Faculteit Wetenschappen

ParaSense: Parallel Corpora for Word Sense Disambiguation

Het gebruik van parallelle corpora voor het automatisch desambigueren van polyseme woorden

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de Wetenschappen: Informatica
aan de Universiteit Gent te verdedigen door
Els LEFEVER

Dit onderzoek werd gefinancierd door het
Onderzoeksfonds van de Hogeschool Gent (HOF).

Promotors:
Prof. Dr. V. Hoste
Prof. Dr. M. De Cock

Gent, 2012

© 2012 Els Lefever
All rights reserved. No parts of this book may be reproduced or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author.

ISBN 978 90 7083 090 8

Cover illustration:
Sittin' on the dock of the bay
by Korneel Detailleur

Noscitur a sociis
You shall know a word
by the company it keeps.
– J. Firth (1957)

Voor Pablo en Ilya

Abstract

This thesis presents a machine learning approach to Word Sense Disambiguation (WSD), the task that consists in selecting the correct sense of an ambiguous word in a given context. We recast the task of disambiguating polysemous nouns as a multilingual classification task. Instead of using a predefined monolingual sense inventory such as WordNet, we use a language-independent framework where the word senses are derived automatically from word alignments on a parallel corpus. As a consequence, the task is turned into a cross-lingual WSD task, that consists in selecting the contextually correct translation of an ambiguous target word.

In order to evaluate the viability of cross-lingual Word Sense Disambiguation, we constructed a lexical sample data set of twenty ambiguous nouns. For the creation of the multilingual sense inventory, we first applied word alignment to a six-lingual parallel corpus and manually clustered the obtained translations by meaning for all target words. The resulting multilingual sense inventory then served as the basis for the annotation of the test data.

The ParaSense WSD system we propose in this thesis presents a truly multilingual classification-based approach to WSD that directly incorporates evidence from four other languages. We built five classifiers with English as an input language and translations in the five supported languages (viz. French, Dutch, Italian, Spanish and German) as classification output. The feature vectors incorporate both local context features as well as translation features that are extracted from the aligned translations. The hypothesis underlying the construction of a multilingual WSD system is that adding translational evidence from multiple languages will be more informative than using only monolingual or bilingual information. We believe it is possible to use the differences between the languages to obtain certain leverage on word meanings and better disambiguate a polysemous word in a given context.

The experimental results confirm the validity of our approach: the classifiers that employ translational evidence constantly outperform the classifiers that only exploit local context information for four out of five target languages, viz. French, Spanish, German and Dutch. Furthermore, a comparison with all systems that participated in a dedicated cross-lingual Word Sense Disambiguation competition revealed that the ParaSense system outperforms all other systems for all five target languages.

As our system extracts all information from the parallel corpus at hand, it is a very flexible and language-independent approach that allows to bypass the knowledge acquisition bottleneck for Word Sense Disambiguation.

Samenvatting

Dit proefschrift gaat over het gebruik van lerende technieken voor de automatische desambiguering van woordbetekenissen, of Word Sense Disambiguation (WSD) in het Engels. Deze taak heeft als doel om de correcte betekenis van een ambigu woord te voorspellen aan de hand van contextuele informatie. De taak om polyseme zelfstandige naamwoorden te desambigueren werd hierbij geherformuleerd als een classificatietaak. Voor het correct toekennen van de juiste betekenis gebruiken we geen vooraf gedefinieerde eentalige betekenislexicons zoals WordNet, maar een taalonafhankelijke aanpak waarbij de woordbetekenissen automatisch afgeleid worden door middel van woordalignatie op een parallel corpus. Bijgevolg wordt de taak omgezet in een meertalige WSD taak, waarbij het de bedoeling is om een contextueel correcte vertaling te voorspellen voor een ambigu woord in een gegeven context.

Om de haalbaarheid van deze meertalige aanpak te testen, hebben we een testcorpus aangelegd voor twintig ambigue zelfstandige naamwoorden. Voor de creatie van de meertalige betekenislexicons werd in een eerste stap automatische woordalignatie toegepast op een zestalig parallel corpus. Vervolgens werden alle geëxtraheerde vertalingen per doelwoord manueel gegroepeerd per betekenis. Het resulterende meertalige betekenislexicon werd daarna gebruikt voor de manuele annotatie van de testdata; per ambigu woord werden vijftig zinnen geselecteerd als testcorpus.

Het ParaSense WSD-systeem dat we beschrijven in dit proefschrift volgt een meertalige classificatiegebaseerde aanpak voor WSD en gebruikt hiervoor infor-

matie uit vier bijkomende talen. Om de aanpak te testen hebben we vijf lerende systemen gebouwd die Engels als brontaal nemen en de contextueel relevante vertaling van het ambigue woord voorspellen in vijf doeltalen (Frans, Nederlands, Italiaans, Spaans en Duits). De informatie die gebruikt wordt voor de constructie van de trainings- en testdata bevat zowel informatie uit de lokale context van het Engelse ambigue woord als informatie die geëxtraheerd wordt uit de corresponderende vertalingen. Voor het Engels-Franse systeem wordt bijvoorbeeld informatie opgehaald uit de Italiaanse, Nederlandse, Spaanse en Duitse vertalingen. Onze motivatie voor een meertalige aanpak komt voort uit de overtuiging dat het gebruik van vertalingen uit verschillende talen informatiever is dan informatie uit één of twee talen. Daarbij wordt uitgegaan van de hypothese dat de verschillen tussen deze talen kunnen aangewend worden om een beter onderscheid tussen de woordbetekenissen te ontdekken, en dat we bijgevolg beter in staat zouden moeten zijn om ambigue woorden op een correcte manier te vertalen in een bepaalde context.

Onze experimentele resultaten bevestigen het nut en de voordelen van onze aanpak: de systemen die meertalige informatie gebruiken behalen betere resultaten dan de systemen die enkel informatie uit de lokale context benutten. Deze conclusie wordt bevestigd voor vier van de vijf doeltalen, namelijk Frans, Spaans, Duits en Nederlands. Verder hebben we de resultaten van onze aanpak ook vergeleken met die van vergelijkbare systemen die hebben deelgenomen aan de cross-linguale WSD-taak in de internationale SemEval-competitie. Daarbij bleek ons ParaSense systeem de beste resultaten te behalen voor alle vijf de doeltalen.

Doordat ons ParaSense systeem alle nodige informatie uit een parallel corpus extraheert, biedt het een heel flexibel en taalonafhankelijk alternatief voor gesuperviseerde systemen die afhankelijk zijn van grote geannoteerde corpora en vooraf samengestelde betekenislexicons.

Acknowledgements

A lot of people have helped and supported me along the long and winding road to my PhD thesis. Finishing my dissertation would not have been possible without their aid.

Foremost, I would like to thank my promotor Prof. Dr. Véronique Hoste, who offered me a very stimulating research environment. I truly admire her perseverance and energy, and the enthusiasm with which she heads the Language and Translation Technology Team. Her encouragements, scientific input and sometimes challenging deadlines forced me to push my limits and achieve goals that I would never have thought reachable. In addition, Véronique also became a dear friend. Thank you so much Véronique.

I'd also like to give special thanks to my other promotor Prof. Dr. Martine De Cock, who gave me the opportunity to obtain this PhD and who carefully revised my publications. Together with Timur, we discovered the intriguing world of Web People Disambiguation, that led to my first A1 journal publication. I am also very grateful to the members of my PhD committee, Prof. Dr. Walter Daelemans, Prof. Dr. Antal van den Bosch, Prof. Dr. Johan De Caluwe and Prof. Dr. Lucia Specia. Their feedback was very valuable and helped me to improve my dissertation. My gratitude is also extended to Oier, who very patiently uncovered the secrets of LSA and SVD, two three-letter acronyms that controlled my thoughts for several weeks and provoked panic attacks and despair.

I warmly thank Korneel Detailleur for the wonderful illustration I used on the cover of my PhD, Gitte for doing the cover layout and Kristien for proofreading the text.

Next, I'd like to give a heartfelt, special thanks to my fellow members of the LT3 team, a bunch of wonderful people that help me love my job. I want to thank Bram, Geert and Sofie for being such nice colleagues, Klaar for sharing laughs and the XLOC courses, Isabelle for her uncurbed optimism and chocolate treats on days when everything went wrong, Bart for his overwhelming insights, intelligent remarks, wonderful pictures and for jolting my green conscience, Peter for the numerous times he assisted me on IT issues and the wonderful pralines he makes and Orphée for proofreading my dissertation and being such a wonderful person. I especially want to thank my office mates Lieve, Marjan and Kathelijne, for sharing the good and bad times, and discussing the joy and sorrow of life.

There have been many other friends, too many to name. Special thanks go to Ellen (and Filip) for introducing me into the magic world of computational linguistics and sharing a lot of time on playgrounds, and to Sandrina for helping me survive university, sanely. A warm thank you to Boris, who has been my "compagnon de route" for so many years, for giving me two exceptional children and for loving me as much as he could. My deepest thanks go to Ruth and Dries, who were my rocks in stormy weather, and made me laugh at times I thought I forgot how to laugh. Thank you so much!

I have reserved the last words of gratitude for the persons I am most attached to. I want to thank my parents and my sister, for their unconditional love and support, for their faith in me, their encouragement and for supporting me in all my pursuits.

Thank you, Ollie, for your tremendous love and for being my best friend. Thank you for teaching me how to enjoy life and being happy again. Thank you, Pablo and Ilya, for being so wonderful and for being the most important part of my life. Thank you.

Ghent, September 2012.

Contents

Part I

Word Sense Disambiguation

CHAPTER 1

Introduction

Natural languages are ambiguous. Ambiguity can be defined as a semantic property of a linguistic expression that occurs "every time a linguistic expression can have more than one distinct denotation" or meaning (Wasow, Perfors and Beaver 2005). Ambiguity is thus not the same as vagueness. Linguistic expressions are called *vague* if "the regions [of the meaning space] they denote do not have perfectly well-defined boundaries" (Wasow et al. 2005). The adjective *tiny* for instance, always means *very small*, but its exact size is not precisely defined in the following examples:

(1) "A mountain is composed of *tiny* grains of earth. The ocean is made up of *tiny* drops of water. Even so, life is but an endless series of little details, actions, speeches, and thoughts. And the consequences whether good or bad of even the least of them are far-reaching."[1]

(2) "I am done with great things and big plans, great institutions and big success. I am for those *tiny*, invisible loving human forces that work from individual to individual, creeping through the crannies of the world like so many rootlets, or like the capil."[2]

[1] Sri Swami Sivananda

[2] William James

Two types of ambiguity are very productive in natural language: lexical ambiguity and syntactic ambiguity.

Syntactic or structural ambiguity occurs whenever a linguistic expression allows more than one syntactic parse. A classic example is “We saw the man with the telescope”, where “with the telescope” can modify either the verb *saw* or the object *man*.

Lexical ambiguity or polysemy occurs when words have multiple intrinsic meanings. One can easily deduce from example (3) that the word *bar* has multiple meanings[3].

(3) Clearly if a wine **bar** were to find that the catering side of the business brought more trade and wished to convert part of the bar area into a restaurant, this would not be a material change and would not require planning permission.

I need a visa and have failed to obtain one, Rosita likewise, and so they put us behind **bars** like prisoners.

They have the whole packet of biscuits, or three or four **bars** of chocolate, or, with our example in the café, a full fried breakfast followed by a Danish pastry.

His knowledge of classical music was very comprehensive – you only had to sing him a snatch of any symphony or concerto and he would be able to identify it immediately, but this isn’t sufficient to get you up there in front of a hundred or more qualified musicians and be able to lead them into the opening **bars** of Beethoven’s 5th, or even the Warsaw Concerto.

Educated at Swanage Grammar School and Cambridge, he pursued a legal career and was called to the **Bar** in 1972.

In common with the Pterophyllum species any fish that shows really intense colour and exaggerated black **bars** should be regarded with suspicion as this is often an indication that they are not long for this world.

When these different meanings are completely distinct from each other, we call them *homographs*, such as the word *bank* that can refer to the *financial institution* or the *river side*. It is often the case, however, that a word can denote more finer-grained sense distinctions that are interrelated or derived from each other. The noun *wood*, for example, has two related senses: (1) a piece of a tree and (2) a geographical area with many trees.

[3] `All examples are extracted from the British National Corpus at http://www.natcorp.ox.ac.uk/.`

Although many frequent words are intrinsically polysemous – the 121 most frequent English nouns have on average 7.8 meanings according to Ng and Lee (1996) – in reality humans easily disambiguate lexical expressions. Computers, on the other hand, have more problems to perform Word Sense Disambiguation (WSD), the task that consists in computationally determining which "sense" of a word is activated by the use of the word in a particular context (Agirre and Edmonds 2006). The Word Sense Disambiguation task is often called AI-complete, which means that the difficulty of the task is comparable to achieving complete natural language understanding (Ide and Véronis 1998).

Different approaches have been proposed to tackle the Word Sense Disambiguation task. Based on the resources used, these approaches are often divided into *knowledge-based*, *supervised* and *unsupervised* approaches. An overview of the main approaches to WSD is presented in Section 2.

The computational WSD task can be defined as a classification task where the possible word senses are the classes, the context of the ambiguous word provides additional evidence, and each occurrence of an ambiguous word is assigned to the correct class based on the available context information that is compared to the context information of the training examples (Lopez de Lacalle 2009). Three major issues, however, arise when applying the more traditional approaches to solving the WSD task.

1. The traditional approaches to WSD start from the hypothesis that words have a fixed set of senses (hereafter referred to as a *sense inventory*) as is the case in dictionaries. Many authors, however, have questioned the use of such an all-purpose and fixed sense inventory for WSD. Agirre and Edmonds (2006) claim that word meaning is, in principle, infinitely variable and context-sensitive, and that it is difficult to divide into sub-meanings or senses. Kilgarriff (1997) proposes to consider word senses as abstractions from clusters of corpus citations, in accordance with how current lexicographers proceed. But even there, the clustering of citations is a subjective task and lexicographers do not always agree on how to divide a dictionary entry into various senses. Moreover, some sense distinctions seem larger than others and there are no clear-cut criteria for lumping two senses together or splitting one sense into different finer-grained senses. It is therefore hard to believe that a single set of (fine-grained) word senses would be appropriate for all different Natural Language Processing applications (Kilgarriff 1997).

2. Both the knowledge-based (i.e. using dictionary information) and supervised (i.e. relying on sense-annotated corpora) approaches require manually constructed lexical resources such as sense inventories, manually-labeled corpora, etc. These resources only exist for a limited number of

languages and are very expensive and time-consuming to build, resulting in the so-called *knowledge-acquisition bottleneck*.

3. For a long time WSD has been considered a separate Natural Language Processing (NLP) task, whereas a lot of researchers argue that WSD is an intermediate task that can help to improve the performance of real NLP applications (Wilks and Stevenson 1996). The following two NLP applications have already shown to benefit from better lexical resolution:

   - **Machine Translation**: Word Sense Disambiguation is crucial for lexical choice in Machine Translation since polysemous words often have different translations depending on their meaning in a particular sentence. Different studies have already revealed significant improvements by integrating a dedicated WSD module in a statistical machine translation framework (Carpuat and Wu 2007, Chan, Ng and Chiang 2007).
   - **Information Retrieval**: a dedicated WSD module can help to distinguish different meanings of the focus word in the documents that result from a given query containing this focus word. If a user enters the query *Java*, for instance, he is mainly interested in the documents related to one of the different meanings of the word (the programming language, the coffee brand or the main island of the republic of Indonesia). Once the meaning of the word is determined in all retrieved documents, the documents could be clustered by meaning in order to present a structured overview to the user. WSD has already been shown to improve cross-lingual IR and document classification in Vossen et al. (2006), Clough and Stevenson (2004) and Agirre, Otegi and Zaragoza (2010).

This thesis presents a multilingual classification-based approach to Word Sense Disambiguation that does not use any information from annotated corpora and external lexical resources, and that starts from the following research hypotheses:

(a) The use of parallel corpora allows to partially clear the knowledge-acquisition bottleneck.

(b) Using translations instead of arbitrarily predefined sense distinctions tackles the sense-granularity problem.

(c) A truly multilingual approach that incorporates information from different languages helps the classifier to further refine the obtained sense distinctions.

(d) Predicting translations instead of abstract sense labels facilitates the integration of the WSD module into practical applications such as Machine Translation or Multilingual Information Retrieval.

We will further elaborate on the motivations and contributions of our approach in Chapter 2 of this dissertation. Section 1.1 gives an overview of the content of this thesis.

1.1 Thesis outline

This thesis, which is nine chapters long, presents research that has been carried out over the last six years. Since most chapters contain a more detailed version of work that has been presented at conferences and published in journals and conference proceedings, this research has benefited greatly from the comments of many anonymous reviewers and discussions at conferences. References to the publications are included in the text throughout this dissertation.

The thesis consists of three main parts. Part I introduces the NLP task of Word Sense Disambiguation and the main approaches that have been proposed to tackle it. Part II describes the architecture of the ParaSense system we propose for WSD, whereas Part III outlines the experimental setup and results. Below is a more detailed overview of the different chapters.

Chapter 2 introduces the Word Sense Disambiguation task, the NLP task that consists in selecting the correct sense of an ambiguous word in a given context. It summarizes the three main approaches that have been used to tackle the WSD problem: (1) knowledge-based approaches, (2) supervised approaches and (3) unsupervised approaches. It also highlights the main shortcomings of the various approaches, being the knowledge-acquisition bottleneck, the rigid sense inventories and the lack of integrating WSD modules in real applications.

Chapter 3 discusses how the WSD problem is transformed into a *cross-lingual* Word Sense Disambiguation (CLWSD) task and presents the ParaSense system. The proposed system tackles the main problems faced by the traditional approaches. The chapter presents a detailed overview of all information sources that were used for solving this CLWSD task. It first discusses the selection and preprocessing of the data sets and continues with the description of the feature vectors, that combine local context information with translational evidence.

Chapter 4 introduces two supervised machine learning algorithms that were used for the experiments: the memory-based learning algorithm as implemented in TIMBL (Daelemans and van den Bosch 2005), and the Support Vector Machine algorithm as implemented in SVMLIGHT (Joachims 1998). It also describes how Genetic Algorithms can be applied to find the optimal parameter settings for the cross-lingual WSD task.

Chapter 5 initiates the experimental part with the description of a dedicated Cross-Lingual WSD benchmark data set. For this data set, a lexical sample data set of 25 ambiguous nouns was created; 5 words for developmental, and 20 words for test purposes. The first part of the chapter explains how the sense inventory was constructed based on the translations of the words in the parallel corpus at hand, whereas the second part of the chapter describes the gold standard annotation process of the trial and test instances.

Chapter 6 describes the experimental setup that was used to evaluate the cross-lingual WSD approach. The ParaSense system takes English as an input language and provides contextually correct translations in five target languages, viz. Italian, Spanish, French, German and Dutch. The chapter summarizes the content of the training and test data sets before it presents the two applied evaluation metrics: the SemEval BEST precision metric and a more straightforward accuracy metric. The chapter concludes by presenting the most frequent translation baseline scores for all five target languages.

All experimental results are discussed in Chapter 7. The chapter starts by presenting an overview of all classification baseline scores when applying the two machine learning algorithms with their default settings on the feature vectors combining English local context and bag-of-words translation features. In addition, it describes the impact on the classification results of two optimization cycles: (1) optimize the feature space by applying latent semantic analysis on the translation features and (2) optimize the memory-based learner's parameter settings by means of a genetic algorithm. The chapter also discusses the contribution of the different translation features to the classification result, which confirms the efficacy of our multilingual approach to WSD.
A set of additional experiments were conducted to investigate the performance of the individual test words and the impact of word alignment errors on the overall classification result. The chapter concludes with positioning the ParaSense system against other state-of-the-art approaches, being all systems that participated in the SemEval-2010 Cross-lingual Word Sense Disambiguation task.

Chapter 8 explores the potential benefits of adding a dedicated cross-lingual WSD module to a statistical machine translation system. For this purpose, we have compared the output of the ParaSense system with the output of two state-of-the-art statistical machine translation systems (Google and Moses) on the French and Dutch lexical sample data set. The latter, more practical, evaluation of the ParaSense system concludes the experimental part of this dissertation.

Chapter 9 summarizes the conclusions of this thesis and presents prospects for future research.

CHAPTER 2

Word Sense Disambiguation

Word Sense Disambiguation (WSD), the task that consists in selecting the correct sense of an ambiguous word in a given context, is a well-researched NLP problem. For a complete overview of the field, we refer to Agirre and Edmonds (2006) and Navigli (2009).

WSD was mentioned for the first time as a computational problem in the early days of Machine Translation. Weaver (1949) stated in his famous Memorandum on Machine Translation that the problem of multiple meanings might be tackled by examining the immediate context:

> If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of words. Fast may mean rapid; or it may mean motionless; and there is no way of telling which.
>
> But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning . . .

Weaver's idea to use the context of the ambiguous word to perform disambiguation remains the central idea of most WSD approaches to date.

Some early work on WSD started during the 1950s. Kaplan (1955), for instance, studied to what extent the immediate context and the number of distinct senses affect the ambiguity caused by a given focus word. Most research on Machine Translation and WSD, however, was abandoned in the 1960s because of the extreme difficulty of the task, which was also mentioned in the unfavorable ALPAC report (Pierce et al. 1966).

WSD research was resurrected in the 1970s when AI approaches, developed for language understanding, were used to solve the WSD task (E.g. Wilks (1975), Rieger and Small (1979)). These approaches heavily rely on a detailed semantic knowledge representation such as manually created selectional restriction rules, lexicons, parsers, etc. Because of the lack of large amounts of digital resources and corpora, however, these approaches could only be tested on very small hand-crafted data sets.

The 1980s saw an exponential growth of research on WSD, mainly because of the emergence of new large-scale lexical resources and corpora. For an overview of the early history of WSD, we refer to Ide and Véronis (1998).

During the 1990s, three major events drastically changed further WSD research (Agirre and Edmonds 2006):

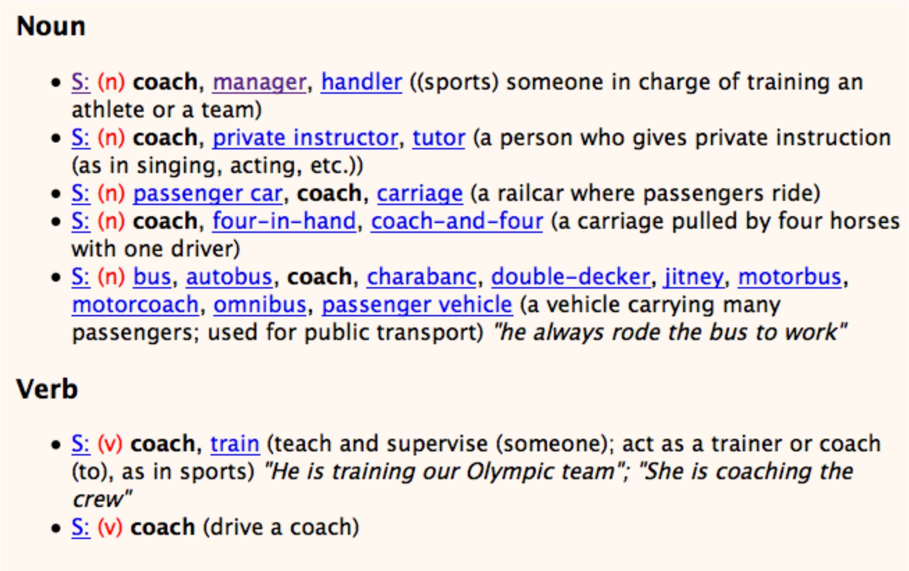
(a) The use of statistics and machine learning methods became predominant in NLP, and consequently in WSD research.

(b) The development of WordNet (Fellbaum 1998), an electronic sense inventory for English that covers in its current version about 155,000 words, organized in sets of synonyms called synsets. WordNet is structured based on hyperonymy, which results in a hierarchical structure for verbs and nouns. Today it is still the most frequently used sense-inventory for manually labeling WSD training and test corpora. Figure 2.1 shows the WordNet[1] synsets for the English word *coach*, whereas Figure 2.2 gives an overview of the different semantic relations that are stored for *coach* in the sense of *bus*.

(c) The organization of the first Senseval competition in 1998. Senseval started as an online competition for Word Sense Disambiguation[2]. The two most popular Senseval tasks were the *all-words task*, where systems are required to provide a sense label for all content words in a given text, and the *lexical sample tasks*, where systems are required to provide sense labels for all instances of a carefully selected sample of words.

