A Generic Approach to Processing Parallel Corpora of the Europarl for Distributional Discourse Patterns

Jolanta Mizera-Pietraszko
Wroclaw University of Technology

Abstract

This work presents a generic approach to building a featured test set in the light of further query processing by multilingual information retrieval systems. Language pair phenomena are found crucial points both in the process of query translation and document retrieval. Therefore, our approach aimed at improvement of machine translation quality is presented on the base of the French aligned to English version of the Europarl corpora. We investigate frequency of the language discourse occurrences commonly used in speech and writing. Linguistic structures are extracted from the corpora together with their representatives in order to create a featured test set of English and French grammatical patterns. In multilingual retrieval process a selection of patterns with the highest frequency simultaneously in these two languages constitutes an indication to a user of such a query formulation that may result in the most relevant system responses. This study of linguistic properties shows how to make the query easier for automatic translation and consequently to improve the system responsiveness.

1. Introduction

This study is partly based on the Author's invention number P387576 registered by the Patent Office of the Republic of Poland on 23.03.2009. It has also been motivated by the approach proposed in the project carried out by the University of Cambridge ESOL Examinations in 2007 [20]. It was aimed at some specific aspects of linguistic analysis of an English language for teaching purposes. Teachers of English were given a set of tests for checking their pragmatic knowledge of the most specific linguistic features and their names, mostly not quite common even amongst the native-speakers. Each test ended with an Answer Key which indicated the linguistic areas needed improvement by the Teacher and therefore how to teach the students in order to avoid the language gap.

The discourse analysis is undertaken both from computational as well as linguistic and cognitive perspectives. From computational linguistics viewpoint discourse analysis is aimed at investigation of language functions along with their forms and interpretation of the meaning in a particular context. Cognitive discourse analysis provides a fresh angle on distribution of linguistic patterns in individual texts. According to Beaugrande (1981) typical discourse of a written text standards should include:

- Cohesion - grammatical correlations between parts of a sentence;
- Coherence - overall order of statements makes the utterance fluent and concise;
- Intentionality - the target reader subconsciously receives the message connotation;
- Acceptability – the utterance addresses a target reader or an audience so that the language should be appropriate for the cultural, educational level as well as the outlook;
- Informativeness – expressing the essence of utterance;
- Situationality – referring to situational circumstances;
- Intertextuality - the interpreters' schemata or inference;

However, these standards convey the linguistic units that have been recognized as crucial points [1] in the process of translation performed by MT (Machine Translation) systems. Hutchins mentions that in an Example-Based translation model the most probable target language equivalent is the one that is the most similar to the source language phrase. Considering translation quality [9], we notice that the outcome, usually called a gist translation, results from the fact that most users are not interested in profiling their queries. Therefore, they prefer natural language to query formulation with the use of e.g. Boolean operators, field restricted search or keywords.

Exploiting the Europarl Parallel Corpus [3], we consider similarities of the two language pairs that is EN-FR and FR-EN in order to concentrate on the linguistic phenomena used at the stage of query formulation that improve translation quality of MT systems. At this stage, using Text Analysis software, we investigate concordances of the grammatical aspects that commonly occur in the corpora. In addition, we take into account the language style and register that influence our findings. Thus, our research question is how to improve performance of the information retrieval systems which involve a database of documents queried by the users. Usually, the documents are parsed into some internal representations. The queries comprise the same rules and hitherto follow the same structure. Those preprocessed documents, which are indexed by the query, constitute the ranked list of the system responses associated by the system self-score. Unlike this traditional concept, we investigate language pair phenomena in the context of the specific structures between the languages finding it a probable distinctive factor which can facilitate the translation process and the cross-language retrieval accordingly. In particular, we associate each member of the subcategory with a specific identifier in such a way that all the grammar structures analyzed in the equivalent context for any of these two languages represent the same identifier. Our featured test set incorporates the speakers' utterances of the Europarl corpora to form the distribution of the linguistic features in parallel collections. The knowledge-based approach proposed constitutes a novel attempt to cross-language information retrieval as an investigation of distinctive factors rather than of language similarities like e.g. interlingua model. Overall, the contribution of this study is two-fold: we propose a holistic approach to explore a language as a whole since the units proposed constitute a complete set of grammatical patterns, and we study each of the language structures accordingly to these distributional patterns.

