. (2006) Using Comparable Corpora to Solve Problems Difficult for Human Translators. *Proceedings of COLING-ACL '06*, Sydney, 17-21 July, Stroudsburg, PA: ACL, 739-746.

Shreve G. M. & Angelone E., eds. (2010) *Translation and Cognition*. Vol. XV, *American Translators Association Scholarly Monograph Series*. Amsterdam/Philadelphia: John Benjamins.

Silverstein C., Henzinger M., Marais H., et al. (1999) Analysis of a Very Large Web Search Engine Query Log. *ACM SIGIR Forum* 33 (1):6-12.

Simard M. (2003) *Mémoires de Traduction Sous-Phrastiques, Département d'informatique et de recherche opérationnelle*. Département d'informatique et de recherche opérationnelle, Université de Montréal. PhD dissertation.

Simard M., Foster G. F. & Perrault F. (1993) *TransSearch: A Bilingual Concordance Tool*. Laval, Quebec: Centre for Information Technology Innovation (CITI). Available from http://rali.iro.umontreal.ca/rali/sites/default/files/publis/sfpTS93e.pdf

Simard M. & Langlais P. (2001) Sub-Sentential Exploitation of Translation Memories. *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September, 335-340. Available from http://mt-archive.info/MTS-2001-Simard.pdf

Simard M. & Macklovitch E. (2005) Studying the Human Translation Process through the *TransSearch* Log-Files. *Proceedings of AAAI Symposium on "Knowledge Collection from volunteer contributors"*, Stanford, Calif., 21-23 March, 70-77.

Sinclair J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sirén S. & Hakkarainen K. (2002) Expertise in Translation. *Across Languages and Cultures* 3 (1):71-82.

Somers H. & Fernandez Diaz G. (2004) Translation Memory vs. Example-Based MT. What's the Difference? *International Journal of Translation* 16 (2):5-33.

Spink A., Jansen B. J., Wolfram D., et al. (2002) From E-Sex to E-Commerce: Web Search Changes. *Computer* 35 (3):107-109.

Spink A., Wolfram D., Jansen M. B. J., et al. (2001) Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology* 52 (3):226-234.

Stamenov M. I. (2009) Cognates in Language, in the Mind and in a Prompting Dictionary for Translation. In *Behind the Mind. Methods, Models and Results in Translation Process Research*, edited by S. Göpferich, A. L. Jakobsen and I. M. Mees. Copenhagen: Samfundslitteratur, 219-251.

Stamenov M. I., Gerganov A. & Popivanov I. D. (2010) Prompting Cognates in the Bilingual Lexicon: Optimizing Access During Translation. *Translation and Cognition Cognition,* edited by G. M. Shreve and E. Angelone. Amsterdam/Philadelphia: John Benjamins, 323-347.

Steinberger R., Pouliquen B., Kabadjov M., et al. (2011) JRC-Names: A Freely Available, Highly Multilingual Named Entity Resource. *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, 12-14 September, 104-110. Available from http://langtech.jrc.ec.europa.eu/Documents/11_RANLP_JRC-Names_Steinberger-et-al.pdf

Strassel S. (2006) *Simple Named Entity Guidelines V6.4*. Philadelphia: Linguistic Data

Consortium. Available from http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.4.pdf (accessed August 2012).

Strauss A. & Corbin J. (1994) Grounded Theory Methodology. An Overview. In *Handbook of Qualitative Research*, edited by N. K. Denzin and Y. S. Lincoln. Thousand Oaks, CA: Sage, 273-285.

Stubbs M. (1995) Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. *Functions of Language* 2 (1):23-55.

Stubbs M. (2002) Two Quantitative Methods of Studying Phraseology in English. *International Journal of Corpus Linguistics* 7 (2):215-244.

Tirkkonen-Condit S. (2000) Uncertainty in Translation Process. In *Tapping and Mapping the Processes of Translation and Interpreting*, edited by S. Tirkkonen-Condit and R. Jääskeläinen. Amsterdam/Philadelphia: John Benjamins, 123-142.

Toury G. (1995) *Descriptive Translation Studies and Beyond*. Amsterdam/Philadelphia: John Benjamins.

Toury G. (2002) What's the Problem with "Translation Problem"? In *Translation and Meaning Part 6*, edited by B. Lewandowska-Tomaszczy and M. Thelen. Maastricht: Hogeschool Zuyd, Maastricht School of Translation and Interpretation, 57-71.

Valli P. (2011) A Study on Concordancing: EU Policy Areas as Translation Problems. In *Translation Studies: Old and New Types of Translation in Theory and Practice*, edited by L. N. Zybatow, A. Petrova and M. Ustaszewski. Bern: Peter Lang, 131-138.

Vinay J.-P. & Darbelnet J. (1995) *Comparative Stylistics of French and English. A Methodology for Translation*. Amsterdam: John Benjamins.

Vincze V., Nagy I. T. & Berend G. (2011) Multiword Expressions and Named Entities in the Wiki50 Corpus. *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, 12-14 September, 289-295. Available from http://www.aclweb.org/anthology-new/R/R11/R11-1040.pdf

Wagner E., Bech S. & Martínez J. M. (2002) *Translating for the European Union Institutions*. Manchester: St. Jerome.

Wang P., Berry M. W. & Yang Y. (2003) Mining Longitudinal Web Queries: Trends and Patterns. *Journal of the American Society for Information Science and Technology* 54 (8):743-758.

Wiechmann D. & Fuhs S. (2006) Concordancing Software. *Corpus Linguistics and Linguistic Theory* 2 (1):107-127.

Wilkinson M. (2011) WordSmith Tools: The Best Corpus Analysis Program for Translators? *Translation Journal* 15 (3). Available from http://translationjournal.net/journal/57corpus.htm

Wilkinson M. (2012) The Best Freeware Corpus Analysis Program for Translators? *Translation Journal* 16 (2). Available from http://www.bokorlang.com/journal/60corpus.htm

Wills W. (1982) *The Science of Translation: Problems and Methods*. Tübingen: Gunter Narr.

Wittgenstein L. (1953) *Philosophical Investigations*. New York: Macmillan.

Wray A. & Perkins M. R. (2000) The Functions of Formulaic Language: An Integrated

Approach. *Language & Communication* (20):1-28.

Wu J.-C., Yeh K. C., Chuang T. C., et al. (2003) TotalRecall: A Bilingual Concordancer for Computer Assisted Translation and Language Learning. *Proceedings of 41st Annual Meeting of the Association of Computational Linguistics - ACL '03, vol.2*, Sapporo, Japan, 7-12 July, Stroudsburg, PA: ACL, 201-204. Available from https://www.aclweb.org/anthology-new/P/P03/P03-2040.pdf

Zetzsche J. (2012) The Year of the Translator. *ATA Chronicle* (May), 31-33.

Zhang Y., Jansen B. J. & Spink A. (2009) Identification of Factors Predicting Clickthrough in Web Searching Using Neural Network Analysis. *Journal of the American Society for Information Science and Technology* 60 (2):1-14.

Zhu C. (1999) UT Once More: The Sentence as the Key Functional Unit of Translation. *Meta* 44 (3):429-447.

APPENDIX

APPENDIX A

The following tables refer to the analysis presented in Chapter 5. Tables A.1 to A.3 contain the same information fields but each table refers to a different subset, i.e. level of analysis: (A.1) full language subset; (A.2) search sessions; (A.3) spot searches.

The first column contains the list of target languages whereas the second shows the size of each language subset in either absolute or relative terms. The columns grouped under the heading 'Institution' show the percentage distribution of each langage across all institutions. The last two columns show the proportion of searches submitted via Quest and the Euramis Portal. At the bottom of each column, some statistics are provided: mean, standard deviation and coefficient of variation.

Table A.4 refers specifically to the session subset and provides the raw number of sessions per language subset. The average session length was then calulated for each language and figures are shown in the third column. The columns grouped under the heading "Search Strategy" show the percentage distribution of the 5 search strategies that could be labeled automatically, i.e. A1, C1, C2, D1, D2, the remaining strings falling under the category "unable". The values are normalized against the total number of sessions for each langauge. In this case, additional statistical measures (range and delta range) are also provided.

TABLE A.1

MAIN SUBSET	Subset size	Institution								Quest	
		Inst_CMS	Inst_EP	Inst_CONSIL	Inst_CDCE	Inst_CDJ	Inst_CDR	Inst_CES	Inst_CDT	Quest_True	Quest_False
BG	33508	61.63%	21.44%	11.53%	1.54%	0.08%	0.96%	2.19%	0.59%	15.82%	84.17%
CS	38064	53.35%	26.12%	11.48%	1.87%	0.06%	5.03%	0.68%	1.38%	70.06%	29.93%
DA	31266	56.24%	21.00%	12.95%	4.17%	0.50%	0.34%	3.93%	0.84%	84.44%	15.55%
DE	47617	71.00%	15.27%	1.33%	1.64%	1.89%	2.10%	4.22%	2.51%	74.75%	25.24%
EL	27812	66.21%	9.60%	5.90%	2.77%	3.12%	3.04%	6.64%	2.68%	86.69%	13.30%
ES	25880	79.89%	8.08%	2.33%	4.76%	0.13%	1.03%	2.80%	0.95%	63.05%	36.94%
ET	47403	56.69%	26.04%	6.70%	3.59%	0.78%	1.96%	3.75%	0.44%	63.05%	36.94%
FI	26765	74.55%	16.14%	1.45%	2.31%	0%	1.05%	3.30%	1.17%	75.54%	24.45%
FR	76049	69.58%	12.38%	4.62%	1.68%	4.77%	1.23%	3.65%	2.05%	70.50%	29.49%
HU	38512	69.31%	22.75%	5.28%	0.97%	0.30%	0.33%	0.42%	0.60%	66.57%	33.42%
IT	29270	72.09%	12.82%	6.29%	1.60%	0.45%	0.80%	2.34%	3.56%	64.75%	35.24%
LT	43942	62.50%	23.64%	5.44%	1.95%	0.59%	2.59%	1.42%	1.84%	43.76%	56.23%
LV	29407	51.29%	16.85%	15.74%	3.24%	0.58%	7.23%	2.83%	2.18%	44.29%	55.70%
NL	23594	65.80%	20.44%	4.26%	2.17%	0.50%	2.34%	4.14%	0.31%	84.48%	15.51%
PL	43431	63.87%	25.75%	2.30%	0.97%	0.19%	4.77%	1.72%	0.39%	83.86%	16.13%
PT	24173	66.66%	16.73%	9.47%	0.51%	0%	0.39%	4.08%	2.12%	33.52%	66.47%
RO	35075	57.58%	24.88%	6.81%	2.71%	0.13%	2.80%	4.06%	0.98%	57.17%	42.82%
SK	30422	58.53%	29.25%	4.91%	3.08%	0%	1.59%	1.13%	1.47%	55.58%	44.41%
SL	38527	51.84%	25.88%	15.02%	0.88%	0.73%	3.21%	1.46%	0.93%	51.43%	48.56%
SV	33826	49.90%	25.04%	12.97%	1.17%	0.29%	7.08%	2.59%	0.93%	75.27%	24.72%
MEAN	24621.35	62.93%	20.01%	7.34%	2.18%	0.75%	2.49%	2.87%	1.40%	63.23%	36.76%
ST_DEV	9545.59	8.41%	6.15%	4.52%	1.15%	1.20%	2.08%	1.51%	0.89%	18.53%	18.53%
CV	0.388	0.134	0.307	0.616	0.527	1.593	0.833	0.526	0.634	0.293	0.504

TABLE A.2

SESSION SUBSET	Subset size	Institution								Quest	
		Inst_CMS	Inst_EP	Inst_CONSIL	Inst_CDCE	Inst_CDJ	Inst_CDR	Inst_CES	Inst_CDT	Quest_True	Quest_False
BG	46.67%	66.29%	17.59%	13.13%	1.23%	0.04%	0.51%	0.90%	0.32%	10.42%	89.57%
CS	34.63%	59.73%	24.32%	8.54%	1.71%	0.05%	4.40%	0.39%	0.86%	60.30%	39.69%
DA	31.78%	61.99%	16.99%	11.61%	4.43%	0.44%	0.35%	2.56%	1.63%	79.99%	20.00%
DE	31.01%	71.38%	15.52%	0.70%	1.59%	2.20%	1.54%	4.07%	3.01%	70.74%	29.25%
EL	33.83%	66.51%	8.63%	5.72%	3.05%	2.93%	3.22%	6.53%	3.41%	86.90%	13.09%
ES	37.13%	80.33%	6.71%	1.82%	6.56%	0.12%	0.88%	2.75%	0.82%	60.69%	39.30%
ET	34.75%	60.70%	24.75%	6.52%	3.16%	0.47%	1.36%	2.72%	0.33%	50.98%	49.01%
FI	35.36%	76.75%	17.64%	0.70%	1.35%	0%	0.48%	2.07%	1.01%	72.81%	27.18%
FR	25.58%	71.74%	11.09%	3.82%	1.71%	4.74%	1.19%	3.93%	1.78%	64.95%	35.04%
HU	36.99%	73.64%	19.65%	5.00%	0.74%	0.20%	0.15%	0.08%	0.53%	60.26%	39.73%
IT	36.63%	72.87%	12.81%	6.22%	1.62%	0.82%	0.42%	1.82%	3.42%	61.85%	38.14%
LT	41.31%	66.28%	24.78%	3.81%	1.63%	0.47%	1.32%	0.82%	0.91%	31.38%	68.61%
LV	38.45%	53.80%	16.25%	15.75%	2.57%	0.28%	7.16%	2.71%	1.48%	36.34%	63.65%
NL	34.86%	70.82%	17.99%	3.89%	1.45%	0.64%	2.12%	2.98%	0.11%	84.52%	15.47%
PL	33.79%	67.40%	25.76%	1.21%	0.83%	0.16%	2.86%	1.49%	0.30%	79.27%	20.72%
PT	40.93%	71.50%	16.09%	8.15%	0.37%	0%	0.13%	2.35%	1.40%	26.28%	73.71%
RO	37.76%	61.35%	24.38%	6.92%	2.95%	0.11%	1.94%	1.68%	0.67%	45.08%	54.91%
SK	39.41%	59.80%	28.93%	4.93%	4.01%	0%	0.95%	0.46%	0.92%	45.23%	54.76%
SL	34.88%	52.54%	27.06%	15.01%	0.60%	0.53%	2.37%	1.15%	0.73%	43.07%	56.92%
SV	35.92%	50.04%	27.91%	11.84%	1.02%	0.21%	6.55%	1.60%	0.84%	68.34%	31.65%
MEAN	36.08%	65.77%	19.24%	6.76%	2.13%	0.72%	1.99%	2.15%	1.22%	56.97%	43.02%
ST_DEV	4.33%	8.18%	6.52%	4.62%	1.53%	1.21%	2.00%	1.51%	0.99%	20.61%	20.61%
CV	0.120	0.124	0.339	0.684	0.719	1.677	1.004	0.703	0.812	0.362	0.479

TABLE A.3

SPOT SUBSET	Subset size	Institution								Quest	
		Inst_CMS	Inst_EP	Inst_CONSIL	Inst_CDCE	Inst_CDJ	Inst_CDR	Inst_CES	Inst_CDT	Quest_True	Quest_False
BG	53.33%	57.55%	24.82%	10.14%	1.82%	0.12%	1.36%	3.31%	0.83%	20.54%	79.45%
CS	65.37%	49.97%	27.07%	13.04%	1.95%	0.07%	5.36%	0.84%	1.66%	75.24%	24.75%
DA	68.22%	53.56%	22.87%	13.57%	4.05%	0.53%	0.34%	4.57%	0.48%	86.51%	13.48%
DE	68.99%	70.84%	15.16%	1.61%	1.66%	1.75%	2.36%	4.30%	2.29%	76.55%	23.44%
EL	66.16%	66.06%	10.10%	6.00%	2.63%	3.22%	2.95%	6.70%	2.31%	86.58%	13.41%
ES	62.87%	79.63%	8.89%	2.63%	3.70%	0.13%	1.11%	2.84%	1.03%	64.44%	35.55%
ET	65.24%	54.57%	26.73%	6.80%	3.82%	0.95%	2.28%	4.30%	0.50%	69.48%	30.51%
FI	64.64%	73.35%	15.31%	1.87%	2.83%	0%	1.37%	3.98%	1.26%	77.03%	22.96%
FR	74.42%	68.84%	12.82%	4.90%	1.67%	4.79%	1.25%	3.56%	2.14%	72.40%	27.59%
HU	63.00%	66.78%	24.57%	5.45%	1.12%	0.36%	0.43%	0.62%	0.64%	70.27%	29.72%
IT	63.37%	71.65%	12.83%	6.33%	1.59%	0.24%	1.03%	2.64%	3.64%	66.43%	33.56%
LT	58.69%	59.84%	22.85%	6.60%	2.17%	0.68%	3.48%	1.84%	2.50%	52.48%	47.51%
LV	61.55%	49.73%	17.24%	15.73%	3.66%	0.77%	7.29%	2.92%	2.62%	49.26%	50.73%
NL	65.03%	63.19%	21.66%	4.47%	2.57%	0.43%	2.47%	4.77%	0.42%	84.44%	15.55%
PL	66.21%	62.07%	25.74%	2.86%	1.04%	0.21%	5.76%	1.84%	0.43%	86.21%	13.78%
PT	59.06%	63.31%	17.18%	10.40%	0.61%	0%	0.58%	5.27%	2.62%	38.54%	61.45%
RO	62.24%	55.29%	25.18%	6.75%	2.57%	0.15%	3.33%	5.51%	1.17%	64.51%	35.48%
SK	60.59%	57.72%	29.45%	4.90%	2.47%	0.01%	2.00%	1.57%	1.84%	62.31%	37.68%
SL	65.10%	51.47%	25.26%	15.03%	1.04%	0.84%	3.66%	1.62%	1.05%	55.92%	44.07%
SV	64.08%	49.82%	23.44%	13.61%	1.25%	0.34%	7.37%	3.15%	0.98%	79.15%	20.84%
MEAN	63.91%	61.26%	20.46%	7.63%	2.21%	0.78%	2.79%	3.31%	1.52%	66.91%	33.08%
ST_DEV	4.33%	8.76%	6.19%	4.52%	1.02%	1.21%	2.15%	1.63%	0.93%	17.08%	17.08%
CV	0.068	0.143	0.303	0.592	0.462	1.551	0.772	0.494	0.609	0.255	0.516

TABLE A.4

SESSION SUBSET	Sessions		Search Strategy (tot#_sessions)					
	tot#_session	Av_length	Cat_A1	Cat_C1	Cat_C2	Cat_D1	Cat_D2	Cat_unable
BG	6029	2.59	38.97%	15.82%	17.26%	1.02%	1.64%	25.26%
CS	5530	2.38	21.37%	22.45%	20.43%	1.89%	2.45%	31.37%
DA	4276	2.32	12.27%	24.08%	24.34%	2.31%	3.22%	33.74%
DE	6406	2.31	17.17%	21.83%	21.90%	2.09%	2.76%	34.23%
EL	4016	2.34	8.16%	26.71%	25.67%	2.01%	3.01%	34.41%
ES	4014	2.39	15.27%	24.36%	22.94%	2.84%	3.13%	31.43%
ET	6977	2.36	17.77%	25.31%	23.54%	1.47%	1.87%	30.01%
FI	4060	2.33	9.13%	25.61%	23.05%	1.77%	2.46%	37.95%
FR	8582	2.27	10.73%	25.94%	22.25%	1.75%	1.85%	37.45%
HU	5892	2.42	20.48%	21.99%	20.60%	2.13%	2.57%	32.19%
IT	4504	2.38	14.09%	23.71%	22.73%	2.35%	3.04%	34.05%
LT	7462	2.43	32.61%	20.97%	18.42%	1.03%	1.32%	25.62%
LV	4740	2.39	23.20%	23.37%	21.35%	1.41%	1.89%	28.75%
NL	3441	2.39	10.69%	23.01%	20.22%	2.49%	3.37%	40.19%
PL	6243	2.35	14.48%	24.93%	22.21%	2.16%	3.10%	33.09%
PT	4025	2.46	18.38%	23.03%	20.96%	1.66%	2.40%	33.54%
RO	5500	2.41	25.01%	20.72%	19.90%	1.96%	2.21%	30.16%
SK	4971	2.41	27.78%	21.98%	18.86%	1.79%	2.27%	27.29%
SL	5661	2.37	18.65%	22.84%	22.27%	1.71%	2.56%	31.95%
SV	5071	2.40	21.10%	21.33%	19.62%	1.87%	2.28%	33.78%
MEAN	5370	2.39	18.87%	23.00%	21.43%	1.89%	2.47%	32.32%
ST_DEV	1326.96	0.07	7.91%	2.40%	2.07%	0.45%	0.57%	3.85%
CV	0.247	0.028	0.419	0.104	0.097	0.239	0.231	0.119
RANGE			38.97-8.16	26.71-15.82	25.67-18.42	2.84-1.03	3.37-1.32	40.19-25.62
DELTA RANGE			30.8	10.88	7.24	1.8	2.04	14.56

APPENDIX B

The following tables refer to the analysis presented in the first part of Chapter 6. Tables B.1 to B.3 contain the same information fields but each table refers to a different subset, i.e. level of analysis: (B.1) full language subset; (B.2) search sessions; (B.3) spot searches.