[1] As stored in the WordNet 3.1 online version at http://wordnetweb.princeton.edu/perl/webwn

[2] http://www.senseval.org/



Noun

- S: (n) **coach**, manager, handler ((sports) someone in charge of training an athlete or a team)
- S: (n) **coach**, private instructor, tutor (a person who gives private instruction (as in singing, acting, etc.))
- S: (n) passenger car, **coach**, carriage (a railcar where passengers ride)
- S: (n) **coach**, four-in-hand, coach-and-four (a carriage pulled by four horses with one driver)
- S: (n) bus, autobus, **coach**, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger vehicle (a vehicle carrying many passengers; used for public transport) *"he always rode the bus to work"*

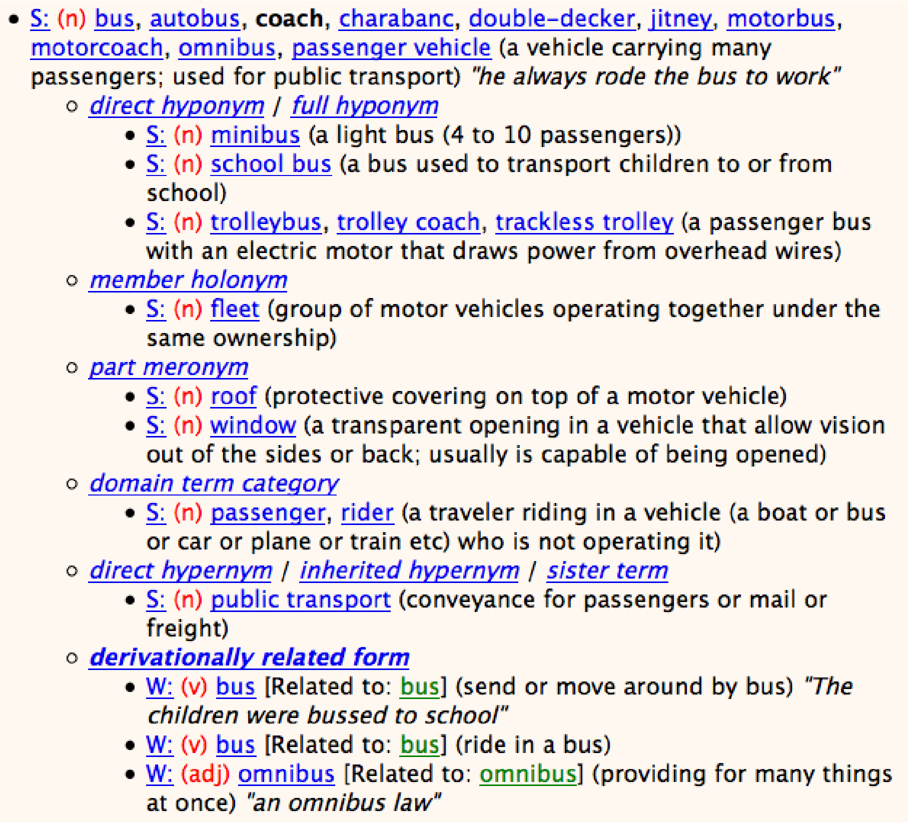
Verb

- S: (v) **coach**, train (teach and supervise (someone); act as a trainer or coach (to), as in sports) *"He is training our Olympic team"; "She is coaching the crew"*
- S: (v) **coach** (drive a coach)

Figure 2.1: Example of WordNet synsets for the English word *coach*.

Not only did Senseval make manually annotated training corpora and benchmark test data available to the WSD community, it also initiated a common WSD evaluation framework. Before, it was difficult to test and compare WSD systems, as there were not many common data sets publicly available (Ng and Lee 1996) and the different researchers did not use a standard evaluation methodology.

The first three Senseval workshops focused on Word Sense Disambiguation, each time growing in the number of languages offered in the tasks and in the number of participating teams. By the fourth workshop, the name was changed to SemEval and the nature of the tasks evolved to also include semantic analysis tasks outside of pure WSD, ranging from the automatic detection of metonymy, time expressions and verb ellipsis to coreference resolution in multiple languages.



- S: (n) bus, autobus, **coach**, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger vehicle (a vehicle carrying many passengers; used for public transport) *"he always rode the bus to work"*
    - ○ *direct hyponym* / *full hyponym*
        - S: (n) minibus (a light bus (4 to 10 passengers))
        - S: (n) school bus (a bus used to transport children to or from school)
        - S: (n) trolleybus, trolley coach, trackless trolley (a passenger bus with an electric motor that draws power from overhead wires)
    - ○ *member holonym*
        - S: (n) fleet (group of motor vehicles operating together under the same ownership)
    - ○ *part meronym*
        - S: (n) roof (protective covering on top of a motor vehicle)
        - S: (n) window (a transparent opening in a vehicle that allow vision out of the sides or back; usually is capable of being opened)
    - ○ *domain term category*
        - S: (n) passenger, rider (a traveler riding in a vehicle (a boat or bus or car or plane or train etc) who is not operating it)
    - ○ *direct hypernym* / *inherited hypernym* / *sister term*
        - S: (n) public transport (conveyance for passengers or mail or freight)
    - ○ ***derivationally related form***
        - W: (v) bus [Related to: bus] (send or move around by bus) *"The children were bussed to school"*
        - W: (v) bus [Related to: bus] (ride in a bus)
        - W: (adj) omnibus [Related to: omnibus] (providing for many things at once) *"an omnibus law"*

Figure 2.2: Example of the different semantic relations that are stored for the first WordNet sense of the English noun *coach*.

In the next sections, we introduce the most important and currently used approaches to WSD. These are often categorized according to the resources used for training the WSD system, being (1) knowledge resources (*Knowledge-based methods*), (2) manually-labeled corpora (*Supervised methods*) and (3) unlabeled corpora (*Unsupervised methods*).

The term *unsupervised* itself is polysemous in WSD research: it can refer to (1) approaches that are not trained on tagged corpora or (2) approaches that do not use manually sense-tagged corpora. We adopted the latter, more strict meaning of *unsupervised*, as is defined by Agirre and Edmonds (2006): *unsupervised* methods are "knowledge-lean approaches that do not require sense-tagged text and do not utilize other manually-crafted knowledge as found in dictionaries or concept hierarchies. These methods are data-driven and language-independent, and rely on the distributional characteristics of unannotated corpora, and translational equivalences in word aligned parallel text".

2.1 Knowledge-based methods

Knowledge-based methods use knowledge resources such as electronic dictionaries, thesauri and lexical knowledge bases to distinguish between the different senses of a word in a given context. Their main advantage is that they cover all polysemous words, whereas corpus-based supervised methods are usually only trained for a restricted set of words. Please refer to Mihalcea (2006) for a more detailed overview of the knowledge-based methods that are briefly described below.

The **Lesk algorithm** (Lesk 1986) calculates the most likely sense of an ambiguous word in context by comparing the context of the input sentence with the dictionary definitions of the ambiguous words. The definition that overlaps most with the input instance is then considered to be the correct sense of the word. Lesk (1986) illustrated his algorithm with the disambiguation of the words *pine* and *cone* in de word pair *pine cone*, based on the following senses for *pine* and *cone* in the Oxford Advanced Learner's dictionary (OALD):

(4) *pine cone*

pine
(1) **seven kinds of evergreen tree with needle-shaped leaves**
(2) pine
(3) waste away through sorrow or illness
(4) pine for something, pine to do something

cone
(1) solid body which narrows to a point
(2) something of this shape, whether solid or hollow
(3) **fruit of certain evergreen trees (fir, pine)**

Since the first definition of *pine* and the third definition of *cone* have the highest overlap in words (three words in common), these meanings will be selected by the Lesk algorithm. Variations on this algorithm have been proposed to reduce the combinatorial explosion in case more than two words have to be combined (Cowie, Guthrie and Guthrie 1992, Vasilescu, Langlais and Lapalme 2004). Related methods use similarity measures on

WordNet glosses instead of standard dictionary definitions (E.g. Banerjee and Pedersen (2002)).

A second group of knowledge-based methods considers **semantic similarity** to find the semantic distance between concepts. Since words that share a common context are usually closely related in meaning, the appropriate senses can be selected by choosing those meanings found within the smallest semantic distance. Most of these methods compute metrics on semantic networks, as in the original methodology proposed by Rada et al. (1989). These semantic similarity methods can address the local context of the focus word (i.e. a couple of words in the immediate context of the focus word or words that are connected to the focus word by syntactic dependencies), or the global context (the entire surrounding text where the word occurs). For an overview of semantic similarity measures, we refer to Budanitsky and Hirst (2001).

Another type of knowledge-based methods are **selectional preference** methods. These restrict the number of meanings of a given focus word based on the likeliness that this meaning can be combined with other words in the sentence. Selectional preferences incorporate information about semantic relations between word classes and concepts. Examples of such semantic constraints are for instance EAT/FOOD and DRINK/LIQUID. These constraints can be used to select the correct sense of a given focus word in case the other senses do not fit the context. The meaning of the word *coach* in the sentence *he drives a coach* will be the COACH_VEHICLE meaning, because the sense of COACH_TRAINER is not compatible with the verb DRIVE. An overview of these methods can be found in Brockmann and Lapata (2003).

Recently, **graph-based** methods have led to good results for WSD (Sinha and Mihalcea 2007, Navigli and Lapata 2007). In a first step, a graph is built whose nodes correspond to senses of the ambiguous word, whereas the edges represent semantic relations between these senses. The task of Word Sense Disambiguation then amounts to finding the most important node for each word (Navigli 2009). Graph-based methods are particularly suited for disambiguating word sequences as they manage to exploit the relations between the different senses in the given context. An interesting graph-based approach is the adaptation of PageRank – a widely used algorithm to compute the ranking of web pages by performing random walks – to perform WSD (Agirre and Soroa 2009).

Finally, also **heuristic methods** can be applied to solve the WSD task. These methods consist of simple rules that assign a particular sense to ambiguous words. One heuristic that is often used as a baseline for WSD evaluation is the *most frequent sense* heuristic. This heuristic starts from a *Zipfian distribution* of word meaning: Zipf (1949) has shown that one meaning of a word is often very frequent in language, while the other meanings show a significant decrease in frequency. Two other heuristics rely on the tendency of a word to preserve its meaning in a given discourse – the *one sense per discourse* heuristic of Gale, Church and Yarowsky (1992b), or in a given collocation – the *one sense per collocation* heuristic of Yarowsky (1993).

2.2 Supervised corpus-based methods

Supervised WSD algorithms are trained on manually sense-tagged corpora. Human annotators apply a sense label to each occurrence of the ambiguous word in the training corpus, and a classifier, or set of rules, is automatically induced from the corpus to predict sense labels for new occurrences of the word.

Initially, machine readable dictionaries, such as the LDOCE (the Longman Dictionary of Contemporary English of Procter (1978)) were often used as sense inventories to annotate occurrences of ambiguous words. Nowadays, WordNet and EuroWordNet (Vossen 1998) have become the most commonly used sense inventories. The most widely used sense-tagged corpus is the SemCor corpus (Landes, Leacock and Tengi 1998), a subset of the English Brown corpus that contains about 700,000 running words. In SemCor, all the words are tagged with their corresponding grammatical class, and more than 200,000 content words are also lemmatized and annotated with the WordNet sense they convey.

Another sense-tagged corpus that is frequently used is the DSO corpus (Ng and Lee 1996). This corpus contains sense-tagged word occurrences for 121 nouns and 70 verbs which are among the most frequent and ambiguous words in English. These occurrences are provided in about 192,800 sentences taken from the Brown corpus and the Wall Street Journal. More recently, sense-tagged corpora for other languages are also being constructed, such as the Dutch SemCor (Vossen et al. 2011) and the Basque SemCor (Agirre et al. 2006).

The main approaches to supervised WSD are often categorized based on the machine learning technique that is used to train the classifier on the manually annotated corpora. For a detailed overview of supervised WSD methods, we refer to Màrquez et al. (2006). They differentiate, amongst others, between:

(a) probabilistic models;
(b) methods based on discriminating rules;
(c) support vector machines and other kernel-based methods;

(d) memory-based methods;

(e) ensemble methods.

In this dissertation, a classification-based approach to WSD will be adopted. For this purpose, we use two of the above-mentioned learning methods, namely support vector machines and memory-based learning for all experiments. We will therefore present a more elaborate overview of the listed supervised learning methods in Chapter 4.

Supervised approaches invariably yielded the best results for various WSD tasks at the different Senseval competitions (Agirre and Edmonds 2006). These supervised approaches face a number of challenges, though, that should be addressed to allow for the construction of efficient and reliable WSD systems for general-purpose applications.

(a) **Lack of sense-tagged corpora**. Training an accurate all-words supervised WSD system requires a huge training corpus that contains enough labeled examples per ambiguous focus word. Different experiments were performed to measure the learning curves of ambiguous words in order to investigate the amount of data required to train a proper WSD system. Agirre and Martínez (2000) studied the learning curves of a set of ambiguous words in the SemCor and DSO corpus, using decision lists as a learning algorithm. Different types of words were selected, based on level of polysemy, frequency, predominance of most frequent sense, etc. They concluded from their experiments that: (1) SemCor does not contain enough training examples to get reliable results and that (2) on the DSO corpus, results seem to stabilize for nouns and verbs at a certain training size[3].

Although sense-tagged corpora are available (SemCor, DSO, SemEval benchmark data sets), these often contain too few training examples to cover all senses of all ambiguous words and hardly exist for languages other than English. Since the manual labeling of training corpora is very expensive and time-consuming, various techniques have been proposed to alleviate the data acquisition issue:

- active learning, a method that reduces the annotation cost by choosing the most informative examples for manual tagging (Fujii, Inui, Tokunaga and Tanaka 2001, Chklovski and Mihalcea 2002, Vossen, Görög, Izquierdo and van den Bosch 2012).
- bootstrapping, where a classifier is trained on a small set of manually-labeled training instances, and in a second step, used to annotate new, unlabeled examples (Mihalcea 2004, Pham, Ng and Lee 2005).

[3] The DSO corpus contains on average 927 examples for nouns and 1,370 examples for verbs.

- the automatic acquisition of training examples from the internet by querying monosemous synonyms of the focus word (Mihalcea and Moldovan 1999, Agirre and Martínez 2004b).

Another way of solving the data acquisition problem is the use of parallel corpora, which will be discussed in more detail in Section 2.3.

(b) **Lack of reliable sense-inventories**. As already mentioned in Chapter 1, the sense divisions in most dictionaries and sense inventories such as WordNet are often based on subjective decisions and too fine-grained to be useful for real world applications (Palmer 1998). When working with very fine-grained senses, the performance of these WSD systems is not able to exceed a certain accuracy (Edmonds and Kilgarriff 2002). Navigli (2009) even considers the dependency on discrete sense inventories as the main bottleneck for integrating WSD in real applications such as Information Retrieval or Text Mining. Furthermore, Specia et al. (2006b) investigated the differences between monolingual and multilingual sense inventories, as well as WordNet and the Portuguese translations for a test set covering eight verbs. They concluded that it is inappropriate to use monolingual sense inventories for disambiguation purposes in an MT context, since there is not a one-to-one relation between the number of fine-grained WordNet senses and the assigned translations.

Besides the questionable reliability of current sense inventories for WSD, there is a desperate lack of electronic sense inventories for languages other than English, even though initiatives have been launched for other languages, amongst which the EuroWordNet database[4]. There is also a line of NLP research that studies the automatic enrichment of knowledge resources such as machine-readable dictionaries and lexicons by extracting collocations and relation triples (Chklovski and Pantel 2004) or topically related words (Agirre et al. 2001) from corpora or from the Web. In a second step, these relation triples need to be disambiguated in order to enrich lexical resources such as WordNet (Navigli 2005).

(c) **Lack of portability to other domains**. An additional limitation of fixed sense inventories is their inability to cope with new words, new usages of words or usages of words in specialized contexts. Several studies have revealed a dramatic performance degradation when a WSD system that is trained on a specific corpus is applied to another corpus or domain (Escudero, Màrquez and Rigau 2000b, Martínez and Agirre 2000). As a consequence, the

[4] EuroWordNet is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the American wordnet for English.

> very costly annotation effort has to be repeated each time a WSD system needs training for a new genre or domain. This problem has been partially tackled by applying domain-driven disambiguation (Gliozzo, Magnini and Strapparava 2004, Buitelaar, Magnini, Strapparava and Vossen 2006), where the sense of a focus word is selected based on the overlap between the domains of the context words and the domain of the target sense. To this end, WordNet synsets have been enriched with domain labels that are organized in a hierarchical structure; BANKING is for instance part of the more general ECONOMY domain, which in turn belongs to the SOCIAL SCIENCE domain[5]. These methods, however, achieve very low recall because they merely use domain information that helps mainly to disambiguate domain words (Navigli 2009).

Since it is unlikely that large quantities of hand-annotated text and robust sense inventories will be available any time soon for a large variety of languages, unsupervised approaches have been introduced as an alternative, because they do not use externally defined lexical resources.

2.3 Unsupervised corpus-based methods

Unsupervised corpus-based methods work directly on raw unlabeled corpora. Instead of assigning word senses from a predefined sense-inventory, they try to distinguish different meanings of the word based on distributional similarity (sense induction methods) or translational equivalence (methods that use word-aligned parallel corpora). By eliminating the very costly human annotation effort, these methods may go some way towards clearing the knowledge acquisition bottleneck for WSD.

2.3.1 Distributional methods

Distributional approaches start from the hypothesis that words occuring in similar contexts will have similar meanings (Harris 1968). These approaches do not categorize words based on a predefined sense inventory, but group occurrences of the word based on their context. In this way, different senses of the word are induced (these methods are often called *sense induction* methods) by clustering occurrences of a word using a context similarity metric (E.g. number of overlapping context words in the sentence). In a second step, a sense-labeling algorithm assigns a sense to each cluster and to new occurrences of the word that are first assigned to one of

[5] These domain labels can be downloaded from http://wndomains.itc.it.

the automatically deduced sense clusters. This two-step process is comparable to how lexicographers work today: first a lot of contexts of the focus word are collected from a corpus and clustered per meaning, before the lexicographer assigns a definition to each cluster (or word meaning) that finally ends up in the dictionary. These methods, which distinguish different meanings of a particular focus word by clustering all of the contexts in which the word occurs are also referred to as **token-based methods**, as opposed to **type-based methods**. The latter methods identify sets (or clusters) of words that are supposed to be semantically related because they are used in similar contexts, such as for instance *line, cord, tie, cable.*

The **context-group discrimination method** of (Schütze 1998) is a good example of this two-step approach to Word Sense Disambiguation. In this algorithm, the instances containing an ambiguous word are first grouped into clusters, where each cluster consists of contextually similar instances. The context of each occurrence of the ambiguous word in the training corpus is represented as a context vector formed from second-order co-occurrence information. Words have second-order co-occurrence with other words if they both occur frequently together with a third word. In case they co-occur with each other, this is called first-order co-occurrence. The clusters themselves are represented by their centroids, which is the average of their elements. A new occurrence of an ambiguous word is subsequently disambiguated by computing the second-order representation of the context, and is assigned to the cluster whose centroid is closest to that representation. Another distributional method was developed by Pedersen and Bruce (1997), who represent the context of each test instance as a vector of features (containing co-occurrence and part-of-speech information) that directly occur near the focus word in that instance (so-called first-order representation). Another difference with the approach of Schütze (1998) is that they select features from the same test data that is being discriminated, whereas Schütze represents contexts in a high dimensional feature space that is created using a separate large corpus (called the *training corpus*).

The automatic labeling of the obtained context clusters remains a considerable challenge for these distributional approaches. One possible approach is to extract a set of meaningful words that are related to the content of the cluster, and consider this bag of words as a more abstract representation of the underlying sense of the word. For instance, a cluster containing occurrences of the ambiguous word *coach* in the sense of *bus*, could comprise words such as *bus, driver, seat, etc.* So far, the best results are obtained by labeling the clusters with information from existing knowledge resources, such as WordNet (McCarthy et al. 2004) .

The unsupervised methods that distinguish between different words senses based on distributional information propose a possible solution for the data-acquisition issue. One important remaining problem is the integration of WSD in real applications such as machine translation. This problem is addressed by those WSD methods that are based on translational equivalence, which are presented in the next section.

2.3.2 Methods based on translational equivalence

A second type of unsupervised methods, which also falls within the remit of this dissertation, use word-aligned bi- or multilingual parallel corpora in order to extract cross-lingual evidence for Sense Discrimination or Word Sense Disambiguation. These methods start from the hypothesis that the different senses of a word might result in different translations in a given target language. Resnik and Yarowsky (2000) showed with a set of empirical studies of translingually-based sense inventories that sense distinctions can indeed be captured by translations into second languages. The authors also revealed that multiple languages are needed to cover all different sense distinctions, as some senses are lexicalized in one particular language but not in the other language. The authors cite as an example the *stake* sense of *interest* which gets a specific translation in German (*Anteil*), but not in Spanish (*interés*), whereas the *benefit* sense of the word is lexicalized in Spanish (*provecho*) and not in German (*Interesse*). In addition, they found that for some sense distinctions, the probability that senses are lexicalized differently across languages seems to correlate with sense granularities: distinctions at the homograph level are lexicalized in 95% of the cases and major sense distinctions in 78%, whereas at the finest sense distinction level, only 52% of the examples were translated differently.

As already illustrated by the *interest* example in the preceding paragraph, there also tends to be parallel polysemy between languages in case the translation preserves the ambiguity of the source language. Another example is the English ambiguous word *movement*, that is translated as *mouvement* in French, both for the *physical motion* as well as for the more abstract *organized activity* sense. As a consequence, this method does not always disambiguate all senses of a polysemous word, but this level of granularity is suitable for a lot of real applications, such as machine translation.

It might also be the case that a given target language draws conceptual distinctions that are not present in the source language. The English word *river* for instance, is translated in French as *fleuve* in case it concerns a big river that flows out into the sea, and is translated as *rivière* when it concerns a smaller tributary river. It is assumed, however, that the context surrounding the monosemous focus word will be informative enough for a classifier to generate a correct translation in the given target language.

Using word-aligned parallel corpora enables one to extract the sense inventory in an automatic way, based on the various translations of a particular word in these corpora. In addition, a sense-tagged corpus based on the induced senses - or translations - can be automatically created and used to train a more traditional supervised classifier (Specia et al. 2005).

Example 5 illustrates four different meanings of the English word *coach* that are all lexicalized in the English-French part of the Europarl corpus (Koehn 2005).

(5) (a) *ENGLISH*: I should like to mention something which was said on the BBC by someone who is not a football referee like you but who is the **coach** of the England team.
FRENCH: J'ai repris des propos qui ont été tenus sur les ondes de la BBC par quelqu'un qui n'est pas arbitre de football comme vous, mais qui est l'**entraîneur** de l'équipe de Grande-Bretagne.

(b) *ENGLISH*: It is no longer the locomotive it once was, it is now the last **coach** in the train.
FRENCH: Elle n'est plus la locomotive, comme c'était le cas jadis, elle est le dernier **wagon** de ce train.

(c) *ENGLISH*: As you know, pensioners and elderly people, who enjoy a spot of sightseeing and visiting the beautiful cities of Europe when they are in good health, often travel by **coach**.
FRENCH: Vous savez que les personnes âgées et les retraités voyagent souvent en **autobus**, car ils aiment, s'ils jouissent d'une bonne santé, faire un brin de tourisme, visiter les belles villes d'Europe.

(d) *ENGLISH*: As for Lunnon, well that used to be three days on the fast stage **coach**, and there are many of us folks who never did get used to they newfangled things like horseless carriages and they motorway things that bring all those foreigners to our neck of the woods.
FRENCH: En fait de Londres, cette ville se trouvait auparavant à trois jours de **diligence** et on est nombreux à ne s'être jamais habitués à leurs nouveaux engins, comme les voitures sans chevaux et ces autoroutes qui ramènent tous ces étrangers

sur notre bout de terre.

Several studies have already underlined the validity of using parallel corpora for sense discrimination (E.g. Ide, Erjavec and Tufiş (2002), Dyvik (2004)) and for bilingual word sense disambiguation (E.g. Gale and Church (1993), Ng, Wang and Chan (2003), Diab and Resnik (2002), Chan and Ng (2005), Apidianaki (2009)).

The "Semantic Mirrors method" developed by Dyvik (2004) discriminates between senses of ambiguous words by "mirroring" translational relations (extracted from a word-aligned and lemmatized parallel corpus) back and forth between two languages. The senses of a word are grouped in semantic fields based on overlapping sets of translations and feature sets assigned to the senses in a field form a semilattice based on inclusion and overlap relations among the feature sets. The final goal of the mirror method is to automatically derive WordNet relations such as synonymy and hyponymy from data extracted from parallel corpora. The underlying hypothesis of the semantic mirrors method is that words that are closely related semantically ought to have strongly overlapping sets of translations, and that words that are more general (”semantically wide”) have a larger set of translations than words with more specific meanings.

Brown et al. (1991) use the idea that different meanings of the word are lexicalized differently across languages. They cite the example of the French verb "prendre" that can be translated as "take" or "make" in English: "prendre une décision" is translated as "make a decision", while "prendre ma voiture" is translated as "take my car". Depending on the translation in the parallel corpus, the two meanings of the word can then be distinguished. They first apply word alignment to identify all translation candidates in the French-English parallel corpus. In a next step, a system is trained to predict a correct translation of the ambiguous focus word by using source-language local context information. Gale, Church and Yarowsky (1992a) follow a similar approach. They first align the English-French Canadian Hansards corpus at sentence level. In a following step, they identify the French sentences that contain a particular translation of the polysemous English word by identifying valid translations of the ambiguous words (called "correspondences") by using association measures such as Mutual Information (Gale and Church 1991b). The English training sentences are tagged with the resulting French translations and the resulting corpus is used to train a WSD classifier.

Diab (2004) presents an unsupervised bootstrapping approach to WSD, which exploits automatically-generated, noisy data for training within a supervised learning framework. The algorithm expects a word-aligned parallel corpus as input and (1) groups all source words that translate to the same focus word, (2) assigns a sense tag to these clusters (based on the word senses' proximity in WordNet, using the similarity measure developed by Resnik (1999)), (3) propagates these to all contexts in the corpus and (4) projects the sense tags from the source language to their corresponding translations in the target language.

Specia, Nunes and Stevenson (2007) propose a hybrid approach that employs an inductive logic programming algorithm to learn disambiguation rules based on corpus-based evidence and several knowledge sources. The system is trained (and tested) to provide a correct Portuguese translation for a set of highly ambiguous English verbs, as verbs seem to benefit more from deeper knowledge sources such as syntactic relation information or selectional restrictions for the different senses of the word. For the construction of the training and test corpus, all instances are labeled with the corresponding translations using a combination of a parallel corpus, statistical information and translation dictionaries. The resulting sense inventory then contains the set of all valid translations of the verb that are included in the corpus.