2. Related Work

A user attempts to query a cross-language, or a multilingual system to acquire his knowledge about some issue hoping to find a lot more information than from a monolingual system. Instead, he retrieves much less than that due to the machine translation quality performed the system. Therefore, this project is a subject to exploratory research aimed at investigation of relationship between translation quality and the cross-language systems responsiveness. We study, in particular, English and French linguistic phenomena for translation efficacy in question answering, information retrieval, search engines and digital libraries [21]. However, in this part of our project we explore the problem from a perspective of processing the Europarl bitexts for specific linguistic relations that limit, to some extent, the query preciseness and affect the information system ranking.

Linguistic, or language-pair phenomena are distributional language patterns in terms of their co-occurrences in texts or verbatim forms.

Discourse analysis of the text in the light of building data collections has been extensively studied by the research groups over around the last fifty years [16][17]. However, some of the most recent advances base on natural language presented in different forms and styles. Joan Bresnan and Ron Kaplan, a psychologist at Harvard formulated a theory known as Lexical Functional Grammar (LFG) which relies on a native-speaker's generative knowledge of the mother tongue modeled by grammatical structures having their formal representations in the form of a net comprising constituent and functional structures (c-structures and f-structures) [19]. The text, which is an example of two complex sentences shown in Fig.1, is split into sentences, formatted and parsed. The syntax f-structures (functional structures) are extracted to which some lemmas and annotations are added. In the last stage, the Dependency Triples are created as presented below.

This is software that takes two strings of space separated words as input and aligns matching words between the two strings. Alignment is done over several stages, where each stage uses different criteria to find candidate matching tokens from the two strings to align.

```
       pron_form : this
pred : be
tense : pres
num : sg
pers : 3
xcomp : subj : pred : pro
               pron_form : this
        pred : software
       num : sg
       pers : 3
       relmod : topicrel : pred : pro
                            pron_form : that
                 coord : 1 : subj : _6580
                             pred : take
                             tense : pres
                             num : sg
                             pers : 3
                             obj : spec : quant : pred : two
                                   pred : string
                                   num : pl
                                   pers : 3
                                   adjunct : 2 : pform : of
                                      obj : adjunct : 3 : pred : space
                                                                 num : sg
                                                                 pers : 3
                                                          4 : pred : separate
                                                                 tense : past
                                                     pred : word
                                                     num : pl
                                                     pers : 3
                             adjunct : 5 : pform : as
                                          obj : pred : input
                                                num : sg
                                                pers : 3
                         6 : subj : _6580
                             pred : aligns
                             tense : pres
                             num : sg
                             pers : 3
                             obj : adjunct : 7 : pred : matching
                                             8 : pform : between
                                                 obj : spec : det : pred : the
                                                              quant : pred : two
                                                       pred : string
                                                       num : pl
                                                       pers : 3
                                   pred : word
                                   num : pl
                                   pers : 3
                 subj : _6580
                 coord_form : and
                 resolved : topicrel
```