The first column contains the list of target languages whereas the second shows the average length of a search query (in number of words) when considering the whole subet (i.e. token count). The columns grouped under the heading "n-grams" show the percentage distribution of each n-gram group (i.e. 1 word, 2-11 words, 11+ words) for each langauge. The last group of columns ("Inst.") provides the distribution of average string length for each language across all institutions. At the bottom of each column, some statistics are provided: mean, standard deviation and coefficient of variation.

Table B.1 has an additional column showing the distribution of average string length per language if string-types are used instead of tokens.

TABLE B.1

MAIN SUBSET	Av_Length_TK	n-grams (TOKEN count)			inst (TOKEN count)								Av_Length_TYPE
		1-grams	2/11-grams	>11-grams	cdce:	cdj:	cdt:	ces:	cdr:	cms:	ep:	consil:	
BG	3.89	13.54%	82.11%	4.35%	2.78	2.61	2.94	2.99	4.35	3.55	4.59	4.75	3.06
CS	3.42	13.05%	84.53%	2.42%	3.16	2.50	2.91	3.12	2.99	3.29	3.49	4.20	2.88
DA	3.36	14.39%	83.65%	1.96%	2.68	3.80	3.15	2.88	2.49	3.25	3.34	4.28	2.96
DE	3.60	12.71%	84.61%	2.68%	2.40	3.95	6.25	3.19	2.50	3.64	3.37	3.04	3.14
EL	3.49	16.52%	80.88%	2.60%	3.15	4.74	2.78	2.72	3.82	3.46	3.25	4.75	3.18
ES	3.35	16.36%	81.48%	2.16%	3.53	4.21	2.47	3.17	2.65	3.34	3.17	4.59	2.93
ET	3.51	14.60%	82.43%	2.97%	3.66	3.77	2.79	2.74	3.04	3.51	3.40	4.43	2.99
FI	3.46	13.17%	84.52%	2.31%	2.63	0.00	2.90	2.97	2.89	3.48	3.65	3.33	3.12
FR	3.56	10.52%	87.52%	1.95%	3.08	2.90	3.16	3.02	3.28	3.60	3.73	4.12	3.14
HU	3.67	12.87%	83.90%	3.22%	3.32	4.26	3.13	2.99	3.18	3.76	3.46	3.50	3.09
IT	3.61	12.57%	84.58%	2.84%	3.16	7.62	2.49	2.82	2.80	3.56	4.18	3.94	3.16
LT	3.57	13.51%	83.47%	3.02%	2.60	3.43	2.63	3.06	2.96	3.51	3.92	3.76	2.83
LV	3.66	14.97%	81.61%	3.42%	2.66	5.50	2.86	3.11	2.77	3.42	4.17	4.68	3.13
NL	3.58	10.86%	86.80%	2.34%	2.75	2.99	2.36	2.69	2.81	3.69	3.55	4.07	3.25
PL	3.44	14.56%	82.94%	2.50%	3.48	2.73	2.59	2.81	2.89	3.27	3.99	3.60	2.87
PT	3.99	13.53%	82.44%	4.03%	3.23	0.00	2.50	3.02	2.40	3.78	5.23	4.11	3.56
RO	3.54	15.97%	81.02%	3.01%	2.92	3.06	2.46	2.90	3.27	3.35	4.14	3.78	2.88
SK	3.36	15.84%	81.78%	2.37%	4.64	3.00	2.61	2.57	3.53	3.16	3.62	3.86	2.73
SL	3.57	14.55%	82.02%	3.43%	3.38	3.43	2.38	3.74	2.79	3.31	4.25	3.56	3.05
SV	3.53	12.35%	85.24%	2.41%	3.04	3.58	3.30	3.46	3.42	3.31	3.89	3.84	3.01
MEAN	3.56	13.82%	83.38%	2.80%	3.11	3.40	2.93	3.00	3.04	3.46	3.82	4.01	3.05
ST_DEV	0.16	1.66%	1.85%	0.65%	0.50	1.65	0.83	0.27	0.48	0.18	0.51	0.48	0.18
CV	0.045	0.120	0.022	0.230	0.161	0.485	0.283	0.089	0.157	0.051	0.133	0.120	0.060

TABLE B.2

SESSION SUBSET	Av_Length TK	n-grams (TOKEN count)			inst (TOKEN count)							
		1-grams	2/11-grams	>11-grams	cdce:	cdj:	cdt:	ces:	cdr:	cms:	ep:	consil:
BG	4.21	8.25%	85.77%	5.98%	2.88	3.33	2.94	2.98	5.10	3.76	5.33	5.25
CS	3.78	8.17%	88.65%	3.18%	3.83	2.17	3.08	4.43	3.36	3.63	3.94	4.64
DA	3.82	8.06%	89.22%	2.71%	3.14	4.34	3.44	3.26	2.60	3.63	4.15	4.82
DE	4.29	6.82%	88.94%	4.25%	2.52	4.30	9.44	3.57	2.71	4.29	3.85	3.31
EL	4.12	9.25%	87.07%	3.67%	3.59	5.82	3.09	3.23	4.78	4.08	3.71	5.88
ES	3.78	9.56%	87.35%	3.08%	3.69	5.33	2.62	3.59	2.67	3.83	3.44	4.63
ET	3.96	8.62%	87.23%	4.15%	4.18	4.75	3.02	2.93	3.57	3.92	4.05	4.39
FI	3.91	7.24%	89.60%	3.16%	2.99	0.00	3.69	3.27	3.82	3.93	3.96	3.29
FR	4.01	5.42%	92.00%	2.58%	3.33	3.30	3.32	3.04	3.81	4.13	3.98	4.58
HU	4.12	7.47%	87.99%	4.54%	3.58	3.69	3.51	3.36	2.68	4.23	3.84	3.78
IT	4.11	7.14%	89.09%	3.77%	3.30	8.13	2.68	3.25	3.44	4.10	4.68	3.90
LT	3.94	7.89%	87.89%	4.22%	2.99	3.91	2.82	3.81	3.64	3.82	4.36	4.23
LV	4.17	9.75%	85.42%	4.83%	2.78	11.75	3.19	3.49	3.42	3.97	4.69	4.98
NL	3.97	6.43%	90.53%	3.05%	3.12	3.09	2.22	2.90	3.11	4.03	3.98	4.49
PL	3.87	8.57%	88.19%	3.24%	3.23	2.87	2.48	2.92	3.23	3.70	4.49	3.89
PT	4.45	7.69%	86.51%	5.80%	4.00	0.00	2.53	3.29	3.77	4.20	5.95	4.34
RO	4.05	8.77%	86.63%	4.60%	3.17	4.36	2.92	3.51	3.59	3.76	4.89	4.42
SK	3.78	8.99%	87.73%	3.28%	4.78	0.00	2.85	3.40	3.57	3.56	4.04	4.36
SL	4.14	9.08%	85.70%	5.22%	4.59	3.99	2.79	4.59	3.04	3.80	5.03	3.91
SV	3.96	7.03%	89.77%	3.20%	3.56	4.24	3.80	3.34	3.77	3.71	4.37	4.25
MEAN	4.02	8.01%	88.06%	3.93%	3.46	3.97	3.32	3.41	3.48	3.90	4.34	4.37
ST_DEV	0.18	1.12%	1.70%	0.99%	0.59	2.69	1.50	0.45	0.64	0.22	0.61	0.61
CV	0.045	0.140	0.019	0.253	0.170	0.678	0.451	0.132	0.184	0.056	0.141	0.141

TABLE B.3

SPOT SUBSET	Av_Length TK	n-grams (TOKEN count)			inst (TOKEN count)							
		1-grams	2/11-grams	>11-grams	cdce:	cdj:	cdt:	ces:	cdr:	cms:	ep:	consil:
BG	3.61	14.70%	81.27%	4.03%	2.72	2.41	2.94	2.99	4.11	3.35	4.14	4.17
CS	3.24	13.53%	84.01%	2.46%	2.85	2.61	2.87	2.80	2.83	3.08	3.28	4.05
DA	3.15	15.15%	83.10%	1.75%	2.44	3.59	2.69	2.78	2.44	3.04	3.05	4.06
DE	3.29	12.58%	85.09%	2.33%	2.35	3.75	4.37	3.03	2.43	3.35	3.14	2.99
EL	3.17	18.01%	79.64%	2.35%	2.89	4.24	2.56	2.47	3.28	3.15	3.05	4.20
ES	3.09	17.65%	80.47%	1.89%	3.37	3.59	2.40	2.92	2.64	3.05	3.05	4.58
ET	3.27	14.82%	82.29%	2.89%	3.44	3.51	2.71	2.68	2.87	3.27	3.08	4.45
FI	3.21	14.57%	83.30%	2.13%	2.53	0.00	2.55	2.88	2.71	3.22	3.45	3.34
FR	3.41	10.47%	87.60%	1.93%	2.99	2.77	3.12	3.01	3.11	3.41	3.66	3.99
HU	3.40	13.74%	83.25%	3.01%	3.22	4.44	2.94	2.97	3.28	3.45	3.28	3.36
IT	3.32	13.67%	83.82%	2.51%	3.08	6.65	2.39	2.65	2.65	3.24	3.89	3.97
LT	3.31	15.05%	82.02%	2.94%	2.39	3.20	2.59	2.83	2.78	3.28	3.59	3.58
LV	3.34	16.25%	80.58%	3.17%	2.60	4.08	2.74	2.89	2.37	3.04	3.86	4.50
NL	3.38	12.02%	85.84%	2.14%	2.64	2.91	2.38	2.61	2.68	3.48	3.37	3.87
PL	3.22	14.96%	82.43%	2.61%	3.59	2.68	2.63	2.76	2.80	3.04	3.74	3.53
PT	3.67	15.50%	81.01%	3.49%	2.91	0.00	2.49	2.94	2.18	3.46	4.76	3.99
RO	3.23	17.18%	80.08%	2.73%	2.74	2.53	2.30	2.78	3.15	3.08	3.71	3.39
SK	3.09	17.12%	80.70%	2.18%	4.50	3.00	2.54	2.41	3.52	2.88	3.36	3.53
SL	3.27	14.94%	82.13%	2.92%	3.00	3.25	2.23	3.41	2.71	3.03	3.79	3.38
SV	3.30	12.94%	84.84%	2.22%	2.80	3.36	3.06	3.49	3.25	3.08	3.57	3.65
MEAN	3.30	14.74%	82.67%	2.58%	2.95	3.13	2.72	2.87	2.89	3.20	3.54	3.83
ST_DEV	0.15	1.94%	2.10%	0.57%	0.50	1.42	0.46	0.26	0.45	0.17	0.43	0.44
CV	0.045	0.131	0.025	0.222	0.170	0.455	0.169	0.092	0.156	0.055	0.121	0.114

APPENDIX C

The following tables refer to the analysis presented in the second part of Chapter 6.

Table C.1 shows the percentage of successful and unsuccessful searches across languages at each level of analysis. In addtion, in the case of search sessions a further specification is provided as to how often the first and second searches of a session were successful (i.e. produced results). In many cases, the first two searches correspond to the full length of the session.

In Table C.2, the small table in the top right corner contains the breakdown of execution time per language but only for the main dataset. Just like C.1, the main table on the page compares the percentage of successful and unsuccessful searches across languages and levels of analysis. The first two columns show the distribution of successful and unsuccessful searches whereas the last three show the distribution of searches for which (i) less than 30, (ii) exactly 30 or (iii) more than 30 results were returned.

Tables C.3 to C.5 and C.6 to C.8 contain the same information fields but each table refers to a different subset, i.e. level of analysis: (C.3/C.6) full language subset; (C.4/C.7) search sessions; (C.5/C.8) spot searches.
Tables C.3 to C.5 show the distribution of searches according to Search Mode (Simple/Advanced), Search Method (Basic/AnyWord/Exact) and Search Direction (Indirect/Reverse/Direct).
Tables C.6 to C.8 provide the following distributions of search metadata and filters: Database, Requesting Institution (DG), Document Type and Number, Maximum Number of Results and Year.

At the bottom of each column, some statistics are provided: mean, standard deviation and coefficient of variation.

TABLE C.1

	MAIN SUBSET		SESSION				SPOT	
	Results_NO	Results_YES	Results_NO	Results_YES	Results1st	Results_2nd	Results_NO	Results_YES
BG	30.04%	69.95%	41.85%	58.14%	39.29%	69.96%	19.70%	80.29%
CS	30.58%	69.41%	49.71%	50.28%	30.05%	64.91%	20.45%	79.54%
DA	31.03%	68.96%	51.89%	48.10%	27.43%	63.71%	21.30%	78.69%
DE	31.07%	68.92%	51.42%	48.57%	27.78%	64.50%	21.93%	78.06%
EL	30.96%	69.03%	52.40%	47.59%	25.17%	64.28%	20.00%	79.99%
ES	31.63%	68.36%	51.40%	48.59%	27.55%	63.68%	19.95%	80.04%
ET	30.43%	69.56%	50.84%	49.15%	26.67%	65.66%	19.56%	80.43%
FI	32.35%	67.64%	52.05%	47.94%	25.59%	64.73%	21.57%	78.42%
FR	33.69%	66.30%	56.76%	43.23%	19.84%	61.69%	25.76%	74.23%
HU	30.54%	69.45%	47.17%	52.82%	33.12%	66.71%	20.77%	79.22%
IT	31.98%	68.01%	51.00%	48.99%	28.17%	64.07%	20.98%	79.01%
LT	33.04%	66.95%	49.79%	50.20%	31.19%	63.47%	21.25%	78.74%
LV	30.84%	69.15%	49.33%	50.66%	29.68%	65.80%	19.30%	80.69%
NL	33.03%	66.96%	51.32%	48.67%	29.23%	62.65%	23.15%	76.84%
PL	29.76%	70.23%	47.94%	52.05%	32.13%	66.79%	20.48%	79.51%
PT	35.98%	64.01%	54.07%	45.92%	25.54%	59.90%	23.43%	76.56%
RO	30.60%	69.39%	49.75%	50.24%	32.63%	62.74%	18.98%	81.01%
SK	29.32%	70.67%	45.36%	54.63%	35.78%	67.99%	18.89%	81.10%
SL	29.26%	70.73%	49.87%	50.12%	29.00%	65.49%	18.22%	81.77%
SV	30.44%	69.55%	47.50%	52.49%	32.63%	66.72%	20.87%	79.12%
MEAN	31.33%	68.66%	50.07%	49.92%	29.42%	64.77%	20.83%	79.16%
ST_DEV	1.64%	1.64%	3.16%	3.16%	4.24%	2.27%	1.77%	1.77%
CV	0.052	0.024	0.063	0.063	0.144	0.035	0.085	0.022