Other WSD systems exploit the multilingual sense alignments in multilingual wordnets, such as the EuroWordNet (Vossen 1998) or BalkanNet (Tufiş, Cristea and Stamou 2004). Tufiş, Ion and Ide (2004) utilize earlier research on word clustering to automatically extract translational equivalents (See Ide et al. (2002)) and combine this research with aligned wordnets. Given two aligned words in a parallel corpus, these words receive the sense label of the synsets of the two words which are mapped by EuroWordNet's interlingual index. This way, the method can be used to automatically sense-tag corpora in several languages at once. Chan and Ng (2005) map WordNet senses in similar definition entries for bilingual English-Chinese dictionaries and gather examples for those entries from a word-aligned parallel corpus.

Several studies have revealed that approaches using bilingual evidence attain state-of-the-art performance for WSD. Ng et al. (2003) use Chinese translations as sense labels for English words in automatically extracted training data from an English-Chinese parallel corpus. They manually decide on valid Chinese translations for the English WordNet senses and test their approach on the SensEval-2 English lexical sample task. The reported accuracy results are comparable to traditional supervised systems trained on manually labeled corpora.

Two major challenges, however, threaten the further development of these methods. First of all, these methods, based on translational equivalence, start from a word-aligned parallel corpus. Although automatic word alignment algorithms achieve reasonable performances (Martin, Mihalcea and

Pedersen 2005), the alignment is not perfect yet and leads to the production of noise in the training corpus, especially for languages with scarce resources. A second challenge consists of the availability of large parallel corpora, which is again problematic for several under-resourced languages. A possible solution to this problem could be the large-scale collection of parallel corpora from the web (Resnik and Smith 2003).

In the next section, we introduce ParaSense, our cross-lingual approach to Word Sense Disambiguation, and highlight some of its advantages compared to the other previously discussed approaches to WSD.

2.4 ParaSense system

The WSD system we propose in this dissertation, the ParaSense system, is a truly multilingual classification-based approach to WSD that, in its current implementation, directly incorporates evidence from four other languages. Instead of using additional lexical resources, we only include information that is extracted from the parallel corpora in hand. To this end, we elaborate on two well-known research ideas:

(a) the possibility to use parallel corpora to extract translation labels and features in an automated way.

(b) the assumption that incorporating evidence from multiple languages into the feature vector will be more informative than a more restricted set of monolingual or bilingual features. This assumption is based on the idea of Dagan, Itai and Schwall (1991) who argued that “two languages are better than one” to select the right sense of an ambiguous focus word. They showed that it was possible to use the differences between languages in order to obtain a certain leverage on word meanings. Furthermore, Ide et al. (2002) demonstrated that the accuracy of sense clustering based on translation equivalents extracted from parallel corpora heavily depends on the number and diversity of the languages in the parallel corpus.

Whereas current WSD approaches, in the best case, only resolve a subset of the main remaining WSD issues, we believe, based on the following assumptions, that the ParaSense Cross-lingual WSD (CLWSD) algorithm proposes an integrated solution:

- Using multilingual unlabeled parallel corpora goes some way towards clearing the data acquisition bottleneck for WSD, because using translations as sense labels excludes the need for manually created sense-tagged corpora and sense inventories such as WordNet or EuroWordNet. Moreover, as there is fairly little linguistic knowledge involved, the framework can be easily deployed for a variety of different languages.

- This approach also deals with the sense granularity problem; finer sense distinctions are only relevant as far as they get lexicalized in different translations of the word. At the same time, the subjectivity problem is tackled that arises when lexicographers have to construct a fixed set of senses for a particular word that should fit all possible domains and applications. In our approach, the use of domain-specific corpora allows to derive sense inventories that are tailored towards a specific target domain or application and to train a dedicated CLWSD system using these particular sense inventories.
- Working immediately with translations instead of more abstract sense labels allows to bypass the need to map (WordNet) senses to corresponding translations. This makes it easier to integrate a dedicated WSD module into real multilingual applications such as machine translation or information retrieval.
- Including evidence from multiple languages, as opposed to existing bilingual WSD methods helps to further refine the obtained sense distinctions. As shown by Resnik and Yarowsky (2000), it is often the case that multiple languages are needed to cover the different sense distinctions. We will illustrate this in Section 8.3, where the impact of the different languages is measured for the overall classification performance. Unlike other multilingual approaches as presented by, for instance, Tufiş, Ion and Ide (2004), we do not use aligned WordNets in the languages under consideration to enhance the WSD classification results. This allows us to present the first real multilingual WSD system that does not use any external lexical resources to perform sense disambiguation.

Part II

The ParaSense System Architecture

CHAPTER 3

Information Sources for Cross-Lingual WSD

In the previous chapter, we reviewed the different approaches to word sense disambiguation and their main shortcomings. Furthermore, we introduced the ParaSense system, our multilingual approach to WSD, and explained how the ParaSense system intends to propose an integrated solution to the most important remaining WSD issues. We will now further elaborate on the ParaSense system architecture, and describe in the following two chapters the information sources and machine learning algorithms that are used to build the system.

We consider the Cross-lingual WSD task as a supervised learning task where the possible senses of an ambiguous word are the classes and each new occurrence of the ambiguous word is assigned to the correct sense class based on disambiguating context information of the ambiguous word. In order to compute the meaning of a word in a given context, we construct a vector space model for each individual word. In this model, training and test documents containing the ambiguous focus word are decomposed into a set of meaningful features. Each instance in the training set (called *feature vector*) thus contains a set of informative *features* and a *class label*, being the translation of the ambiguous word in a given target language. For new instances, the classifier then has to predict a correct class label by measuring the distance between the test feature vector and all training feature vectors in the vector space or by applying the model that is derived from the training data.

This chapter describes the informative features that are used by the ParaSense system to perform Cross-Lingual Word Sense Disambiguation. Instead of using a predefined monolingual sense inventory, such as WordNet, we use a language-independent framework where the word senses are derived automatically from word alignments on a parallel corpus. As information sources, both local context information as well as translational evidence is integrated in the feature vectors that are used to train and test the Cross-lingual WSD classifier.

Section 3.1 discusses the linguistic preprocessing of the data that was used to construct the training and test instances. In section 3.2, we discuss the information used to construct the feature vectors, whereas section 3.3 describes the way in which the classification label was selected for all instances.

3.1 Preprocessing of the Data

We based our data set on the Europarl parallel corpus[1], which is extracted from the proceedings of the European Parliament (Koehn 2005). Parallel corpora contain texts in different languages that are translations of each other. We experimented with 6 languages from the Europarl corpus, viz. English (our source language), Dutch, French, German, Italian and Spanish. We want to remark, however, that our approach is language- and corpus-independent; it can be applied to any parallel corpus for any given target language.

We selected a set of polysemous English nouns and then preprocessed all Europarl sentences containing these nouns and their aligned translations in the other five languages. The decision to first focus our research on nouns was inspired by the fact that various papers report better disambiguation results for nouns than for adjectives or verbs (Mihalcea 2002, Yarowsky 2000). In addition, we also assume that more noise will be introduced in the training data for verbs due to erroneous word alignment. Therefore it seemed more logical to deliver a proof of concept for the most manageable part-of-speech category, viz. the nouns.

In further research, we would like to investigate the portability of our ParaSense system to the other part-of-speech categories. Several studies already revealed major differences between nouns and verbs in language acquisition (Gentner 1982), or how both grammatical categories map to the physical world (Roy and Reiter 2005). Gentner (1981) argues that nouns typically lexicalize perceptual information that is conflated into

[1] http://www.statmt.org/europarl/

concrete objects, whereas verbs encode relations between these objects, which makes them "less tightly constrained by the perceptual world". An important consequence is that the meanings of verbs and other relational terms vary a lot cross-linguistically, while nouns show to be more stable in translation[2]. It would be interesting to compare these observations with cross-lingual disambiguation results for both part-of-speech categories.

3.1.1 Selection of Polysemous English Nouns

A sample set of polysemous English words was selected and used for all experiments. Three criteria were applied for the selection of these words:

(a) The word is polysemous and should have at least three WordNet senses.

(b) The word should occur at least 50 times in the Europarl corpus in order to have a minimum set of training instances. An additional manual check was carried out in order to ensure that at least two different senses of the word were represented in the Europarl corpus.

(c) The resulting set of words should be a subset of the words that are used in the Cross-Lingual Lexical Substitution task (Sinha et al. 2009). Since we intended to also create a benchmark test set for these words, only words that were also used in other semantic disambiguation test sets were chosen in order to allow researchers to easily test their systems on different benchmark sets.

Table 3.1 lists the 20 polysemous words that were selected, together with their number of WordNet senses and frequency in the Europarl Corpus.

3.1.2 Sentence Alignment

A first processing step needed to prepare parallel corpora for automatic processing is sentence alignment. Sentence alignment is the process of finding corresponding text chunks at sentence level in parallel texts. These alignments can be 1:1 alignments (1 sentence in the first language corresponds to exactly 1 sentence in the other language), 1:many alignments (1 sentence in the first language corresponds to more than 1 sentence in the second language) or many:1 alignments (more than 1 sentence in the first language corresponds to 1 sentence in the second language). Other

[2] Gentner (1981) set up an experiment where English texts are first translated into another language, and then translated back to English by another bilingual speaker. It turns out that decidedly more of the original nouns than verbs appear in the second translation.

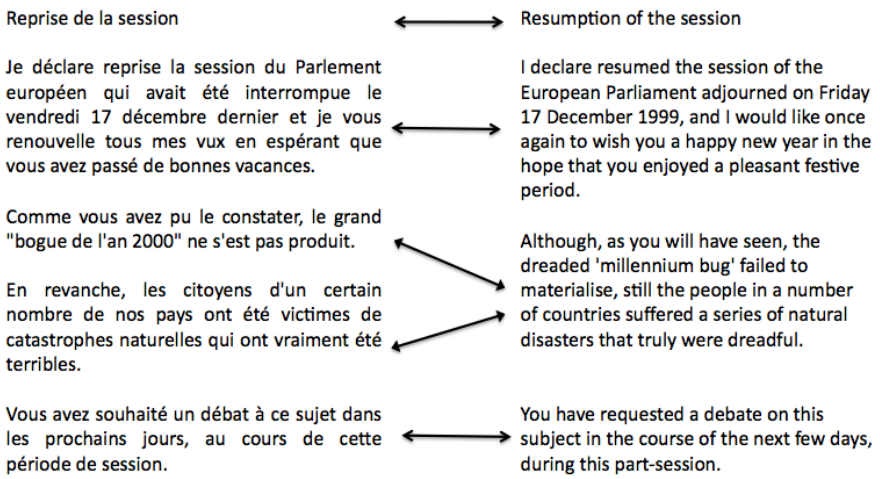
Word	**Number of WN senses**	**Occurrences in Europarl**
coach	5	73
education	6	4557
execution	7	536
figure	13	2663
job	13	7844
letter	5	1873
match	9	384
mission	5	1432
mood	3	118
paper	7	3735
post	11	1638
pot	9	81
range	9	1608
rest	7	2304
ring	9	206
scene	10	344
side	12	4207
soil	4	294
strain	11	172
test	6	1617

Table 3.1: Overview of the twenty polysemous words together with their number of WordNet senses and Europarl frequency information.

possible combinations are zero alignments (when no translational equivalence can be found for a sentence) and many-to-many alignments where several sentences in the first language correspond to several sentences in the second language. Figure 13 illustrates the sentence alignment process for an extract of the French-English Europarl corpus.

The alignment of sentences in the Europarl corpus is done with an implementation of the algorithm by Gale and Church (1991a). The Gale and Church algorithm is a sentence-length-based approach to sentence alignment; it starts from the assumption that long sentences tend to be translated by long sentences, and short sentences by short sentences. A probabilistic score is assigned to each corresponding pair of sentences and these scores are used in a dynamic programming[3] set-up to find the maximum

[3] *Dynamic programming* was first introduced by Bellman (1957) and refers to a class of algorithms that try to solve complex problems by breaking them down into simpler subproblems. The overall solution for the complex problem is then reached by combining the solutions to the various subproblems.



Reprise de la session

Resumption of the session

Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances.

I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.

Comme vous avez pu le constater, le grand "bogue de l'an 2000" ne s'est pas produit.

En revanche, les citoyens d'un certain nombre de nos pays ont été victimes de catastrophes naturelles qui ont vraiment été terribles.

Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.

Vous avez souhaité un débat à ce sujet dans les prochains jours, au cours de cette période de session.

You have requested a debate on this subject in the course of the next few days, during this part-session.

Figure 3.1: Example of 1:1 and 2:1 sentence alignments (extracted from the French-English part of Europarl).

likelihood alignment of sentences. Other sentence alignment algorithms are based on word correspondences (E.g. Kay and Röscheisen (1993)) and start from the assumption that corresponding sentences contain words that are translations from each other. These algorithms combine sentence and word alignment in order to improve the performance of both processes.

We only considered the 1-1 sentence alignments between English and the five other languages[4] in order to obtain a real six-lingual parallel corpus (see Tufiş, Ion and Ide (2004) for a similar strategy). This resulted in a sentence-aligned subcorpus of Europarl that contains 884,603 sentences per language.

In a next step, we selected all English sentences containing one of the twenty ambiguous target nouns and their aligned translations in the five target languages after which the resulting sentences were linguistically preprocessed.

To summarize, the following steps were taken to select the Europarl data that was used to train the ParaSense system:

[4] This six-language sentence-aligned subsection of Europarl can be downloaded from `http://lt3.hogent.be/semeval`.

(a) Starting point: five bilingual sentence-aligned corpora (English-Spanish, English-German, English-Italian, English-Dutch, English-French) containing each around 1,300,000 sentence pairs.

(b) Selection of all 1-1 sentence alignments from the five bilingual corpora that share overlapping English sentences. This results in a subcorpus of Europarl containing 884,603 sentences.

(c) Selection of all English sentences, together with their aligned translations in the five target languages, that contain one of the twenty ambiguous test words. This results in a training corpus of 35,686 sentences (per language).

3.1.3 Shallow Linguistic Analysis

In order to store relevant information for our ambiguous focus words, we first had to take the following shallow linguistic analysis steps:

(a) **Tokenisation**
Tokenisation consists in splitting off punctuation from the adjoining words. This might seem a trivial task at first sight, but a couple of issues should be solved for the different languages, such as abbreviations (where the periods do not have to be separated from the preceding letters), acronyms, language specific apostrophe rules (E.g. in English, the genitive mark *'s* should not be split, in French a lot of determiners and pronouns have contracted variants such as *l', s', m', n', j', t'*), etc.

(b) **Part-of-Speech tagging**
During part-of-speech tagging, a grammatical category or *part-of-speech* code is assigned to each orthographic token.

(c) **Chunking**
During chunking, syntactically related words are combined into *chunks* or non-overlapping shallow parses. An example of such a chunk is a Noun Phrase (NP) containing a head (noun) and optional modifiers such as determiners and adjectives. Other examples of chunks are Verb Phrases (VP) and Adverbial Phrases (ADVP). Example (6) shows an example of a chunked English sentence.

(6) [It] NP — [is] VP — [no longer] ADVP — [the locomotive] NP — [it] NP — [once] ADVP — [was] VP — [,] O — [it] NP — [is] VP — [now] ADVP — [the last coach] NP — [in the train] PP — [.] O

(d) **Lemmatisation**
The process of lemmatisation generates the base form or *lemma* for each orthographic token. For verbs, this base form is usually the infinitive, whereas for the other grammatical categories, the base form corresponds to the *stem* of the word, i.e. the word without inflectional endings. In order to perform correct lemmatisation, part-of-speech information is needed to disambiguate word forms with multiple lemmas. This is illustrated in the French word form *sens*, which can be a noun (stem: *sens*) or a verbal form (stem: *sentir*).

All English sentences were preprocessed by means of the memory-based shallow parser (MBSP) of Daelemans, Buchholz and Veenstra (1999) that performs tokenisation, part-of-speech tagging, text chunking and lemmatisation. The aligned translations were preprocessed by means of the Treetagger tool (Schmid 1994) that performs tokenisation and outputs part-of-speech (PoS) and lemma information. For the aligned translations, we used PoS-tags and lemma information to select all content words that were stored as bag-of-words features in their lemmatized form. For the English sentences, both grammatical and chunk information for the local context was explicitly stored in the feature vector, as these features have shown to work well for the disambiguation of polysemous nouns (Agirre and Edmonds 2006). We refer to Section 3.2 for a more detailed description of the feature vector construction. Table 3.2 shows the English MBSP output and Table 3.3 the French Treetagger output for the two aligned Europarl sentences that are shown in Example 7. The first column of the tables contains the word form, the second column the part-of-speech information and the third column contains the lemma of the word. The MBSP output also contains a fourth column with additional chunk information.

(7) It is no longer the locomotive it once was, it is now the last coach in the train.

Elle n'est plus la locomotive, comme c'était le cas jadis, elle est le dernier wagon de ce train.

Word	**Part-of-Speech**	**lemma**	**Chunk info**
It	PRP	it	I-NP
is	VBZ	be	I-VP
no	RB	no	I-ADVP
longer	RBR	long	I-ADVP
the	DT	the	I-NP
locomotive	NN	locomotive	I-NP
it	PRP	it	B-NP
once	RB	once	I-ADVP
was	VBD	be	I-VP
,	,	,	O
it	PRP	it	I-NP
is	VBZ	be	I-VP
now	RB	now	I-ADVP
the	DT	the	I-NP
last	JJ	last	I-NP
coach	NN	coach	I-NP
in	IN	in	I-PP
the	DT	the	I-NP
train	NN	train	I-NP
.	.	.	O

Table 3.2: MBSP output for the English sentence containing one word per line and the accompanying part-of-speech, lemma and chunk information.

Word	**Part-of-Speech**	**lemma**
Elle	PRO:PER	elle
n'	ADV	ne
est	VER:pres	être
plus	ADV	plus
la	DET:ART	le
locomotive	NOM	locomotive
,	PUN	,
comme	KON	comme
c'	PRO:DEM	ce
était	VER:impf	être
le	DET:ART	le
cas	NOM	cas
jadis	ADV	jadis
,	PUN	,
elle	PRO:PER	elle
est	VER:pres	être
le	DET:ART	le
dernier	ADJ	dernier
wagon	NOM	wagon
de	PRP	de
ce	PRO:DEM	ce
train	NOM	train
.	SENT	.

Table 3.3: Treetagger output for the French sentence containing one word per line and the accompanying part-of-speech and lemma information.

3.2 Selection of Informative Features

For the feature vector construction, we combined a set of local context features that were extracted from the English sentence and a set of bag-of-words features that were extracted from the aligned translations in the four other languages that are not the target language of the classifier. We created two flavors of the translation features: a set of binary bag-of-words features and a set of *latent semantic* translation features that resulted from applying Latent Semantic Analysis to the content words of the aligned translations.

The idea to enrich the more commonly used local context features with multilingual translational evidence starts from the assumption that incorporating evidence from multiple languages into the feature vector will be more informative than only using monolingual or bilingual features. The working hypothesis we adopted for all experiments is that the differences between the different languages that are integrated in the feature vector will enable us to refine the obtained sense distinctions and that adding more languages will improve the classification results accordingly.

3.2.1 Local Context Features

We extracted the same set of local context features from both the English training and test instances. The linguistically preprocessed English instances were used as input to build a set of commonly used WSD features (Agirre and Edmonds 2006):

- features related to the **focus word itself** being the word form of the focus word, the lemma, part-of-speech and chunk information.
- **local context features** related to a window of three words preceding and following the focus word containing for each of these words their full form, lemma, part-of-speech and chunk information.

The motivation to incorporate local context information for a seven-word window containing the ambiguous focus word is twofold. Firstly, we assume that the immediate context of the ambiguous focus word will be more efficient in capturing compound and collocation information for the ambiguous focus word. Secondly, previous research has shown that a classifier using this specific set of features performs very well for WSD (Hoste et al. 2002). An interesting extension to this set of source language features would be to include bag-of-words features extracted from a larger context of the ambiguous focus word.

As an example, we list the set of local context features that were extracted for the following training instance for the word **coach**:

(8) It is no longer the locomotive it once was, it is now the last **coach** in the train.

(a) features focus word: coach coach NN I-NP
(b) features context word -3: now now RB I-ADVP
(c) features context word -2: the the DT I-NP
(d) features context word -1: last last JJ I-NP
(e) features context word +1: in in IN I-PP
(f) features context word +2: the the DT I-NP
(g) features context word +3: train train NN I-NP

3.2.2 Translation Features

In addition to the commonly deployed local context features, we also extracted a set of binary bag-of-words (BOW) features from the aligned translations that are not the target language of the classifier. For the French classifier for instance, we extract bag-of-words features from the Italian, Spanish, Dutch and German aligned translations. The difference between these features and more traditional BOW vector spaces (Lund and Burgess 1996), where the BOW space records a function of co-occurrence between focus words and a predefined set of content words, is that we consider a set of translation features instead of monolingual context words.

Per ambiguous focus word, a list of all content words (nouns, adjectives, adverbs and verbs) that occurred in the linguistically preprocessed aligned translations of the English sentences containing this word, were extracted. Each content word then corresponds to exactly one binary feature per language. For the construction of the translation features for the training set, we used the Europarl aligned translations. The extraction of the binary translation features is illustrated by Example (9) that lists two English sentences for the word *pot* and their aligned translations in German and Italian. Table 3.4 lists the extracted lemmatized German content words and the corresponding binary translation features per sentence, whereas Table 3.5 lists the translation terms and bag-of-words features for Italian.

(9) English:

Sentence 1: Our Europe, that melting **pot** of cultures, languages and people, is possible thanks to free movement and study programmes.

Sentence 2: Macao, as has already been said, has always been a melting **pot** of cultures and of new meetings of cultures, of religions too, and has always been a territory where peace, tranquillity and coexistence between peoples of the most diverse ethnic backgrounds have reigned.

German:

Sentence 1: Unser Europa, das durch die **Vermischung** der Kulturen, der Sprachen, der Menschen gekennzeichnet ist, fördernn wir durch die Freizügigkeit und durch Studienprogramme.

Sentence 2: Macau war, wie hier ja schon gesagt wurde, stets ein **Umschlagplatz** der Kulturen, einer Begegnungsstätte der Zivilisationen und auch der Religionen, und immer war es auch ein Territorium, in dem Frieden und Ruhe herrschten und die verschiedenartigsten Volksgruppen zusammenlebten.

Italian:

Sentence 1: La nostra Europa, quel **crogiolo** di culture, lingue e persone, è possibile grazie alla libera circolazione e ai programmi di studio.

Sentence 2: Macao, come è stato detto, è sempre stata un **crogiolo** di culture, civiltà e religioni, una regione in cui le etnie più diverse convivono in pace e serenità.

	Europa Vermischung Kultur Sprache Mensch kennzeichnen sein fördern Freizügigkeit Studienprogramm Macau hier ja schon sagen werden stets Umschlagplatz Begegnungsstätte Zivilisation auch Religion immer Territorium Frieden Ruhe herrschen verschiedenartig Volksgruppe zusammenleben														
sentence 1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sentence 2	0	0	1	0	0	0	1	0	0	0	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 3.4: Binary bag-of-words translation features for the two German training instances listed in Example (9).

	Europa crogiolo cultura lingua persona essere possibile grazie libero circolazione programma studio Macao dire sempre civiltà religione regione etnia più diverso convivere pace serenità										
sentence 1	1	1	1	1	1	1	1	1	1	1	1	0
	0	0	0	0	0	0	0	0	0	0	0	0
sentence 2	0	1	1	0	0	1	0	0	0	0	0	0
	1	1	1	1	1	1	1	1	1	1	1	1

Table 3.5: Binary bag-of-words translation features for the two Italian training instances listed in Example (9).

As we do not dispose of similar aligned translations for the test instances for which we only have the English test sentences at our disposal, we had to adopt a different strategy. We decided to use the Google Translate API[5] to automatically generate translations for all English test instances in the five target languages. This automatic translation process can be done using whatever machine translation tool, but we chose the Google API because of its easy integration into our programming code.

Online machine translation tools have already been used before to create artificial parallel corpora that were used for NLP tasks such as for instance Named Entity Recognition (Shah et al. 2010). Similarly, Navigli and Ponzetto (2010) used the Google Translate API to enrich BabelNet, a wide-coverage multilingual semantic network, with lexical information for all languages.

Once the automatic translations were generated, we preprocessed them in the same way as we did for the aligned training translations. Subsequently, we selected all content words from these aligned translations and constructed the binary bag-of-words features per language.

We decided to also include the translation of the ambiguous focus word itself in the set of binary translation features. Although it would be possible to filter the Google translation of the ambiguous word by applying word alignment, we would certainly introduce noise due to false word alignments and potentially remove the wrong translation from the bag-of-words list in those cases. We performed a small test in order to verify whether the classifier was biased by a correct or wrong Google translation for the ambiguous focus word. Table 3.6 and 3.7 list the Spanish and Italian Google translations of the ambiguous word *post* for the 50 test sentences. The third column indicates which test instances for *post* received a wrong translation label by the French classifier. We refer to Appendix B for an overview of the 50 English input sentences containing the ambiguous focus word *post*.

The tables for this given focus word clearly show that there is no one-to-one relation between correct/wrong Google translations in the closely related languages and a correct/wrong translation label generated by the ParaSense system. The classifier thus seems robust to translation errors made by Google and does not seem biased by a correct or wrong Google translation for the ambiguous focus word in the set of bag-of-words translation features.

[5] `http://code.google.com/apis/language/`

	Google: Italian	Google: Spanish	Output ParaSense
1	post: wrong	correo	
2	messaggio: wrong	poste	wrong
3	posto: wrong	puesto	
4	posto	puesto	
5	post: wrong	puesto	
6	postale	correo	
7	postazione	punto de parada	wrong
8	posto	puesto	
9	valido	puesto	
10	posta	correo	
11	post: wrong	correo	
12	posto	puesto	
13	funzione	puesto	
14	posto: wrong	puerto	
15	messaggio: wrong	puesto	
16	segnaletico: wrong	poste	wrong
17	posta	correo	
18	inviare: wrong	cargo	
19	posta	post: wrong	
20	postale	postal	
21	punto	etapa: wrong	wrong
22	post: wrong	puesto: wrong	
23	posto: wrong	puesto	
24	punto: wrong	punto de parada	wrong
25	posto	puesto	

Table 3.6: French and Spanish Google translations and French classification scores for the first 25 test sentences containing the ambiguous focus word *post*.