Figure 1. F-structure Tree of the input text

Comparison of speech segments with a hierarchical structure of a sample discourse aimed at location of the emphasized parts of the utterances according to the segment category and level is proposed in [10]. In the lexical database WordNet, developed under direction of G. Miller, the groups of word forms called synsets represent variety of concepts interlinked into semantic and syntactic co-references navigated by a browser. The options in the pull-down menu include example sentences with the word submitted to the browser glosses, database locations, lexical file information or the sense keys or numbers [11]. During the MT Summit XI [12], a related research presents a similar approach to prepare data for translation indicating a vital role of removing duplication of documents, division of the text into the segments and its conversion into the format required by the MT system. Parts of discourse require knowledge about phonetics, morphology, syntax, semantics, pragmatics and discourse defined as linguistic units larger than a single utterance [13]. Another work proposes a manual categorization of newspaper information adopting a certainty model of level(absolute, higher, moderate and low), perspective (representing either the author's or a reported viewpoint, or alternatively the third party's viewpoint that is directly or indirectly involved in the event), focus (abstract or factual information-based) and time (past, present, or future) [14]. The model gives very promising results in comparison to others. Language discourse is also analyzed in the view of building a dialog memory system that interprets the user's query as an interactive communication with the system. The notion of discourse relies on semantic representations of the expressions denoted by a matching algorithm [15]. Preprocessing source language requires knowledge about properties and meaning of the word (morphology level - stemmer), relations between words (syntax level - wrapper), conceptual knowledge about the sentence meaning on the whole (semantic level - parser) or in some cases making complex sentence from two simple ones (merger) [4]. In this work we propose to develop a holistic approach to discourse referenced in this section by considering all the possible grammar structure types from the natural language perspective.

2.1. The Featured Test Set

Apart from the ordinary rules there are some crucial points that most MT systems usually fail to perform[5]. These include some grammar rules, long or more complex sentences, terminology, phonology and many more that even the professional translators have difficulties in interpretation.

Our featured test set is built in conformity with the ESOL project aimed at discourse analysis for teaching purposes [20]. It covers the following areas: lexis, morphology, coherence and cohesion. Some structures like e.g. morphology belong to more than one category to be compliant with the main areas, so that we agreed to analyze each of the areas as a whole in the text. In other words, the corpus is processed for one area every time.

The figure below shows the categories within the four areas called here linguistic features.

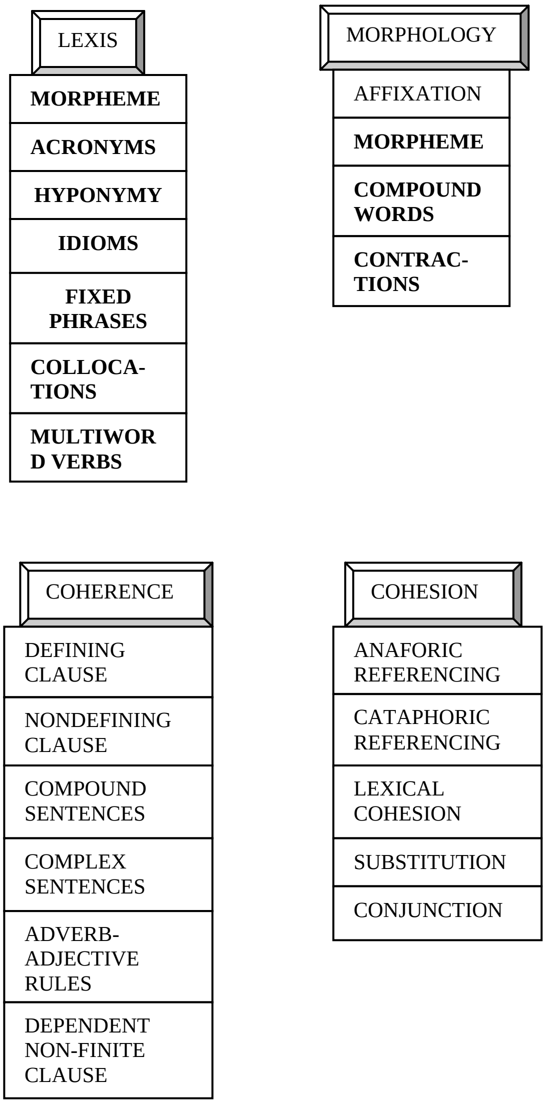


Figure 2. Grammatical areas covered by the discourse analysis of the text

At the beginning, our test consists of 31 parts being the linguistic features in the Figure above, each of which belongs to one of the four categories presented and thus is processed separately. The analysis includes six adverb-adjective rules which when added to the number of our subcategories makes the total of thirty-one structure kinds.