TABLE C.2

RESULTS	MAIN SUBSET					SESSION SUBSET					SPOT SUBSET				
	Results_NO	Results_YES	Res_1-29	Res_30	Res_30+	Results_NO	Results_YES	Res_1-29	Res_30	Res_30+	Results_NO	Results_YES	Res_1-29	Res_30	Res_30+
BG	30.04%	69.95%	28.81%	37.90%	3.23%	41.85%	58.14%	20.91%	34.44%	2.78%	19.70%	80.29%	35.73%	40.93%	3.63%
CS	30.58%	69.41%	28.25%	39.68%	1.47%	49.71%	50.28%	19.84%	29.14%	1.30%	20.45%	79.54%	32.71%	45.27%	1.55%
DA	31.03%	68.96%	26.66%	41.80%	0.50%	51.89%	48.10%	19.60%	28.08%	0.41%	21.30%	78.69%	29.95%	48.19%	0.54%
DE	31.07%	68.92%	27.76%	40.99%	0.17%	51.42%	48.57%	20.76%	27.75%	0.06%	21.93%	78.06%	30.90%	46.94%	0.21%
EL	30.96%	69.03%	28.19%	40.83%	0%	52.40%	47.59%	21.23%	26.35%	0%	20.00%	79.99%	31.74%	48.24%	0%
ES	31.63%	68.36%	27.81%	40.28%	0.26%	51.40%	48.59%	19.90%	28.51%	0.17%	19.95%	80.04%	32.48%	47.24%	0.31%
ET	30.43%	69.56%	26.46%	41.97%	1.12%	50.84%	49.15%	19.18%	29.27%	0.68%	19.56%	80.43%	30.33%	48.74%	1.35%
FI	32.35%	67.64%	27.57%	39.52%	0.54%	52.05%	47.94%	20.10%	27.37%	0.46%	21.57%	78.42%	31.66%	46.17%	0.58%
FR	33.69%	66.30%	29.11%	36.12%	1.06%	56.76%	43.23%	19.77%	22.75%	0.69%	25.76%	74.23%	32.32%	40.72%	1.18%
HU	30.54%	69.45%	28.11%	40.77%	0.57%	47.17%	52.82%	21.64%	30.61%	0.55%	20.77%	79.22%	31.90%	46.73%	0.58%
IT	31.98%	68.01%	29.03%	34.78%	4.19%	51.00%	48.99%	21.16%	24.42%	3.40%	20.98%	79.01%	33.58%	40.76%	4.65%
LT	33.04%	66.95%	28.18%	36.27%	2.48%	49.79%	50.20%	21.69%	25.86%	2.65%	21.25%	78.74%	32.75%	43.61%	2.37%
LV	30.84%	69.15%	27.80%	39.44%	1.89%	49.33%	50.66%	19.86%	29.44%	1.35%	19.30%	80.69%	32.77%	45.69%	2.22%
NL	33.03%	66.96%	27.94%	38.94%	0.07%	51.32%	48.67%	21.38%	27.20%	0.08%	23.15%	76.84%	31.47%	45.30%	0.07%
PL	29.76%	70.23%	26.32%	43.68%	0.22%	47.94%	52.05%	20.63%	31.21%	0.20%	20.48%	79.51%	29.23%	50.04%	0.23%
PT	35.98%	64.01%	26.74%	34.38%	2.89%	54.07%	45.92%	18.59%	25.14%	2.18%	23.43%	76.56%	32.40%	40.80%	3.38%
RO	30.60%	69.39%	27.28%	39.44%	2.67%	49.75%	50.24%	20.30%	27.32%	2.61%	18.98%	81.01%	31.51%	46.79%	2.70%
SK	29.32%	70.67%	26.39%	42.39%	1.87%	45.36%	54.63%	20.57%	32.00%	2.05%	18.89%	81.10%	30.18%	49.15%	1.76%
SL	29.26%	70.73%	26.29%	42.42%	2.00%	49.87%	50.12%	18.62%	29.77%	1.72%	18.22%	81.77%	30.39%	49.22%	2.16%
SV	30.44%	69.55%	27.71%	39.37%	2.46%	47.50%	52.49%	21.79%	27.29%	3.40%	20.87%	79.12%	31.03%	46.14%	1.94%
MEAN	31.33%	68.66%	27.62%	39.55%	1.48%	50.07%	49.92%	20.38%	28.20%	1.34%	20.83%	79.16%	31.75%	45.83%	1.57%
ST_DEV	1.64%	1.64%	0.90%	2.57%	1.22%	3.16%	3.16%	0.97%	2.72%	1.17%	1.77%	1.77%	1.46%	3.01%	1.32%
CV	0.052	0.024	0.032	0.065	0.823	0.063	0.063	0.047	0.096	0.876	0.085	0.022	0.046	0.066	0.840

TABLE C.3

MAIN SUBSET	Search mode		Search Method			Direction		
	Mode_S	Mode_A	Basic_(S.mode/Qst)	Any_(A.mode/Qst)	Exact_(A.mode)	Dir_I(A/S.mode)	Dir_R(A.mode)	Dir_D(A.mode)
BG	44.98%	55.01%	85.27%	9.78%	4.93%	99.29%	0%	0.71%
CS	78.91%	21.08%	95.38%	3.24%	1.36%	98.94%	0.33%	0.73%
DA	92.50%	7.49%	89.34%	10.35%	0.30%	89.64%	0%	10.36%
DE	91.44%	8.55%	95.51%	3.97%	0.51%	95.47%	2.42%	2.11%
EL	93.33%	6.66%	91.51%	7.37%	1.11%	99.63%	0%	0.37%
ES	88.84%	11.15%	94.08%	3.01%	2.89%	95.47%	0%	4.53%
ET	79.03%	20.96%	88.46%	10.48%	1.05%	95.95%	0%	4.05%
FI	88.63%	11.36%	95.49%	4.20%	0.29%	99.15%	0%	0.85%
FR	87.44%	12.55%	96.50%	2.63%	0.86%	98.52%	0.04%	1.44%
HU	84.62%	15.37%	89.16%	9.21%	1.62%	99.17%	0.70%	0.13%
IT	79.74%	20.25%	96.97%	2.22%	0.79%	93.51%	0.10%	6.39%
LT	55.68%	44.31%	88.60%	7.12%	4.27%	99.89%	0%	0.11%
LV	66.68%	33.31%	93.75%	4.18%	2.06%	99.99%	0%	0.01%
NL	97.99%	2.00%	94.03%	5.57%	0.38%	100.00%	0%	0%
PL	95.23%	4.76%	93.41%	5.62%	0.96%	98.26%	0%	1.74%
PT	67.40%	32.59%	93.94%	5.28%	0.76%	95.35%	0%	4.65%
RO	65.47%	34.52%	93.74%	4.11%	2.14%	97.21%	0%	2.79%
SK	73.05%	26.94%	93.79%	3.92%	2.27%	99.20%	0%	0.80%
SL	71.68%	28.31%	91.26%	6.55%	2.17%	87.67%	0%	12.33%
SV	85.15%	14.84%	91.54%	5.00%	3.45%	99.76%	0.07%	0.17%
MEAN	79.39%	20.60%	92.59%	5.69%	1.71%	97.10%	0.18%	2.71%
ST_DEV	14.06%	14.06%	3.10%	2.59%	1.32%	3.48%	0.55%	3.49%
CV	0.177	0.682	0.034	0.456	0.775	0.036	3.020	1.286

TABLE C.4

SESSION SUBSET	Search mode		Search Method			Direction		
	Mode_S	Mode_A	Basic_(S.mode/Qst)	Any_(A.mode/Qst)	Exact_(A.mode)	Dir_I(A/S.mode)	Dir_R(A.mode)	Dir_D(A.mode)
BG	36.97%	63.02%	82.22%	12.61%	5.16%	99.35%	0%	0.65%
CS	72.04%	27.95%	94.03%	4.22%	1.73%	99.27%	0.27%	0.46%
DA	89.57%	10.42%	91.19%	8.37%	0.43%	94.89%	0%	5.11%
DE	89.66%	10.33%	95.27%	4.14%	0.57%	95.88%	1.63%	2.49%
EL	93.47%	6.52%	90.81%	7.93%	1.24%	99.84%	0%	0.16%
ES	88.45%	11.54%	93.77%	2.64%	3.57%	94.59%	0%	5.41%
ET	72.22%	27.77%	89.84%	8.69%	1.46%	97.14%	0%	2.86%
FI	87.68%	12.31%	94.87%	4.89%	0.23%	99.58%	0%	0.42%
FR	83.63%	16.36%	95.96%	2.78%	1.24%	99.13%	0%	0.87%
HU	79.75%	20.24%	83.65%	14.02%	2.31%	98.90%	0.93%	0.17%
IT	77.01%	22.98%	95.63%	3.26%	1.10%	92.01%	0.12%	7.87%
LT	43.46%	56.53%	86.09%	9.01%	4.88%	99.96%	0%	0.04%
LV	60.58%	39.41%	93.31%	4.01%	2.67%	99.98%	0%	0.02%
NL	97.73%	2.26%	94.26%	5.10%	0.63%	100.00%	0%	0%
PL	93.01%	6.98%	92.81%	5.99%	1.18%	98.15%	0%	1.85%
PT	65.81%	34.18%	93.03%	6.07%	0.88%	93.94%	0%	6.06%
RO	53.83%	46.16%	91.66%	5.31%	3.01%	97.09%	0%	2.91%
SK	66.56%	33.43%	91.88%	4.88%	3.22%	99.46%	0%	0.54%
SL	64.50%	35.49%	92.58%	5.14%	2.27%	86.04%	0%	13.96%
SV	79.52%	20.47%	89.43%	6.79%	3.77%	99.84%	0.12%	0.04%
MEAN	74.77%	25.22%	91.61%	6.29%	2.08%	97.25%	0.15%	2.59%
ST_DEV	16.94%	16.94%	3.80%	3.05%	1.46%	3.56%	0.41%	3.57%
CV	0.227	0.672	0.042	0.485	0.704	0.037	2.654	1.377

TABLE C.5

SPOT SUBSET	Search mode		Search Method			Direction		
	Mode_S	Mode_A	Basic_(S.mode/Qst)	Any_(A.mode/Qst)	Exact_(A.mode)	Dir_I(A/S.mode)	Dir_R(A.mode)	Dir_D(A.mode)
BG	52.00%	47.99%	87.95%	7.31%	4.73%	99.24%	0%	0.76%
CS	82.55%	17.44%	96.10%	2.72%	1.16%	98.65%	0.39%	0.96%
DA	93.86%	6.13%	88.48%	11.28%	0.23%	85.48%	0%	14.52%
DE	92.24%	7.75%	95.61%	3.90%	0.48%	95.22%	2.90%	1.88%
EL	93.25%	6.74%	91.86%	7.08%	1.04%	99.52%	0%	0.48%
ES	89.07%	10.92%	94.27%	3.23%	2.48%	96.01%	0%	3.99%
ET	82.65%	17.34%	87.72%	11.44%	0.83%	94.94%	0%	5.06%
FI	89.16%	10.83%	95.83%	3.82%	0.33%	98.88%	0%	1.12%
FR	88.75%	11.24%	96.68%	2.58%	0.73%	98.22%	0.06%	1.72%
HU	87.47%	12.52%	92.39%	6.38%	1.22%	99.42%	0.49%	0.09%
IT	81.31%	18.68%	97.74%	1.62%	0.62%	94.59%	0.08%	5.33%
LT	64.28%	35.71%	90.36%	5.79%	3.84%	99.81%	0%	0.19%
LV	70.49%	29.50%	94.02%	4.29%	1.67%	100.00%	0%	0%
NL	98.13%	1.86%	93.90%	5.83%	0.26%	100.00%	0%	0%
PL	96.37%	3.62%	93.71%	5.43%	0.84%	98.38%	0%	1.62%
PT	68.51%	31.48%	94.58%	4.74%	0.67%	96.42%	0%	3.58%
RO	72.54%	27.45%	95.00%	3.38%	1.61%	97.34%	0%	2.66%
SK	77.26%	22.73%	95.03%	3.30%	1.66%	98.95%	0%	1.05%
SL	75.53%	24.46%	90.55%	7.32%	2.12%	88.92%	0%	11.08%
SV	88.31%	11.68%	92.72%	4.00%	3.26%	99.66%	0.03%	0.31%
MEAN	82.19%	17.80%	93.23%	5.27%	1.49%	96.98%	0.20%	2.82%
ST_DEV	12.00%	12.00%	2.92%	2.65%	1.24%	3.81%	0.65%	3.82%
CV	0.146	0.674	0.031	0.502	0.834	0.039	3.292	1.356

TABLE C.6

MAIN SUBSET	Database (A/S.mode)		DG	DocType	DocNum	MaxResults		Year
	DB_YES_S	DB_YES_A	DG_YES(A.mode)	Type_YES(A.mode)	Num_YES(A.mode)	Max_<=30	Max_>30(A.mode)	Year_YES(A.mode)
BG	2.34%	2.44%	3.11%	1.38%	0.04%	91.16%	8.83%	20.56%
CS	1.23%	4.63%	2.44%	0%	0%	95.86%	4.13%	19.48%
DA	0.81%	0.17%	0.63%	0%	0%	98.64%	1.35%	1.53%
DE	0.59%	15.97%	5.12%	0%	0%	99.43%	0.56%	7.92%
EL	0.05%	2.10%	0.64%	0%	0%	100.00%	0%	1.94%
ES	1.24%	0.41%	18.32%	0%	0%	99.35%	0.64%	14.75%
ET	0.61%	0.12%	1.77%	0.01%	0%	96.92%	3.07%	33.43%
FI	0.14%	0.42%	2.99%	0%	0%	98.53%	1.46%	7.92%
FR	0.96%	1.14%	7.36%	0%	0.01%	97.14%	2.85%	8.38%
HU	0.10%	2.14%	10.26%	0.01%	1.26%	97.58%	2.41%	19.19%
IT	0.61%	5.04%	3.42%	0.06%	0.06%	89.69%	10.30%	10.33%
LT	3.39%	1.08%	2.72%	0%	0.01%	91.98%	8.01%	54.37%
LV	0.23%	0.67%	1.05%	0.02%	0%	95.31%	4.68%	18.23%
NL	2.38%	3.17%	11.44%	0%	0%	99.84%	0.15%	11.44%
PL	0.15%	8.22%	11.36%	0.14%	0.33%	99.29%	0.70%	48.01%
PT	3.32%	0.35%	2.89%	0%	0%	92.28%	7.71%	3.66%
RO	0.42%	1.02%	2.01%	0%	0%	90.71%	9.28%	37.41%
SK	1.24%	4.34%	2.42%	0.03%	0%	94.77%	5.22%	55.91%
SL	0.13%	0.60%	0.54%	0%	0%	94.41%	5.58%	12.49%
SV	0.01%	0.11%	4.56%	0%	0.01%	92.44%	7.55%	10.41%
MEAN	1.00%	2.71%	4.75%	0.08%	0.09%	95.77%	4.22%	19.87%
ST_DEV	1.06%	3.78%	4.69%	0.31%	0.29%	3.40%	3.40%	16.93%
CV	1.062	1.397	0.986	3.724	3.326	0.036	0.805	0.852

TABLE C.7

SESSION SUBSET	Database (A/S.mode)		DG	DocType	DocNum	MaxResults		Year
	DB_YES_S	DB_YES_A	DG_YES(A.mode)	Type_YES(A.mode)	Num_YES(A.mode)	Max_<=30	Max_>30(A.mode)	Year_YES(A.mode)
BG	3.28%	2.66%	3.17%	1.48%	0.05%	90.75%	9.24%	19.97%
CS	2.18%	3.20%	2.74%	0%	0%	95.16%	4.83%	22.27%
DA	1.02%	0.19%	0.48%	0%	0%	98.66%	1.33%	1.15%
DE	1.01%	16.18%	5.11%	0%	0%	99.47%	0.52%	7.01%
EL	0.04%	3.58%	0.65%	0%	0%	100.00%	0%	1.46%
ES	1.39%	0.45%	25.51%	0%	0%	99.29%	0.70%	20.37%
ET	0.57%	0.13%	1.77%	0.02%	0%	97.12%	2.87%	35.43%
FI	0.19%	0.17%	3.60%	0%	0%	98.48%	1.51%	10.54%
FR	1.29%	0.78%	9.55%	0%	0%	97.17%	2.82%	8.01%
HU	0.16%	2.49%	10.88%	0.03%	0.97%	97.30%	2.69%	21.84%
IT	0.54%	6.98%	4.38%	0.04%	0%	88.06%	11.93%	11.44%
LT	5.57%	1.03%	2.42%	0%	0%	89.58%	10.41%	52.31%
LV	0.35%	0.94%	1.09%	0.04%	0%	94.77%	5.22%	19.05%
NL	1.77%	3.76%	13.44%	0%	0%	99.81%	0.18%	13.97%
PL	0.28%	10.92%	9.26%	0.09%	0.19%	99.18%	0.81%	47.31%
PT	3.40%	0.44%	3.63%	0%	0%	92.25%	7.74%	4.31%
RO	0.64%	1.25%	2.42%	0%	0%	87.67%	12.32%	40.37%
SK	2.60%	5.46%	2.96%	0.02%	0%	93.05%	6.94%	56.91%
SL	0.20%	0.73%	0.73%	0%	0%	93.08%	6.91%	14.08%
SV	0.01%	0.16%	4.38%	0%	0.04%	87.28%	12.71%	12.78%
MEAN	1.32%	3.08%	5.41%	0.09%	0.06%	94.91%	5.08%	21.03%
ST_DEV	1.45%	4.14%	5.96%	0.33%	0.22%	4.40%	4.40%	16.76%
CV	1.094	1.347	1.103	3.825	3.488	0.046	0.865	0.797

TABLE C.8

SPOT SUBSET	Database (A/S.mode)		DG	DocType	DocNum	MaxResults		Year
	DB_YES_S	DB_YES_A	DG_YES(A.mode)	Type_YES(A.mode)	Num_YES(A.mode)	Max_<=30	Max_>30(A.mode)	Year_YES(A.mode)
BG	1.76%	2.18%	3.05%	1.28%	0.03%	91.51%	8.48%	21.23%
CS	0.79%	5.84%	2.18%	0%	0%	96.23%	3.76%	17.11%
DA	0.72%	0.15%	0.76%	0%	0%	98.64%	1.35%	1.83%
DE	0.41%	15.84%	5.13%	0%	0%	99.42%	0.57%	8.47%
EL	0.05%	1.36%	0.64%	0%	0%	100.00%	0%	2.17%
ES	1.15%	0.39%	13.83%	0%	0%	99.39%	0.60%	11.24%
ET	0.63%	0.11%	1.77%	0%	0%	96.82%	3.17%	31.72%
FI	0.11%	0.58%	2.61%	0%	0%	98.56%	1.43%	6.29%
FR	0.85%	1.31%	6.26%	0%	0.01%	97.13%	2.86%	8.57%
HU	0.07%	1.81%	9.67%	0%	1.54%	97.74%	2.25%	16.68%
IT	0.64%	3.66%	2.74%	0.08%	0.11%	90.63%	9.36%	9.54%
LT	2.35%	1.15%	3.06%	0.01%	0.02%	93.67%	6.32%	56.66%
LV	0.17%	0.44%	1.01%	0%	0%	95.65%	4.34%	17.54%
NL	2.71%	2.79%	10.13%	0%	0%	99.86%	0.13%	9.79%
PL	0.08%	5.56%	13.42%	0.19%	0.47%	99.35%	0.64%	48.70%
PT	3.27%	0.28%	2.33%	0%	0%	92.31%	7.68%	3.18%
RO	0.32%	0.78%	1.60%	0.01%	0.01%	92.56%	7.43%	34.39%
SK	0.48%	3.26%	1.90%	0.04%	0%	95.89%	4.10%	54.96%
SL	0.10%	0.50%	0.39%	0%	0%	95.12%	4.87%	11.26%
SV	0.02%	0.07%	4.73%	0%	0%	95.33%	4.66%	8.09%
MEAN	0.83%	2.40%	4.36%	0.08%	0.11%	96.29%	3.70%	18.97%
ST_DEV	0.95%	3.60%	4.17%	0.29%	0.35%	2.93%	2.93%	17.22%
CV	1.144	1.500	0.955	3.552	3.223	0.030	0.792	0.907

TABLE C.9

MAIN SUBSET	ExecTime			
	Timeout(-1)	Time_0	Time_1-3	Time_>3
BG	0.66%	50.93%	34.32%	14.73%
CS	0.39%	52.34%	37.62%	10.02%
DA	0.03%	54.10%	37.00%	8.88%
DE	0.09%	55.65%	36.91%	7.42%
EL	0.02%	53.48%	37.66%	8.84%
ES	0.21%	55.77%	37.37%	6.84%
ET	0.29%	52.70%	38.24%	9.05%
FI	0.09%	54.61%	39.28%	6.09%
FR	0.14%	59.26%	34.03%	6.69%
HU	0.32%	52.90%	38.94%	8.15%
IT	0.20%	54.69%	34.63%	10.66%
LT	0.83%	54.54%	32.92%	12.52%
LV	0.51%	51.90%	36.36%	11.73%
NL	0.04%	54.24%	38.44%	7.31%
PL	0.21%	52.60%	39.48%	7.91%
PT	0.14%	58.16%	34.99%	6.84%
RO	0.66%	52.07%	34.16%	13.75%
SK	0.77%	50.85%	37.06%	12.08%
SL	0.27%	51.62%	39.82%	8.54%
SV	0.08%	55.16%	35.99%	8.84%
MEAN	0.30%	53.88%	36.76%	9.34%
ST_DEV	0.26%	2.23%	2.03%	2.48%
CV	0.861	0.041	0.055	0.265

APPENDIX D

The following tables refer to the analysis presented in Chapter 7.