	Google: Italian	Google: Spanish	Output ParaSense
26	incarico	puesto	
27	postazione	puesto	
28	first-past-the-post: wrong	first-past-the-post	wrong
29	first-past-the-post: wrong	first-past-the-post	wrong
30	messaggio: wrong	cargo	
31	posta	correo	
32	posto	puesto	
33	post: wrong	mensaje: wrong	
34	post: wrong	mensaje: wrong	
35	posto	puesto	
36	post: wrong	mensaje: wrong	wrong
37	post	puesto: wrong	
38	posto: wrong	otro: wrong	wrong
39	posto	puesto	
40	messaggio: wrong	punto	wrong
41	posto	puesto	
42	posto: wrong	puesto	wrong
43	postale	correo	
44	postazione	puesto	
45	posto: wrong	puesto	
46	posto	puesto	
47	posta	correo	
48	posto	puesto	
49	postale	correo	
50	post: wrong	mensaje: wrong	

Table 3.7: French and Spanish Google translations and French classification scores for the last 25 test sentences containing the ambiguous focus word *post*.

3.2.3 Latent Semantic Translation Features

Manual inspection of the binary translation features that were introduced in Section 3.2.2 revealed two potential issues with this type of bag-of-words features:

(a) The translation features result in very sparse feature vectors where only a small amount of the binary translation features has a positive value per instance.

(b) It is often the case that synonyms occur in instances denoting the same meaning of the ambiguous focus word. These synonyms, however, are considered as two completely different words, as only exact overlap of lexical units is taken into account when measuring the similarity of these binary features. It is, for instance, clear that both sentences denote the same meaning of the word *coach* in Example (10), although there is no exact lexical overlap between the content words of the two sentences.

(10) English:

Sentence 1: I should like to mention something which was said on the BBC by someone who is not a football referee like you but who is the coach of the England team.

Sentence 2: When a person other than the sportsman or woman has taken part , such as a club, an association, a federation, a doctor, a coach, etc., that person should also be subject to a penalty, exactly equal to that given to the sportsman or woman, because we should not forget that the person with the shortest working life will always be the sportsman or woman.

In order to tackle both issues, we made an alternative version of the translations features by applying Latent Semantic Analysis on the set of bag-of-words translation features.

Latent Semantic Analysis

Latent Semantic Analysis (Landauer and Dumais 1997, Landauer, Foltz and Laham 1998) once again starts from the distributional hypothesis that words that are close in meaning will occur in similar contexts. In order to compare distributions of different words, first a term-document matrix is created where the rows represent unique words, the columns stand for the documents and the cells contain the word counts per document. As is done by Agirre, Lopez de Lacalle and Martínez (2005), we build one matrix per ambiguous focus word and use instances instead of full documents.

In order to normalize the term frequencies in our feature-by-instance matrix, we applied the TF-IDF (term frequency – inverse document frequency) weighting scheme that computes the relative frequency of the term in the document (instance in our case) compared to the frequency of the term in the entire document corpus (Salton 1989). Given a document collection *D*, a word *w*, and an individual document *d in D*:

$$W_{w,d} = f_{w,d} \cdot \log(|D|/f_{w,D})$$

where $f_{w,d}$ equals the number of times *w* appears in *d*, $|D|$ is the size of the corpus and $f_{w,D}$ equals the number of documents in *D* in which *w* appears (Berger et al. 2000).

In a next step, the singular value decomposition (SVD) technique is applied to construct a condensed representation of the feature space by reducing its dimensionality, which makes it possible to infer much deeper (*latent semantic*) relations between features.

Hence, SVD has some attractive characteristics that can be exploited to optimize the informativeness of the very sparse bag-of-words translation features:

(a) The translation features are highly dimensional and, as a result, very sparse - the predominant value being zero. SVD reduces the high dimensionality of the feature vectors by keeping the most relevant information. This way we can both deal with *data redundancy* (similar features will be collapsed in the same dimension) and apply some kind of *smoothing* by removing non-informative features.

(b) SVD is capable of capturing *latent* and higher-order associations between terms. Consequently, it is capable of finding hidden associations between synonyms of different instances.

Example 11 illustrates this by listing three English training instances containing the polysemous word *ring* together with their aligned Dutch translation. Although there is no lexical overlap between the first two Dutch sentences, it is clear that both sentences denote the *criminal combination of persons* sense, whereas the third sentence refers to *a circular line or figure*. When considering the binary translation features overlap, sentences 1 and 2 will have low similarities, but it is to be expected that SVD will find correlations between the semantically related features from sentences 1 and 2.

(11) (a) *English*: I should also like to add that these two texts focus, in particular, on strengthening the framework of criminal law in order to fight organised **rings** of facilitators.

Dutch: Ter verduidelijking wil ik er nog aan toevoegen dat het er in deze twee teksten voornamelijk om gaat het strafrechtelijk kader te versterken om te kunnen optreden tegen **netwerken** voor mensensmokkel.

(b) *English*: That figure has now risen to 800000, and the well-organised criminal slave trading **rings** for that is what I call them do not shrink from trafficking in children as well.

Dutch: Dit aantal is nu gestegen naar 800.000, en de goed georganiseerde criminele **organisaties** van slavenhandelaars, zoals ik deze lieden graag wil noemen, deinzen er niet voor terug om ook kinderen te verhandelen.

(c) *English*: It is mainly due to the lack of information among sportsmen and women, and the report therefore proposes that there should be an indicator on the boxes of pharmaceutical products, consisting of five Olympic **rings** and a traffic light.

Dutch: Deze is hoofdzakelijk het gevolg van een gebrekkige voorlichting aan de sportlieden. In het verslag wordt dan ook voorgesteld om de farmaceutische producten te voorzien van een duidelijk etiket met vijf Olympische **ringen** en een verkeerslicht.

If we now consider the two most important dimensions that result from the SVD reduction, we indeed see in Table 3.8 that the first two sentences are much more correlated than the third sentence, which is characterized by very different values. SVD is indeed capable of finding correlations between terms that are semantically close and collapses them into the same dimension in the new representation.

	Sentence 1	Sentence 2	Sentence 3
dim_1	1.321	1.233	3.243
dim_2	-0.507	-0.861	1.295

Table 3.8: Three Dutch training instances for *ring* represented in the reduced semantic space that results from SVD. The first two sentences are highly correlated, whereas the third sentence shows very different values.

In the next section, we describe the mathematical foundations of singular value decomposition, which is the core tool of Latent Semantic Analysis.

Singular Value Decomposition

The singular value decomposition (SVD) technique is a dimensionality reduction technique that is capable of finding correlations between the different features. To this end, SVD decomposes a given $m \times n$ term-by-document – or in our case feature-by-instance – matrix A into the product of three new matrices:

$$A = USV$$

where

- U is the $m \times r$ matrix whose columns are orthogonal eigenvectors of A, also called the left singular vectors[6];
- S is a diagonal $r \times r$ matrix whose diagonal elements are the r singular values of A, that are represented in descending order;
- V is the $r \times n$ matrix whose columns are orthogonal eigenvectors of AA, also called the right singular vectors;
- $V\{T$ is the transpose of V.

The matrices U and V thus represent terms and documents in a new space, where U contains the terms represented in the latent space (rows of A) and V contains the context in the latent space (columns of A).

Figure 3.2 illustrates the singular value decomposition of the $m \times n$ matrix X. The matrix S is the diagonal matrix containing exactly r singular values, where r is called the *rank*[7] of X.

A more intuitive explanation of SVD consists in viewing SVD as a "process where the axes are rotated in the n-dimensional space. The largest

[6] A is the *transpose* of matrix A, or the matrix which is formed by turning all the rows of A into columns and vice-versa.

[7] The rank of a matrix is the number of linearly independent rows or columns in the matrix.

$$\overset{X}{\underset{m\times n}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn}\end{pmatrix}}} = \overset{U}{\underset{m\times r}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr}\end{pmatrix}}} \overset{S}{\underset{r\times r}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr}\end{pmatrix}}} \overset{V^{\mathsf{T}}}{\underset{r\times n}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn}\end{pmatrix}}}$$

Figure 3.2: Singular value decomposition of the $m \times n$ matrix X with rank r.

variation among the documents are represented along the first axis, the second largest variation along the second dimension and so forth until the last singular value" (Lopez de Lacalle 2009).

When computing the SVD of a matrix, it is desirable to reduce its dimensions by keeping its first k singular values. Since these are arranged in descending order along the diagonal of S, and this arrangement is retained when constructing U and V, keeping the first k singular values is equivalent to keeping the first k rows of S and V and the first k columns of U. The most important dimensions that result from the SVD reduction are supposed to represent *latent semantic dimensions* or the most important concepts related to the instances and terms.

By selecting the top k singular values, we obtain a dimensionality reduction, or a *latent semantics* representation of matrix A, where noisy dimensions have been removed. This matrix A is referred to as the *rank-k approximation* of A or the *reduced SVD* of A. The low-rank matrix approximation consists in solving the problem of approximating a matrix A with another *truncated* matrix $A\{*$ which has a specific rank k. This is known as the Eckart–Young Low Rank Approximation theorem (Eckart and Young 1936).

More formally, by selecting the first k singular values from S and the first k columns of U and V, and by replacing the rest of the singular values by zero, we obtain the rank k approximation of A in the following way:

$$A_k = U_k S_k V_k$$

In this way, we pass from the original vector space defined by A to the k dimensional reduced space A_k or the *latent semantic space.* By representing the features and instances in a low-dimensional vector space, words with similar distributional patterns are projected into the same dimension.

Latent Semantic Translation Features

After reducing the dimensionality of our training matrix to A_k, we mapped our training and test data in the newly constructed latent space. In order to do so, we applied the LSI method of Deerwester et al. (1990), which allows to project the vectors in the low-dimensional space and collapses words with similar distributions into the same dimension of the reduced space. We built one unique feature-by-instances matrix and selected 50 dimensions of the latent space (cfr. Lopez de Lacalle (2009) for a similar approach), meaning that we selected the first k (50) columns from the S and U matrices. We thus mapped each row and column – respectively corresponding to a feature and instance – in the k-dimensional space[8].

Next, we projected the training and test vectors $\overrightarrow{t}$ into the reduced space by applying the following transformation:

$$\overrightarrow{t}_k = \overrightarrow{t} U_k S_k \{-1$$

As a result, the sparse feature vector $\overrightarrow{t}$ turns into a dense feature vector $\overrightarrow{t}_k$ in the low-dimensional space.

To summarize, here is a list of the different steps that were taken to convert the binary bag-of-words translation features into a set of *Latent semantic translation features*:

(a) Construct a feature-by-instances matrix per focus word per language, containing the bag-of-words translations as features.

(b) Apply TF-IDF to the feature-by-instances matrix in order to normalize the feature weights.

(c) Apply singular value decomposition to the normalized feature-by-instances matrix.

(d) Select the 50 first dimensions.

(e) Project the training and test features to the new (reduced) semantic space.

For the construction of the latent semantic translation features, we used the Gensim software package (Rehůřek and Sojka 2010), a set of robust and efficient python scripts to realize semantic modeling from plain text corpora[9].

The next section describes how the classification labels were selected for all instances in the training base.

[8] We can map each row (features) and column (instances) to the k-dimensional space because the space is defined by the k principal eigenvectors (corresponding to the largest eigenvalues) of AA and A.

[9] The Gensim toolbox can be freely downloaded from http://radimrehurek.com/gensim/.

3.3 Selection of the Classification Label

In our cross-lingual WSD approach, the classification label for a given training instance corresponds to the translation of the ambiguous focus word in the aligned translation in the target language.

In order to detect all relevant translations for the twenty ambiguous focus words, we ran the unsupervised statistical word alignment method GIZA++ (Och and Ney 2003) with its default settings on the selected English Europarl sentences and their aligned translations. Statistical word alignment is based on the assumption of co-occurrence: words that are translations of each other co-occur more often than random in aligned sentence pairs. GIZA++ is an efficient implementation of the IBM models (Brown, Della Pietra, Della Pietra and Mercer 1993) that starts from unannotated (raw) data from a large sentence-aligned corpus. The simplest IBM model (IBM Model One) is a lexical model that only takes the frequencies of the words in the source and target sentences into account. The more advanced IBM models also take into consideration word order (*distortion*) and 1-to-many alignments (*fertility*). Figure 3.3 shows an example of the word alignment output for an English-French sentence pair in the Europarl corpus.

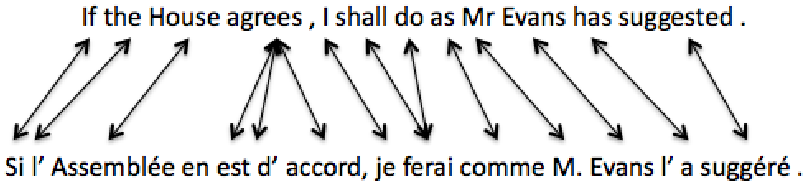

If the House agrees , I shall do as Mr Evans has suggested .
Si l' Assemblée en est d' accord, je ferai comme M. Evans l' a suggéré .

Figure 3.3: Example of the word alignment output for an English-French sentence pair.

The obtained word alignment output for the ambiguous word was then considered to be the classification label for the training instances for a given classifier (e.g. the French translation resulting from the word alignment is the label that is used to train the French classifier). This way we obtained all class labels (or oracle translations) for all training instances for our five classifiers (English as an input language and French, German, Dutch, Italian and Spanish as target languages).

We created two flavors of the training data:

(a) **Non-verified labels**

The first training set contains the automatically generated word alignment translations as labels. A post-processing step was applied to these translations in order to automatically filter leading and trailing determiners and prepositions from the GIZA++ output. It is, for instance, often the case that an English noun corresponds to a determiner and noun in the romance languages, as is illustrated in the following English-French training sentence for the word *execution*:

(12) English: That is why we urge President Clinton to grant clemency in this case and a moratorium on federal **executions** .

French: Nous prions donc le Président Clinton de faire preuve de clémence dans ce cas et d' instaurer un moratoire sur **les exécutions** fédérales .

(b) **Manually verified labels**

For the creation of the second training set, we manually verified all word alignment correspondences of the ambiguous words. This manual verification step is described in detail in Section 5.1.1. The second set-up gives an idea of the upper-bound performance in case of perfect word alignment output for the ambiguous nouns under consideration.

CHAPTER 4

Machine learning of Cross-Lingual WSD

In Section 2.2, we briefly touched upon the different machine learning algorithms used for WSD. In this chapter, we further elaborate on these learning methods and pay special attention to the two learning algorithms that were used for our Cross-lingual WSD system: Memory-based Learning (Section 4.1.1) and Support Vector Machines (Section 4.1.2). We selected the Support Vector Machines as a state-of-the-art machine learning technique that derives a model from the training data, and Memory-based Learning as state-of-the-art machine learning algorithm that does not build a model, but keeps all training instances in memory.
Section 4.2 describes how the algorithm parameter settings can be optimized by means of Genetic Algorithms.

4.1 Learning methods for WSD

In this section, we introduce some of the main approaches to supervised WSD, based on the machine learning technique that is used to train the classifier (Màrquez et al. 2006).

Probabilistic models are statistical models that usually estimate a set of probabilistic parameters that express the conditional or joint probability distributions of categories (sense labels) and contexts (information extracted from the local or global context of the ambiguous word). These parameters can then be used to assign to each new example the particular sense category that maximizes the conditional probability of a category given the observed context features (Màrquez et al. 2006). Examples of statistical algorithms that have been successfully applied to WSD are the *Naive Bayes algorithm* (Gale et al. 1992a, Leacock, Chodorow and Miller 1998, Escudero, Màrquez and Rigau 2000c) and the *Maximum Entropy approach* (Suárez and Palomar 2002).

A second group of supervised methods are based on **discriminating rules**. These methods (grouping decision lists and decision trees) use selective rules associated with each word sense; for a new instance to be classified, one or more rules that match the context information of the focus word are selected, after which the corresponding sense is assigned to the ambiguous word.
Decision lists can be considered as ordered lists of weighted *if-then-else* rules. They have been successfully applied to WSD by Yarowsky (1995), Kilgarriff and Palmer (2000) and Martínez, Agirre and Màrquez (2002).
Decision trees represent classification rules by a branching tree structure, where each branch of the tree represents a rule, and a sense label is assigned when a leaf of the tree (or terminal node) is reached.
Some methods (such as *AdaBoost*) are based on **rule combination** and linearly combine many simple and, sometimes, less accurate classification rules into a strong classifier. Escudero, Màrquez and Rigau (2000a) showed that such a boosting algorithm works particularly well for the WSD classification task.

More recently, **ensemble methods** have been applied to solve WSD (See for instance Klein et al. (2002) and Florian et al. (2002)). These ensemble methods combine several classifiers of different nature in order to improve the overall disambiguation performance. In this way, a classifier is built that overcomes the weaknesses of the individual learning algorithms. The individual classifiers can be combined in different ways. *Majority voting* outputs the sense label which has the majority of votes (or class predictions) by the different classifiers. In case of equal votes, a random choice can be made or else the ensemble does not output a sense label. The *Probability mixture* method takes into account the probability scores that are returned by the different classifiers: the probability scores are normalized and summed up by sense, and the sense label that gets the highest overall score is output by the ensemble.

In our Cross-lingual WSD system, two other machine learning algorithms were used, a Memory-based learner and Support Vector Machines, which are presented in the following two sections.

4.1.1 Memory-based Learning

Memory-based learning, also known as exemplar-based, or instance-based, learning, is a learning method based on the similarity of examples. The most popular memory-based learning method is the k **Nearest Neighbor (kNN)** method. During the training phase, this method stores all training examples – together with their sense label – in memory (hence called memory-based learning). At classification time, a previously unseen test example is presented to the system and the algorithm looks for the k most similar examples or *nearest neighbors* in memory and performs an "average" of their senses to predict a class label. In order to measure the distance between the new occurrence and the examples in memory, a similarity metric is used. As there is no generalization regarding the training data and induction is delayed to runtime, this strategy is often referred to as *lazy learning*. Authors like Daelemans, van den Bosch and Zavrel (1999) argue that an example-based method is eminently suited to NLP tasks, because it does not perform any kind of generalization on the data and therefore does not forget exceptions, which are very important in NLP. Ng and Lee (1996) presented the first work on kNN for WSD, and various other papers reported good results for WSD using a kNN system, such as for instance Hoste et al. (2002) and Escudero et al. (2000c).

In our ParaSense system, we used the memory-based learning (MBL) algorithms implemented in TIMBL (Daelemans and van den Bosch 2005), which has successfully been deployed in previous WSD classification tasks (Hoste et al. 2002). TIMBL is an implementation of the IB1 algorithm (Aha, Kibler and Albert 1991), with as main difference the definition of the k value. In TIMBL the value of k refers to *k-nearest distances* rather than *k-nearest examples*. This is done because several examples in memory can be equally similar to a new instance. So, instead of choosing one at random, all examples at the same distance are added to the set of nearest neighbors (Daelemans and van den Bosch 2005).

Figure 4.1 illustrates how general kNN works. The classification of the green point depends on the number of chosen neighbors k: when taking 3 neighbors, the red triangles seem closest, but when considering 5 neighbors, one might prefer to take the class label of the blue squares.

In other words, given a set of training instances in memory (x_1, y_1) (x_2, y_2) ...(x_n, y_n), the classification task at run time consists in finding the closest x_i for a new data point x_q. Three components are crucial to perform this classification step: (1) a distance metric, (2) k or the num-

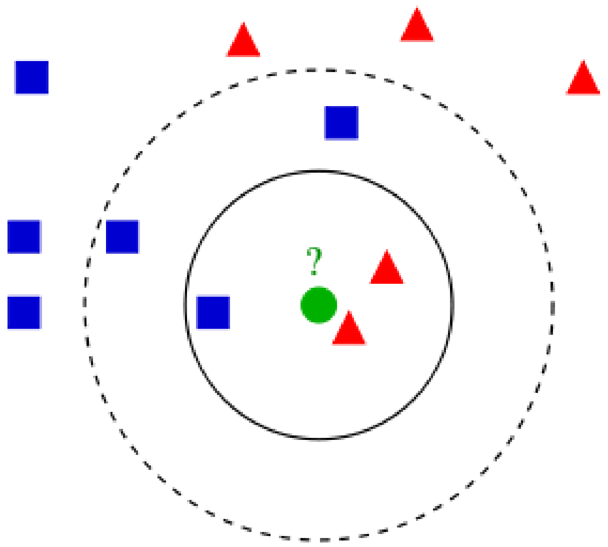


Figure 4.1: Example of k-NN classification

ber of nearest neighbors that is considered and (3) a model to extrapolate from the nearest neighbors.

- **Distance metric**.

  For classification purposes, the class label of the most *similar* instances in memory are taken as the class label for the newly presented instance. To this end, we need to define a distance metric that expresses how similar two instances are or how far x_q and x_i are. The most basic distance metric for symbolic features is the **overlap metric**, which is also implemented as the default distance metric in TIMBL. This metric states that the distance Δ between x_q and x_i can be defined as the sum of the distances δ between all n features:

  $$\Delta(x_q, x_i) = \sum_{i=1}^{n} \delta(x_{qi}, x_{ii})$$

  As the overlap metric only calculates the number of matching and mismatching features, all features are considered equally important. It is, however, often the case that some features will be more informative for the classification than others. In this case, feature selection or feature weighting will be required.

- **Nearest Neighbors**.

  The *k* nearest neighbors are the *k* training instances in memory which are *nearest* to the test instance to be classified. The class label of the nearest neighbors is used as classification for the new test instance. In the original kNN implementation (Cover and Hart 1967), the test instance receives the class label of the most common category among the nearest neighbors. Since the Euclidean distance is used in case of continuous feature vectors, it rarely happens that nearest neighbors are exactly equidistant. This is not the case for discrete and symbolic features, where the overlap is 0 or 1. Therefore, TIMBL implemented *k* as the number of *nearest distances* instead of nearest neighbors. Hoste et al. (2002) have shown that no single value of *k* works best for all data sets. It is, therefore, important to determine the value of *k* experimentally for different data sets.

- **Model to extrapolate from the nearest neighbors**.

  TIMBL uses **majority voting** as the default to decide on the classification label of a new test instance: all nearest neighbors are considered equally important, and their most frequent class is taken as the class label for the new instance.

We only mentioned the default settings of TIMBL in this section, but TIMBL offers a large range of parameter settings for distance metrics, weighting methods, *k* parameter, class weights, etc., that can all have a major impact on the classification results. This is why we decided to run optimization experiments on our training data by means of a Genetic Algorithm. We will expand on this in Section 4.2.

4.1.2 Support Vector Machines

A support vector machine (SVM) is a supervised learning system that is based on the principle of Structural Risk Minimization from the Statistical Learning Theory (Vapnik 1998). For a comprehensive introduction to support vector machines, we refer to Cristianini and Shawe-Taylor (2000). Support vector machines learn, in their most basic form, a linear *hyperplane* that separates two categories of training examples. SVM classifiers seek an optimal separating hyperplane, where the *margin*, being the minimal distance from the separating hyperplane to the nearest data points, is maximal. The data points that are at the margin are called the *support vectors*. New data points are mapped in the data space and their category is predicted based on the side of the separating hyperplane – or decision boundary – on which they are located. The distance between the data point and the hyperplane tells us something about the classification certainty. A lot of WSD systems using SVMs performed very well, among which Strapparava, Gliozzo and Giuliano (2004), Agirre and Martínez (2004a) and Lopez de Lacalle (2009).

Figure 4.2 illustrates the geometrical intuition about the maximal margin hyperplane in a two-dimensional space: the hyperplane separates the stars and circles in such a way that the margin between the support vectors (blue stars and circle) and the hyperplane is maximal.

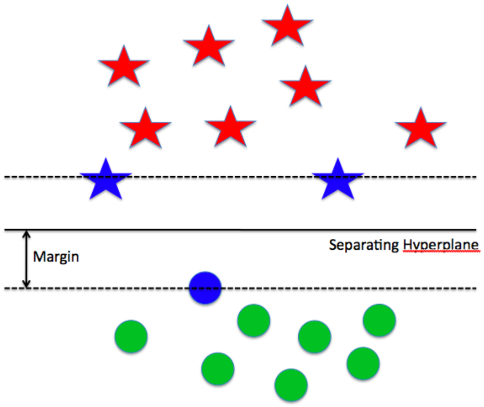


Figure 4.2: Example of SVM classification

More formally, let X be the feature space containing a data set with instances $(x_1, y_1), \ldots, (x_n, y_n)$ with $y_i \in \{-1, 1\}$. In SVM classification, the separation of the two classes (provided that they are linearly separable) is done by means of a maximum margin hyperplane $H_{w,b}$ defined by the equation:

$$\langle w, x\rangle + b = 0$$

where $\langle ., .\rangle$ stands for the inner product of two vectors, $w \in X$ is a vector orthogonal to the hyperplane and $b/||w||$ is the distance from the hyperplane to the point of origin[1]. The hyperplane $H_{w,b}$ is chosen in such a

[1] The distance of the hyperplane $H_{w,b}$ to an arbitrary point x is $|\langle w, x\rangle + b|/||w||$.

way that $\langle x_i, w\rangle + b \geq 1$ if $y_i = 1$ and $\langle x_i, w\rangle + b \leq -1$ if $y_i = -1$. This can be summarized by the constraint

$$y_i \cdot (\langle x_i, w\rangle + b) \geq 1$$

for all $i = 1, \ldots, n$.