At first, every feature contains a model sentence before processing the Europarl. However after that, the set will be expanded by the numeral phrases extracted from the corpora depending upon the variety of structures belonging to one particular type. Thus, the total number of sentences or phrases is not defined.

3. Discourse Analysis of the Europarl Corpora Content

In this section we present the Europarl corpora and our framework for future processing the two language versions. In addition, we explain the measures used to determine the generic structure of the databases. The corpus has been built from the Proceedings of the European Parliament by the research group from University of Southern California in 2002. The newest version of the corpus was released in 2003. It contains speakers' utterances published in 11 languages including English from a period of 1996 to 2003. For detailed description see[1].

```
Reprise de la session
Je déclare reprise la session du. Parlement européen...
<CHAPTER ID=1>
Resumption of the session
<SPEAKER ID=1 NAME="President">
I declare resumed the session of the European Parliament adjourned on….
<P>
```

Figure 3. A standard format of fr-en.tgz file with metadata and the source French text

The size of the French corpus aligned to English is 103 MB. It is available both text as well as sentence-aligned and contains 69,700 speaker turns. They are divided into 273,174 paragraphs of 2-5 sentences each. The paragraphs consist of 746,147 sentence-pairs that include 20,893,546 words altogether[2]. The sentence alignment is based on the Church and Gale algorithm[3] which provides probabilistic score for the corresponding paragraphs and then sentences counted as a difference of their lengths and its variance in characters as units.

$$\varphi_t(e, f) = e - f \exp \frac{c_f}{c_e})\sqrt{6.8 \times f}$$

probabilistic score with 6.8 as a variance and

an expected value of a number of characters in two languages

$$D(c_i, \varphi_t) = -\log P(c \,|\varphi_t) = -\log 2\{1 - P(|\varphi_t|\}$$

distance measure between two translations

P($|\varphi_t|$ - probability with standardized normal distribution of

independed characters with a magnitude $|\varphi_t|$

$f(c_i) = \{c|C_1, c_2, ..., c_n)$ each character can match

some characters in a target language

Bayes Theorem

$$\mathrm{P}(c_i|\varphi_t) = P(\varphi_t|c_i) \times P(c_c) / P(\varphi_t)$$

Thus, the input file is structured into paragraphs called hard regions and the sentences called soft regions which are divided into words. The score is calculated in order to find the maximum sentence alignment function which, in turn determines whether to delete the corresponding region {0-1} or {1,0}, substitute it according to the rules provided by the system {1-1}, expand it {1-2} or merge {2-2}. For any language pair the vast majority of sentences match one to one.

$$\underset{\{s,t\}}{arg\ min}\ f(s,t) \in \{s \mathrm{I} \forall t : f(s) \cong f(t)\} \qquad (\ \mathbf{1})$$

s – a sentence length in a source language

t - a sentence length in a target language

The function *f(s.t)* is defined as a minimum distance between the sentences – one being a translation of the other one. Assuming maximum sentence match 2-2 the function *f(s, t)* is extended to four arguments as follows:

$f(s_1, t_1, 0, 0)$ substitution of s_1 *with* t_1 **(2)**

$f(s_1, 0, 0, 0)$ deletion of s_1

$f(0, t_1, 0, 0)$ insertion of t_2

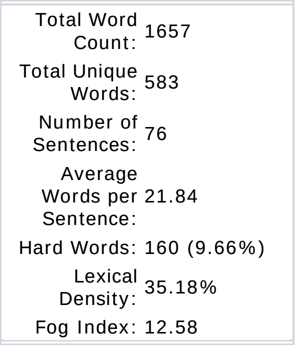
$f(s_1, t_1, s_2, 0)$ contraction of s_1, s_2 to t_1

$f(s_1, t_1, 0, t_2)$ expansion of s_1 to t_1 and t_2

$f(s_1, t_1, s_2, t_2)$ merger of $s_1 + t_1$ and $s_2 + t_2$ and match of both of them