Table D.1 contains a breakdown of the raw amount of string-types and string-tokens for each n-gram group (from 1-gram to 15-grams, individually). Based on these values, the type-token ratio was calculated, which can be found in the third line. The fourth line contains the percentage of hapaxes (i.e. strings with frequency of 1) for each n-gram group. The percentage clearly increases as the string becomes longer.

Tables D.2 to D.4 contain the same information fields but each table refers to a different subset, i.e. level of analysis: (D.2) full language subset; (D.3) search sessions; (D.4) spot searches. The small table in the top corner of the page provides the percentage of strings for each language for which a match with a domain descriptor was found and the second column shows the percentage of the strings with multiple matches. The large table gives an overview of the distribution of matches for each domain across all languages.

At the bottom of each column, some statistics are provided: mean, standard deviation and coefficient of variation.

Due to multiple matches, the actual distribution of each field is tricky to calculate using the previous tables. Table D.5 provides a different perspective of analysis so as to obtain a more precise distribution of descriptor fields. The fist column contains the descriptor fields. The second column contains the distribution of matches for those strings where only one descriptor matched the string content. The remaining columns show at which position in a multiple match the field was matched. This table should give a better idea of the order in which matches are found (the script works on a first-come-first-matched principle). For example, field 32 (Education and Communication) only appears as first match, whereas field 24 (Finance) tends to be matched more often as second. Others, like fields 00 (Eurojargon) and 10 (European Communities), are (much) more frequent as third match rather than as first match. This strongly suggests that these last fields tend to occur in combination with others. The last column shows a more accurate distribution of the descriptor fields which confirms the previous findings as to which domains are most common.

TABLE D.1

Freq	1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	8-grams	9-grams	10-grams	11-grams	12-grams	13-grams	14-grams	15-grams
TOTAL TYPES	24685	92461	75180	53410	32051	19048	12130	7970	5582	4099	3061	2349	1910	1470	1179
TOTAL TOKEN	103412	251211	151000	91113	49521	28277	17569	11803	7648	5578	4322	3282	2517	1980	1504
TYPE/TOKEN	0.24	0.37	0.50	0.59	0.65	0.67	0.69	0.68	0.73	0.73	0.71	0.72	0.76	0.74	0.78
Hapax %	46.42	57.78	68.41	73.56	77.59	78.51	80.02	80.73	82.66	83.34	82.75	83.23	82.88	82.93	84.31

TABLE D.2

MAIN SUBSET	Matches	
	EV_YES	EV_MULTI/ev_yes
BG	58.12%	37.90%
CS	60.97%	37.85%
DA	57.21%	34.79%
DE	58.21%	34.47%
EL	54.92%	34.85%
ES	54.35%	33.27%
ET	57.07%	34.86%
FI	57.87%	33.02%
FR	53.44%	33.20%
HU	60.52%	37.05%
IT	57.86%	34.27%
LT	58.55%	35.17%
LV	57.68%	36.93%
NL	57.07%	35.78%
PL	59.28%	36.55%
PT	58.50%	38.44%
RO	59.94%	39.07%
SK	59.83%	37.87%
SL	57.39%	36.98%
SV	58.52%	36.07%
MEAN	57.87%	35.92%
ST_DEV	1.93%	1.83%
CV	0.033	0.051

MAIN SUBSET	EuroVoc Categories/EV_YES																					
	00	04	08	10	12	16	20	24	28	32	36	40	44	48	52	56	60	64	66	68	72	76
BG	12.20%	14.42%	4.58%	20.07%	14.38%	5.51%	6.40%	13.79%	10.63%	9.24%	1.47%	7.27%	4.93%	3.07%	4.91%	5.67%	1.58%	5.61%	1.94%	3.16%	4.48%	1.87%
CS	11.70%	15.51%	5.81%	24.26%	16.12%	6.74%	8.16%	15.40%	12.12%	9.46%	1.35%	8.43%	5.89%	3.14%	5.89%	3.51%	1.44%	6.56%	3.01%	3.91%	6.31%	2.85%
DA	10.90%	14.37%	4.14%	17.86%	12.71%	4.78%	6.77%	12.27%	8.27%	7.97%	1.00%	6.72%	3.62%	2.41%	3.88%	2.91%	1.08%	5.07%	2.22%	2.88%	4.18%	1.57%
DE	10.04%	14.17%	5.68%	25.11%	20.62%	7.23%	11.73%	18.81%	12.29%	12.25%	1.58%	10.71%	6.15%	4.07%	6.38%	5.08%	1.67%	8.50%	4.29%	4.43%	6.06%	2.14%
EL	11.05%	13.07%	3.39%	12.89%	9.72%	4.07%	5.82%	11.09%	7.30%	7.41%	0.84%	6.84%	2.79%	2.88%	3.59%	3.37%	1.28%	4.49%	1.92%	2.62%	3.39%	1.36%
ES	11.96%	12.45%	2.49%	10.49%	8.45%	4.52%	6.00%	10.39%	7.03%	7.21%	0.69%	5.66%	3.70%	2.38%	2.99%	2.57%	1.29%	3.92%	1.20%	2.36%	3.10%	1.16%
ET	10.78%	13.95%	6.00%	24.59%	16.93%	8.24%	11.12%	17.89%	13.75%	12.08%	1.69%	10.05%	5.30%	4.65%	6.04%	5.46%	1.90%	9.85%	3.99%	4.45%	6.31%	2.58%
FI	10.42%	13.49%	2.84%	12.16%	9.45%	4.45%	6.87%	11.84%	6.53%	6.89%	0.96%	6.34%	3.19%	2.98%	3.21%	3.38%	0.91%	5.19%	1.96%	2.71%	3.26%	1.02%
FR	10.18%	14.24%	7.23%	35.09%	29.52%	11.31%	18.28%	27.51%	17.22%	17.68%	2.21%	15.73%	9.05%	5.64%	9.70%	7.28%	2.88%	12.17%	4.73%	6.64%	9.06%	2.94%
HU	11.49%	13.90%	5.05%	23.28%	15.86%	6.21%	9.97%	16.55%	11.02%	10.54%	1.28%	8.63%	5.17%	3.31%	5.44%	5.68%	2.14%	7.66%	2.90%	3.36%	6.04%	1.81%
IT	10.86%	13.50%	3.53%	15.36%	12.15%	4.23%	7.57%	11.60%	8.33%	7.11%	0.99%	6.49%	4.25%	2.31%	3.72%	3.44%	1.32%	4.92%	1.99%	2.93%	4.10%	1.10%
LT	11.27%	13.65%	5.81%	23.90%	16.70%	8.12%	9.78%	19.56%	12.55%	11.38%	1.48%	10.05%	6.00%	3.58%	6.36%	4.72%	1.65%	7.01%	3.60%	3.92%	6.67%	2.47%
LV	11.53%	15.26%	4.48%	17.07%	11.39%	4.90%	5.83%	10.85%	8.54%	8.12%	0.92%	6.28%	3.97%	3.03%	3.94%	3.48%	1.47%	5.12%	2.21%	2.93%	4.48%	1.64%
NL	11.15%	13.46%	2.84%	12.75%	10.50%	3.42%	5.60%	9.43%	5.96%	6.51%	0.79%	5.20%	3.02%	2.39%	2.97%	2.80%	0.94%	3.60%	1.27%	2.46%	3.01%	1.21%
PL	11.74%	14.88%	6.01%	24.75%	16.46%	7.54%	9.90%	16.67%	13.98%	11.84%	1.49%	8.99%	5.17%	3.96%	7.00%	6.16%	1.66%	7.30%	3.73%	3.94%	6.70%	2.09%
PT	12.05%	15.26%	3.54%	13.01%	9.88%	4.06%	5.75%	9.31%	8.67%	5.80%	0.76%	5.71%	3.28%	2.03%	3.42%	3.39%	1.41%	3.56%	1.85%	2.22%	3.95%	1.59%
RO	11.79%	13.64%	5.27%	23.21%	14.55%	6.29%	8.13%	14.47%	11.06%	9.94%	1.03%	8.33%	5.52%	3.02%	5.53%	4.82%	1.58%	5.19%	2.81%	3.31%	5.90%	2.30%
SK	11.59%	15.07%	4.51%	18.28%	12.50%	5.35%	7.41%	12.47%	8.53%	7.67%	0.93%	6.71%	4.36%	2.61%	4.71%	3.84%	1.29%	5.48%	2.61%	2.88%	4.82%	1.83%
SL	11.32%	15.28%	5.63%	22.34%	14.89%	6.47%	8.84%	15.51%	12.33%	9.50%	1.34%	8.21%	5.22%	2.80%	5.45%	3.76%	1.84%	6.35%	2.97%	3.14%	5.67%	2.20%
SV	11.13%	16.40%	5.66%	19.35%	13.39%	5.31%	7.94%	12.91%	10.42%	8.16%	1.18%	6.67%	4.51%	3.04%	4.88%	2.94%	1.00%	5.05%	2.40%	3.24%	4.61%	1.89%
MEAN	11.26%	14.30%	4.72%	19.79%	14.31%	5.94%	8.39%	14.42%	10.33%	9.34%	1.20%	7.95%	4.76%	3.17%	5.00%	4.21%	1.52%	6.13%	2.68%	3.37%	5.10%	1.88%
ST_DEV	0.60%	0.97%	1.30%	6.06%	4.74%	1.89%	2.97%	4.34%	2.91%	2.76%	0.38%	2.41%	1.44%	0.87%	1.66%	1.31%	0.46%	2.16%	0.98%	1.00%	1.54%	0.57%
CV	0.053	0.068	0.275	0.306	0.331	0.319	0.354	0.301	0.282	0.296	0.319	0.304	0.304	0.275	0.332	0.310	0.300	0.352	0.366	0.297	0.302	0.302

TABLE D.3

SESSION SUBSET	Matches	
	EV_YES	EV_MULTI
BG	59.68%	39.20%
CS	62.82%	40.02%
DA	58.58%	36.72%
DE	61.01%	37.14%
EL	57.76%	37.25%
ES	54.85%	33.85%
ET	58.46%	36.33%
FI	59.33%	33.11%
FR	54.91%	34.21%
HU	62.92%	39.22%
IT	59.84%	36.24%
LT	59.32%	36.50%
LV	59.17%	37.99%
NL	57.09%	37.93%
PL	60.71%	37.49%
PT	60.45%	40.81%
RO	60.75%	41.22%
SK	62.12%	39.51%
SL	59.74%	40.65%
SV	61.19%	37.39%
MEAN	59.54%	37.64%
ST_DEV	2.21%	2.28%
CV	0.037	0.061

SESSION SUBSET	EuroVoc Categories																					
	00	04	08	10	12	16	20	24	28	32	36	40	44	48	52	56	60	64	66	68	72	76
BG	11.87%	14.35%	4.73%	19.89%	14.94%	5.67%	6.44%	15.02%	10.48%	9.28%	1.83%	7.16%	4.84%	3.42%	4.71%	6.92%	1.81%	5.92%	1.71%	3.34%	3.93%	1.53%
CS	11.40%	15.38%	5.02%	21.35%	13.99%	5.96%	6.90%	12.76%	10.39%	7.28%	1.02%	8.09%	4.90%	3.04%	5.92%	3.55%	1.42%	5.38%	2.70%	3.28%	4.96%	2.06%
DA	9.58%	14.27%	4.62%	20.18%	14.94%	5.37%	7.66%	14.35%	8.93%	8.81%	1.49%	7.26%	3.84%	2.86%	4.10%	3.14%	1.45%	5.89%	3.02%	3.45%	4.34%	1.71%
DE	9.26%	14.52%	4.11%	19.63%	15.73%	5.28%	8.83%	13.30%	8.49%	8.81%	1.17%	7.96%	4.21%	2.93%	4.28%	3.81%	1.39%	6.19%	3.49%	3.16%	4.32%	1.39%
EL	9.93%	13.72%	4.87%	17.49%	12.75%	5.17%	7.56%	14.60%	10.04%	9.73%	1.15%	8.24%	3.12%	4.19%	4.61%	5.11%	1.72%	5.66%	2.59%	3.45%	4.39%	1.56%
ES	10.88%	12.63%	3.09%	14.15%	12.36%	7.64%	9.35%	15.53%	10.90%	9.99%	1.04%	7.18%	5.63%	2.75%	3.79%	3.18%	2.14%	5.70%	1.27%	3.20%	4.07%	1.17%
ET	10.31%	13.80%	4.57%	17.75%	12.65%	5.64%	8.88%	13.57%	9.88%	8.66%	1.51%	7.74%	3.50%	3.88%	4.66%	4.68%	1.49%	7.77%	2.97%	3.24%	4.03%	1.70%
FI	10.00%	14.60%	3.38%	15.74%	11.73%	5.39%	9.09%	15.25%	7.42%	8.54%	1.26%	8.01%	3.41%	4.32%	3.89%	5.05%	1.33%	6.32%	2.67%	3.25%	4.41%	1.08%
FR	9.88%	15.06%	3.25%	18.09%	14.71%	5.41%	9.25%	13.46%	8.00%	8.34%	1.01%	6.96%	3.64%	2.73%	5.12%	4.13%	1.53%	6.17%	2.69%	3.23%	3.99%	1.17%
HU	10.79%	14.13%	3.94%	20.58%	13.14%	5.39%	9.47%	14.44%	9.87%	8.82%	1.02%	6.79%	4.37%	2.55%	4.64%	5.93%	1.88%	6.39%	2.51%	2.67%	4.81%	1.13%
IT	10.20%	13.82%	4.37%	18.57%	15.36%	4.94%	9.05%	14.18%	9.20%	7.65%	1.18%	7.16%	4.70%	2.78%	4.76%	4.59%	1.77%	5.48%	2.72%	3.66%	4.61%	0.90%
LT	10.64%	13.86%	4.35%	18.68%	12.67%	6.27%	7.90%	15.45%	10.22%	8.73%	1.13%	7.57%	4.95%	2.48%	4.97%	3.80%	1.34%	5.34%	2.93%	3.15%	4.33%	1.58%
LV	10.94%	15.75%	5.45%	20.37%	13.66%	5.33%	6.78%	13.06%	10.49%	9.11%	1.22%	6.90%	4.61%	3.61%	4.51%	4.72%	2.21%	5.82%	2.66%	3.39%	4.79%	1.62%
NL	9.70%	13.75%	4.32%	19.22%	16.49%	4.76%	8.66%	13.94%	8.49%	9.94%	1.12%	7.70%	4.27%	4.17%	4.55%	4.40%	1.68%	4.93%	1.59%	4.15%	4.30%	1.42%
PL	11.51%	15.19%	4.46%	18.83%	12.93%	5.63%	8.19%	12.97%	10.98%	9.70%	1.25%	6.52%	3.95%	3.25%	4.75%	4.55%	1.48%	5.71%	3.07%	3.47%	4.51%	1.32%
PT	11.83%	16.93%	5.26%	18.20%	14.29%	5.61%	7.65%	12.55%	12.69%	7.99%	0.95%	7.62%	4.68%	2.82%	4.76%	5.06%	2.15%	4.94%	2.95%	3.24%	5.43%	2.35%
RO	10.62%	13.68%	5.40%	22.55%	14.75%	5.90%	7.91%	14.92%	10.48%	9.17%	0.95%	7.85%	5.11%	2.89%	5.02%	5.14%	1.76%	5.04%	2.48%	3.50%	4.93%	1.88%
SK	11.21%	14.90%	4.76%	19.94%	14.18%	6.15%	8.86%	14.16%	9.48%	8.32%	1.03%	6.78%	4.76%	3.42%	5.03%	4.56%	1.51%	5.86%	2.91%	3.20%	4.82%	1.96%
SL	11.96%	15.12%	5.39%	21.88%	13.75%	5.66%	8.18%	14.38%	11.84%	8.76%	1.35%	7.17%	5.11%	2.41%	5.52%	3.15%	1.68%	5.72%	2.90%	2.74%	4.99%	1.64%
SV	9.86%	17.74%	5.97%	19.22%	13.85%	5.03%	8.56%	13.31%	10.19%	7.94%	1.10%	6.38%	4.51%	2.83%	5.40%	2.71%	0.87%	4.78%	2.89%	3.41%	4.38%	1.89%
MEAN	10.62%	14.66%	4.57%	19.12%	13.94%	5.61%	8.26%	14.06%	9.92%	8.78%	1.19%	7.35%	4.41%	3.17%	4.75%	4.41%	1.63%	5.75%	2.64%	3.31%	4.52%	1.55%
ST_DEV	0.82%	1.18%	0.76%	1.97%	1.24%	0.61%	0.89%	0.91%	1.27%	0.74%	0.22%	0.54%	0.66%	0.59%	0.52%	1.02%	0.32%	0.67%	0.54%	0.31%	0.39%	0.37%
CV	0.077	0.081	0.167	0.103	0.089	0.109	0.108	0.065	0.128	0.084	0.184	0.074	0.149	0.187	0.110	0.232	0.197	0.116	0.203	0.092	0.087	0.236

TABLE D.4

SPOT SUBSET	Matches	
	EV_YES	EV_MULTI
BG	56.95%	36.70%
CS	60.06%	36.58%
DA	56.60%	33.91%
DE	56.98%	33.21%
EL	53.51%	33.51%
ES	54.10%	32.84%
ET	56.36%	33.99%
FI	57.23%	32.96%
FR	52.95%	32.86%
HU	59.19%	35.64%
IT	56.73%	33.08%
LT	58.13%	34.10%
LV	56.82%	36.17%
NL	57.27%	34.53%
PL	58.73%	35.97%
PT	57.27%	36.77%
RO	59.53%	37.72%
SK	58.43%	36.53%
SL	56.15%	34.92%
SV	57.06%	35.21%
MEAN	57.00%	34.86%
ST_DEV	1.86%	1.56%
CV	0.033	0.045