In the extreme case, i.e. when $y_i \cdot (\langle x_i, w\rangle + b) = 1$, the distance between x_i and the hyperplane is $1/||w||$. The distance between training instances from different classes, and hence positioned on either side of the hyperplane, is therefore at least $2/||w||$. The maximum margin hyperplane $H_{w,b}$, i.e. the one that generalizes best over unseen data, thus has to be chosen in such a way that the margin that separates examples from different classes is as large as possible, in other words, $2/||w||$ needs to be maximized, which implies minimizing $||w||$.

It is, however, possible to obtain a *dual representation* or *Lagrangian formulation* in which data points are considered in the form of dot products between vectors (Lopez de Lacalle 2009). In order to solve the optimization problem in its dual form, we have to find the Lagrange multipliers $\alpha_i \geq 0$ $(i = 1, \ldots, n)$ that maximize:

$$L(\alpha_1, \ldots, \alpha_n) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$$

Subject to (for any $i = 1, \ldots, n$):

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0$$

In order to classify a new point x, one then has to determine the sign of

$$\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x\rangle + b$$

In case $\alpha_i = 0$ then instance x_i has no influence on w, in case $\alpha_i > 0$, the instance x_i is used to calculate w and is called a *support vector*.

It is, however, often the case that the classes in the data set are not linearly separable. It is then possible with SVMs to map the data space X in a highly dimensional space H through a non-linear mapping ϕ, where the data points can again be separated linearly. Let us consider an instance

$x \in X$, then its mapping to H will be $x := \phi(x)$ and the inner product between instances x and x' is denoted as:

$$k(x, x') = \langle x, x' \rangle$$

with function k being a *kernel* on $X \times X$. As the kernel functions can be implemented independently from the classification algorithm itself, different kernels can be considered for different types of feature spaces. Examples of commonly used kernel functions are *linear*, *polynomial*, *radial basis function* and *sigmoid* kernel functions.

For our experiments, we used the SVMLight toolbox of Joachims (1998) that implements Vapnik's (1998) Support Vector Machine and that contains an efficient multi-class support vector machine algorithm. This implementation uses the *soft margin variant* of the SVM method. As it is not always possible to construct a hyperplane that divides the data space linearly, the soft margin method allows classification errors during training in order to construct a more efficient hyperplane. In this way, the soft margin variant permits a trade-off between training errors and the maximization of the margin, for which parameter C must be estimated in the SVMLight implementation. We conducted experiments with SVMLight in two different set-ups: the first set-up applies the classfier with its default settings, whereas for the second set-up we used and optimized value of the C parameter. Several studies (E.g. Guo et al. 2007) present experiments with different kernel functions and parameter settings for the WSD task, and report good results for the linear kernel with trade-off parameter $C = 1.0$. Therefore, we decided to also use the multi-class implementation of SVMLight with the linear kernel and $C = 1.0$ as an optimized version of the classifier.

4.2 Parameter Optimization by means of Genetic Algorithms

As already mentioned in Section 4.1.1, most machine learning algorithms have a wide range of parameters that can be tuned to optimize the classification results. In order to find the optimal weights for the most important algorithm parameters, Genetic Algorithm (GA) optimization can be applied to the training data set. Genetic algorithms have for a long time been used to find solutions for optimization problems in large search spaces where exhaustive search in the search space is computationally not feasible. For a detailed overview of the functioning of genetic algorithms, we refer to Holland (1975), Goldberg (1989) and Mitchell (1996). Hoste et al. (2002) already showed that optimizing the parameter settings per focus word by means of a GA leads to an overall significant improvement of the classification results for WSD. Although the GA optimization process could possibly lead to overfitting on the training data, the results reported on the test data give a realistic idea of the performance improvement obtained through optimization, since the test data were selected from another corpus than the one we trained on (Cfr. Chapter 5 for a more detailed overview of the test and training corpus construction).

Genetic algorithms are search methods that are inspired by Darwin's theory of evolution – and its central idea of natural selection – and by genetics in biological systems. They start with a population of candidate solutions, called *individuals*, for a given search problem and explore different areas of the search space in parallel. Based on the Darwinian principle of "survival of the fittest", these candidate solutions are then combined to find the optimal, or better, solutions. In the case of parameter optimization, the individuals contain possible values for the classifier parameter settings. These individuals are represented as bit strings of fixed length, called *chromosomes* or *genomes*, and a possible value of a bit is called an *allele*. An essential part of the genetic algorithm is the Darwinian-based *fitness function* that judges the quality of the obtained solutions and decides which individual will survive into the next generation. In a next step, new individuals are combined using procedures of *mutation* and *crossover*. Figure 4.3 shows the general scheme of a genetic algorithm that consists of three central principles: selection, recombination and mutation.

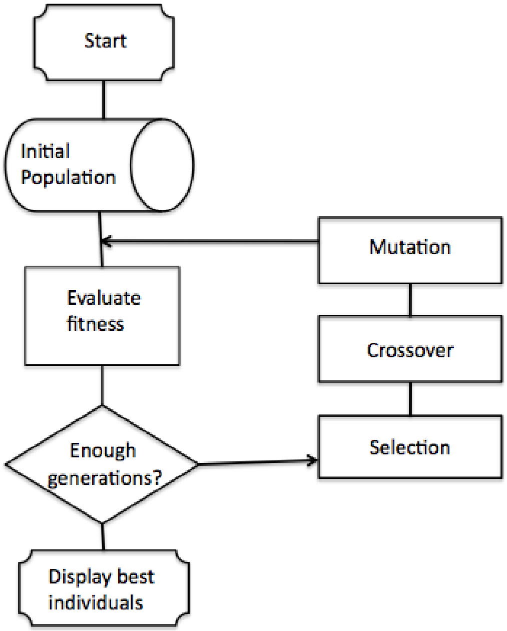


Figure 4.3: Graphical representation of an optimization procedure by means of a genetic algorithm.

- **Fitness-based selection**.

  The fitness function assesses the quality of a given solution in the evaluation process or its "fitness" for solving the problem in hand. In our setup, the fitness function evaluates the selected parameter settings with respect to the classification accuracy. After the fitness assignment, the **selection** process selects the fittest individuals that will produce offspring. Popular selection techniques are *proportional or roulette wheel selection* (Goldberg 1989) where the selection probability of an individual is determined by its fitness divided by the sum of fitnesses, and *truncation selection* (Crow and Kimura 1970) where individuals are sorted according to their fitness and only the best individuals are selected as parents. For our experiments, we applied the *tournament selection* (Goldberg and Deb 1991). For this selection method, an individual is selected by randomly picking a predefined number of individuals from the population. In a next step, the best individual from this group is chosen as the parent. This process is then repeated as many times as there are individuals to be selected.

- **Mutation**.

  In nature, mutation leads to the creation of a new individual by making a subtle change to some part of the chromosome. If mutation results in a stronger individual, then this individual will tend to pass the changed gene onto his/her offspring. Similarly, in GA optimization, chromosomes are mutated to breed new individuals in order to combine fit solutions while maintaining diversity in the population. The mutation process generates new chromosomes by slightly altering the genes of a parent. In binary code genes this mutation is performed by changing gene codes from 0 to 1 or the other way around. It is important that the mutation rate is moderate, otherwise it will have a negative effect on the fitness of the overall population. Figure 4.4 gives a graphical example of the mutation process.

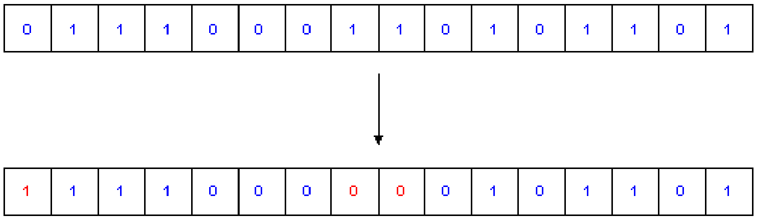


Figure 4.4: Example of mutation applied to binary genes.

- **Crossover**.

  The crossover process generates new individuals by randomly exchanging segments of two parents' chromosomes. The combination of the parents' chromosomes is done by selecting one (or more) crossover points, which split the chromosomes in different segments. The crossover reproduction happens with a certain probability, called the crossover rate which varies between 0 (no crossover) and 1 (crossover always applies). Figure 4.5 illustrates a case of one-point crossover, where the parents are split at a selected point in their chromosomes and the chromosome of the offspring is created by combining the parts of each of the parents' chromosomes[2].

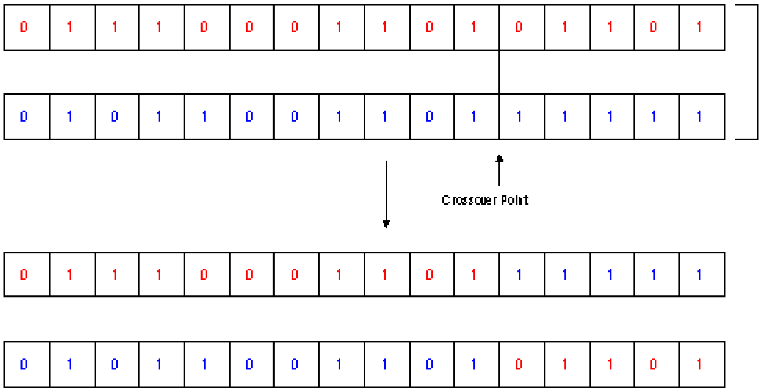


Figure 4.5: Example of one-point crossover.

For our experiments, we used the GAGrid implementation by Vereeken (2012). This is a **generational** GA, where the new population is generated using mutation and crossover on the fittest individuals that have been selected from the old population. At the end of each generation, the entire old population is replaced by the new offspring. Another type of GA are **steady-state** algorithms, where the fittest individuals are used to create offspring replacing the weakest individuals. This latter type sequentially replaces the individuals by their offspring during generation.

GAGrid is a GA implementation based on JGAP (Java Genetic Algorithm Package)[3] that provides basic genetic algorithms. It also uses the JCGrid toolkit[4] and thus enables grid computing. As a result, tasks can be distributed over a cluster of computer nodes. As both packages are written in Java, the application is platform-independent. We applied the generational GA with its standard representation and default settings. This was done in the knowledge that the optimization problem we seek to solve through GA optimization also applies to the parameter settings of the GA itself. Optimization of these parameters falls, however, outside the scope of this dissertation.

The following GA settings were used for all experiments:

[2] Both the mutation and crossover figures are copied from http://www.schatten.info/info/ga/genetic.html.

[3] textithttp://jgap.sourceforge.net/

[4] textithttp://jcgrid.sourceforge.net/

maximum number of generations	20
population size	50
crossover type	uniform crossover
crossover rate	0.02
selection type	tournament selection
selection size	13
mutation rate	0.2
preserve the fittest individual	True
min_delta_fitness	0.001

The *min_delta_fitness* value is a double indicating an absolute minimum value for the difference between fitness values between two generations.

Part III

Experiments and Evaluation

CHAPTER 5

Construction of a Cross-lingual WSD Benchmark Data Set

In order to evaluate the viability of Cross-lingual WSD (CLWSD), we constructed a lexical sample data set of 25 ambiguous English nouns. This data set was also used for the SemEval-2010[1] "Cross-Lingual Word Sense Disambiguation" task, in which systems had to provide translations of the ambiguous target nouns in five supported languages (viz. Dutch, French, German, Spanish and Italian). For a detailed description of the SemEval task, we refer to Lefever and Hoste (2010). We first released a **trial set** of five ambiguous nouns (*bank, movement, occupation, passage, plant*) to all participants of the CLWSD task in order to give them an idea of the format and content of the real test data and to give participating teams the opportunity to develop a system tailored to this specific data format. Thereafter, a **test set** of twenty nouns (*coach, education, execution, figure, job, letter, match, mission, mood, paper, post, pot, range, rest, ring, scene, side, soil, strain* and *test*) was released for the real evaluation campaign.

For the creation of the hand-tagged gold standard, we retrieved all translations of a given polysemous English noun from the parallel corpus and clustered them by meaning. Section 5.1 describes in detail how the sense inventory was constructed, whereas Section 5.2 describes the annotation

[1] `http://www.http://semeval2.fbk.eu/semeval2.php`

process of the trial and test instances[2].

5.1 Construction of the sense inventory

The document collection which served as the basis for the gold standard sense inventory was the six-lingual sentence-aligned subcorpus of Europarl that was described in detail in Section 3.1.2. This subcorpus contains 884,603 English sentences and their aligned translations in Dutch, French, English, Spanish and Italian. We selected from this subcorpus all English sentences containing one of the 25 ambiguous focus nouns and their aligned translations in the five target languages, resulting in a sentence-aligned corpus containing 46,840 sentences per language (35,686 sentences containing one of the ambiguous test words and 11,154 sentences containing one of the ambiguous trial words).

After the selection of these sentences, the following two steps were taken for the trial and test data in order to obtain a multilingual sense inventory:

(a) word alignment of the sentence-aligned data for the extraction of possible translations for the selected ambiguous nouns, followed by a manual evaluation of these alignments.

(b) manual clustering by meaning (per focus word) of the resulting translations.

The resulting multilingual sense inventory served as the basis for the annotation of both the trial and test data. For their selection of one or more contextually correct translations of a given English focus word, the annotators were only allowed to choose between the translations present in the multilingual sense inventory. We return to this annotation process in Section 5.2.

5.1.1 Word Alignment

In order to detect the possible translations for the set of ambiguous nouns, we adopted the same approach as for the selection of the classification label (cfr. Section 3.3), and performed statistical word alignment on the 46,840 selected Europarl sentences by means of GIZA++.

An example of these word alignments (marked in bold) for the word *mood* is given in the sentences below.

[2] The resulting gold standard for the trial and test instances, as well as the evaluation script can be downloaded from http://lt3.hogent.be/semeval/SemEval_2010/

(13) *SOURCE*: In the course of the debate , I shall , of course , follow carefully the points of view put forward and interpret the **mood** of the Assembly before the continued discussion to establish the EU ' s position .

DUTCH: Ik zal uiteraard tijdens het debat nauwkeurig volgen wat er aan standpunten naar voren wordt gebracht , en de **stemming** peilen met het oog op de verdere discussie waarin de EU haar standpunt zal bepalen .

GERMAN: Ich werde selbstverständlich im Fortgang dieser Aussprache genau die hier vorgetragenen Standpunkte verfolgen und im Hinblick auf die weitere Debatte zur Positionsbestimmung der EU die **Stimmungslage** ausloten .

FRENCH: Je suivrai bien entendu avec attention les points de vue formulés au cours du débat , et je m' efforcerai de saisir l' **atmosphère** avant la suite de la discussion , qui aura pour objet d'établir la position de l' UE .

SPANISH: Obviamente , durante el debate seguiré atentamente los criterios que se presenten y captaré la **disposición de ánimo** para el próximo debate , en el que se fijará la postura de la Unión .

ITALIAN: Naturalmente nel corso della discussione ascolterò con particolare attenzione i vari punti di vista e ne trarrò le debite **conclusioni** ai fini del dibattito in corso sulla definizione della posizione dell' Unione europea .

As example (13) clearly illustrates, one single focus word can lead to multiword translations, such as *disposición de ánimo* in Spanish; and to compounds, such as *Stimmungslage* in German. In both cases, we keep the multipart translation as a valid translation suggestion.

One could argue that the CLWSD task, that aims to predict a correct translation for an ambiguous focus word, would benefit from considering larger text chunks than isolated words as a translation unit. This could enable the system to better process fuzzy links and compound translations. We decided, however, to select ambiguous words instead of phrases for the following reasons:

- It was our objective to create a lexical sample of words as a benchmark set for CLWSD. Therefore, we followed the procedure that is usually applied for the creation of the more traditional WSD lexical sample tasks:

- the ambiguous focus word (E.g. *bank*) is tagged in the input sentence. In case the word is part of a compound, only the head is tagged:

  E.g.
  ... savings <head> bank </head> ...
  ... central <head> bank </head> ...
  ... City <head> bank </head> ...

- the WordNet labels that are used to tag the training corpus sometimes refer to a compound containing the ambiguous target word. There are, for instance, separate WordNet labels for "bank" in the meaning of "savings bank", "letter" in the "varsity letter" meaning, "department of education" as separate meaning of "education", etc. This way, we follow the more traditional lexical sample approach that tags isolated words in the input, and sometimes assigns compound meanings to the ambiguous focus word.

- We conceived the CLWSD task as an unsupervised task where participants are not provided with a sense-tagged training corpus. In case we would consider working with phrases, participants would then have to decide themselves upon the phrase boundaries for the creation of the training data. As a consequence, it would almost be impossible to organize a shared lexical sample task for CLWSD where participating systems are evaluated based on the same gold standard for a predefined set of ambiguous words.

- Working with phrases instead of isolated words would reduce the number of training instances per phrase and certainly result in sparseness in the training data, which is particularly problematic for those ambiguous focus words that have low frequencies in the corpus already (E.g. *coach, pot*).

Another valid approach would be to apply decompounding on the translation labels. In this case, a generic decompounding module is needed that can be applied for all languages, in order to make sure that the differences in classification performance across languages is not due to different decompounding accuracy. In order to reduce the effect of error percolation to a minimum, we opted to only use word alignment, and no other preprocessing such as decompounding, on the data. Further research is required to test the positive impact of decompounding on the word alignment and to measure the error percolation it possibly provokes for the WSD task. In addition, we believe that the compound translations contain very strong local context information that might be useful to generate an appropriate translation in multilingual applications.

All GIZA++ alignment links for the ambiguous focus words were manually verified in the six languages. The human annotators (one per language) were instructed to correct wrong word alignments and assign a "NULL" link to words for which no valid translation could be identified. While checking the word alignment output, the annotators were also asked to provide additional information in a dedicated remarks section for the four specific remark categories as illustrated below:

(a) the translation is a *compound* that corresponds to an English multiword

(14) *SOURCE*: By the same token , we should praise the **Green Paper** on a European strategy for a sustainable , competitive and secure energy supply .
GERMAN: Ebenso loben sollten wir das **Grünbuch** "Eine europische Strategie für nachhaltige , wettbewerbsfähige und sichere Energie".

Paper	Grünbuch

Remarks: compound Green Paper

(b) there is a *fuzzy link* between the focus word and its translation. Fuzzy links denote translation-specific shifts such as paraphrases or divergent translations, where there is no exact translational correspondence between the source and target words.

(15) *SOURCE*: That agreement is the **test** of whether Europe is on the move .
DUTCH: Die overeenstemming is **bepalend** voor de vraag of Europa op koers ligt .

test	bepalend

Remarks: fuzzy link: the test of - is bepalend voor (English: "determines")

(16) *SOURCE*: I have in mind a voluntary agreement, according to which 30% or 50% of the funds would be used for research, training and **further education**, and for new infrastructure .

SPANISH: Pienso en una disposición voluntaria según la cual el 30% o el 50% de los recursos se emplearía para fines de formación y **perfeccionamiento** como para la creación de nuevas infraestructuras .

education perfeccionamiento

Remarks: fuzzy link: further education - perfeccionamiento (English: "improvement, perfectioning")

(17) *SOURCE*: If FIFA had done its **job**, the Commission would not have needed to get involved in this .
SPANISH: Si la FIFA hubiese hecho lo que **debía**, la Comisión no se habría visto obligada a actuar .

job debía

Remarks: fuzzy link: its job - lo que debía (English: "the right thing")

(c) there is a *tokenisation* problem (e.g. the English focus word is part of a hyphenated compound whereas only the focus word itself should be considered)

(18) *SOURCE*: Finally , it is vital that we accept the committee ' s amendments on **execution-only** business .
FRENCH: Enfin , il est vital que nous acceptions les amendements de la commission concernant les opérations de simple **exécution** .

execution exécution

Remarks: tokenisation (execution-only)

(d) the focus word is used with a different part-of-speech tag (other than *noun*) and therefore marked as *wrong input*, meaning that it should not be considered for building up the sense inventory.

(19) *SOURCE*: I agree fully with Mr Hatzidakis that the Stabilisation and Association Agreement helps us to **coach** the country towards European standards as regards electoral reform .
ITALIAN: Condivido senza riserve lopinione dellonorevole Hatzidakis , secondo cui laccordo di stabilizzazione e di associazione ci aiuta ad **avvicinare** il paese ai requisiti europei in materia di riforma elettorale .

coach avvicinare

Remarks: wrong input (PoS)

Table 5.1 presents an overview of the frequency of all remark categories (expressed in percentages of the total amount of instances) for the trial data. A manual inspection of these remarks per category revealed a number of trends. First of all, compound translations tend to occur very frequently in German and Dutch (up to 73% for German), and much less frequently in the romance languages. Secondly, the number of fuzzy links also varies considerably between the different words and different languages. Finally, the number of tokenisation and wrong input problems is very limited.

Tables 5.2 and 5.3 list the frequencies of the two most interesting remark categories, the fuzzy links and compounds categories for the twenty words in the test data (divided alphabetically across the two tables). The figures show tendencies similar to the trial words: fuzzy links percentages range from $\leq$ 5% fuzzy links across all languages for the words *education* and *execution*, to $>$ 20% fuzzy links for words such as *pot*, *mood* (32.6% fuzzy links in Dutch), *range* (63.9% fuzzy links in Dutch) or *strain* (42.44% fuzzy links in German). Whereas these fuzzy links are partly caused by compound translations, we can observe a similar, yet less pronounced, tendency in Italian, Spanish and French. We observed that these words have generally been translated more freely. They are often paraphrased, or have more or less free translational correspondences, typically resulting in fuzzy or even NULL links. Example (20) illustrates this in a training instance containing the word **mood**:

(20) *SOURCE*: We are **in a listening mood** , but we would also like to contribute to improving human rights in the world .

DUTCH: We **willen graag luisteren** , maar we willen ook graag bijdragen aan een verbetering van de mensenrechten in de wereld .

GERMAN: Wir sind **auf Zuhören eingestellt** , wurden aber auch gern dazu beitragen , die Lage bei den Menschenrechten in der Welt zu verbessern .

FRENCH: Nous sommes **enclins à l'écoute** , mais nous voulons également contribuer à améliorer la situation des droits de l'homme dans le monde .

SPANISH: Estamos **a la escucha** , pero también nos gustaría contribuir a mejorar los derechos humanos en el mundo .

> *ITALIAN*: Siamo **pronti ad ascoltare** , ma vorremmo anche contribuire al miglioramento della situazione dei diritti umani nel mondo .

The considerable proportion of compound translations also results in a higher number of different translations for Dutch and German, which has important consequences for the multilingual WSD task the data set was designed for. In multilingual WSD systems, the sense label typically consists of a translation, whereas in more traditional WSD approaches, the label consists of a sense picked from a predefined sense inventory such as WordNet. As a consequence, the multilingual WSD systems for Dutch and German will have a broader set of classes (or translations) to choose from, which makes the WSD task more complicated. Figure 5.1 illustrates this by listing the number of different translations (or classes in the context of WSD) for all trial and test words.

However, the part of the compound that does not lexicalise the ambiguous word often carries important information for disambiguating the ambiguous word. If we consider, for instance, the German and Dutch translations of *flower pot*, *Blumentopf* and *bloempot*, respectively, the “Blumen-” and “bloem-” parts restrain the sense of *pot* to a very large extent.

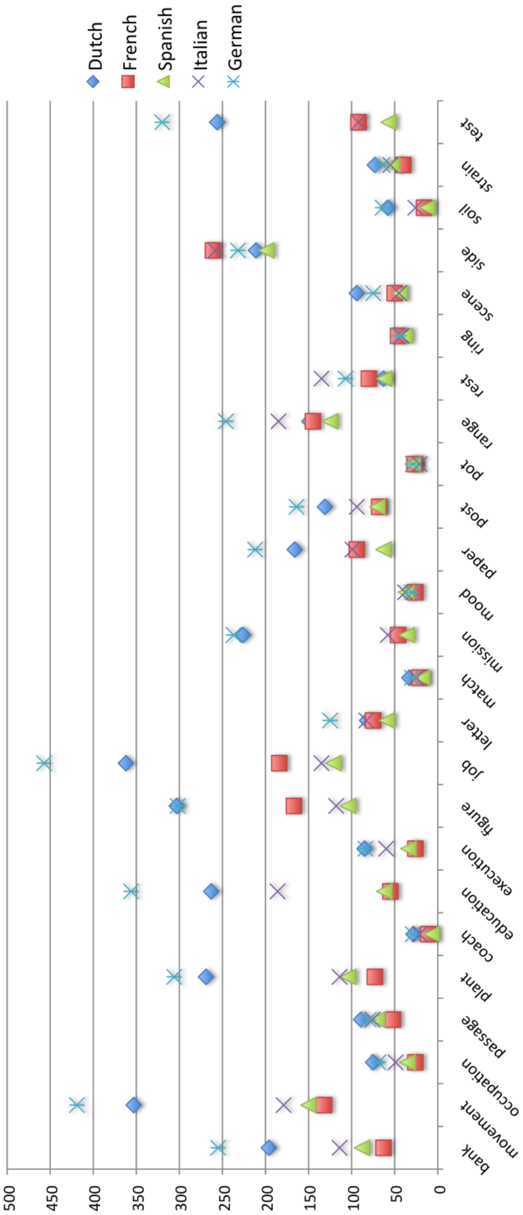


Figure 5.1: Number of different translations per word for Dutch, French, Spanish, Italian and German.