Having acknowledged the low error rate of around 4% in the sentence alignment, we process the parallel texts to establish occurrences of the linguistic aspects of grammar shown in Figure 1. Apart from some obvious rules like inversion of noun and adjective in English and French, we need to find out others that occur the most often in natural language. The procedure is carried out in the following steps [13]:

- Corpus splitting into two separate English and French corpora
- Extraction of the speaker's turn files from the corpora
- Conversion of the files into the formats required by our software
- Extraction of the chunks that fulfill the test set assumption presented



Total Word Count:	1657
Total Unique Words:	583
Number of Sentences:	76
Average Words per Sentence:	21.84
Hard Words:	160 (9.66%)
Lexical Density:	35.18%
Fog Index:	12.58

Figure 4. A simple analysis of the word types in the file extracted from the English corpora

[1] Available at http://www.isi.edu./~koehn/publications/europarl/

[2] Koehn Philipp: Europarl: Multilingual Corpus for Evaluation of Machine Translation, Draft

[3] Available at people.csail.mit.edu/koehn/publications/de-news/church_and_gale.ps

The figure shows statistics of a simple fragment of an English file with the readability level.

Lexical Density is measured as the proportion of the lexis to the total number of words. The text is meant difficult when the measure reaches 60-70%,

and relatively easy for reading with the density measure of 40%.

$$LD = \frac{\sum_{i=1}^{n} w_i}{N_w} \times 100[\%] \qquad (3)$$

w_i - a word different from others

N_w - total number of words in the text

The Fox Index is calculated as a sum of an average number of words in a sentence and the percentage of at least three-syllable words multiplied by 0.4. In case of technical documentation the measure is around 10 and for the professional prose, 18[4].

$$F(I) = \{ \sum_{s=1}^{n} w_s + w_{3s}[\%] \} \times 0.4 \qquad (4)$$

w_s - number of words calculated as a mean of sentecne length

w_{3s} - number of at least three - sylable words in the text in [%]

Thus, regarding these two measures, this is a clearly readable fragment of the text. The measures give us an insight into the way in which we need to process it as well as the techniques of extracting the linguistic features.

3.1. Processing the Corpora for Data

Our procedure is performed separately for every linguistic subcategory. The average number of words in the file is 26,000 which can be recalculated into 150,000 tokens.

The displays of acronym concordances using the KWIC technique (Keywords in Context [5]) revealed that all the words "can't" and "don't" have been changed to their full forms according to the formal register adopted by the European Parliament. The fact has been approved by the report describing the preparation procedure of the corpus [4].

Due to the number of the categories listed in section four, we present our approach to only one of them called morpheme as a smallest linguistic unit that assigns a semantic meaning to the word. Therefore, we construct bilingual concordances for use of the following morphemes: -ed,- ble, -ly, -al, -ory, -tion, -ity and -less.

```
536 ed
2  of the European Parliament adjourned on
   Friday 17 December
3   in a number of countries suffered a
   series of natural disasters
```

Fig 5. Concordances of morpheme "ed" followed by their line number in the file

In the next step, we build a list of keywords presented here in green in order to remove the words with suffix "ed" witch are not morphemes and to find only the selected keywords' frequencies.

```
5 accepted
 72  and which was also explicitly accepted
     by President Prodi , who
174   . /They have either been accepted or
     transposed with no
193   position the Council has accepted six
     of Parliament 's ten
241   all of the amendments accepted by the
     Commission and
359   and Social Affairs but not accepted by
     the Committee on Regional

1 accomplished
671     if  anything  at  all  is  to  be
     accomplished there . /For example
2 accused
 31  for Alexander Nikitin to be accused of
     criminal activity and
699  If we do not wish to stand accused of
     pursuing a cost – intensive
```

Fig 6. Concordances of the keywords in context extracted from the same Speaker's Turn file