SPOT SUBSET	EuroVoc Categories																					
	00	04	08	10	12	16	20	24	28	32	36	40	44	48	52	56	60	64	66	68	72	76
BG	12.73%	14.42%	4.42%	20.16%	13.80%	5.33%	6.32%	13.25%	10.72%	9.16%	1.18%	7.22%	5.00%	2.73%	5.07%	4.49%	1.36%	5.29%	2.13%	2.98%	4.96%	2.17%
CS	11.82%	15.55%	4.79%	19.78%	13.25%	5.47%	6.80%	13.14%	10.02%	8.28%	1.23%	6.46%	4.96%	2.40%	4.38%	2.60%	1.08%	5.56%	2.42%	3.27%	5.46%	2.57%
DA	11.51%	14.41%	4.45%	19.08%	13.29%	5.11%	7.22%	13.04%	9.02%	8.61%	0.93%	7.32%	3.99%	2.50%	4.28%	3.18%	1.03%	5.34%	2.12%	2.97%	4.65%	1.69%
DE	10.53%	13.99%	3.93%	16.67%	13.87%	4.97%	7.95%	13.35%	8.69%	8.50%	1.08%	7.19%	4.36%	2.82%	4.57%	3.45%	1.06%	5.86%	2.78%	3.08%	4.22%	1.54%
EL	11.61%	12.69%	4.01%	15.83%	12.18%	5.18%	7.34%	14.04%	8.88%	9.28%	1.03%	8.92%	3.79%	3.37%	4.55%	3.84%	1.58%	5.75%	2.36%	3.26%	4.28%	1.82%
ES	12.50%	12.32%	3.64%	14.73%	11.29%	5.41%	7.66%	13.84%	9.03%	9.97%	0.89%	8.23%	4.81%	3.61%	4.33%	3.78%	1.56%	5.24%	1.88%	3.29%	4.40%	1.86%
ET	11.06%	14.01%	4.17%	17.66%	11.91%	6.08%	7.51%	12.67%	9.89%	8.70%	1.06%	6.83%	3.98%	3.04%	4.17%	3.51%	1.29%	6.71%	2.81%	3.17%	4.81%	1.94%
FI	10.89%	12.81%	3.66%	14.99%	11.93%	5.68%	8.35%	14.96%	8.62%	8.70%	1.18%	7.76%	4.33%	3.41%	4.10%	3.77%	1.03%	6.61%	2.34%	3.47%	3.90%	1.39%
FR	10.27%	13.93%	3.53%	16.35%	13.93%	5.41%	8.57%	13.28%	8.33%	8.51%	1.13%	7.69%	4.58%	2.68%	4.47%	3.25%	1.32%	5.70%	2.11%	3.16%	4.46%	1.49%
HU	11.94%	13.73%	4.37%	18.72%	13.30%	5.05%	7.60%	13.60%	8.78%	8.78%	1.10%	7.40%	4.27%	2.89%	4.47%	4.00%	1.71%	6.39%	2.36%	2.88%	5.18%	1.74%
IT	11.10%	13.30%	3.85%	17.09%	13.12%	4.80%	8.48%	13.10%	9.80%	8.49%	1.12%	7.68%	4.99%	2.57%	3.98%	3.56%	1.35%	5.75%	2.01%	3.18%	4.77%	1.49%
LT	11.71%	13.45%	4.42%	17.62%	12.59%	6.04%	7.02%	14.53%	8.96%	8.51%	1.11%	7.58%	4.23%	2.86%	4.68%	3.40%	1.17%	5.26%	2.57%	2.82%	5.55%	2.07%
LV	11.66%	14.91%	4.93%	19.07%	12.68%	5.80%	6.63%	12.31%	9.35%	9.44%	0.94%	7.52%	4.51%	3.39%	4.52%	3.50%	1.35%	5.91%	2.45%	3.33%	5.35%	2.04%
NL	11.82%	13.27%	3.99%	17.98%	14.44%	5.04%	7.78%	13.68%	8.66%	9.09%	1.19%	7.50%	4.41%	3.07%	4.15%	3.84%	1.18%	5.34%	1.95%	3.23%	4.38%	1.91%
PL	11.90%	14.67%	4.57%	18.60%	12.15%	5.72%	7.09%	12.77%	10.32%	8.53%	1.08%	6.81%	3.87%	2.85%	5.56%	4.69%	1.13%	5.40%	2.67%	2.70%	5.34%	1.71%
PT	12.36%	14.00%	4.57%	17.65%	13.07%	5.55%	8.08%	13.29%	11.36%	7.96%	1.13%	7.92%	4.37%	2.76%	4.65%	4.37%	1.77%	4.86%	2.23%	2.91%	5.42%	2.06%
RO	12.55%	13.58%	4.55%	20.81%	12.66%	5.77%	7.27%	12.72%	10.08%	9.21%	0.99%	7.51%	5.10%	2.73%	5.17%	4.03%	1.26%	4.64%	2.67%	2.78%	5.78%	2.27%
SK	11.81%	15.15%	4.85%	19.27%	12.79%	5.42%	7.27%	12.95%	8.86%	8.11%	0.98%	7.28%	4.58%	2.34%	5.04%	3.78%	1.29%	5.84%	2.71%	2.98%	5.38%	1.95%
SL	10.89%	15.36%	4.71%	18.41%	12.75%	5.71%	7.56%	13.42%	10.29%	8.14%	1.09%	7.26%	4.30%	2.49%	4.38%	3.40%	1.59%	5.51%	2.45%	2.78%	4.99%	2.10%
SV	11.90%	15.57%	5.31%	18.91%	12.74%	5.34%	7.35%	12.45%	10.27%	8.07%	1.18%	6.53%	4.38%	3.07%	4.42%	2.99%	1.05%	5.07%	2.03%	3.04%	4.62%	1.83%
MEAN	11.63%	14.06%	4.34%	17.97%	12.89%	5.44%	7.49%	13.32%	9.50%	8.70%	1.08%	7.43%	4.44%	2.88%	4.55%	3.67%	1.31%	5.60%	2.35%	3.06%	4.90%	1.88%
ST_DEV	0.67%	0.95%	0.48%	1.65%	0.78%	0.35%	0.60%	0.66%	0.84%	0.52%	0.10%	0.56%	0.38%	0.36%	0.40%	0.50%	0.23%	0.53%	0.29%	0.21%	0.52%	0.29%
CV	0.057	0.068	0.110	0.092	0.060	0.065	0.080	0.050	0.088	0.060	0.088	0.076	0.086	0.124	0.087	0.137	0.176	0.095	0.122	0.070	0.107	0.155

TABLE D.5

EV FIELD	Single EV match	Eurovoc multiple matches												TOTAL FIELD	Ratio on tot EV
		Match 1	Match 2	Match 3	Match 4	Match 5	Match 6	Match 7	Match 8	Match 9	Match 10	Match 11	Match 12-18		
00	43.50%	10.89%	28.87%	12.08%	3.26%	1.01%	0.26%	0.09%	0.03%	0%	0.002%	0%	0.002%	46820	7.43%
04	38.25%	42.98%	16.01%	2.05%	0.51%	0.15%	0.05%	0.01%	0.01%	0%	0%	0%	0%	59965	9.51%
08	27.97%	32.87%	30.73%	7.00%	1.11%	0.23%	0.08%	0%	0%	0%	0%	0%	0%	18403	2.92%
10	30.64%	2.91%	37.41%	18.78%	7.21%	2.13%	0.61%	0.20%	0.07%	0.03%	0.004%	0.001%	0.001%	77092	12.23%
12	45.09%	27.10%	22.40%	4.35%	0.81%	0.19%	0.04%	0.01%	0%	0%	0%	0%	0%	55732	8.84%
16	36.34%	16.47%	32.51%	9.60%	3.13%	1.31%	0.41%	0.13%	0.05%	0.04%	0.004%	0%	0%	23126	3.67%
20	46.51%	20.41%	25.67%	5.62%	1.33%	0.34%	0.08%	0.03%	0%	0.01%	0%	0.003%	0%	32694	5.19%
24	51.53%	9.75%	29.34%	7.26%	1.55%	0.45%	0.10%	0.02%	0.01%	0%	0%	0%	0%	56784	9.01%
28	42.99%	43.04%	12.24%	1.45%	0.23%	0.04%	0.00%	0%	0%	0%	0%	0%	0%	40222	6.38%
32	51.76%	48.16%	0.08%	0.00%	0.00%	0.00%	0%	0%	0%	0%	0%	0%	0%	36375	5.77%
36	48.02%	25.18%	20.77%	5.01%	0.87%	0.13%	0.02%	0%	0%	0%	0%	0%	0%	4733	0.75%
40	44.84%	37.59%	14.96%	2.08%	0.44%	0.07%	0.01%	0.01%	0%	0%	0%	0%	0%	30820	4.89%
44	44.43%	11.88%	29.49%	9.04%	3.42%	1.14%	0.36%	0.19%	0.04%	0.02%	0%	0%	0%	18521	2.94%
48	49.50%	26.80%	20.20%	3.10%	0.36%	0.03%	0.01%	0%	0%	0%	0%	0%	0%	12329	1.96%
52	46.46%	21.78%	24.85%	5.28%	1.26%	0.28%	0.08%	0.01%	0%	0%	0%	0%	0.005%	19478	3.09%
56	45.93%	12.73%	31.31%	7.04%	2.36%	0.32%	0.15%	0.15%	0.01%	0%	0%	0%	0.006%	16411	2.60%
60	51.50%	16.13%	24.34%	6.18%	1.51%	0.25%	0.07%	0.02%	0%	0%	0%	0%	0.017%	5909	0.94%
64	43.23%	35.10%	18.91%	2.45%	0.27%	0.03%	0.01%	0.00%	0%	0%	0.004%	0%	0%	23877	3.79%
66	47.07%	5.91%	35.37%	8.83%	2.08%	0.53%	0.12%	0.04%	0.03%	0.01%	0%	0%	0.010%	10440	1.66%
68	52.19%	34.60%	11.43%	1.41%	0.26%	0.04%	0.06%	0%	0%	0%	0%	0%	0.007%	13475	2.14%
72	37.96%	26.14%	27.24%	7.07%	1.14%	0.37%	0.08%	0.01%	0%	0%	0%	0%	0%	19884	3.15%
76	36.19%	4.72%	29.13%	17.63%	9.26%	1.92%	0.52%	0.26%	0.23%	0.14%	0%	0%	0%	7329	1.16%
														630419	100.00%
Total Raw Matches	268708	149880	149880	43707	12978	3689	1032	355	124	51	6	2	7	630419	

APPENDIX E

This section contains examples of the 100 most frequent strings (in absolute terms) for each n-gram length, from 1-grams to 15-grams.

The strings come with their respective frequency count. The longer the strings, the lower the frequency count. To generate the following frequency lists, all strings were lowercased before counting the occurrences.

Freq	1-gram	Freq	1-gram	Freq	2-gram	Freq	2-gram
195	tfeu	57	take-up	199	europe 2020	67	programming period
162	erdf	55	tftp	172	economic governance	66	financial framework
142	eeas	55	branding	171	european semester	66	european council
132	esma	55	redress	159	impact assessment	66	fiscal consolidation
123	cip	55	ets	152	digital agenda	64	background information
115	eafrd	55	workshop	149	flagship initiative	64	public authorities
114	benchmark	55	efsa	135	smart regulation	64	reasoned opinion
112	roadmap	55	review	128	innovation union	64	cohesion fund
109	stakeholders	55	etuc	127	resource efficiency	63	market failure
107	networking	54	recovery	124	legal certainty	63	corporate governance
105	accountability	54	newsroom	122	capacity building	63	trade repositories
103	csdp	54	profiling	117	delegated acts	62	lisbon treaty
101	ict	54	flagship	109	general court	62	energy mix
99	scoreboard	54	baseline	109	social inclusion	62	peer review
98	background	54	monitoring	109	infringement procedure	61	carbon leakage
95	enforcement	53	entrepreneurship	103	food security	61	ecosystem services
93	governance	53	gmes	100	task force	61	implementing acts
90	leverage	52	contractor	95	mid-term review	61	economic recovery
90	enpi	52	teu	94	regulatory framework	60	establishment plan
89	mainstreaming	52	wto	93	gender equality	59	schengen area
86	redd	52	ipcc	91	sovereign debt	59	low-carbon economy
86	unfccc	51	cross-compliance	89	public procurement	59	steering committee
84	oecd	51	repository	87	short selling	59	business continuity
80	expertise	50	bottlenecks	83	high representative	59	cohesion policy
79	derivatives	50	fao	82	carbon footprint	58	higher education
78	grant	50	osce	82	good governance	58	audit trail
73	eea	50	sustainability	79	eastern partnership	58	emerging economies
73	equity	49	conditionality	77	cloud computing	57	raise awareness
73	recast	49	infringement	76	where appropriate	57	negotiating directives
71	omnibus	49	policy-making	76	better regulation	57	senior management
71	fp7	49	discharge	75	asset relief	56	spatial planning
70	ownership	49	cfp	75	policy makers	56	criminal justice
69	biodiversity	48	esf	75	law enforcement	56	staff regulations
66	olaf	48	rationale	74	best practice	56	shared management
66	follow-up	48	esrb	74	financial regulation	56	payment appropriations
63	overarching	48	predictability	74	legislative proposal	55	policy framework
63	scope	48	alignment	73	energy efficiency	55	best practices
63	gdp	47	cfsp	73	single market	55	public consultation
63	resilience	47	features	73	executive summary	54	southern corridor
62	cost-effective	47	frontload	73	learning mobility	54	bank resolution
62	stakeholder	47	unodc	72	as amended	54	supporting documents
62	eco-innovation	47	lng	72	otc derivatives	53	compliance costs
61	eulex	47	notwithstanding	72	flagship initiatives	53	outermost region
60	screening	47	livestock	72	stockholm programme	53	citizens initiative
60	regulation	46	mission	71	track record	53	smart growth
59	referral	46	ilo	71	infringement proceedings	53	criminal law
59	fisheries	46	modalities	70	development aid	53	health check
59	website	46	mapping	68	progress report	53	smart grids
58	employability	46	eagf	68	digitally driven	52	venture capital
57	procurement	46	cybercrime	68	budget line	51	water management

Freq	3-gram	Freq	3-gram
205	single market act	48	european banking authority
180	europe 2020 strategy	48	gross national income
151	memorandum of understanding	47	in terms of
132	terms of reference	47	sound financial management
132	level playing field	46	european supervisory authority
124	millennium development goals	46	sugar restructuring fund
122	credit default swaps	45	european investment bank
107	lifelong learning programme	45	maximum sustainable yield
94	credit rating agencies	45	european development fund
92	enterprise europe network	44	eu citizenship report
92	joint research centre	44	national reform programmes
91	annual growth survey	44	specific supply arrangements
88	impact assessment board	44	energy action plan
83	european vacancy monitor	44	counterparty credit risk
82	youth in action	43	on the ground
82	development cooperation instrument	43	civil society organisations
79	ordinary legislative procedure	43	marie curie actions
76	value for money	42	economies of scale
74	european social fund	42	sustainability impact assessments
74	covenant of mayors	42	in line with
73	european neighbourhood policy	42	education and training
71	european skills passport	41	name and shame
70	call for proposals	41	production guarantee fund
69	eu 2020 strategy	41	public private partnerships
69	multiannual financial framework	41	staff working document
65	rule of law	41	european heritage label
64	rules of procedure	41	in due course
62	passenger name record	40	special purpose vehicle
62	common fisheries policy	40	declaration on honour
62	research executive agency	40	draft amending budget
62	water framework directive	40	biodiversity action plan
58	small business act	39	internal audit capability
57	code of conduct	39	regulatory technical standards
56	european research area	39	excessive deficit procedure
56	european digital agenda	38	law enforcement agencies
56	state of play	38	financial supervision package
55	joint technology initiatives	38	multi-annual financial framework
54	court of justice	38	european refugee fund
54	college of commissioners	38	in this context
54	corporate social responsibility	38	research framework programme
54	european anti-fraud office	38	health technology assessment
54	naked short selling	37	marine knowledge 2020
53	access to justice	37	financial transaction tax
52	credit default swap	37	subsidiarity monitoring network
51	external action service	37	early warning system
51	single european sky	37	single payment scheme
50	law enforcement authorities	37	european tourism day
50	european fisheries fund	37	intellectual property rights
49	social market economy	37	security of supply
49	service level agreement	36	growth for jobs

Freq	4-gram	Freq	4-gram
231	youth on the move	34	economic and financial committee
204	european external action service	34	exercise of the delegation
128	digital agenda for europe	34	european business test panel
125	european court of justice	34	european higher education area
123	stability and growth pact	33	youth in action programme
119	european regional development fund	33	economic and social committee
105	european data protection supervisor	33	extractive industries transparency initiative
100	text with eea relevance	33	european year of volunteering
90	charter of fundamental rights	32	as a general rule
86	commission staff working document	31	european securities market authority
79	european systemic risk board	31	food and veterinary office
71	european asylum support office	30	committee on legal affairs
70	state of the union	30	subject to the provisions
68	audiovisual media services directive	29	committee of the regions
65	partnership and cooperation agreement	29	letters of formal notice
60	risk sharing finance facility	28	general block exemption regulation
57	treaty on european union	28	international civil aviation organisation
56	european court of auditors	28	internal market information system
55	union for the mediterranean	28	european maritime safety agency
53	common european asylum system	28	river basin management plans
53	european consensus on development	28	new skills and jobs
50	travel smarter live better	27	democratic republic of congo
49	vocational education and training	27	competitiveness and innovation programme
48	letter of formal notice	27	type-approval authorities expert group
48	european economic recovery plan	27	erasmus for young entrepreneurs
48	in the light of	27	open method of coordination
46	second strategic energy review	27	economic and monetary union
45	european financial stability facility	27	seventh research framework programme
45	has decided as follows	26	political and security committee
43	european globalisation adjustment fund	26	protection of personal data
42	climate and energy package	26	justice and fundamental rights
42	european food safety authority	26	education at a glance
42	accra agenda for action	26	europe 2020 monitoring platform
42	european statistical system committee	26	maritime affairs and fisheries
41	your first eures job	26	business case for diversity
41	energy efficiency action plan	25	conference of committee chairs
41	european day of languages	25	european aviation safety agency
40	european financial stabilisation mechanism	25	communication on smart regulation
40	information and communication technologies	25	security of food supply
40	services of general interest	25	european equal pay day
39	radio spectrum policy programme	24	structural and cohesion funds
39	state of the art	24	european progress microfinance facility
39	convention on biological diversity	24	europe aid cooperation office
38	with a view to	24	commission staff working paper
37	combined heat and power	24	league of arab states
36	has adopted this decision	24	european neighbourhood policy instrument
36	knowledge and innovation communities	24	european agenda for culture
36	regulatory procedure with scrutiny	24	maritime awareness and risks
35	instrument for pre-accession assistance	24	marine strategy framework directive
35	on a case-by-case basis	23	alternate of the chairperson