	Dutch	**French**	**Spanish**	**Italian**	**German**
bank (total: 4029 instances)					
Compound	31.0%	3.4%	0.7%	11.0%	73.0%
Fuzzy link	0.6%	4.4%	0.8%	0.8%	1.4%
Tokenisation	1.7%	1.8%	0.1%	0.1%	1.3%
Wrong Input	0.3%	3.1%	0.3%	0.3%	2.3%
movement (total: 4221 instances)					
Compound	20.9%	2.1%	1.0%	4.1%	66.0%
Fuzzy link	4.1%	2.6%	1.7%	1.0%	4.1%
Tokenisation	0.3%	0.6%	0.0%	0.4%	0.6%
Wrong Input	0.0%	0.0%	0.0%	0.0%	0.0%
occupation (total: 633 instances)					
Compound	8.5%	0.3%	0.0%	8.0%	10.1%
Fuzzy link	9.1%	7.6%	0.5%	1.6%	6.6%
Tokenisation	1.6%	30.6%	0.2%	1.7%	2.4%
Wrong Input	0.0%	0.0%	0.0%	0.0%	0.0%
passage (total: 237 instances)					
Compound	3.4%	0.8%	0.4%	9.7%	1.3%
Fuzzy link	19.0%	19.8%	11.8%	3.4%	18.1%
Tokenisation	0.0%	2.5%	0.0%	0.0%	0.0%
Wrong Input	0.0%	0.0%	0.0%	0.0%	0.0%
plant (total: 1631 instances)					
Compound	46.8%	7.9%	7.8%	22.4%	54.8%
Fuzzy link	2.6%	8.3%	1.3%	1.0%	4.5%
Tokenisation	0.5%	1.1%	0.0%	1.3%	0.8%
Wrong Input	1.3%	2.6%	1.5%	1.3%	2.0%

Table 5.1: Percentages of remark categories per word in the trial data

	Dutch	**French**	**Spanish**	**Italian**	**German**
coach (total: 73 instances)					
Compound	26.0%	2.7%	1.4%	20.6%	38.4%
Fuzzy link	12.3%	5.5%	0.0%	6.9%	13.7%
education (total: 4557 instances)					
Compound	17.0%	0.0%	0.5%	4.3%	25.7%
Fuzzy link	2.0%	0.0%	1.9%	0.4%	4.8%
execution (total: 536 instances)					
Compound	6.0%	0.0%	1.9%	2.8%	10.6%
Fuzzy link	2.8%	4.5%	2.1%	0.7%	4.1%
figure (total: 2663 instances)					
Compound	9.9%	0.0%	2.5%	0.0%	10.9%
Fuzzy link	2.0%	0.0%	4.0%	1.3%	6.1%
job (total: 7844 instances)					
Compound	7.9%	0.1%	1.2%	0.1%	10.0%
Fuzzy link	4.6%	0.0%	2.5%	0.8%	8.1%
letter (total: 1874 instances)					
Compound	7.5%	0.8%	1.0%	1.3%	14.5%
Fuzzy link	7.7%	2.1%	6.1%	2.6%	8.4%
match (total: 384 instances)					
Compound	10.2%	0.5%	1.3%	1.0%	7.6%
Fuzzy link	6.3%	1.0%	1.0%	2.3%	7.0%
mission (total: 1432 instances)					
Compound	33.0%	0.1%	0.1%	0.8%	38.6%
Fuzzy link	8.2%	0.6%	3.6%	0.3%	3.5%
mood (total: 118 instances)					
Compound	5.1%	1.7%	5.1%	0.8%	2.5%
Fuzzy link	14.4%	2.5%	17.8%	9.2%	26.3%
paper (total: 3735 instances)					
Compound	70.7%	0.2%	1.7%	1.4%	73.6%
Fuzzy link	5.7%	0.1%	2.4%	1.9%	2.2%

Table 5.2: Percentages of compounds and fuzzy links per word for the first ten test words

	Dutch	**French**	**Spanish**	**Italian**	**German**
post (total: 1638 instances)					
Compound	14.3%	0.6%	5.5%	1.4%	25.5%
Fuzzy link	32.6%	1.0%	7.3%	2.3%	8.7%
pot (total: 81 instances)					
Compound	18.5%	18.3%	8.6%	4.9%	22.2%
Fuzzy link	14.8%	9.8%	18.5%	8.5%	24.7%
range (total: 1608 instances)					
Compound	5.0%	0.5%	3.0%	1.1%	9.3%
Fuzzy link	63.9%	0.2%	7.5%	2.4%	22.2%
rest (total: 2304 instances)					
Compound	12.9%	0.1%	0.5%	0.7%	10.6%
Fuzzy link	23.8%	0.8%	5.0%	3.2%	12.3%
ring (total: 206 instances)					
Compound	24.8%	1.5%	6.8%	2.4%	22.8%
Fuzzy link	17.5%	9.2%	18.0%	2.4%	16.0%
scene (total: 345 instances)					
Compound	12.2%	0.0%	12.2%	0.0%	13.1%
Fuzzy link	12.8%	10.5%	14.0%	8.1%	17.2%
side (total: 4207 instances)					
Compound	5.6%	0.0%	4.8%	1.1%	13.1%
Fuzzy link	28.6%	0.0%	8.7%	5.8%	15.6%
soil (total: 294 instances)					
Compound	19.1%	0.7%	0.3%	1.4%	18.0%
Fuzzy link	1.4%	2.0%	1.4%	2.0%	9.5%
strain (total: 172 instances)					
Compound	10.5%	2.9%	5.2%	0.0%	9.9%
Fuzzy link	18.0%	14.0%	19.8%	11.1%	42.4%
test (total: 1617 instances)					
Compound	33.3%	0.6%	0.3%	0.9%	40.3%
Fuzzy link	13.8%	4.0%	5.2%	2.6%	7.3%

Table 5.3: Percentages of compounds and fuzzy links for the last ten test words

Word Alignment Performance

We evaluated the performance of the automatically generated word alignments against our manually validated word alignment reference. A straightforward measure for doing this is the F-score, which combines precision and recall. The following formulas were used to calculate precision, recall and F-score on all word-to-word links for our focus words, with R referring to the reference set of manually generated alignments and A referring to the automatic alignments generated by the system:

$$Precision = \frac{|A \cap R|}{|A|} \tag{5.1}$$

$$Recall = \frac{|A \cap R|}{|R|} \tag{5.2}$$

$$F{-}score = \frac{2 \cdot Precison \cdot Recall}{Precision + Recall} \tag{5.3}$$

Table 5.4 lists the average precision, recall and F-score for all trial and test words on all five language pairs (with English as the source language, and the other five languages as target languages). The scores also contain additional standard deviation figures between brackets. For all languages, a similar word alignment performance can be observed with F-scores ranging between 76% and 82% and standard deviations between 10% and 13%.

	Precision	Recall	F-score
Spanish	79.32 (12.7)	85.67 (10.1)	81.74 (10.1)
French	75.48 (12.8)	83.64 (9.4)	79.12 (10.8)
Italian	76.68 (13.2)	77.38 (14.1)	76.38 (12.7)
Dutch	77.04 (12.4)	80.66 (11.3)	78.47 (11.0)
German	75.69 (13.2)	81.07 (9.6)	78.07 (10.9)

Table 5.4: Word alignment performance averaged across all twenty-five ambiguous words, complemented with standard deviation information.

When we focus on the single words, however, we observe considerable performance differences. Table 5.5 gives an overview of the precision, recall and F-scores for all individual words in Dutch and shows F-scores ranging from as low as 48.7% for *post* to as high as 93.4% for *bank*. In general, we see that word alignment performance seems to be related to the number of compound and fuzzy translations of a given word. For the

	Precision	Recall	F-Score
coach	75.30	81.30	78.20
education	88.70	92.00	90.36
execution	86.10	75.70	80.60
figure	84.70	86.50	85.60
job	85.00	86.30	85.60
letter	90.70	90.60	90.60
match	67.50	56.10	61.20
mission	71.50	85.30	77.87
mood	69.10	77.00	72.90
paper	80.90	91.70	86.00
post	42.10	57.60	48.70
pot	64.70	79.50	71.46
range	74.00	92.40	82.20
rest	78.80	92.30	85.10
ring	52.50	72.40	60.95
scene	76.60	74.90	75.70
side	70.10	90.30	79.00
soil	88.90	85.10	87.00
strain	66.50	65.00	65.70
test	78.60	70.60	74.40
bank	92.40	94.30	93.40
movement	93.10	90.70	91.90
occupation	91.30	84.30	87.60
passage	78.30	63.10	69.90
plant	78.70	81.40	80.00

Table 5.5: Dutch word alignment performance expressed in precision, recall and F-score for all individual trial and test words.

word *ring*, for example, 24.8% of the Dutch translations and 22.8% of the German translations were compound translations.

In order to substantiate these word alignment figures, we also assessed the quality of the manual correction of the word alignments by calculating inter-annotator agreement on a sample of 6,500 sentences. We again used the formulas of (5.1), (5.2) and (5.3) to calculate precision, recall and F-score, with R now referring to the set of word-to-word alignments of the first annotator and A referring to the set of alignments that were verified by the second annotator. The resulting inter-annotator agreement reaches an F-score of 95.1% which allows us to consider the word alignment correction process as being reliable.

5.1.2 Manual Clustering

After manual verification, the resulting translations were clustered per meaning for evaluation purposes. A first objective was to make the annotation job easier and more efficient. We assumed it would be very difficult for the annotators to scroll through the entire list of valid translation candidates to select the three most appropriate translations. The choice of the correct translation for the word used in a given context is more restricted once the annotator has decided on the appropriate sense cluster. In addition, we also wanted to gain some insights into the overlap of the obtained sense clusters with the sense distinctions made in existing dictionaries. Furthermore, the sense inventory might also be a useful resource for the evaluation of unsupervised sense-clustering algorithms in future experiments.

For the clustering of the translation labels, the following steps were followed:

- The translations were coupled across the languages on the basis of unique sentence IDs.
- We created a matrix containing the five translation labels per sentence, and also stored the full sentence per language. As it is sometimes the case that Europarl sentences denoting the same meaning of the ambiguous focus word contain the same five translations in the supported languages, we created a unique list of these translation combinations. Table 5.6 shows a sample of the translation matrix for the word *bus*. As sentences 2 and 3 share the same target translations in the five supported languages, only one of the two sentences will be kept in the final list for manual inspection. This list was the starting point for the creation of the clustering for a given ambiguous focus word.

ID	**Dutch**	**Italian**	**French**	**German**	**Spanish**
1	bustoerisme	corriera	autobus	Bustourismus	autocar
2	busvervoer	pullman	autocar	Busreise	autocar
3	busvervoer	pullman	autocar	Busreise	autocar

Table 5.6: Sample of the translation matrix for the ambiguous focus word *bus*.

- All unique translation combinations were manually inspected and grouped by meaning. The resulting clusters were organized in two levels in which:
  - the top level reflects the main sense categories. If we take, for instance, the word *coach*, we have four top level clusters: (1) (sports) manager/handler meaning, (2) bus, (3) carriage and (4) part of a train.
  - the subclusters represent the finer sense distinctions. In case translations denote more specific meanings that cannot be used for more general usages of the word, a subcluster is created. As a case in point, we can take the word *pot*, that has seven top level clusters, of which the first sense refers to *pot* in the "container" sense. This first top level sense has in turn three subclusters referring to (1) the more general "container" or "cooking vessel" sense, (2) the "mixture" sense as it is used in *melting pot* and (3) the "flower pot" meaning of the word.
- Translations that correspond to English multiword units were identified and in case of non-apparent compounds (i.e. not marked with a hyphen), the different compound parts were separated by §§ in the clustering file.
- Finally, all clustered translations were manually lemmatized.

Table 5.7 exemplifies such a manual clustering for the word *pot*. Additional examples of the manual clustering can be found in Appendix C which presents clustering tables for *coach*, *execution* and *figure*.

Since the clustering task was a very labour-intensive process, it was only performed by one single annotator and revised by a second one. The clustering of the translations was primarily meant as an aid for the annotation of the trial and test instances, and not included in the evaluation of the system classification output. We are well aware that grouping all translations into clusters and subclusters is a subjective task that brings up the issue of sense granularity and the arbitrary division into sense distinctions as performed by lexicographers (Cfr. Section 1). We will show, however, in Section 5.9, that there is a fairly high consensus on the cluster choice for the annotation of the test and trial sentences.

In order to gain some insights into the sense coverage by the Europarl corpus, we also compared the Europarl translations and corresponding clustering results with the sense distinctions made in monolingual[3] and

[3] WordNet, Longman Dictionary of Contemporary English (LDOCE), Van Dale synoniemenwoordenboek, Duden Synonymworterbuch, le Grand Robert.

English	Dutch	German	French	Spanish	Italian
1. Cooking Vessel, pot					
1.1. General					
	pot	Topf	chaudron	caldero	calderone
	potje	Mann	marmite	olla	vaso
		Töpfchen	pot	cazuela	vasetto
honey pot	potje	Honig§§glas	pot	tarro	vasetto
1.2. Mixture					
	geheel	Fass	Magma	magma	magma
	hoop	Topf	pot	olla	calderone
	ketel	Glas	sac	saco	vasetto
pot pourri	hutspot	Potpourri	fourre-tout	variopinto	accozzaglia
	potpourriachtig		pot-pourri	popurrí	guazzabuglio
melting pot	smelt§§kroes	Schmelz§§tiegel	melting-pot	crisol	crogiolo
	mengsel	Gemisch	enchevêtrement	cúmulo	melting-pot
1.3. Flower pot					
	pot	Blumen§§topf	pot	olla	vaso
	bloem§§pot			maceta	
2. Marijuana					
	blowen	kiffen	herbe	maría	canna
	hasj	Cannabis			
3. batch, great deal of					
	hoop	Topf	sac	saco	calderone
	pot	Pot	montagne	cazuela	mucchi
pot of money	aardig	Reibach	cagnotte	cantidad	risparmi
4. funding					
	bedrag	Geld	cagnotte	ayuda	somma
	fonds	Fonds	fonds	fondo	fondo
	kas	Topf	poste	beneficio	cassa
5. take a pot shot					
	onder vuur nemen	Schuss	tirer à vue	quemarropa	attaccare
6. toilet					
	po	Toiletteneimer	vase de nuit	orinal	vaso
7. Pol Pot					
	Pot	Pot	Pot	Pot	Pot
		Pot-Regime			
		Pol-Pot-Truppe			

Table 5.7: Part of the manual clustering for the word **pot**

bilingual dictionaries[4] in a number of Master's Theses (Boden 2010, Claus 2011, DeConinck 2010, Santy 2008). These studies all lead to similar conclusions:

- In general, the main sense distinctions are covered in both the dictionaries and the Europarl corpus.
- Overall, the explanatory (monolingual) dictionaries contain more and finer sense distinctions than the clustering based on the Europarl translations, while the translation dictionaries contain fewer meanings than the Europarl clustering.
- Different sense distinctions are covered by different dictionaries. If we compare for instance the LDOCE with the meanings listed in the Van Dale English-Dutch, we see that the "point of support" meaning of the word *rest* is not covered in the LDOCE, whereas the monolingual dictionaries distinguish in general more and finer sense distinctions than the bilingual dictionaries. This observation confirms the statement of Kilgarriff (1997) that "any working lexicographer is well aware that, every day, they are making decisions on whether to lump or split senses that are inevitably subjective: frequently, the alternative decision would have been equally valid."
- Since the Europarl corpus contains British English, typical American English meanings are lacking from the corpus. For example, the word *letter* receives the following meaning in the LDOCE dictionary: "(AmE) A large cloth letter that you sew onto a jacket, given as a reward for playing in a school or college sports team". Another example is the American English "cheapest type of seats on a plane or train" meaning of *coach* that is listed in the LDOCE.
- Sometimes dictionaries do not contain fine sense distinctions which are very domain-specific and related to the political context of the Europarl corpus. The corpus, however, does not contain very fine or specialized sense distinctions not belonging to the domain of the corpus. If we take for instance the word *letter*, both the LDOCE and the corpus contain the first main meaning, namely *missive* ("a written message addressed to a person or organization"), whereas only the Europarl corpus contains the finer sense distinctions of *mail, exchange of letters, legal document/agreement/law*. The clustering for *side*, on the other hand, contains the meaning *of secondary importance/on the side*, but does not contain the various finer sense distinctions that are listed in the LDOCE[5].

[4] Van Dale English-French, English-Dutch, English-Spanish, English-German and Oxford Dictionary English-Spanish

[5] (1) "used to say that someone does work in addition to their regular job", (2) "secretly,

- Since the language used in the corpus reflects current language usage, old-fashioned meanings are not to be found in the corpus. The LDOCE lists, for instance, the “farming as a job or way of life” meaning of *soil*, which is not present in the corpus. The same observation can be made for informal or spoken language, since the Europarl register is very formal in general. The LDOCE lists, for instance, the (spoken) “television station” meaning of *side*, which is lacking in the Europarl corpus. New meanings or word usages, however, are sometimes covered by the corpus and not (yet) by the dictionaries. An example of such new word usage is *anthrax letter*.

The above-mentioned observations clearly acknowledge the influence of the corpus domain on the extracted sense distinctions. Europarl vocabulary is often related to politics, economics, legal matters, etc and is, in general, very formal. In order to optimize the performance of the ParaSense system for very different domains, it would therefore be recommended to add parallel data from these specialized domains to the training corpus.

5.2 Annotation of trial and test instances

The resulting sense inventory was used to annotate the sentences in the trial and test set. For the construction of the trial set, three annotators (per language) labeled 20 sentences per ambiguous noun, whereas for the test set, 50 sentences per ambiguous word were annotated, which amounts to 1100 sentences in total. Both trial and test sentences were extracted from the JRC-ACQUIS Multilingual Parallel Corpus[6] and the British National Corpus[7]. While manually selecting the sentences, we tried to cover a wide range of the different meanings of the ambiguous focus word. As we conceived the CLWSD task as an unsupervised task, we encouraged participants to use the Europarl corpus to train their system. Therefore, we decided to select the trial and test sentences from a corpus other than the one that was used for the creation of the sense inventory. As a consequence, there is no perfect overlap between the domain and vocabulary usage of the sense inventory and the trial and test sentences. We do believe, however, that the Europarl corpus is generic enough to generate a sense inventory that covers most of the meanings of our set of ambiguous focus words. This assumption is confirmed by the comparison we carried

and dishonestly or illegally” and (3) “food that is served on the side, is ordered with the main dish in a restaurant, but is nog usually part of that dish”.

[6] http://wt.jrc.it/lt/Acquis/

[7] http://www.natcorp.ox.ac.uk/

out between the Europarl translations and the meanings that are listed in the monolingual and bilingual dictionaries (Cfr. Section 5.1.2).

For the annotation of the ambiguous focus words in the sentences, we proceeded in the following way: the annotators were asked to (a) pick the contextually appropriate sense cluster and to (b) choose their three preferred translations from this cluster, without being guided towards a choice for more coarse- or fine-grained clusters. In case they could not find three appropriate translations, they were also allowed to provide fewer. These translations were used to assign frequency weights to the gold standard translations per sentence. Example (21) below shows the annotation result in both French and German for an English test sentence containing the word *pot*. The translations are derived from the (simplified) sense inventory displayed in Table 5.7.

(21) Bring them in one at a time and show them a series of articles such as a **pot**, a plate, a flask and a small stool.

French Cluster: 1.1
French 1: chaudron
French 2: marmite
French 3: pot

Spanish Cluster: 1.1
Spanish 1: caldero
Spanish 2: olla
Spanish 3: cazuela

For each sentence, the gold standard that results from the manual annotation contains a set of translations enriched with frequency information (reflecting the number of times a given translation was chosen by the annotators). The format of both the input file and gold standard is similar to the format that was used for the SemEval Cross-Lingual Lexical Substitution task (Sinha, McCarthy and Mihalcea 2009). Table 5.8 lists the six-language gold standard for the test sentence in example (21).

Language	gold standard translations and frequency weights
French	marmite 2;boîte 1;pot 1;seau 1;chaudron 1;
Dutch	pot 3;potje 3;jampotje 1;
Italian	pentola 3;vaso 3;calderone 3;
Spanish	olla 3;cazuela 2;caldero 2;tarro 1;
German	Töpfchen 3;Topf 3;Flakon 1;Eimer 1;

Table 5.8: Gold standard for the ambiguous word *pot* for the sentence in example (21)

5.2.1 Cluster Agreement

Table 5.9 illustrates the agreement on the appropriate sense cluster for all test words for French. The first two columns represent the average number of clusters (fine-grained subclusters) and top level clusters per sentence, and the annotator consensus scores can be gleaned from the last two columns. The agreement scores simply represent the percentage of sentences (out of 50 for the test data) for which all annotators agree on the fine-grained cluster (column 4) or on the top cluster (column 5). The results show that for some words, there is fairly little consensus when also incorporating the subclusters, but they also reveal a clear top cluster consensus, which is directly reflected in the percentage of sentences on which all annotators agree.

Two words seem to cause more disagreement with respect to the choice of the correct sense cluster: *occupation* and *movement*. The latter words are more abstract words, where the boundaries are less clear cut between the different senses. If we take *mission* as an example, we distinguished seven different senses of the word in the clustering file:

(a) organized (military/humanitarian) operation (E.g. UN mission)

(b) special assignment given to one person or a group of people (E.g. election observation mission)

(c) group of representatives or delegates (E.g. trade mission)

(d) project, effort, attempt, goal, objective

(e) statement

(f) sense of mission

(g) religious mission

In the gold standard annotation files, we noticed a lot of hesitation between the first three meanings of the word: for 5 sentences there is hesitation between sense (a) and (b), for 6 sentences between sense (b) and (c), for 4

Word	**Avg # cl**	**Avg # top cl**	**% cl cons**	**% top cl cons**	**Word**	**Avg # cl**	**Avg # top cl**	**% cl cons**	**% top cl cons**
bank	1.15	1.00	85	100	mood	1.64	1.26	42	74
movement	1.70	1.30	40	70	paper	1.12	1.02	88	98
occupation	1.65	1.50	40	55	post	1.18	1.08	82	92
passage	1.20	1.10	80	90	pot	1.24	1.08	80	94
plant	1.45	1.10	60	90	range	1.10	1.10	90	90
coach	1.10	1.10	90	90	rest	1.18	1.02	84	98
education	1.16	1.10	84	90	ring	1.22	1.16	80	86
execution	1.58	1.22	48	78	scene	1.26	1.14	78	86
figure	1.30	1.04	70	96	side	1.24	1.16	76	84
job	1.22	1.20	80	82	soil	1.24	1.06	84	94
letter	1.14	1.06	86	94	strain	1.16	1.08	84	92
match	1.20	1.08	80	92	test	1.20	1.18	82	84
mission	1.46	1.46	58	58					

Table 5.9: Overview of the annotator consensus for French for all 25 ambiguous words

sentences between (a) and (b) and for 2 sentences between the senses (a), (b) and (c). Although there was enough translational evidence to distinguish between the first three senses of the word, these sense distinctions seemed to be less intuitive for the annotators.

CHAPTER 6

Experimental Setup

This chapter describes the experimental set-up that was used to evaluate the cross-lingual WSD approach. We built five classifiers per ambiguous target word with English as an input language and translations in the five supported languages as classification output. To evaluate the classifiers, we used the sense-tagged corpus and test set we described in Chapter 5. For the creation and testing of the ParaSense system, we used the set of 20 ambiguous target words and disregarded the trial words[1].

The chapter starts with a brief description of the train and test datasets in Section 6.1. Section 6.2 introduces the two evaluation metrics that were used; the BEST precision and recall metric, and a straightforward accuracy measure. Section 6.3 lists the most frequent translation baselines per language for all words in the benchmark data set.

6.1 Training and Test Data

This section briefly reviews the size and contents of the training and test sets that were used for all experiments.

[1] As stated in Chapter 5, the set of trial words was exclusively used within the framework of the SemEval competition to give participating teams an idea of how the real test data would look like that was released later on.

Training Data

In order to train five classifiers (one per target language), one training corpus for each of the 20 ambiguous focus words was constructed, comprising all English Europarl sentences, each time containing the focus word and the aligned translations in the five target languages. Table 6.1 presents an overview of the number of training instances and the number of classification labels per language for each ambiguous focus word.

Table 6.1 reveals a large variation between both the number of training instances, varying between 63 for *pot* and 7531 for *job*, and for the level of ambiguity that is reflected by the number of classification labels per language. Considerable inter-language differences can also be noticed with respect to the number of classification labels. Spanish contains in average 61.5 translation labels per focus word, whereas in German and Dutch, the sense inventory contains respectively 160.4 and 134.35 different translations (averaged over all ambiguous focus words). The higher figures for Dutch and German can be explained by the large number of compounds in both Germanic languages.

The average number of training instances per classification label for all test words are listed in Table 6.2. These have been enriched with the overall mean of training instances per translation and standard deviation information for all five target languages. As expected, Dutch and German have in average less training instances per classification label, due to their larger sense inventories per ambiguous focus word. Furthermore, the large standard deviation figures highlight major differences in the average number of training instances per classification label. Most of the ambiguous words can be linked to at least 10 training examples per class on average. This is not the case, however, for all test words under consideration. Words like *mood, pot, ring* and *strain* only occur twice or three times per class label on average.

Test Data

For the construction of the test dataset, we created a lexical sample for each of the 20 test words. The lexical sample contains 50 English sentences (per word), which were extracted from the BNC. All instances were manually annotated per language, which resulted in a set of gold standard translation labels per instance. For a detailed description of the test dataset, we refer to Chapter 5.

	Training Instances	French Trans	Italian Trans	Spanish Trans	Dutch Trans	German Trans
coach	66	11	21	7	28	29
education	4380	55	186	62	263	356
execution	489	26	60	34	85	84
figure	2298	167	118	104	303	302
job	7531	184	135	121	362	457
letter	1822	75	83	59	82	125
match	109	21	25	18	33	30
mission	1390	46	58	36	226	237
mood	100	26	38	37	35	30
paper	3650	94	99	63	166	212
post	998	68	94	70	131	164
pot	63	27	21	29	27	29
range	1428	145	185	125	149	246
rest	1739	80	135	62	63	107
ring	143	46	42	38	42	45
scene	284	50	45	44	94	75
side	3533	261	258	199	211	232
soil	287	16	26	12	58	64
strain	134	40	55	53	73	64
test	1368	92	92	57	256	320
mean	1590.6	76.5	88.8	61.5	134.35	160.4
standard deviation	1919.96	64.05	64.66	45.71	102.63	127.82

Table 6.1: Number of training instances and classification labels for all twenty test words in all five target languages (viz. French, Italian, Spanish, Dutch and German), enriched with mean and standard deviation information.