The aim of the limitation is to focus on the verbs only. For every category we build statistics of the keywords to find the highest probability of the term category in natural language. In other words, we consider only the most common occurrences of the category taken from our test set. In figure 5 we extract only words with frequency 6-19. We notice that the word frequency is in inverse proportion to the number of words.

```
Word      Number      Cumulative
Frequency of Words    Vocabulary
   1       139          139
   2        42         181
   3        20         201
   4        14         215
   5        10         225
   6         4         229
   7         1         230
   8         1         231
   9         4         235
   10        1         236
   11        1         237
   16        2         239
   19        1         240

Cumulative   Percentage  Percentage
Word Count   Vocabulary  Word Count
   139       4.92210      25.93284
   223       6.40935      41.60448
   283       7.11756      52.79851
```

[4] TextStat – Simple Text Analysis Tool, version 1.52, 2002, available at http://www.niederlandistik.fu-berlin.de/textstat/TextSTAT-Doku-EN.html

```
    339        7.61331      63.24627
    389        7.96742      72.57463
    413        8.10907      77.05224
    420        8.14448      78.35821
    428        8.17989      79.85075
    464        8.32153      86.56716
    474        8.35694      88.43284
    485        8.39235      90.48507
    517        8.46317      96.45522
    536        8.49858     100.00000
```

Figure 7. Statistics of the morpheme keywords extracted from the British corpus

Afterwards, we move to the French version of the same file (The Speaker's Turn) and start the procedure from the very beginning. The French verbs ending with "-u", or "-i" are translations of the English verbs ending with "-ed". The figure below shows two keywords ending with "-u".

```
10 niveau
237 les choses en main au niveau politique .
/Il est imp
290  vue  de  la  reconnaissance  au  niveau
européen des certificats
292 nessaire d ' harmoniser au niveau de la
Communauté les conditions
311  Il  faut  tenir  compte  du  niveau  des
difficultés que les
479  des  définitions  reconnues  au  niveau
international , quand cela
505  rement  ,  maintenir  un  haut  niveau  de
compétence technologique
742 son avis au Conseil réuni au niveau des
chefs d ' État ou de
824 transmis au Conseil réuni au niveau des
chefs d ' État ou de
868 changements de procédure au niveau de
notre règlement pourraient
880 même pied que le Conseil au niveau du
comité de conciliation

10 tenu
 82 qu ' il a également été tenu compte des
questions qui touchent
139 est une nécessité , compte tenu de leur
rôle structurant dans
235  les embouteillages  . /Compte  tenu  des
très bons résultats de
312  -- ce qui est possible compte tenu de l
' analyse proposée --
477 faire remarquer que , compte tenu de l '
utilisation limitée
480  t hors d ' usage . /Compte tenu de ces
remarques , et tout
562  et spatiale a toujours tenu une place
importante et dans
564  . /Je pense qu ' il a été tenu compte
dans ce cas de tous
812  et des citoyennes  . /Compte  tenu  de
cette situation intenable
   879 dernière minute , compte tenu de la
nature même des négociations
```

Figure 8. Concordances of the keywords in context extracted from the French Speaker's Turn file

Results presented in figure 8 indicate that word "niveau" is an adverb, not a verb so we remove it from the keyword list as well as all the words not being verbs. The matches are specified in alphabetical order.

For testing occurrences of coherence in these two corpora we built a list of linking words, while cohesion features were studied by searching concordances of "which" and "that" using the KWIK method.

```
question , as expressed in the resolutions
which it has adopted , clear
one  , there is still no Dutch channel  ,
which is what I had requested
the Vice - President , Mrs de Palacio  ,
which appeared in a Spanish a

series of natural disasters that truly were
dreadful
My question relates to something that will
come up on Thursday and which
an objection of that kind to what is merely
a report
```

Fig. 9. Extraction of cohesion features from the English corpora.

The text fragments have been selected to show the relative pronouns in different contexts that they occur in the corpora which is a basement into selection of the contexts for our linguistic features.