Freq	5-gram	Freq	5-gram
100	smart sustainable and inclusive growth	22	universal declaration of human rights
84	task force on economic governance	21	economic and financial affairs council
77	competitiveness and innovation framework programme	21	state of the art defence
73	european economic and social committee	21	your voice on europe 2020
70	european neighbourhood and partnership instrument	21	small business act for europe
70	european securities and markets authority	20	european system of financial supervision
56	eu charter of fundamental rights	20	re-use of public sector information
52	new skills for new jobs	20	effective administrative and judicial redress
50	european convention on human rights	20	consistency with the other policies
49	state of the union address	20	aid scheme for damaged fodder
45	civilian planning and conduct capability	20	justice fundamental rights and citizenship
37	european network for rural development	19	climate change mitigation and adaptation
36	common foreign and security policy	19	better training for safer food
36	european energy programme for recovery	19	intergovernmental panel on climate change
35	territories with specific geographical features	19	eurojustice network of european prosecutors-general
34	former yugoslav republic of macedonia	19	trans-european transport network executive agency
33	european consensus on humanitarian aid	19	international public sector accounting standards
32	common security and defence policy	18	by means of delegated acts
32	common consolidated corporate tax base	18	country policy and institutional assessment
31	treaty establishing the european community	18	prudential treatment of re-securitization operations
29	call for expression of interest	18	international day for the elderly
29	european research council executive agency	18	legislative financial statement for proposals
28	european security and defence college	18	integration of people with disabilities
28	should therefore be amended accordingly	18	european forum for primary care
28	in a spirit of compromise	18	innovative medicines initiative joint undertaking
27	central finance and contracting unit	17	summary of the impact assessment
27	small arms and light weapons	17	en route air navigation services
27	working party on the environment	17	european system of financial supervisors
27	committee of european securities regulators	17	by means of implementing acts
26	as a matter of priority	17	same or equally effective treatment
26	european court of human rights	17	derivatives markets future policy actions
26	european master s in translation	17	soft open method of coordination
26	as a matter of urgency	17	african caribbean and pacific countries
26	european security and defence policy	17	framework convention on climate change
25	president of the european council	17	as a matter of principle
25	high level group on milk	16	magnesia bricks production defence coalition
25	re-launch of the single market	16	treaty on the european union
24	european geostationary navigation overlay service	16	european platform for roma inclusion
24	council of the european union	16	european court of justice rulings
24	european voluntary humanitarian aid corps	16	consent of the european parliament
24	social services of general interest	16	trade control and expert system
24	state of the european union	16	southeast europe energy efficient fund
23	european groupings of territorial cooperation	16	food security related social transfers
23	7th framework programme for research	16	world organisation for animal health
23	european road safety action plan	16	united states conference of mayors
23	working party for schengen matters	16	investing in digitally driven growth
22	european committee of social rights	16	decentralised energy production current barriers
22	comprehensive economic and trade agreement	15	consultative commission on industrial change
22	save where explicitly provided otherwise	15	it should be noted that
22	paris declaration on aid effectiveness	15	efficient partnership in cohesion policy

Freq	6-gram	Freq	6-gram
53	european network and information security agency	14	committee on economic and monetary affairs
53	european insurance and occupational pensions authority	13	european qualifications framework for lifelong learning
49	european agricultural fund for rural development	13	implementing measures for the members statute
46	european institute of innovation and technology	13	advisory group on climate change financing
43	agenda for new skills and jobs	13	international dialogue on peacebuilding and statebuilding
36	prevention of and fight against crime	13	taxation and customs union of the
33	executive agency for competitiveness and innovation	13	en route air navigation services tariffs
32	global monitoring for environment and security	13	united nations convention on biological diversity
30	european week of regions and cities	13	revenue arising from the commission contribution
29	area of freedom security and justice	13	council of europe convention on cybercrime
29	preservation and management of natural resources	13	inspection of tissue and cell procurement
27	conditions of employment of other servants	13	responding to the challenges of globalisation
26	executive agency for health and consumers	12	delegation of powers to the commission
24	task force on the economic governance	12	the economics of ecosystems and biodiversity
23	education audiovisual and culture executive agency	12	european bank for reconstruction and development
22	world day against the death penalty	12	jakarta charter on business on biodiversity
22	international organisation of vine and wine	12	produced in an appellation of origin
21	institute for prospective and technological studies	12	managing vibrations and noise at work
20	deep and comprehensive free trade agreement	12	state of the european union speech
20	european fund for south east europe	12	designating the european capital of culture
20	committee on payment and settlement systems	12	tar sands and land use changes
20	european information sharing and alert system	12	european livestock and meat trading union
19	agreement on anti-subsidies and countervailing measures	12	committee for the prevention of torture
19	for the purposes of this regulation	12	eu strategy for the black sea
19	high level expert group on milk	12	weight of votes in the council
19	european credit transfer and accumulation system	12	central asia research and education network
19	european union agency for fundamental rights	12	statute of the court of justice
19	fuel cells and hydrogen joint undertaking	12	providing security in a changing world
18	agreement on subsidies and countervailing measures	12	environment and natural resources thematic programme
18	un framework convention on climate change	11	committee on employment and social affairs
18	united nations provisional central product classification	11	strategic report on the 2007-2013 programmes
18	eu citizens rights the way forward	11	commissioner for maritime affairs and fisheries
18	industrial policy for the globalisation era	11	impact assessments in the eu institutions
17	quartet envoy to the middle east	11	general court of the european union
17	european marine observation and data network	11	annual accounts of the european union
17	transforming europe for a post-crisis world	11	mission of iceland to the eu
16	water for peace peace for water	11	western and central pacific fisheries commission
16	third internal market for energy package	11	printing and reproduction of recorded media
16	committee on standards and technical regulations	11	red sea-dead sea water conveyance concept
16	general agreement on tariffs and trade	11	ljubljana centre for excellence in finance
16	kosovo under un security council resolution	11	regulating financial services for sustainable growth
16	organisation for economic cooperation and development	11	register of the commission expert groups
15	employment social affairs and equal opportunities	11	general affairs and external relations council
15	forest law enforcement governance and trade	11	regulated access to next generation access
15	european agricultural guidance and guarantee fund	11	community plan of action for sharks
14	united nations convention to combat desertification	11	european forum on forward looking activities
14	organisation for economic co-operation and development	11	failures in protection of human rights
14	united nations international maritime dangerous goods	11	article 29 data protection working party
14	technical platform for cooperation on health	11	council of the baltic sea states
14	council of european municipalities and regions	11	association of european public health schools

Freq	7-gram
51	european instrument for democracy and human rights
40	court of justice of the european union
38	international covenant on civil and political rights
32	international day for the eradication of poverty
24	body of european regulators for electronic communications
24	dismantling the obstacles to eu citizens rights
23	committee for medicinal products for human use
23	a new strategy for the single market
22	committee on internal market and consumer protection
21	strategy for equality between women and men
21	reducing emissions from deforestation and forest degradation
21	halting the loss of biodiversity by 2010
20	setting up an eu rapid response capability
19	rapid alert system for food and feed
19	strategy for smart sustainable and inclusive growth
18	convention on the rights of the child
17	organisation for security and cooperation in europe
16	subsidiary body for scientific and technological advice
16	aid to producer groups for preliminary recognition
16	social and family affairs humanitarian assistance scheme
16	european conference of postal and telecommunications administrations
16	manual on the measurement of volunteer work
15	commissioner for education culture multilingualism and youth
15	recommendation from the commission to the council
15	treaty on the functioning of the eu
15	european regional and local health authorities network
15	learning the languages of the neighbouring countries
15	charter of fundamental rights of the eu
14	rapid alert system for feed and food
14	police and judicial cooperation in criminal matters
14	european contact network of spam enforcement agencies
14	rules of procedure of the european parliament
14	position of the council at first reading
14	european centre for disease prevention and control
14	working party on cooperation in criminal matters
13	framework programme for research and technological development
13	surveillance committee of the european anti-fraud office
13	wto agreement on subsidies and countervailing measures
13	office for democratic institutions and human rights
13	european commissioner for research innovation and science
12	mountain regions islands and sparsely populated areas
12	treaty establishing the european atomic energy community
12	procedures for implementation of the financial regulation
12	protocol relating to the status of refugees
12	blueprint for a north sea offshore grid
12	international criminal tribunal for the former yugoslavia
12	promoting the learning mobility of young people
12	european society for mental health and deafness
11	marrakesh agreement establishing the world trade organization
11	committee on the elimination of racial discrimination

Freq	7-gram
11	laws and customs of war on land
11	code of conduct on national administrative practices
11	accelerating the transformation of europe through innovation
11	community of european railways and infrastructure companies
11	parliamentary assembly of the council of europe
11	scientific committee on health and environmental risks
11	matching skills to the needs of industry
11	borders and border control customs cooperation visa
11	international centre for missing and exploited children
11	international cooperation humanitarian aid and crisis response
10	code of conduct on division of labor
10	copenhagen criteria for accession to the eu
10	international center for trade and substainable development
10	strengthening the legal framework of the osce
10	information and training measures for workers organisations
10	european broadband investing in digitally driven growth
10	global energy efficiency and renewable energy fund
10	european capitals and cities of sport federation
10	report on better lawmaking covering the year
10	democratic front for the liberation of rwanda
10	all information provided is true and accurate
10	directorate general for economic and financial affairs
10	handbook on integration for policy-makers and practitioners
10	a secure europe in a better world
10	organisation of the black sea economic cooperation
10	treaty protocol on services of general interest
10	vienna convention on diplomatic and consular relations
10	united nations framework convention on climate change
10	scientific technical and economic committee for fisheries
10	group of governors and heads of supervision
9	framework agreement between the eu and libya
9	united nations universal declaration on human rights
9	commissioner for the internal market and services
9	seamless pipes and tubes of stainless steel
9	secretariat of the convention of biological diversity
9	special panels for serious crimes in dili
9	an industrial policy for the globalisation era
9	rome statute of the international criminal court
9	directorate-general for humanitarian aid and civil protection
9	to authorise the commission to open negotiations
9	european year of combating violence against women
9	convention on the law of the sea
9	not yet published in the official journal
9	identifying single market bottlenecks and missing links
9	setting up a common european asylum system
9	member of the committee of the regions
9	environment partnership of the black sea synergy
9	quality framework for public and social services
9	teeb for local and regional policy makers
9	specific programme on fundamental rights and citizenship

Freq	8-gram
121	treaty on the functioning of the european union
52	acting in accordance with the ordinary legislative procedure
39	charter of fundamental rights of the european union
28	unlocking the potential of cultural and creative industries
27	high representative for foreign affairs and security policy
26	committee on civil liberties justice and home affairs
26	council of bars and law societies of europe
22	towards adequate sustainable and safe european pension systems
21	european network of civil aviation safety investigation authorities
21	action plan on enhancing the security of explosives
19	staff regulations of officials of the european communities
17	european code of conduct for clearing and settlement
17	subject to its conclusion at a later date
17	communication on the relaunch of the single market
16	delivering a single market to consumers and citizens
16	framework convention for the protection of national minorities
16	global health responding to the challenges of globalisation
16	convention on the rights of persons with disabilities
15	call for expression of interest in the appointment
15	development fund for the electronics and information industry
15	a strategy for smart sustainable and inclusive growth
15	european centre for the validation of alternative methods
14	european association of craft small and medium-sized enterprises
14	parliamentary assembly of the union for the mediterranean
14	existing provisions in the area of the proposal
14	international council for the exploration of the sea
13	committee on women s rights and gender equality
13	european centre for the development of vocational training
13	committee of european insurance and occupational pensions supervisors
13	judicial cooperation in criminal matters and police cooperation
13	treaty establishing the european coal and steel community
13	european agreement on transfer of responsibility for refugees
12	european year for combating poverty and social exclusion
12	european framework cooperation for security and defence research
12	measures for the protection of the aquatic environment
12	portal on learning opportunities throughout the european space
12	food and agriculture organisation of the united nations
12	governance tools and policy cycle of europe 2020
12	section for agriculture rural development and the environment
11	until the commission has adopted a final decision
11	is not subject to any conflict of interest
11	central american small and light weapons control programme
11	concerted work strategy and practical measures against cybercrime
11	communication on the re-launch of the single market
11	europeans development aid and the millennium development goals
11	concept on strengthening eu mediation and dialogue capacities
11	treatment resulting from its schedule of specific commitments
11	free movement of people in the schengen area
11	having regard to article 63 3 and 4
11	guide on the application of the leader axis

Freq	8-gram
11	joint european resources for small and medium-sized enterprises
11	future involvement of the european union in space
11	technical committee on the free movement of workers
11	working party on internal and external fisheries policy
11	a review of this regulation shall be made
11	european commissioner for education culture multilingualism and youth
11	council of the notariats of the european union
11	towards adequate sustainable and safe european pension system
10	taxation and customs union discharge audit and anti-fraud
10	advisory committee on safety and health at work
10	employment social policy health and consumer affairs council
10	high court of justice of england and wales
10	union of the capitals of the european union
10	competitive automotive regulatory system for the 21st century
10	waterborne transport on both domestic and international routes
10	within the limits of the union s competences
10	un committee for the elimination of racial discrimination
10	how poverty affects women in the european union
10	european credit system for vocational education and training
10	international covenant on economic social and cultural rights
10	intergovernmental science-policy platform on biodiversity and ecosystem services
10	high level advisory group on climate change financing
10	territorial dialogue for smart sustainable and inclusive growth
9	eu programme for the prevention of violent conflicts
9	towards a new energy strategy for europe 2011-2020
9	sixth progress report on economic and social cohesion
9	european fund for the integration of third-country nationals
9	field guide to the main languages of europe
9	paris memorandum of understanding on port state control
9	public-private partnership on enhancing the security of explosives
9	failures in protection of human rights and justice
9	assessment of restrictions on palestinian water sector development
9	actions applicable to cbrn prevention detection and response
9	expert group for technical advice on organic production
9	below the indicative allocation in various member states
9	unlocking the potential of creative and cultural industries
9	joint assistance to support projects in european regions
9	products from cloned animals in the food chain
9	transport applications of the global navigation satellite systems
8	s n diethylamino methyl alpha ethyl-2-oxo-1-pyrrolidineacetamide l tartrate
8	international agreement on the protection of personal data
8	un convention on the rights of the child
8	standard minimum rules for the treatment of prisoners
8	decentralised renewable energy sources embedded in local settings
8	kosovo under un security council resolution 1244 1999
8	working party on data protection and information exchange
8	international commission for the conservation of atlantic tunas
8	eu rapid information system for dangerous consumer products
8	eu code of conduct on division of labor
8	international air agreements under the treaty of lisbon

Freq	9-gram
51	standing committee on the food chain and animal health
26	un convention on the rights of persons with disabilities
23	europe 2020 strategy for smart sustainable and inclusive growth
21	strategic framework for european cooperation in education and training
18	commissioner for international cooperation humanitarian aid and crisis response
17	subject to its possible conclusion at a later date
17	united nations convention on the rights of the child
16	improving the implementation of the stability and growth pact
16	is not bankrupt and is not being wound up
16	invitation to submit a best and final binding offer
15	national society for the prevention of cruelty to children
15	the commission should be empowered to adopt delegated acts
14	un special representative on sexual violence in armed conflicts
14	code of conduct of the stability and growth pact
14	agreement in the form of an exchange of letters
14	protocol on access and benefit sharing of genetic resources
14	having regard to the consent of the european parliament
13	aviation security with a special focus on security scanners
13	regional policy contributing to smart growth in europe 2020
13	assembly of the states parties to the rome statute
12	the face of female poverty in the european union
12	competitiveness and innovation framework programme entrepreneurship and innovation programme
12	directorate general for employment social affairs and equal opportunities
11	2 3-methyl-4 2 2 2-trifluoroethoxy 2-pyridinyl methyl thio 1h-benzimidazole
11	how public procurement can underpin the europe 2020 priorities
11	advisory committee for the coordination of social security systems
11	joint european support for sustainable investment in city areas
11	europol heads of high tech crime units task force
10	eu concept on strengthening eu mediation and dialogue capacities
9	community support for economic reform programmes and structural adjustment
9	un committee for the elimination of the racial discrimination
9	strengthening capacities to respond to crises and security threats
9	committee of the permanent representatives of each member state
9	ad hoc open-ended working group on access and benefit-sharing
9	framework programme for research technological development and demonstration activities
9	regional policy contributing to sustainable growth in europe 2020
9	to address additional financing needs of the iter project
9	translation centre for the bodies of the european union
9	single market delivering smart sustainable and inclusive economic growth
8	shared competence between the eu and the member states
8	enhancing economic policy coordination for stability growth and jobs
8	european urban agenda and its future in cohesion policy
8	appointing the members of the group for technical advice
8	adapting and promoting the social dialogue at community level
8	third country state-owned enterprises in eu public procurement markets
8	system of quality control used for purchasing power parities
8	special committee on the financial economic and social crisis
8	integration as a driver for development and social cohesion
8	convention on environmental impact assessment in a transboundary context
8	chemical biological radiological and nuclear cbrn centres of excellence

Freq	9-gram
8	short selling and certain aspects of credit default swaps
7	defense committee of the seamless stainless steel tubes industry
7	guidelines for the regulation of computerized personal data files
7	results of the 2009 public consultation on learning mobility
7	un high level meeting on the millennium development goals
7	office of the high representative vice-president of the commission
7	universal periodic review at the un human rights council
7	after transmission of the proposal to the national parliaments
7	commissioner for human rights of the council of europe
7	plan of action on gender equality in development cooperation
7	globally harmonised system of classification and labelling of chemicals
6	european youth campaign against racism xenophobia antisemitism and intolerance
6	eu high representative for foreign affairs and security policy
6	gdp and beyond measuring progress in a changing world
6	code of conduct of the growth and stability pact
6	eu macro-structural bottlenecks to growth at the national level
6	areas under the national fisheries jurisdiction of third countries
6	strictly limited to the purposes of the original transfer
6	convention on the conservation and management of pollock resources
6	annual report on monitoring the application of community law
6	programme of options specifically relating to remoteness and insularity
6	code of practice of the european fruit juice association
6	applicability of article 101 to multilateral interbank-payments in sdd
6	report on the implementation of the european social partners
6	treaty on the functioning of the european union tfeu
6	national objectives and commitments towards the europe 2020 objectives
6	right to interpretation and to translation in criminal proceedings
6	technical committee of the international organization of securities commissions
6	including killing a settler and her son in 1996
6	shall enter into force on the day of adoption
6	on a new digital agenda for europe 2015 eu
5	by weight of nitrogen in relation to ammonium nitrate
5	procedure for determining and verifying declared noise emission values
5	when is the proposal likely to come into effect
5	spectacles goggles and the like corrective protective or other
5	oecd arrangement on guidelines for officially supported export credits
5	hereby take note of the commission s intention to
5	clean urban transport including modal shift towards public transport
5	ownership harmonisation alignment managing for results and mutual accountability
5	committee for the environment public health and food safety
5	european regulators group for electronic communications networks and services
5	presentation of committee activities in 2009 by policy sectors
5	green paper on the future of eu budget support
5	committee on monetary financial and balance of payments statistics
5	most serious crimes of concern to the international community
5	for certain fish stocks and groups of fish stocks
5	manual of diagnostic tests and vaccines for terrestrial animals
5	social dialogue social rights working conditions adaptation to change
5	european supervisory authority european insurance and occupational pensions authority
5	mandate for the trilogue on the 2011 draft budget