	French	Italian	Spanish	Dutch	German
coach	5	3.1	9.4	2.4	2.3
education	79.6	23.5	70.6	16.7	12.3
execution	18.8	8.1	14.4	5.8	5.8
figure	13.8	19.5	22.1	7.6	7.6
job	40.9	55.8	62.2	20.8	16.5
letter	24.3	22	30.9	22.2	14.6
match	5.2	4.4	6.1	3.3	3.6
mission	30.2	24	38.6	6.2	5.9
mood	3.8	2.6	2.7	2.9	3.3
paper	38.8	36.9	57.9	22	17.2
post	14.7	10.6	14.3	7.6	6.1
pot	2.3	3	2.2	2.3	2.2
range	9.8	7.7	11.4	9.6	5.8
rest	21.7	12.9	28	27.6	16.3
ring	3.1	3.4	3.8	3.4	3.2
scene	5.7	6.3	6.5	3	3.8
side	13.5	13.7	17.8	16.7	15.2
soil	17.9	11	23.9	4.9	4.5
strain	3.4	2.4	2.5	2.3	2.1
test	14.9	14.9	24	5.3	4.3
mean	18.37	14.29	22.47	9.86	7.63
standard deviation	18.42	13.42	20.57	8.15	5.45

Table 6.2: Average number of training instances per classification label for all twenty test words, mean and standard deviation information per target language.

6.2 Evaluation Metrics

As evaluation metrics, we used both the BEST precision and recall metric, as well as a straightforward accuracy measure.

6.2.1 The BEST Precision and Recall Metric

The BEST precision and recall metric was introduced by McCarthy and Navigli (2007) in the framework of the SemEval-2007 competition. The metric takes into account the frequency weights of the gold standard translations: translations that were picked by different annotators received a higher associated frequency which is incorporated in the formulas for calculating precision and recall. For the BEST precision and recall evaluation, the system can propose as many guesses as the system believes are correct, but the resulting score is divided by the number of guesses. In this way, systems that output many guesses are not favored and systems can maximize their score by guessing the most frequent translation from the annotators.

Precision and recall were originally used in the information retrieval domain: precision refers to the fraction of retrieved documents that are relevant, while recall refers to the fraction of relevant documents found by the search engine. In our case, precision refers to the number of correct translations in relation to the total number of translations generated by the system, while recall refers to the number of correct translations generated by the classifier. As our ParaSense system predicts a translation label for all sentences in the test set, precision and recall will give identical results.

The following variables are used for the BEST precision and recall formulas. Let H be the set of annotators, T the set of test words and h_i the set of translations for an item $i \in T$ for annotator $h \in H$. Let A be the set of words from T where the system provides at least one answer and a_i the set of guesses from the system for word $i \in A$. For each i, we calculate the multiset union (H_i) for all h_i for all $h \in H$ and for each unique type (res) in H_i that has an associated frequency ($freq_{res}$). Equation 6.1 lists the BEST precision formula, whereas Equation 6.2 lists the formula for calculating the BEST recall score:

$$Precision = \frac{\sum_{a_i:i\in A} \frac{\frac{\sum_{res\in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|A|} \tag{6.1}$$

$$Recall = \frac{\sum_{a_i:i\in T} \frac{\frac{\sum_{res\in a_i} freq_{res}}{|a_i|}}{|H_i|}}{|T|} \tag{6.2}$$

The BEST precision and recall metrics allow to evaluate systems that generate multiple translations, which was permitted in the framework of the SemEval competition. In our experiments with the ParaSense system, however, we chose for a more strict approach where exactly one translation label is generated by the system. Example (22) lists the Dutch gold standard and fictitious system output for a test instance containing the word *coach*:

(22) Gold standard: coach.n.nl :: bus 3; autobus 3; touring car 2; toerbus 1;
System output: coach.n :: bus; toerbus;

If we take into account the associated frequency weights in the gold standard, the system's credit in the numerator of the BEST precision and recall formula would be:

$$\frac{\frac{3+1}{2}}{9} = 0.2222$$

6.2.2 The Accuracy Metric

The second metric we use is a straightforward accuracy measure that divides the number of correct answers by the total amount of test instances. In this case, we do not take into account the frequency weights that are associated with the gold standard translations - every gold standard translation has an associated weight of "1". The resulting credit is still divided by the number of system outputs though, which leads to the following formula for calculating accuracy:

$$Accuracy = \frac{\sum_{a_i:i\in T} \frac{\sum_{res\in a_i}}{|a_i|}}{|T|} \tag{6.3}$$

If we reconsider example 22, we obtain the following system credit in the numerator of the accuracy formula:

$$\frac{1+1}{2} = 1$$

It is clear from the example that the accuracy metric is a more relaxed metric than the BEST precision and recall metric. As the gold standard frequencies are not taken into account, systems are not rewarded for picking the "best possible translation" for a novel instance; all gold standard translations are considered equally important. In addition, the resulting credit is not divided by the number of responses from annotators.

6.3 Most Frequent Translation Baselines

As a first baseline, we selected the most frequent lemmatized translation that resulted from the automated word alignment (GIZA++) for all ambiguous nouns in the training data. This baseline is inspired by the most frequent sense baseline often used in WSD evaluations. As already mentioned in Section 2.1, Zipf (1949) has shown that one meaning of the word is often very frequent in language, while the other meanings show a significant decrease in frequency. The main difference between the most frequent sense baseline and our baseline is that the latter is corpus-dependent: we do not take into account the overall frequency of a word as it would be measured based on a large general purpose corpus, but calculate the most frequent sense (or translation in this case) based on our training corpus.

Table 6.3 presents an overview of the BEST precision scores, whereas Table 6.4 lists the accuracy baselines per language. An important remark concerning the BEST precision and recall metrics is that both of these metrics result in identical scores in case the system outputs a translation suggestion for all test instances. As this is the case for the five baselines, where all test items are assigned the most frequent translation in the target language, we only list the BEST precision scores, knowing that the BEST recall scores are identical in this case.

Although the two metrics do not show identical results, there are similar trends to be noticed. In general, French and Spanish achieve better results, whereas Dutch and German yield lower scores. For Italian, the results show a more nuanced picture. On the one hand, Italian results are similar to the other two romance languages, but on the other hand, a couple of worse performing words (*post, pot, side*) have a negative impact on the overall average score. A second observation is that some words seem to be "easier" to tackle in all considered languages (*education, mission, soil*), whereas other words such as *coach, match* and *paper* appear to be more problematic in all supported languages. The consistently good scores for words such as *mission* can sometimes be explained by the existence of a more general translation which can be used for various senses of the words. For the word *mission*, this is *mission* in French, *missión* in Spanish,

ish, *missione* in Italian, *missie* in Dutch and *Mission* in German. Some of the badly performing words suffer from frequent fuzzy translations and their resulting erroneous word alignments (E.g. In Dutch, the word *match* is most frequently aligned with the preposition *met*), whereas other words are assigned the most frequent translation which is heavily biased by the domain of the training corpus. To illustrate this, we refer to the Dutch and German baseline translations for the word *paper*, i.e. *witboek* and *Weissbuch* respectively. These are translations of the English compound *white paper* which occurs very frequently in the Europarl proceedings.

BEST Precision Baseline scores per test word

	French	Italian	Spanish	Dutch	German
coach	9.54	7.36	18.59	7.81	12.62
education	31.94	21.56	32.85	15.04	19.87
execution	39.63	27.58	37.94	16.74	9.17
figure	17.60	11.57	19.40	13.36	11.25
job	19.53	17.94	20.12	9.89	6.67
letter	33.43	34.64	17.05	15.06	5.58
match	14.38	10.16	13.52	0.00	0.00
mission	40.18	30.45	41.19	22.31	25.13
mood	17.11	9.18	9.07	22.29	32.22
paper	2.24	2.59	4.67	4.00	1.52
post	23.34	8.06	18.58	14.37	12.51
pot	28.30	0.00	6.00	33.60	4.22
range	4.73	5.11	4.73	4.89	3.11
rest	19.92	19.73	18.71	23.70	9.39
ring	17.84	14.91	13.04	26.82	25.52
scene	21.47	22.82	26.23	9.36	0.51
side	8.22	0.00	10.21	20.55	27.38
soil	32.91	22.70	34.10	25.84	27.36
strain	11.11	10.73	12.48	13.46	14.83
test	31.21	26.25	34.55	15.89	14.31

BEST Precision Average Baseline score per language

Average	20.71	15.17	19.65	15.75	13.16

Table 6.3: BEST precision baseline scores for all five languages.

Accuracy Baseline scores per test word					
	French	Italian	Spanish	Dutch	German
coach	38	34	42	38	26
education	100	86	100	78	94
execution	96	86	94	60	36
figure	54	46	56	50	48
job	70	82	86	52	48
letter	96	76	56	50	42
match	40	36	42	0	0
mission	100	94	100	96	98
mood	74	40	46	94	100
paper	8	12	10	4	4
post	78	40	58	70	54
pot	52	0	6	76	20
range	34	30	28	22	20
rest	66	62	56	58	48
ring	58	36	52	60	80
scene	72	80	80	46	4
side	50	0	58	88	92
soil	98	88	96	100	98
strain	40	40	36	52	50
test	92	84	100	94	84
Average Accuracy Baseline score per language					
Average	65.8	52.6	60.1	59.4	52.3

Table 6.4: Accuracy baseline scores for all five languages.

CHAPTER 7

Experimental Results

This chapter presents an overview of all experiments that were conducted using the ParaSense Cross-lingual WSD system. With these experiments, we aimed to examine the validity of our multilingual classification-approach to WSD and answer the research questions that were formulated in the introduction of this dissertation: (1) does it help to incorporate translational evidence in the feature vector to obtain more accurate predictions of translation labels and (2) to what extent do the classification results improve by adding evidence from multiple languages to the feature vectors?

Section 7.1 lists the classification results when both machine learning algorithms are applied with their default settings to the combined local context and binary translation feature sets. In a first optimization cycle, we measured the performance differences when applying latent semantic analysis to the bag-of-words translation features. The resulting scores are presented in Section 7.2. A second step concerns optimizing the algorithm parameter settings. Section 7.3.1 discusses the optimization experiments that were performed by means of a genetic algorithm in order to tailor the TIMBL parameter settings to the CLWSD task. A global overview of the experimental results for the default systems and the systems exploiting optimized features and algorithm settings per language is presented in Section 7.3.2.

Furthermore, a set of additional experiments were conducted in order to better understand the functioning of the ParaSense system. Section 7.4 discusses the contribution of the different translation features to the overall classification result. The results confirm our hypothesis that adding translational evidence helps the classifier to correctly disambiguate our set of target nouns. The multilingual classifier clearly outperforms the classifier which only uses local context information. In addition, the classifier that incorporates all four languages achieves very good classification scores. To summarize, the multilingual approach appeared to constantly achieve good classification results in four of the five target languages, viz. French, Spanish, Dutch and German. For Italian, however, our approach did not yield the expected results, i.e. adding translational evidence did not seem to improve the classification results. Section 7.5 investigates the classification scores for the individual test words per language, while Section 7.6 reports on the performance differences when using manually verified translation labels instead of fully automatically generated translation labels. We conclude this chapter with a comparison of the ParaSense system with all systems that participated in the SemEval Cross-lingual Word Sense Disambiguation task.

7.1 Classification Results using Local Context and Binary Translation Features

In a first set of experiments, we investigated the viability of our cross-lingual WSD approach by comparing both classifiers in their default set-up to the most frequent translation baseline. The system classification results are listed in two tables. Table 7.1 gives an overview of the BEST precision scores, whereas Table 7.2 shows the more straightforward accuracy figures. Both tables list the scores averaged over all twenty test words for the most frequent translation baseline and the ParaSense system that combines the English local context features and the binary bag-of-words translation features. Results are listed for both machine learning algorithms, TIMBL and SVMLight.

For the sake of completeness, we also listed the individual accuracy scores per test word for both classifiers: Table 7.3 lists the TIMBL accuracy scores when using the English local context features and binary translation features, whereas Table 7.4 presents the accuracy results when using SVMLight on the same feature sets.

	French	Italian	Spanish	Dutch	German
Most Frequent Translation Baseline					
Baseline	0.207	0.152	0.197	**0.158**	0.132
Classification results for TIMBL					
Local context features + binary translation features	**0.222**	**0.174**	**0.210**	0.150	**0.140**
Classification results for SVMLIGHT					
Local context features + binary translation features	0.162	0.121	0.172	0.097	0.085

Table 7.1: BEST precision scores averaged over all twenty test words for both machine learning algorithms applied with their default settings.

	French	Italian	Spanish	Dutch	German
Most Frequent Translation Baseline					
Baseline	0.658	0.526	0.601	**0.594**	0.523
Classification results for TIMBL					
Local context features + binary translation features	**0.691**	**0.593**	**0.621**	0.560	**0.524**
Classification results for SVMLIGHT					
Local context features + binary translation features	0.521	0.412	0.518	0.368	0.358

Table 7.2: Accuracy scores averaged over all twenty test words for both machine learning algorithms applied with their default settings.

The classification results show that only the TIMBL classifier succeeds in beating the most frequent translation baseline, with the exception of Dutch. Beating the baseline seems a fair challenge for Dutch, though, as the most frequent translation baseline for Dutch already performs rather well compared to the other Germanic language (viz. German).

The detailed results per focus word show a large variety in classification performance both within the set of focus words in a particular language as well as between the different target languages. We notice, for instance, performance differences of 16% between French and German for both classifiers. In general, French and Spanish obtain the best classification scores, followed by Italian that yields more moderate results for both machine learning algorithms. Both classifiers have most problems to predict correct translations for Dutch and German. As illustrated by figure 5.1, these two languages have a lot of compound translations in their sense inventory, which drastically increases the number of translation labels the classifier has to choose from.

With respect to the classification scores for the individual test words, very large differences can be noticed for all languages. For French, for instance, we obtain accuracy scores varying between 0.30 (*range*) and 0.98 (*soil*) for timbl and between 0.06 (*scene*) and 1.00 (*education*) for SVMLight. We will investigate these differences in more detail below.

	French	Italian	Spanish	Dutch	German
coach	0.38	0.36	0.42	0.08	0.28
education	0.90	0.86	0.94	0.74	0.52
execution	0.96	0.86	0.94	0.42	0.38
figure	0.64	0.40	0.62	0.50	0.42
job	0.66	0.70	0.74	0.42	0.58
letter	0.92	0.74	0.64	0.64	0.56
match	0.40	0.36	0.42	0.08	0.14
mission	0.96	0.90	0.94	0.94	0.90
mood	0.64	0.40	0.14	0.84	1.00
paper	0.92	0.82	0.94	0.78	0.66
post	0.76	0.56	0.56	0.68	0.54
pot	0.46	0.36	0.18	0.56	0.72
range	0.30	0.34	0.22	0.22	0.26
rest	0.88	0.78	0.76	0.64	0.66
ring	0.34	0.08	0.20	0.04	0.20
scene	0.72	0.76	0.78	0.48	0.36
side	0.38	0.44	0.66	0.84	0.94
soil	0.98	0.88	0.98	0.98	0.98
strain	0.70	0.54	0.40	0.52	0.06
test	0.92	0.72	0.94	0.80	0.32
Average	0.691	0.593	0.621	0.56	0.524

Table 7.3: Timbl accuracy results for the ParaSense flavor containing binary translation features.

	French	Italian	Spanish	Dutch	German
coach	0.36	0.32	0.42	0.28	0.40
education	1.00	0.86	1.00	0.78	0.14
execution	0.96	0.86	0.94	0.32	0.72
figure	0.54	0.28	0.56	0.46	0.48
job	0.48	0.22	0.72	0.24	0.34
letter	0.96	0.76	0.58	0.54	0.60
match	0.40	0.34	0.42	0.66	0.30
mission	0.98	0.82	0.96	0.10	0.08
mood	0.48	0.18	0.20	0.52	0.48
paper	0.16	0.36	0.18	0.12	0.18
post	0.42	0.32	0.44	0.08	0.64
pot	0.18	0.18	0.22	0.30	0.24
range	0.34	0.08	0.16	0.22	0.02
rest	0.66	0.60	0.48	0.56	0.48
ring	0.14	0.00	0.40	0.00	0.06
scene	0.06	0.38	0.06	0.08	0.20
side	0.26	0.12	0.10	0.24	0.74
soil	0.98	0.88	0.96	1.00	0.98
strain	0.52	0.22	0.56	0.36	0.04
test	0.54	0.46	1.00	0.50	0.04
Average	0.521	0.412	0.518	0.368	0.358

Table 7.4: SVMLight accuracy results for the ParaSense flavor containing binary translation features.

7.2 Optimization of the Feature Space

As already mentioned in Chapter 3, the set of binary bag-of-words translation features results in very sparse feature vectors, as only a limited set of translations is aligned with a particular training instance. In addition, overlap between these translation features is based on exact lexical match, meaning that synonyms are not considered as overlapping features. In order to tackle these problems, we decided to apply latent semantic analysis (LSA) to the bag-of-words translation features. As explained in Section 3.2.3, LSA uses singular value decomposition, a dimensionality reduction technique that is capable of discovering correlations between the different features. For the creation of our latent semantic translation features, we selected the 50 best SVD dimensions, as was done by Lopez de Lacalle (2009).

Table 7.5 gives an overview of the BEST precision scores, whereas Table 7.6 shows the more straightforward accuracy figures for both classifiers, TIMBL and SVMLIGHT. Both tables list the scores averaged over all twenty test words for the most frequent translation baseline and two flavors of the ParaSense system: one flavor that combines the English local context features and the binary bag-of-words translation features, and another flavor that combines the local context features and the latent semantic translation features. In addition, we also present detailed TIMBL classification results for all individual test words in Table 7.7, and SVMLIGHT accuracy scores in Table 7.8.

All classification results show that for both classifiers, the ParaSense system that combines the English local context features with a set of binary translation features outperforms the system that incorporates latent semantic translation features in the feature vector. We will further elaborate on this observation in Section 7.3.2.
Apparently, the classifier does not benefit from performing a dimensionality reduction on the sparse feature vectors. If we consider French, the approach does work for some words with a very small training base, i.e. performance improvements of 8% for *coach* and *match* and up to 20% for *side*. On the other hand, we observe performance drops of 12% (*pot*) and 44% (*strain*) for other words trained on an equally small feature base. It might be the case that dimensionality reduction does not work well for the CLWSD task because certain word meanings are only represented by very few training examples. It is to be expected that the distinctive bag-of-words features that are contained in these training examples will probably not end up in the most important dimensions, and might get filtered out in the end. As already mentioned in this thesis, exceptions can be important for NLP tasks, and removing outliers from the training

data might be harmful for the classification results (Daelemans, van den Bosch and Zavrel 1999).

In the next section, we will investigate the impact of parameter optimization on both flavors of the ParaSense system, one flavor incorporating the binary translation features and the other flavor using the latent semantic translation features.

	French	Italian	Spanish	Dutch	German
Most Frequent Translation Baseline					
Baseline	0. 207	0.152	0.197	**0.158**	0.132
Classification results for timbl					
Local context features + binary translation features	**0.222**	**0.174**	**0.210**	0.150	**0.140**
Local context features + latent semantic translation features	0.192	0.161	0.205	0.130	0.121
Classification results for SVMlight					
Local context features + binary translation features	0.162	0.121	0.172	0.097	0.085
Local context features + latent semantic translation features	0.160	0.110	0.162	0.088	0.080

Table 7.5: best precision scores averaged over all twenty test words for both machine learning algorithms applied with their default settings.

	French	Italian	Spanish	Dutch	German
Most Frequent Translation Baseline					
Baseline	0.658	0.526	0.601	**0.594**	0.523
Classification results for timbl					
Local context features + binary translation features	**0.691**	**0.593**	**0.621**	0.560	**0.524**
Local context features + latent semantic translation features	0.588	0.532	0.603	0.455	0.449
Classification results for SVMlight					
Local context features + binary translation features	0.521	0.412	0.518	0.368	0.358
Local context features + latent semantic translation features	0.510	0.387	0.501	0.354	0.347

Table 7.6: Accuracy scores averaged over all twenty test words for both machine learning algorithms applied with their default settings.

	French	Italian	Spanish	Dutch	German
coach	0.46	0.44	0.54	0.30	0.22
education	0.78	0.78	0.88	0.64	0.44
execution	0.80	0.62	0.72	0.60	0.54
figure	0.46	0.44	0.66	0.34	0.36
job	0.62	0.46	0.62	0.40	0.48
letter	0.82	0.70	0.62	0.56	0.58
match	0.48	0.50	0.46	0.46	0.46
mission	0.82	0.78	0.86	0.52	0.64
mood	0.52	0.28	0.28	0.40	0.40
paper	0.66	0.62	0.72	0.56	0.56
post	0.70	0.38	0.54	0.38	0.54
pot	0.34	0.30	0.44	0.44	0.38
range	0.28	0.44	0.20	0.28	0.16
rest	0.80	0.66	0.74	0.64	0.58
ring	0.30	0.24	0.40	0.16	0.42
scene	0.46	0.56	0.64	0.50	0.38
side	0.58	0.50	0.68	0.64	0.64
soil	0.88	0.86	0.88	0.78	0.74
strain	0.26	0.40	0.46	0.26	0.14
test	0.74	0.68	0.72	0.24	0.32
Average	0.588	0.532	0.603	0.455	0.449

Table 7.7: Timbl accuracy results per test word for the ParaSense flavor containing latent semantic translation features.

	French	Italian	Spanish	Dutch	German
coach	0.38	0.32	0.42	0.30	0.42
education	1.00	0.86	1.00	0.78	0.18
execution	0.96	0.86	0.94	0.32	0.74
figure	0.54	0.28	0.56	0.32	0.48
job	0.24	0.22	0.72	0.22	0.32
letter	0.96	0.76	0.56	0.50	0.60
match	0.34	0.36	0.30	0.52	0.16
mission	1.00	0.78	0.54	0.14	0.08
mood	0.48	0.02	0.28	0.52	0.42
paper	0.16	0.34	0.18	0.16	0.26
post	0.54	0.20	0.44	0.04	0.56
pot	0.04	0.10	0.24	0.06	0.22
range	0.34	0.06	0.16	0.22	0.02
rest	0.64	0.62	0.52	0.56	0.48
ring	0.26	0.04	0.44	0.00	0.06
scene	0.06	0.24	0.06	0.26	0.18
side	0.22	0.06	0.14	0.54	0.70
soil	0.98	0.88	0.96	1.00	0.98
strain	0.52	0.34	0.56	0.20	0.04
test	0.54	0.40	1.00	0.42	0.04

Table 7.8: SVMlight accuracy results per test word for the ParaSense flavor containing latent semantic translation features.

7.3 Parameter Optimization

This section presents the classification results that are obtained for both machine learning algorithms when optimizing the parameter settings. For the timbl classifier, we applied a genetic algorithm to find the optimal parameter values, whereas for SVMlight we experimented with parameter settings that have shown to work well for word sense disambiguation (Guo, Che, Hu, Zhang and Liu 2007).

7.3.1 GA experiments

As mentioned in Section 4.1.1, classifiers are initialized with different parameter settings that can be optimized to improve the classification scores. In order to optimize these algorithm parameters, we used the genetic algorithm as described in Section 4.2 with its default settings. For a detailed overview of the timbl parameter settings, we refer to the timbl User Manual (Daelemans et al. 2002). Since we intended to gain some insights into the optimal parameter settings for all test words in all five languages, GA optimization was performed on the training data containing the *latent semantic translation* feature set. The latent semantic feature vectors are of reasonable length (228 features in total), whereas the binary feature vectors contain tens of thousands of features. Running the GA with the latter training data set would turn the optimization experiments into a very computationally demanding and time-consuming task.

The following timbl parameter settings were tuned by performing 10-fold cross validation on the training data:

- **Algorithm** (parameter -a)
  - 0: the IB1 algorithm, which is the default kNN algorithm. This algorithm usually leads to more accuracy at the expense of efficiency.
  - 1: the IGTREE algorithm, that is a fast heuristic decision-tree-based approximation of IB1.
- **Feature weighting** (parameter -w)
  - 0: there is no feature weighting; all features are accorded equal importance.
  - 2: *Information Gain* weighting looks at each feature in isolation and measures its contribution to the knowledge of the correct class label.
  - 1: *Gain ratio* feature weighting. Gain ratio (Quinlan 1993) is a normalized version of the Information Gain weighting (Information Gain normalized by the entropy of the feature values).

  - 3: *Chi-squared* weighting (White and Liu 1994): feature selection measure based on the chi-squared statistics, that is not affected by the Gain Ratio bias towards features with more values.
  - 4: *Shared Variance* weighting is a chi-square-based measure that corrects for the degrees of freedom.

- **Number of nearest neighbors** used for extrapolation (parameter -k): we varied the number of k between 1 and 11.
- **Type of class voting weights** that are used to extrapolate from the nearest neighbors (parameter -d):
  - Z: *majority voting*. All neighbors have equal weights in the voting process (default setting).
  - IL: *Inverse Linear* weighting (Dudani 1976) estimates a neighbor with smaller distance more heavily than one with a greater distance: the nearest neighbor gets a weight of 1, the furthest neighbor a weight of 0 and the other weights are scaled linearly to the [-1,1] interval.
  - ID: *Inverse distance* weighting proposes a small variation of the Inverse Linear weighting, where a small constant is used in the weighting formula to avoid division by zero (Wettschereck, Aha and Mohri 1997).