3.2. Statistical Analysis of the Corpora Linguistic Features

The section is an attempt to follow the statistical approach presented in [6]. Therefore, for each category we process the whole parallel corpora to recognize the language features and in order to evaluate the translation quality. The Europarl corpora translation is based on the IBM statistical model 4 and tested with an ICI decoder.

The BLUE score of the translation quality result is 0.2787 for the French-English version and 0.2555 for the English-French version respectively. The metric is computed as a product of the weighted geometric mean of the corpus length multiplied by its brevity score [7]. The BLUE precision is measured in uni-grams (words) or bi-grams (phrases) in the document [5].

$$Br_p = e^{(1-r/c)} \text{ where } c \le r \quad \textbf{(5)}$$

Br_p - brevity penalty

c - total length of the coprus

r - total length of the sentences of the closest length to the candidate sentence length

At this stage, our featured test set has been expanded by adding the phrases extracted from the corpora to the sample examples for each feature. To avoid ambiguity, the frequency of linguistic features is presented for both English and French versions in the overall percentage that they occur in each corpus.

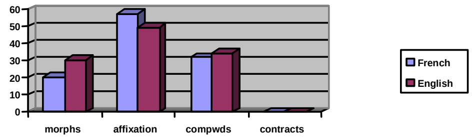


Figure 11. Morphological analysis of the Europarl corpora

Morphemes are analyzed in relation to lexis and then in morphological context as seen in figures 10 and 11. Compound words were extracted by hyphen and, in the next step, selected accordingly.

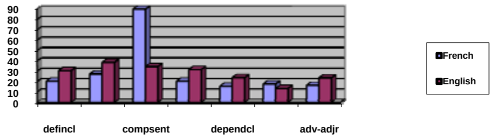


Figure 12. Coherence in the Europarl corpora

Coherence was extracted from the corpora by selecting commas and analyzing their surroundings in case of clauses and compound sentences, or searching for relative pronoun "that/which" in definite clauses. We considered some adjective-adverb rules like irregular adverbs, qualitative + classifying adjectives, and collocations adverb + adjective. Figure 12 shows the results.

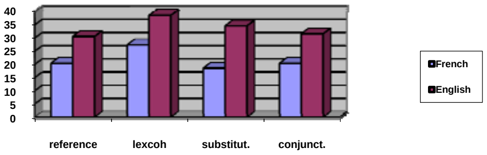


Figure 13. Cohesion in the Europarl corpora

Cohesive devices are recognized by Halliday and Hasan [8] as referencing, substitution, lexical cohesion, ellipsis and conjunction. At this point, the corpora was searched for words "one/ones/it", "after/before", "and/but", "however" 'and/ because" (mnemonic FANBOYS). The statistics are illustrated in figure 13.

4. Conclusion

In our featured test set we have collected examples of some linguistic structures from the ESOL project as the linguistic units representatives and added a number of the sentences from the Europarl corpora respectively. For evaluation of a machine translation system each feature can be processed separately in particular to test both the model and the algorithm that performs the grammar rule being analyzed.

Based on some online translations we notice that the linguistic structures that achieve the best evaluation score may be eliminated from our further analysis aimed at improving up-to-date translation working models, in particular Interlingua in its hybridized forms. Since the backbone of the model is a Latin language, the translation quality depends on the similarity of the language pair to the Latin grammar rules. However, it should be noted that the present trend is towards expanding translation models to the language pairs that differ significantly from some Latin language rules like e.g. Romance, Indo-European, Slavic, Semitic languages or other language groups.

Our findings show that both language style and the formal register required e.g. in the European Parliament determine co-occurrences of the some structures like compound sentences, affixation and collocations so that the knowledge-sensitive translation models perform them better.

Regarding the linguistic units analyzed, only those with the comparable frequency in these two languages simultaneously achieve the highest translation score which results in relevance closer to the monolingual systems respectively. As presented in figures 10, 11 and 12 these units are idioms, contractions and non-defining clauses.