Freq	10-gram
21	opportunities and challenges for european cinema in the digital era
17	the text of the agreement is attached to this decision
17	conditions of employment of other servants of the european communities
16	combating sexual abuse sexual exploitation of children and child pornography
16	cultural artistic and scientific knowledge for preservation access and retrieval
15	vice president of the european commission for the digital agenda
15	european charter on local and regional services of general interest
14	a new trade policy for europe under the europe2020 strategy
13	europe 2020 a strategy for smart sustainable and inclusive growth
13	principles for good international engagement in fragile states and situations
13	active inclusion for young people with disabilities or health problems
12	guidelines on closure of assistance 2000-2006 from the structural funds
11	eu comprehensive approach to the united nations security council resolutions
11	towards a strengthened network and information security policy in europe
11	as adapted to the eea agreement by protocol 1 thereto
11	open method of coordination in social protection and social inclusion
10	the eu and central asia strategy for a new partnership
10	communication on the strategy for equality between women and men
10	international treaty on plant genetic resources for food and agriculture
10	holdings undergoing restructuring due to a reform of a cmo
9	green paper on territorial cohesion turning territorial diversity into strength
9	european association for the co-ordination of consumer representation in standardisation
9	inclusion social policy aspects of migration streamlining of social policies
9	high level group on the competitiveness of the agro-food industry
9	towards an eu strategy on the rights of the child
8	to support economic reforms and help georgia improve debt sustainability
8	guide to taking account of social considerations in public procurement
8	international trade policy in the context of climate change imperatives
8	the european parliament and the council of the european union
8	european elections 2009 awareness raising activities of the european commission
8	regions 2020 an assessment of future challenges for eu regions
8	resolution setting out the grounds on which it is based
8	ad-hoc open ended working group on access and benefit sharing
8	the economic significance of the global loss of biological diversity
8	communication and information resource centre for administrations businesses and citizens
8	council decision establishing the organisation and functioning of the eeas
8	advantages granted with the object of establishing a customs union
7	meeting our critical needs for growth and jobs in europe
7	convention for the protection of human rights and fundamental freedoms
7	except for products falling under annex i of the treaty
7	seventh framework programme for research technological development and demonstration activities
7	strategic report 2010 on the implementation of the programmes 2007-2013
7	results of consultations with the interested parties and impact assessments
7	european foundation for the improvement of living and working conditions
7	wto agreement on the application of sanitary and phytosanitary measures
7	internal rules on the recruitment of officials and other servants
7	man of the year 2009 of central and eastern europe
7	national of one of the member states of the communities
7	pending the completion of the procedures for its formal conclusion
7	directive on the application of patients rights in cross-border healthcare

Freq	10-gram
7	why does action have to be taken by the eu
7	international federation of national red cross and red crescent societies
7	this means that the document is for internal use only
7	un convention on environmental impact assessment in a transboundary context
7	convention on the elimination of all forms of racial discrimination
7	an open and secure europe serving and protecting the citizens
7	products of the milling industry malt starches inulin wheat gluten
7	joint research center s institute for prospective and technological studies
6	minimum standards on procedures for granting and withdrawing refugee status
6	human rights social and environmental standards in international trade agreements
6	communication on adapting and promoting social dialogue at community level
6	executive body for the convention on long-range transboundary air pollution
6	report on the evaluation of the 2009 european parliament elections
6	interim partnership agreement between the ec and the pacific states
6	rules governing the holding of meetings simultaneously with plenary sittings
6	local and regional sustainable energy models in their socio-economic context
6	takes note of the commission statements annexed to this resolution
6	laying down harmonisation conditions for the marketing of construction products
6	memorandum of understanding with the u s conference of mayors
6	conference of the parties to the convention on biological diversity
6	protocol no 30 to the treaty establishing the european community
6	which in case of split stays will be calculated cumulatively
6	statistical usage of the register other than for the census
6	directive on the control of major-accident hazards involving dangerous substances
6	the right control is carried out at the right time
6	data and knowledge in deep-sea fisheries in the high seas
6	green paper towards adequate sustainable and safe european pension systems
6	on the sixth progress report on economic and social cohesion
6	united nations office of the high commissioner for human rights
5	agenda for skills and jobs and investment in life-long learning
5	register of the commission expert groups and other similar entities
5	improving the economic governance and stability framework of the union
5	commission letter to the member state via its permanent representation
5	recognise the basic right of a person to sufficient resources
5	united nations convention on the rights of persons with disabilities
5	other means would not be adequate for the following reason
5	the obligation of professional secrecy shall apply to all persons
5	protocol on the privileges and immunities of the european communities
5	thematic programme for environment and sustainable management of natural resources
5	reaffirming the free movement of workers rights and major developments
5	equality between man and women action against discrimination civil society
5	pan-european data network dedicated to the research and education community
5	road wheels and parts and accessories thereof for industrial assembly
5	countervailing measures on dynamic random access memory chips from korea
5	contribution of the cor to the un climate change conference
5	pursuant to which the commission submitted the proposal to parliament
5	high level group on the competitiveness of the agro-food chain
5	within the deadlines set out in the notice of initiation
5	relations of the european union and the gulf cooperation council
5	community guidelines for state aid in the agriculture and forestry

Freq	11-gram
54	high representative of the union for foreign affairs and security policy
35	after transmission of the draft legislative act to the national parliaments
33	european convention for the protection of human rights and fundamental freedoms
29	guidelines for the examination of state aid to fisheries and aquaculture
21	convention on the elimination of all forms of discrimination against women
20	international convention on the elimination of all forms of racial discrimination
20	advisory group on the food chain and animal and plant health
13	establishing the organisation and functioning of the european external action service
12	aquaculture inland fishing processing and marketing of fishery and aquaculture products
11	non acceptance of the appropriate measures for the fisheries insurance scheme
11	financial regulation applicable to the general budget of the european communities
11	motion for a resolution to wind up the debate in plenary
11	protection of individuals with regard to automatic processing of personal data
11	inclusion social policy aspects of migration streamlining of social policies unit
10	report from the commission to the european parliament and the council
10	community guidelines for state aid in the agriculture and forestry sector
10	wto understanding on rules and procedures governing the settlement of disputes
10	innovative approaches for chronic illnesses in public health and healthcare systems
10	recent developments and future priorities in the area of home affairs
9	in order to address safety issues identified in the scientific conclusions
9	impact assessments in the eu institutions do they support decision making
9	tar sands and land use changes in the fuel quality directive
9	member states shall determine how such reference is to be made
8	temporary capital tax gains tax relief on sales of cultivated land
8	alignment to the treaty on the functioning of the european union
8	confederation of the food and drink industries of the european union
8	stronger involvement of local and regional authorities in the europe2020 strategy
8	signing of the renewal of the european parliament s environmental policy
8	assessment of potential and promotion of new generation of renewable technologies
8	census of the roma on the basis of ethnicity in italy
8	streamlining financial rules and accelerating budget implementation to help economic recovery
7	encourage the participation of young people in democratic life in europe
7	recitals 3 to 10 of the provisional regulation are hereby confirmed
7	concrete and forward-looking measures to improve the social integration of roma
7	united nations charter for the maintenance of international peace and security
7	eu citizenship report 2010 dismantling the obstacles to eu citizens rights
7	guidance note on article 55 of council regulation ec n¬∞1083 2006
7	european community s development and external assistance policies and their implementation
7	laying down the rules and general principles concerning mechanisms for control
7	thematic working group of the european evaluation network for rural development
7	measures against discrimination based on inter alia racial or ethnic origin
6	on the supplementary supervision of insurance undertakings in an insurance group
6	territorial pact of regional and local authorities on europe 2020 strategy
6	framework decision on the mutual recognition of decisions imposing financial penalties
6	european union high representative for the common foreign and security policy
6	community initiative programme interreg iiia for the development of cross-border cooperation
6	state party to the rome statute of the international criminal court
6	procedure with associated committees rule 50 of the rules of procedure
6	releasing the amounts secured by way of the provisional duties imposed
6	citizenship report including analysis and remedies for obstacles to free movement

Freq	11-gram
6	fisheries dispute opposing the eu to iceland and the faroe islands
6	androulla vassiliou the european commissioner for education culture multilingualism and youth
6	development aspects of the international day for the eradication of poverty
6	community framework for state aid for research and development and innovation
6	where the codecision procedure applies the european parliament and council may
5	may not be processed automatically unless domestic law provides appropriate safeguards
5	on the acceleration of the rate of attainment of the treaty
5	to address the consequent outflow of patients due to the implementation
5	support to digitisation of cinemas is necessary to safeguard cultural diversity
5	work for an environment conducive to creativity innovation and cultural expression
5	where powers are conferred upon the commission to adopt delegated acts
5	this decision shall enter into force on the day of adoption
5	thematic programme for the environment and sustainable management of natural resources
5	european charter for equality of women and men in local life
5	science and society programme under the seventh framework programme for research
5	academic network for legal studies on immigration and asylum in europe
5	legal and factual background and pleas in law adduced in support
5	proliferation of weapons of mass destruction and their means of delivery
5	declaration on the enhancement of the european security and defence policy
5	child and family policy education gender equality employment housing and healthcare
5	annual report on the implementation of the instrument for pre-accession assistance
5	summary of responses and how they have been taken into account
5	vice-president viviane reding eu commissioner for justice fundamental rights and citizenship
5	thematic strategy for the environment and sustainable management of natural resources
4	european partnership for the anticipation of change in the automotive sector
4	europe 2020 a european strategy for smart sustainable and inclusive growth
4	monitoring committee of the parliamentary assembly of the council of europe
4	directorate general for economic and financial affairs of the european commission
4	commission outlines action plan to boost europe s prosperity and well-being
4	programme to support the further development of an integrated maritime policy
4	implementing measures for the statute for members of the european parliament
4	any such measures shall respect articles 2 2 and 3 3
4	rate of the duty applicable to the net free-at-union frontier price
4	reduction of workers exposure to the risk of work-related musculo-skeletal disorders
4	calling for a moratorium on the use of the death penalty
4	results of the monitoring of dioxin levels in food and feed
4	the national authorities plan to involve have involved local regional authorities
4	other means would not be adequate for the following reason s
4	accounting and office bookkeeping not relating to the parliamentary assistance allowance
4	restrictions on imports or exports or any measures having equivalent effect
4	annual report from the european commission on the instrument for stability
4	security challenges in the sahel region and the horn of africa
4	successes and challenges after 10 years of economic and monetary union
4	on the conservation of natural habitats and wild fauna and flora
4	the appointment will be for a fixed term of five years
4	please insert the date of entry into force of this regulation
4	caspar cultural artistic and scientific knowledge for preservation access and retrieval
4	presidential committee of the parliamentary assembly of the council of europe
4	italian national agency for new technologies energy and sustainable economic development

Freq	12-gram
32	having regard to the opinion of the european economic and social committee
30	international guidelines for the management of deep-sea fisheries in the high seas
29	having regard to the treaty on the functioning of the european union
21	convention on the protection and promotion of the diversity of cultural expressions
17	high representative of the european union for foreign affairs and security policy
15	overview of information management in the area of freedom security and justice
14	it shall be published in the official journal of the european union
11	towards a europe free from tobacco smoke policy options at eu level
11	europe 2020 flagship initiative innovation union transforming europe for a post-crisis world
10	directive on minimum standards on procedures for granting and withdrawing refugee status
10	the economics and ecosystems and biodiversity for national and international policy markets
10	completion of programme for enterprises improvement of the financial environment for smes
9	guidelines governing the protection of privacy and trans-border flows of personal data
9	draft general budget of the european union for the financial year 2011
9	it shall take effect immediately or at a later date specified therein
9	european convention on the legal status of children born out of wedlock
8	un convention on the elimination of all forms of discrimination against women
8	as a consequence of the entry into force of the lisbon treaty
8	this decision shall enter into force on the date of its adoption
8	directives on minimum standards for the qualification and procedures of asylum seekers
8	climate change assessments review of the processes and procedures of the ipcc
7	plan of action on gender equality and women s empowerment in development
7	european convention for the protection of human rights and fundamental freedoms echr
7	distribution of food products to the most deprived persons in the union
7	conference of the parties to the united nations convention on biological diversity
7	forging a durable framework to combat terrorism within the rule of law
7	convention on international trade in endangered species of wild fauna and flora
7	strengthen and broaden international regional national and local activities to conserve biodiversity
7	pilot project to encourage conversion of precarious work into work with rights
7	distinguishing number of the country which has granted extended refused withdrawn approval
7	should be consistent with the principles of the proposed water framework directive
7	vice-president of the commission high representative of the union for foreign affairs
6	in a manner sufficient to present the legal basis for the complaint
6	challenges for european film heritage from the analogue and the digital era
6	making the difference strengthening capacities to respond to crises and security threats
6	common information sharing environment for the surveillance of the eu maritime domain
6	community guidelines on state aid for rescue and restructuring firms in difficulty
6	provides a valuable instrument for bolstering union-wide cooperative efforts in this field
6	conservation and management of straddling fish stocks and highly migratory fish stocks
6	economic and social development of mountain regions islands and sparsely populated areas
6	climate change the new industrial policies and ways out of the crisis
6	guidance on european court of justice rulings on the rights of students
6	title xii of the treaty on the functioning of the european union
6	this decision shall enter into force on the day of its adoption
6	budget heading 04 03 03 01 on industrial relations and social dialogue
5	commission communication on the future involvement of the european union in space
5	appointments shall be made for all grades of the ad function group
5	on tar sands and land use changes in the fuel quality directive
5	forum on sex discrimination in access to insurance and related financial services
5	frames and mountings for spectacles goggles or the like and parts thereof

Freq	12-gram
5	evaluation of the management of h1n1 influenza in 2009-2010 in the eu
5	in view of the renewal of the committee s term of office
5	efta surveillance authority raises no objections to the following state aid measure
5	international dialogue on peacebuilding and statebuilding between fragile states and development partners
5	legislative framework to address the dangers associated with precursors should be developed
5	european agency for the management of operational cooperation at the external borders
5	pending the completion of the necessary procedures for its entry into force
5	appropriations of an administrative nature financed from the envelop of specific programs
5	code of conduct on complementarity and division of labour in development policy
5	the commission shall make a report in respect of the delegated powers
5	the fishing opportunities for community fishing vessels for certain deep-sea fish stocks
5	community support for economic reform programmes and structural adjustment review and prospects
5	concerning codes of conduct regarding commercial communications for inappropriate food and beverages
5	combating illegal fishing at the global level the role of the eu
4	hague programme on strengthening freedom security and justice in the european union
4	united kingdom uk film council distribution and exhibition initiatives digital screen network
4	the repetition of similar services or works entrusted to the contractor awarded
4	provided that the formal steps for participation are completed in due time
4	code of conduct for the members of the european court of auditors
4	on the green paper on territorial cohesion turning territorial diversity into strength
4	utmost account of the commission s recommendation justifying any departure from it
4	the fishing opportunities for eu fishing vessels for certain deep-sea fish stocks
4	guidelines on cooperation between the hearing officer and the trade investigation services
4	production of renewable energy as a regional development policy in rural areas
4	organisation of the working time of persons performing mobile road transport activities
4	the protocol on the application of the principles of subsidiarity and proportionality
4	in accordance with article 18 1 of the financial regulation any revenue
4	education and training in the context of poverty reduction in developing countries
4	all rights of the producer and owner of the work reproduced reserved
4	on council s position on draft general budget of the european union
4	the role of civil society in the fta between eu and india
4	laying down a performance scheme for air navigation services and network functions
4	representatives of the peoples of the states brought together in the community
4	unless otherwise specified the provisions in force concerning customs duties shall apply
4	fair revenues for farmers a better functioning food supply chain in europe
4	cohesion policy strategic report 2010 on the implementation of the programmes 2007-2013
4	parts of machines appliances and instruments of heading 8427 n e c
4	regulation laying down the rules and general principles concerning mechanisms for control
4	concerning the rights of passengers when travelling by sea and inland waterway
4	making the bank or financial institution or the third party stand as
4	standards and guidelines for quality assurance in the european higher education area
4	having regard to the fact that the commission has made a proposal
4	encouraging the development of youth exchanges and of exchanges of socio-educational instructors
4	the case for the eu citizen state of play and way forward
3	selected items from the communication on the re-launch of the single market
3	an initiative to unleash the potential of young people to achieve smart
3	as a consequence the control system provides insufficient assurance on farmer compliance
3	on the basis of the scientific conclusions set out in annex ii
3	the elements for which that power may be exercised should be defined
3	jha counsellors are invited to a meeting with the following provisional agenda

Freq	13-gram
31	unesco convention on the protection and promotion of the diversity of cultural expressions
13	framework convention for the protection of national minorities of the council of europe
12	council decision establishing the organisation and functioning of the european external action service
11	leipzig charter on sustainable european cities and territorial agenda of the european union
10	eu plan of action on gender equality and women s empowerment in development
10	internal rules on the implementation of the general budget of the european union
9	report on the application of existing instruments and proposal for the new system
8	laying down specific measures for agriculture in the outermost regions of the union
8	it shall not affect the validity of the delegated acts already in force
8	eu manufacturing industry what are the challenges and opportunities for the coming years
8	handbook for the processing of visa applications and the modification of issued visas
7	equal treatment between persons irrespective of religion belief disability age or sexual orientation
7	the role of social protection as economic stabiliser lessons from the current crisis
7	directive of the european parliament and the council on attacks against information systems
6	parts of trailers semi-trailers and other vehicles not mechanically propelled n e c
6	promoting youth access to the labour market strengthening trainee internship and apprenticeship status
6	strategy against proliferation of weapons of mass destruction and their means of delivery
6	action plan for the implementation of the legal framework for electronic public procurement
6	conference of community and european affairs committees of parliaments of the european union
6	on the financial regulation applicable to the general budget of the european communities
6	a better understanding of our seas and oceans to boost competitiveness and growth
6	heading s of the multiannual financial framework and expenditure budget line s affected
6	panel on food additives flavourings processing aids and materials in contact with food
5	contribution of biodiversity and ecosystems to the achievement of the millennium development goals
5	harms public health and the environment by destroying ozone in the upper atmosphere
5	report on the implementation of the european social partners framework agreement on telework
5	legislative dossiers on the protection of the eu financial interests blocked in council
5	by means of implementing acts in accordance with article 291 of the treaty
5	interest rate applied by the european central bank to its main refinancing operations
5	the stockholm programme an open and secure europe serving and protecting the citizens
5	global approach to transfers of passenger name record pnr data to third countries
5	no state aid within the meaning of article 49 of the eea agreement
5	on the protection of individuals with regard to the processing of personal data
5	regulation no 37 of the economic commission for europe of the united nations
5	demographic change and its consequences for the future cohesion policy of the eu
5	public consultation on opportunities and challenges for european cinema in the digital era
4	policy options for progress towards a european contract law for consumers and businesses
4	having regard to the undertaking given by the council representative by letter of
4	financial supervision package briefing by the rapporteurs following the conclusion of trialogue negotiations
4	communication on relations with the complainant in respect of infringements of community law
4	parliament s position on the 2011 draft budget as modified by the council
4	fixing the fishing opportunities for certain fish stocks and groups of fish stocks
4	implementation of enhanced cooperation procedure regarding law applicable to divorce and legal separatio
4	protocol to the convention for the protection of human rights and fundamental freeedoms
4	no state aid within the meaning of article 61 of the eea agreement
4	mediation support group consisting of representatives of the council secretariat and the commission
4	shall be registered in the community register of medicinal products under number s
4	commission recommendation on the active inclusion of people excluded from the labour market
4	communication on options for an eu vision and target for biodiversity beyond 2010
4	establishing the thematic strategy for the environment and sustainable management of natural resources

Freq	13-gram
4	on the independence integrity and accountability of the national and community statistical authorities
4	applicability of article 81 of the ec treaty to multilateral interbank-payments in sdd
4	council decision 1990 424 eec as amended by council decision 2006 965 ec
4	thematic strategy for the environment and sustainable management of natural resources including energy
4	convention of 25 june 1991 implementing the schengen agreement of 14 june 1985
4	and as a part a comprehensive and consistent drive to combat social exclusion
4	does not go beyond what is necessary in order to achieve those objectives
4	measures adopted pursuant to this article shall not interfere with member states competences
4	this directive is addressed to the member states in accordance with the treaties
4	framework convention on the protection of national minorities of the council of europe
3	geneva act to the hague agreement concerning the international registration of industrial designs
3	measures aimed at a reinforced and high level network and information security policy
3	eeas staff should be provided with adequate common european professional development and training
3	public service broadcasting in the digital era the future of the dual system
3	un convention against torture and other cruel inhuman or degrading treatment or punishment
3	eu high representative of the european union for foreign affairs and security policy
3	action will better achieve the objectives of the proposal for the following reasons
3	always been vigilant as regards the adoption of anticompetitive behaviour by market players
3	major stages of the decision-making process between the commission and the other institutions
3	having regard to the opinion of the european economic and social committee of
3	reply of the commission to the special report of the court of auditors
3	special representatives to the un secretary general on sexual violence in armed conflict
3	premise of the principles of true and fair view and substance over form
3	the situation of roma and on freedom of movement in the european union
3	view to embarking on a comprehensive debate on the current discharge procedure system
3	as regards the aid granted in the framework of the german alcohol monopoly
3	where a declaration for release for free circulation is presented in respect of
3	if no comments have been received by the general secretariat of the council
3	on the census of the roma on the basis of ethnicity in italy
3	a permanent asset for each member state and the european union at large
3	encouraging participation in decision-making and speeding up the response to their specific needs
3	the procedure under article 107 of directive 2001 83 ec has been started
3	invitation to submit comments pursuant to article 108 2 of the tfeu treaty
3	there was no need for consultation of interested parties or for external expertise
3	the accounts are kept in euro on the basis of the calendar year
3	the non-confidential version of the decision will be made available under the case
3	conduct procurement of goods services and works in accordance with the detailed provisions
3	council framework decision 2002 475 jha of 13 june 2002 on combating terrorism
3	illicit manufacturing of and trafficking in firearms their parts and components and ammunition
3	agreement between the european community and the swiss confederation in the audiovisual field
3	was granted a derogation from the rules of origin laid down in regulation
3	together with the annexes and the declaration made unilaterally by the european union
3	2008 international guidelines for the management of deep-sea fisheries in the high seas
3	bright bars and iron castings tubes pipes and fittings and other iron castings
3	it is alleged that the situation of the union industry as described above
3	seeking to implement an emissions trading system that applies to other contracting states
3	they are eligible for cover in the event of sickness in a statutory
2	lays down rules and criteria for the selection and implementation of those projects
2	prolong the authorisation of the aid temporarily found compatible with the internal market
2	the commission may adopt by means of delegated acts in accordance with article

Freq	14-gram
16	may be revoked at any time by the european parliament or by the council
14	on lessons learned from the a h1n1 pandemic health security in the european union
12	member states did not take their responsibility to implement effective control and sanction systems
11	committee on the protection of individuals with regard to the processing of personal data
10	failures in protection of human rights and justice in the democratic republic of congo
9	minimising the likelihood of cbrn incidents occurring and limiting their consequences should they materialise
9	regulated competition between regular passenger services and for cabotage haulage operations by non-resident hauliers
9	convention on the conservation and management of pollock resources in the central bering sea
8	the institution which objects shall state the reasons for objecting to the delegated act
7	directive on environmental liability with regard to the prevention and remedying of environmental damage
7	programme of options specific to the remote and insular nature of the outermost regions
7	stocktaking study on the measures and actions taken in council of europe member states
7	the delegation of powers shall be automatically extended for periods of an identical duration
7	transport applications of the global navigation satellite systems short and medium term eu policy
6	these indicators shall be a critical element in the determination of appropriate supervisory actions
6	investigating and remedying the abuse of power by large supermarkets operating in the eu
6	on the assessment of the effects of certain plans and programmes on the environment
6	supervising those financial institutions that are not subject to the supervision of competent authorities
6	european research and innovation potential in the light of the challenges posed by globalisation
6	community guidelines for state aid in the agriculture and forestry sector 2007 to 2013
6	shall bring into force the laws regulations and administrative provisions necessary to comply with
6	priorities and outline of a new eu policy framework to fight violence against women
6	on the protection and respect for sexual and human rights of women and children
6	expenditure related to staff in active employment of information society and media policy area
5	if specific legislation is adopted the scope of this legislation should be adapted accordingly
5	president of the european investment bank eib and chairman of its board of directors
5	draft declaration in relation to supervisory powers on credit rating agencies and other areas
5	the facility shall be used for sustainable energy projects in particular in urban settings
5	access to information public participation in decision-making and access to justice in environmental matters
5	declares in his her own name or as legal representative of the company organisation
5	subject to its possible conclusion at a later date the agreement should be signed
5	convention for the protection of individuals with regard to automatic processing of personal data
5	must enjoy an independence allowing them to perform their duties free from external influence
5	recommending the development and use of ecosystem-based approaches to climate change adaptation and mitigation
5	please specify the name of the case and the case number in all correspondence
4	insurance provided by two or more direct insurance service providers on a joint basis
4	in addition the committee has also recommended as temporary measures to be implemented immediately
4	report on racism and xenophobia in the member states of the eu in 2009
4	the general direction of the european efforts towards an increased network and information security
4	face special challenges because of their location at the external border of the union
4	european commission s green paper on towards adequate sustainable and safe european pension systems
4	aviation activities in the scheme for greenhouse gas emission allowance trading within the community
4	eu plan of action on gender equality and women s empowerment in development 2010-2015
4	fair revenues for farmers a better functioning of the food supply chain in europe
4	eu code of conduct on complementarity and the division of labour in development policy
4	feasibility of establishing a roster of experts in mediation process and related thematic areas
4	ad-hoc open-ended working group on review of implementation of the convention on biological diversity
4	state aid temporary rules adopted as a response to the economic and financial crisis
4	discrimination on the basis of ethnic origin or race has no place in europe
4	strengthening international organisations and non state actors with a crisis prevention or response mandate

Freq	14-gram
4	the external cost charge may be related to the cost of traffic-based air pollution
4	governing board of the european foundation for the improvement of living and working conditions
4	the basic amount to be taken into account for calculating the reduced agricultural components
4	situation of the jordan river with special regard to the lower jordan river area
3	capacity building for decommissioning of nuclear facilities for redirection of former iraqi wmd scientists
3	the regulation constitutes a first step towards a legal frame for european political parties
3	products originating in and imported into the shall be subject to the mfn tariff
3	have the replacement published for information in the official journal of the european union
3	it has been prepared in line with the monitoring requirements specified in annex vii
3	the relationship between trade measures and environmental measures in order to promote sustainable development
3	treaty of peace between the state of israel and the hashemite kingdom of jordan
3	weaknesses in the systems set up in the member states for lifelong learning programme
3	on the european system of national and regional accounts in the community of 2010
3	replies to written questions put to the council by members of the european parliament
3	with attached fittings suitable for conducting gases or liquids for use in civil aircraft
3	stable prices sound public finances and monetary conditions and a sustainable balance of payments
3	the adoption of the proposal will lead to the repeal of the existing legislation
3	european convention for the prevention of torture and inhuman or degrading treatment or punishment
3	a lamp facing in a forward direction used to make the vehicle more easily
3	there is therefore a clear need for a community framework of national control systems
3	draft amending budget n¬∞5 2010 of the european union for the financial year 2010
3	dipartimento della sicurezza pubblica per la sicurezza informatica e la protezione delle infrastrutture critiche
3	proposal for a directive on criminal law protection of the community s financial interests
3	oslo guidelines for the use of military and civil defence assets in disaster relief
3	partnership and cooperation agreement between the european communities and their member states and turkmenistan
3	guarantee is limited to a percentage of the ceiling of the credit lines authorised
3	laying down the procedures for the exercise of implementing powers conferred on the commission
3	implementing the principle of equal treatment between persons irrespective of racial or ethnic origin
3	proposed interim measures for the freezing and disclosure of debtors assets in cross-border cases
3	as regards distribution of food products to the most deprived persons in the union
3	regions which suffer from severe and permanent natural or demographic handicaps including islands regions
3	guidance note on article 55 of council regulation ec n¬∞1083 2006 revenue generating projects
3	growing of cereal grains hard and soft wheat rye barley oats maize rice etc
3	written records detailing the identity of the biocidal product or active substance labelling data
3	council conclusions on access to eurodac by member states police and law enforcement authorities
3	within the framework of its exclusive competence in the negotiations of bilateral fishing agreements
3	second optional protocol to the to the international covenant on civil and political rights
3	thereinafter referred to as the financial framework the necessary adjustments will be made accordingly
3	professional training for the judiciary and mutual understanding of other member states legal systems
3	sufficient resources and social assistance to live in a manner compatible with human dignity
3	with a view to bringing an end to the situation of excessive government deficit
3	we the children end-decade review of the follow-up to the world summit for children
3	on the execution in the european union of orders of freezing property and evidence
3	subject of a judgment which has the force of res judicata for either fraud
3	european social charter and the related recommendations of the european committee of social rights
3	where a declaration for release for free circulation is presented in respect of imports
3	the tse road map 2 a strategy paper on transmissible spongiform encephalopathies for 2010-2015
3	demonstrate an ability to work in a third language prior to their first promotion
2	ad hoc working party on the cooperation and verification mechanism for bulgaria and romania
2	retail banking and banking services to smes and other corporate customers in the uk

Freq	15-gram
13	vice-president of the commission high representative of the union for foreign affairs and security policy
12	towards establishing the common information sharing environment for the surveillance of the eu maritime domain
9	this regulation shall be binding in its entirety and directly applicable in all member states
7	council of europe convention on the protection of children against sexual exploitation and sexual abuse
7	having called on interested parties to submit their comments pursuant to the provisions cited above
7	enhancing economic policy coordination for stability growth and jobs tools for stronger eu economic governance
7	regulation no 13 of the economic commission for europe of the united nations un ece
6	establishing the thematic strategy for the environment and sustainable management of natural resources including energy
6	on the protection of personal data processed in the framework of police and judicial cooperation
6	state of play and future synergies for increased effectiveness between erdf and other structural funds
5	memorandum of understanding signed at the ninth meeting of the conference of the parties cop9
5	having regard to the treaty on the functioning of the european union and in particular
5	agreement on implementation of article vi of the general agreement on tariffs and trade 1994
5	stocktaking study on the measures and actions taken in council of europe member states 2006
5	resolution of the european council of 17 june 1997 on the stability and growth pact
5	oslo guidelines on the use of military and civil defence assets in international disaster relief
5	accompanying measures for sugar protocol countries in favour of the republic of trinidad and tobago
5	better functioning food supply chain in europe and the high level expert group on milk
4	your request specifying the relevant information should be sent by registered letter or fax to
4	parliament s position on the 2011 draft budget as modified by the council all sections
4	council regulation on administrative cooperation and combating fraud in the field of value added tax
4	agreement concerning the conservation and management of straddling fish stocks and highly migratory fish stocks
4	communication from the commission on the application of state aid rules to public service broadcasting
4	decision 2010 437 eu establishing the organisation and functioning of the european external action service
4	report on the functioning of the present eu regime on civil procedural law across borders
4	action plan to strengthen the commission s supervisory role under shared management of structural actions
4	cases where an egf application under article 2 b is submitted by a member state
4	completion of programme for enterprises improvement of the financial environment for small and middle-sized enterprises
4	the situation of roma eu citizens moving to and settling in other eu member states
4	this deployment plan shall inter alia determine the number of inspections to be carried out
3	assessment of the effects of certain plans and programmes on the environment and public consultation
3	claim for extension or payment of a bond financial guarantee or indemnity of whatever form
3	to support the development of innovative ict-based content services pedagogies and practice for lifelong learning
3	presidential committee of the political groups of the parliamentary assembly of the council of europe
3	agreement on port state measures to prevent deter and eliminate illegal unreported and unregulated fishing
3	the infrastructure charge shall be based on the principle of the recovery of infrastructure costs
3	where a declaration for release for free circulation is presented in respect of imports of
3	that the council with the consent of the parliament adopts this decision on the conclusion
3	amending the interinstitutional agreement of 17 may 2006 on budgetary discipline and sound financial management
3	flood risk assessment flood mapping and the production and implementation of flood risk management plans
3	annexes and the declaration made unilaterally by the european union attached to the final act
3	strategic reports of 2010 by the commission on the implementation of the cohesion policy programmes
3	additional protocol to the agreement between the european economic community and the republic of iceland
3	improve transparency and regulatory oversight of over-the-counter derivatives in an internationally consistent and non-discriminatory way
3	by failing to adopt or to notify the authority of the measures necessary to implement
3	declaration on the rights of persons belonging to national or ethnic religious and linguistic minorities
3	unless the european parliament and the council decide otherwise on a proposal of the commission
3	the contribution of eu policies promoting equality between men and women in combating youth crime
3	framework agreement on the reduction of workers exposure to the risk of work-related musculo-skeletal disorders
3	local and regional cooperation to protect the rights of the child in the european union

Freq	15-gram
3	on the signature of the agreement in the form of the exchange of letters between
3	vice president of the european commission and high representative for foreign affairs and security policy
3	capturing impacts of leader and of measures to improve quality of life in rural areas
3	development of a european system for exchanging information on asylum migration and countries of origin
3	european potential in research and innovation in the light of the challenges posed by globalisation
3	council conclusions on the mdgs for the un plenary meeting in new york and beyond
3	protection of individuals with regard to automatic processing of personal data and additional protocol 181
3	they are credited for the amount being claimed from the moment the demand is received
3	european public debate and welcomes the commission s proposal for a regulation on this matter
3	2003 eu strategy against proliferation of weapons of mass destruction and their means of delivery
3	consultations the member state concerned decides to grant an authorisation it shall inform the other
3	that to the best of his her knowledge all information provided is true and accurate
3	a legal vacuum that would be created by excluding such a food from the regulation
2	isas are written in the context of an audit of financial statements by an auditor
2	this directive does not go beyond what is necessary in order to achieve those objectives
2	such services have to go beyond mere databases listing available businesses and offer a comprehensive
2	this draft regulation was approved by the european statistical system committee in written procedure on
2	setting out the requirements for accreditation and market surveillance relating to the marketing of products
2	shall be taken into account in all activities undertaken by the parties under this agreement
2	adoption of the decision on the conclusion of the agreement by the council was delayed
2	decision to initiate the procedure was published in the official journal of the european union
2	value must at least be equal to any reference price fixed or to be fixed
2	presentations by the commission dg empl and the organisation for economic co-operation and development oecd
2	sources of radiation or the presence of moving or fixed objects shall not interfere with
2	were at the time of the facts imported into the european union free of import
2	commission regulation ec no 1126 2008 of 3 november 2008 adopting certain international accounting standards
2	agreement on the text in view of its transmission to the european parliament for assent
2	proposal on the organisation of the working time of persons performing mobile road transport activities
2	notify the competent authority of any material changes affecting the conditions for the initial authorisation
2	support for integrated territorial development strategies of a pilot nature based on a bottom-up approach
2	butter stored in france due to the very small quantity stored in that member state
2	with a view to bringing the situation of an excessive deficit to an end by
2	to set their national targets for reducing the number of people at risk of poverty
2	androulla vassiliou the european commissioner responsible for sport has paid tribute to juan antonio samaranch
2	in order to address safety issues identified in the scientific conclusions set out in annex
2	legislative financial statement for proposals having a budgetary impact exclusively limited to the revenue side
2	strategy for the economic and social development of mountain regions islands and sparsely populated areas
2	in fact the investigation established that overall this price difference ranged in the rip between
2	concerning the right of deduction of vat borne on certain types of means of transport
2	sixth protocol to the european convention for the protection of human rights and fundamental freedoms
2	eusec rd congo is the only structure fully dedicated to reforming the military security sector
2	swords cutlasses bayonets lances and similar arms and parts thereof and scabbards and sheaths therefor
2	contains for sickness cash benefits a protection period during which sickness cash benefits are granted
2	following significant changes in the policy guidelines which led to the conclusion of this protocol
2	conclusion of the partnership and cooperation agreement between the european communities and their member states
2	an effective implementation of cross compliance needs verification of respect of obligations at farmers level
2	if he declines the post offered to him he shall retain his right to reinstatement
2	council position at first reading with a view to the adoption of a regulation o
2	laying down the obligations of operators who place timber and timber product on the market
2	ad hoc working group on further commitments for annex i parties under the kyoto protocol

THE END