The approach presented indicates that only detailed distinction of grammar rules allows the designer to evaluate MT systems more efficiently and to indicate the improvements required in order to achieve possibly human like translation quality. Our method is efficient and can be ported to other language pairs.

In our further study we plan to concentrate on the translation models in order to analyze the performance of the system components in relation to the linguistic structures presented in this report.

5. Acknowledgments

The project is supported by a research grant "GRANT – Support for Research Work by Scholarships for Doctoral Students" undertaken under the Program entitled Working Human Capital, Priority VIII, Activity 8.2 Transfer of Knowledge, Subfield 8.2.2 Regional Strategies of Innovation, awarded from the Regional Marshal Office and sponsored in 85% by the European Social Fund, in 7.5% by the National Budget and in 7.5% by the Regional Marshal Office.

6. References

[1] Hutchins J.,: “Machine Translation History”, *Encyclopedia of language and Linguistics*, Second Edition, vol. 7 pp.375-383, Oxford, Elsevier, (2006)

[2] White J.: “Toward an Automated Task-Based MT Evaluation Strategy”, *Proceedings of the Workshop on MT Evaluation at LREC*, (2000)

[3] Boitet Ch.: “Factors for Success (and failure) in MT”, *Fifth MT Summit, Luxemburg*, (1995)

[4] Munning Ch., Schutze H.: Foundations and Statistical NLP, The MIT Press, Cambridge, Massachusetts, London, (2000)

[5] Ratliff E.: “Me Translate Pretty One day”, *WIRED*, issue 14.12, (2004)

[6] Geer D.: “Statistical MT Gains Respect”, *IEEE Computer*, (2005)

[7] Callison- Burch C.: “Re-evaluating the role of BLUE in MT Research”, *11th Conference of the European Chapter of the Association for Computational Linguistics*, CACL 2006

[8] Halliday M.Rugayia H.: “Cohesion in English”, *Longman*, (1976)

[9] Lewis D.: PC-Based MT; “An Illustration of Capabilities in Response to Submited Test Sentence”, *MT Review No 12, the Periodical of the Natural Language Translation*, Specialist Group of the British Computer Society, Issue No 12, (2001)

[10] Lisa J. Stifelman: “A Discourse Analysis Approach to Structured Speech”, *the AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*. March 27-29, Stanford University

.

[11] Ch. Fellbaum: *WordNet An Electronic Lexical Database*, ISBN: 13-978-0-262-06197-1, The MIT Press, (1998), www.wordnet.princeton.edu

[12] Christopher Cieri, Stephanie Strassel, Meghan Lammie Glenn, Lauren Friedman “Linguistic Resources in Support of Various Evaluation Metrics” *MT Summit XI*, Workshop on Automatic Procedures in MT Evaluation, Copenhagen, September 9-14, (2007)

[13] Daniel Jurafsky, James H. Martin : “Speech and Language Processing: An Introduction to Natural Language Processing”, *Computational Linguistics and Speech Recognition*, Second Edition, Prentice-Hall, (2009)

[14] Victoria Rubin, Elizabeth D. Liddy, Noriko Kando: *Certainty Identification in Texts: Categorization Model and Manual Tagging Results*, Springer, Dordrecht, The Netherlands. (2006)

[15] Christian Hying, Betreuer: Ulrich Heid Jan van Kuppevelt,Sunna Torge: "A Context Manager Exploiting Discourse Structure”, *Diplomarbeit*, *IMS, Sony Stuttgart*, (2002)

[16] Levine Philip, Scollon Ron:*Discourse Technology – Multimodal Discourse Analysis*, Georgetown University Press, pp.240, ISBN: 9781589011014, 2004.

[17]Deborah Schiffrin, Deborah Tannen, Heidi E. Hamilton (Eds.), *Handbook of Discourse Analysis:* Blackwell Publishers, 872 pages,2003.

[18] Beaugrande, R., and W. U. Dressler 1981, *Introduction to text linguistics.* London: Longman. 270 pages. Location: Dallas SIL Library 410 B 375.