Figure 7.1 illustrates the variance in the evolving fitness scores for all test words in Dutch. The box-and-whisker plots show the minimum, first quartile, median, third quartile, the maximum and the outliers[1] for the GA fitness scores in Dutch. A box plot is displayed for each test word. The first and third quartile are displayed as the bottom and top of the box, the median as a horizontal stroke through the box. In general, we notice a large variance in fitness scores for the different parameter settings combinations. The figure clearly illustrates that most words need to evolve for a long time in order to reach the median score. The third quartile, which shows less variation in the fitness scores, reflects the additional evolutionary stages it takes to produce the best individuals. In addition, a large variance in the maximum fitness scores can be noticed across the different test words. This variance will be further discussed in Section 7.5.

Table 7.9 lists the selected TIMBL parameters per word per language. With respect to the selected parameters, general tendencies can be noticed across words and languages. *Gain ratio* and the *shared variance weighting* are the optimal feature weighting techniques. Furthermore, with respect to the type of class voting weights, we can observe that the default *majority voting*, and to a minor extent the *inverse distance weighting*, give the

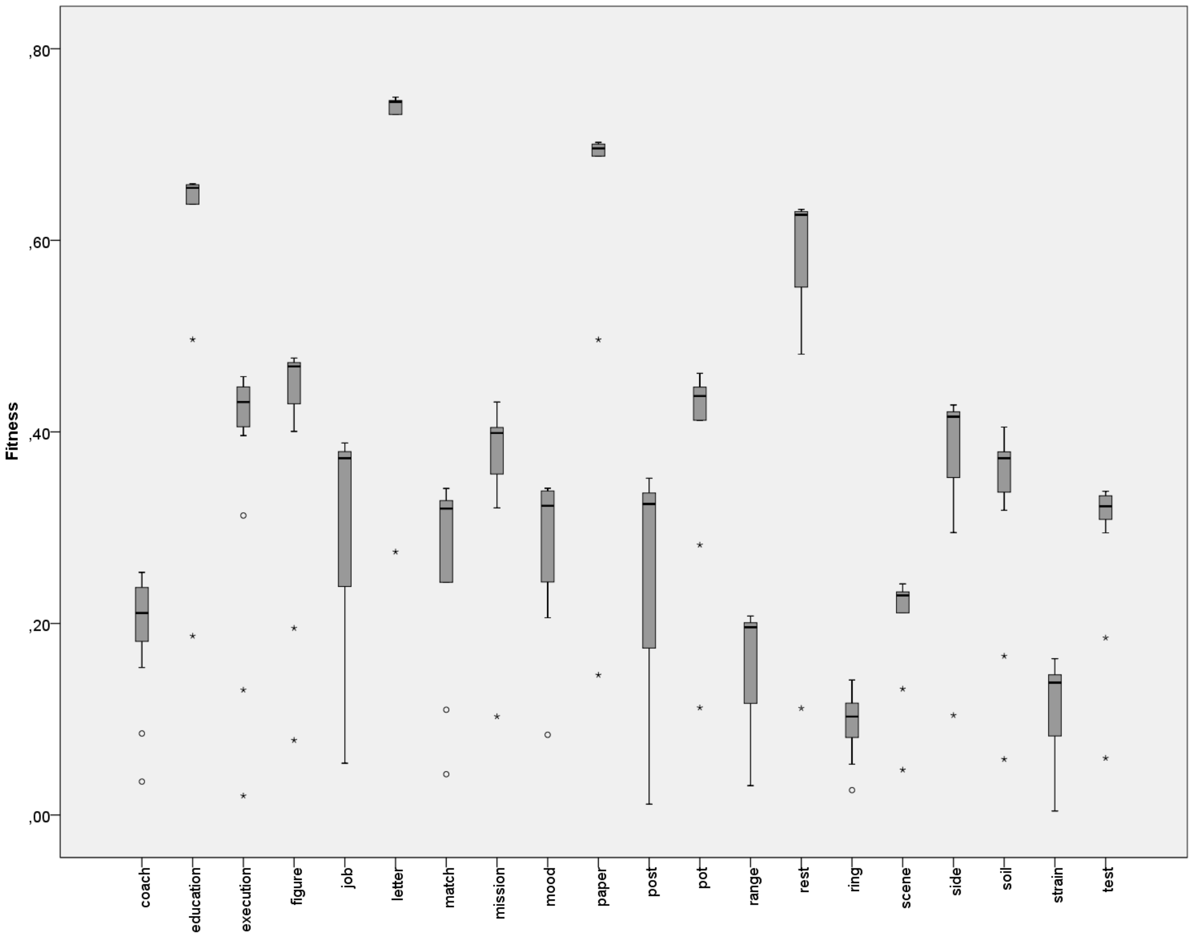
[1] The outliers are the scores which lie more than 3 times outside the interquartile range.

best classification results. Considering the different selected values of k, the number of nearest distances taken into account, we observe a general use of high k values, in average ranging between 8 and 11. Finally, the *IB1 algorithm* is shown to work best for all languages, except for some words in German.

The resulting optimized TIMBL parameters were used for all experiments using the latent semantic feature data. For the data containing the binary features, we took the most frequently used parameter settings per language. As the latter data also contain binary values instead of numeric values for the translation bag-of-words features, we replaced the dedicated numeric distance metric by the *Jeffrey divergence metric*. Jeffrey divergence is a statistical dissimilarity metric that can be used to compute the distance between class distributions of two values of the same feature and is said to work well for very sparse feature vectors (Daelemans et al. 2009). In addition, it has proven to perform well in preliminary CLWSD experiments that were performed on the trial data set (Lefever and Hoste 2011).

We fully acknowledge that more optimization could have been performed on both the feature space and the parameter settings, including research such as performing joint feature selection and parameter optimization by means of the GA (Daelemans, Hoste, De Meulder and Naudts 2003) or applying the GA to the binary bag-of-words feature set, experimenting with different SVM kernels and accompanying parameter settings, etc. However, advanced and far-reaching research on optimization falls outside the scope of this dissertation.

Figure 7.1: Box-and-whisker plots representing the GA fitness scores for all test words in Dutch.



	French				Italian				Spanish				Dutch				German			
	w	a	d	k	w	a	d	k	w	a	d	k	w	a	d	k	w	a	d	k
coach	1	0	ID	8	4	0	ID	8	2	0	IL	7	1	0	ID	8	0	0	ID	7
education	0	0	IL	6	4	0	Z	8	1	0	Z	10	3	0	Z	10	1	0	Z	11
execution	3	0	Z	5	3	0	ID	9	2	0	Z	7	1	0	Z	8	0	0	IL	9
figure	0	0	ID	5	0	0	IL	8	1	0	ID	8	1	0	ID	11	2	0	Z	11
job	1	0	ID	11	1	0	Z	7	4	0	Z	9	0	0	IL	8	4	0	ID	11
letter	4	0	ID	11	1	0	ID	5	4	0	Z	10	4	0	Z	10	4	0	Z	11
match	0	0	IL	7	0	0	ID	7	3	0	IL	10	2	0	Z	11	4	0	ID	10
mission	1	0	ID	7	1	0	Z	5	1	0	ID	9	4	0	ID	9	1	0	Z	11
mood	2	0	Z	6	2	0	ID	9	4	0	Z	10	1	0	Z	10	0	0	Z	8
paper	1	0	ID	5	1	0	IL	9	1	0	IL	9	1	0	ID	10	1	0	Z	9
post	4	0	Z	11	1	0	Z	11	4	0	ID	10	4	0	Z	11	4	0	Z	10
pot	0	0	IL	9	2	0	ID	4	4	0	Z	6	3	0	Z	11	4	0	Z	9
range	1	0	ID	11	0	0	IL	8	4	0	Z	8	1	0	Z	8	4	0	ID	8
rest	1	0	Z	4	1	0	IL	8	1	0	Z	4	1	0	ID	4	1	0	Z	9
ring	1	0	Z	11	3	0	Z	8	4	0	Z	8	0	0	Z	9	1	1	Z	1
scene	4	0	ID	9	2	0	Z	11	3	0	ID	10	4	0	Z	5	4	0	ID	11
side	4	0	Z	6	4	0	Z	3	4	0	ID	6	4	0	ID	10	1	0	IL	11
soil	3	0	ID	10	0	0	Z	6	3	0	Z	9	0	0	ID	7	1	1	Z	1
strain	0	0	IL	8	0	0	Z	8	2	0	Z	9	3	0	Z	8	2	1	Z	1
test	3	0	Z	10	4	0	ID	11	3	0	Z	9	3	0	Z	6	0	0	Z	4

Table 7.9: Selected TIMBL parameter settings for the algorithm (parameter a), feature weighting (parameter w), number of nearest neighbors used for extrapolation (parameter k) and type of class voting weights (parameter d) per language for the training data containing local context and latent semantic features.

7.3.2 Overall classification results

This section presents the overall classification results for the CLWSD benchmark test set that was described in Chapter 5. Table 7.10 gives an overview of the BEST Precision scores, whereas Table 7.11 presents the accuracy figures. For both metrics, the scores are averaged across all twenty test words per language. A number of observations can be made with respect to these overall classification results.

Translation Features The classification results show that the ParaSense system incorporating the latent semantic translation features does not outperform the system using the binary translation features, except for the Dutch translations generated with the SVMLIGHT classifier and the Italian labels that are predicted by TIMBL.

From a qualitative point of view, however, we discovered possible benefits of applying a latent semantic reduction on the bag-of-words translation features. As has been mentioned before, the most frequent translation (MFT) is very predominant in the CLWSD task. In order to have an idea of how much the classifier is biased towards the MFT, we measured the number of sentences in Dutch for which the TIMBL classifier predicted this translation. However, as the MFT sometimes is the contextually best translation for a given test instance, we also measured the number of occurrences of the MFT in the gold standard set. For the ParaSense system containing the latent semantic translation features, on average 53% of the sentences are attributed the MFT, a number that is comparable to the gold standard translations where the annotators labeled 49.70% of the sentences with the MFT. On average, 7.85 different translation labels were predicted by the TIMBL classifier per test word for Dutch. In the ParaSense system containing the binary translation features, however, 71.2% of the test sentences receive the MFT classification label, and only 2.3 different translation labels are predicted per test word. If we take for instance the predicted labels for *soil*, the system containing the latent semantic translation features generates 4 different translation labels: *bodem* (27 sentences), *grond* (7 sentences), *bodemerosie* (2 sentences) and *grondgebied* (14 sentences), whereas the system incorporating the binary bag-of-words features only generates the most frequent translation (*bodem*).

To conclude, the system incorporating the binary translation features outperforms the system using the latent semantic translation features. Further analysis reveals that the ParaSense system with the latent semantic features is less biased towards predicting the most frequent translation label, which results in a more varied set of more precise translations for the ambiguous words under consideration. However, since the MFT is such a strong baseline, the classifier using latent semantic translations generates more varied but also more incorrect translation labels.

Machine Learning Algorithm The scores clearly confirm the importance of the parameter settings for both classifiers. Although most machine learning algorithms use sensible default settings, previous research has underlined the importance of tuning the parameter settings for particular tasks (Daelemans et al. 2003). This conclusion is confirmed by our experiments where the baseline scores for both algorithms are outperformed considerably by the same classifiers when applying optimized parameters. Table 7.11 shows, for instance, performance increases up to 16.4% for TIMBL and up to 7.8% for SVMlight for the German ParaSense system incorporating latent semantic translation features. As we performed the GA experiments on the training data containing these latent semantic translation features, it seems logical that running TIMBL with the optimized parameters results in the largest performance improvement for this flavor of the ParaSense system.

A second observation concerns the choice of the machine learning algorithm for this particular task. As can be noticed in the overall results, the TIMBL classifier clearly outperforms the SVM classifier. If we consider the ParaSense system containing the latent semantic translation features, we observe performance differences of 15% for French, 24.1% for Italian, 16.3% for Spanish, 26.4% for Dutch and 19.8% for German. TIMBL easily beats the most frequent translation baseline for both flavors of the ParaSense system, while SVMLIGHT is not able to outperform the baseline. As Support Vector Machines usually perform very well on highly dimensional sparse data, we assume that the TIMBL classifier outperforms SVMLIGHT for this task because of the additional parameter optimization we applied (parameter tuning by means of the GA and use of the Jeffrey divergence metric that is said to perform very well for sparse feature vectors). As for SVMLIGHT, we used settings that have shown to work well for WSD, but we did not extensively investigate the performance of different possible SVM kernels and accompanying parameter settings for the CLWSD task. This would have led us too far from the core subject of this dissertation, namely investigating the viability of cross-lingual WSD by means of parallel corpora. As there is still a lot of room for improvement through optimization, we cannot draw final conclusions with regard to TIMBL outperforming SVMLIGHT for the sparse WSD feature vectors.

Given the considerable performance differences between the two machine learning approaches, though, and the satisfying scores we obtained with TIMBL, we decided to only use the latter Memory-based Learning implementation for the rest of the CLWSD experiments.

	French	Italian	Spanish	Dutch	German
Most Frequent Translation Baseline					
Baseline	0.207	0.152	0.197	0.158	0.132
Classification results for SVMLIGHT					
Default SVMLIGHT local context + latent semantic translation features	0.160	0.110	0.162	0.088	0.080
Default SVMLIGHT local context + binary translation features	0.162	0.121	0.172	0.097	0.085
Optimized SVMLIGHT local context + latent semantic translation features	0.177	0.117	0.176	0.098	0.101
Optimized SVMLIGHT local context + binary translation features	0.193	0.129	0.185	0.101	0.112
Classification results for TIMBL					
Default TIMBL local context + latent semantic translation features	0.192	0.161	0.205	0.130	0.121
Default TIMBL local context + binary translation features	0.222	0.174	0.210	0.150	0.140
Optimized TIMBL local context + latent semantic translation features	0.229	**0.189**	0.232	0.180	0.159
Optimized TIMBL local context + binary translation features	**0.246**	0.177	**0.235**	**0.188**	**0.171**

Table 7.10: Overall classification results expressed in BEST precision scores averaged over all twenty test words.

	French	Italian	Spanish	Dutch	German
Most Frequent Translation Baseline					
Baseline	0.658	0.526	0.601	0.594	0.523
Classification results for SVMLight					
SVMLight Baseline local context + latent semantic translation features	0.510	0.387	0.501	0.354	0.347
SVMLight Baseline local context + binary translation features	0.521	0.412	0.518	0.368	0.358
Optimized ParaSense local context + latent semantic translation features	0.554	0.393	0.533	0.380	0.415
Optimized ParaSense local context + binary translation features	0.591	0.413	0.552	0.364	0.436
Classification results for timbl					
timbl Baseline local context + latent semantic translation features	0.588	0.532	0.603	0.455	0.449
timbl Baseline local context + binary translation features	0.691	0.593	0.621	0.560	0.524
Optimized ParaSense local context + latent semantic translation features	0.704	**0.634**	0.696	0.644	0.613
Optimized ParaSense local context + binary translation features	**0.752**	0.624	**0.705**	**0.684**	**0.668**

Table 7.11: Overall classification results expressed in accuracy scores averaged over all twenty test words.

7.4 Contribution of the translation features

The main novelty of our ParaSense system lies in the application of a multilingual approach to perform WSD, as opposed to the more classical approach that only uses monolingual local context features. Consequently, we also ran a set of additional experiments to examine the contribution of the different translation features to the WSD performance. The accuracy figures for all five target languages for a varying number of translation features are listed in five tables: Table 7.12 shows the accuracy figures for French, Table 7.13 for Dutch, Table 7.14 for German, Table 7.15 for Spanish and Table 7.16 for the Italian classifier.

The scores clearly confirm the validity of our hypothesis: the classifiers using translational evidence constantly achieve better results than the ones that merely use English local context features for four target languages, viz. French, Spanish, German and Dutch. For Italian, however, adding translation features does not result in better classification scores, and even provokes a performance decrease in most cases. Only the classifier that incorporates the Spanish translation features outperforms the classifier that merely uses English local context features. Manual inspection of the Italian classification output did not reveal a valid explanation for the deviant behavior of the Italian classifier employing translational evidence. More test data will be needed to confirm or contradict the obtained results for Italian.

For French, the other two romance languages seem to contribute most: the classifier that uses Italian, Spanish and Dutch bag-of-words features achieves the best score (75.70%), and all other classifiers incorporating Italian and Spanish obtain good results. The classifier that only uses German and Dutch translations obtains the lowest scores (71.90%). For Spanish and Italian, similar trends can be noticed. For Spanish, the classifiers that use the French and Italian translation features yield good results, whereas the classifier that combines local context features with German and Dutch translation features obtains the worst results. For Italian, the classifier that incorporates the Spanish translation features outperforms all other classifiers.

For Dutch and German, the interpretation of the scores is less straightforward. For Dutch, the Spanish-German combination achieves the best result (69.60%), but using the Spanish and French translation features also results in good classification results. For German, the classifiers that incorporate the Dutch translation features obtain good results; nevertheless the best score is achieved by the classifier using the Spanish translations.

In general, the scores confirm our initial hypothesis: the classifier using all translation features obtains good classification results and even outperforms the average scores taken across the classifiers using less translation features for four out of five target languages (viz. French, Spanish, German and Dutch). The best scores, however, are obtained by classifiers incorporating translations from particular languages. For French, Italian and Spanish, the other romance languages seem to contribute most to the classification result. For Dutch and German, utilizing translations from the other germanic language also leads to good classification scores. In order to draw final conclusions with regard to the contribution of the different languages, it would be very interesting to test whether the hypothesis also holds when the translation feature set is extended to more distant language families.

As it is not possible to know in advance which translation features will contribute most for a particular target language, we are convinced that using all translations results in a more flexible and language-independent approach that has proven to achieve good classification results for the supported languages in the cross-lingual WSD task.

	French
Baseline	0.658
All four translation features	
Italian, Spanish, German, Dutch	0.752
Three translation features	
Italian, Spanish, German	0.751
Spanish, German, Dutch	0.744
Italian, German, Dutch	0.754
Italian, Spanish, Dutch	0.757
Average	0.751
Two translation features	
Spanish, German	0.746
Italian, German	0.751
German, Dutch	0.719
Italian, Spanish	0.753
Spanish, Dutch	0.736
Italian, Dutch	0.749
Average	0.742
One translation feature	
German	0.737
Spanish	0.755
Italian	0.751
Dutch	0.732
Average	0.743
No translation features	
None	0.734

Table 7.12: Accuracy figures for French for a varying number of translation features including the other four languages viz. Italian, Spanish, German and Dutch.

	Dutch
Baseline	0.594
All four translation features	
Italian, Spanish, German, French	0.684
Three translation features	
Italian, Spanish, German	0.678
Spanish, German, French	0.681
Italian, German, French	0.667
Italian, Spanish, French	0.681
Average	0.677
Two translation features	
Spanish, German	0.696
Italian, German	0.674
German, French	0.677
Italian, Spanish	0.679
Spanish, French	0.688
Italian, French	0.666
Average	0.680
One translation feature	
German	0.686
Spanish	0.687
Italian	0.675
French	0.683
Average	0.682
No translation features	
None	0.642

Table 7.13: Accuracy figures for Dutch for a varying number of translation features including the other four languages viz. Italian, Spanish, German and French.

	German
Baseline	0.523
All four translation features	
Spanish, Italian, Dutch, French	0.668
Three translation features	
Dutch, Spanish, French	0.664
French, Italian, Dutch	0.667
Dutch, Italian, Spanish	0.677
French, Spanish, Italian	0.664
Average	0.668
Two translation features	
Spanish, French	0.659
French, Dutch	0.665
Spanish, Dutch	0.663
French, Italian	0.652
Spanish, Italian	0.672
Italian, Dutch	0.677
Average	0.665
One translation feature	
Italian	0.658
Spanish	0.679
Dutch	0.658
French	0.667
Average	0.665
No translation features	
None	0.642

Table 7.14: Accuracy figures for German for a varying number of translation features including the other four languages viz. French, Spanish, Italian and Dutch.

	Spanish
Baseline	0.601
All four translation features	
German, Italian, Dutch, French	0.705
Three translation features	
Dutch, Italian, German	0.700
French, Italian, German	0.701
Dutch, German, French	0.693
French, Dutch, Italian	0.707
Average	0.700
Two translation features	
German, French	0.691
French, Dutch	0.699
German, Dutch	0.684
French, Italian	0.706
German, Italian	0.701
Italian, Dutch	0.708
Average	0.698
One translation feature	
Italian	0.708
German	0.689
Dutch	0.686
French	0.686
Average	0.692
No translation features	
None	0.674

Table 7.15: Accuracy figures for Spanish for a varying number of translation features including the other four languages viz. French, Italian, German and Dutch.

	Italian
Baseline	0.526
All four translation features	
Spanish, German, Dutch, French	0.624
Three translation features	
Dutch, Spanish, German	0.622
French, German, Dutch	0.632
French, German, Spanish	0.629
French, Spanish, Dutch	0.637
Average	0.630
Two translation features	
Spanish, German	0.634
French, German	0.633
German, Dutch	0.628
French, Spanish	0.647
Spanish, Dutch	0.629
French, Dutch	0.631
Average	0.633
One translation feature	
German	0.629
Spanish	0.659
Dutch	0.620
French	0.644
Average	0.638
No translation features	
None	0.652

Table 7.16: Accuracy figures for Italian for a varying number of translation features including the other four languages viz. French, Spanish, German and Dutch.

7.5 Experimental results for the individual test words

In addition to the overall score averaged over all twenty test words, we also examined the individual scores for all test words. Figure 7.2 shows the accuracy figures for the timbl classifier using the latent semantic translation features, while Figure 7.3 shows the accuracy figures for the system containing the binary translation features. The scores are listed for all individual test words in the five supported languages.

Both figures show similar curves, except for some exceptions where accuracy figures are very different between the two systems (E.g. *execution* in Dutch and *ring* in German). In addition, the scores also show similar trends across languages. If we compare the language curves in Figure 7.2, they follow similar paths, except for some specific words that perform much better in one particular language. Examples of these outliers are *figure* in Spanish, *mood* in German and *pot* in Dutch. The reason for this behavior might be due to the fact that these words have more generic translations in these respective languages (viz. *figura*, *Stimmung* and *pot*) that cover different meanings of the ambiguous focus word.

It is also clear from the individual test scores that some words (e.g. *coach, figure, match, range*) are particularly hard to disambiguate, while others obtain very high scores (e.g. *rest*). As we already mentioned, the almost perfect scores for some words can be attributed to a very generic translation which accounts for all senses of the word even though there might be more suitable translations for each of the senses depending on the context. Because the manual annotators were able to select three good translations for each test instance, the most generic translation is often part of the gold standard translations. This is also reflected in the high baseline scores for these words. For the words performing badly in most languages, an inspection of the training data properties revealed two possible explanations for the poor classification results. Firstly, there seems to be a link with the number of training instances, corresponding to the frequency of the word in the training corpus. As is shown in Table 6.1, both for *coach* and *match* – two words that consistently perform badly across languages – there are very few training examples in the corpus. We can assume that adding more training data for these particular words could enable the classifier to predict more accurate translation labels. Secondly, the ambiguity or number of valid translations per word in the training data also seems to play a role in the classification results. Both *figure* and *range* appear very hard to classify correctly, and both words are very ambiguous, with no fewer than 167 and 145 translations, respectively, to choose from in French.

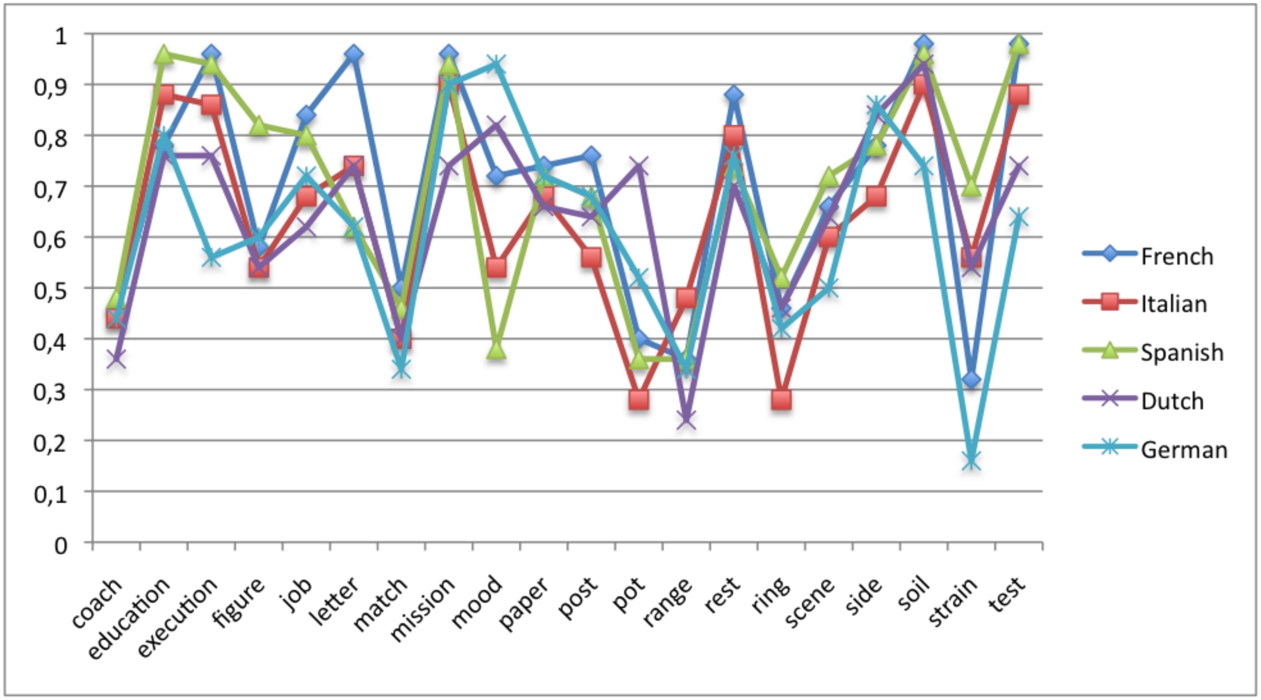


Figure 7.2: Accuracy figures for the ParaSense system incorporating latent semantic translation features for all twenty test words in all five supported languages.

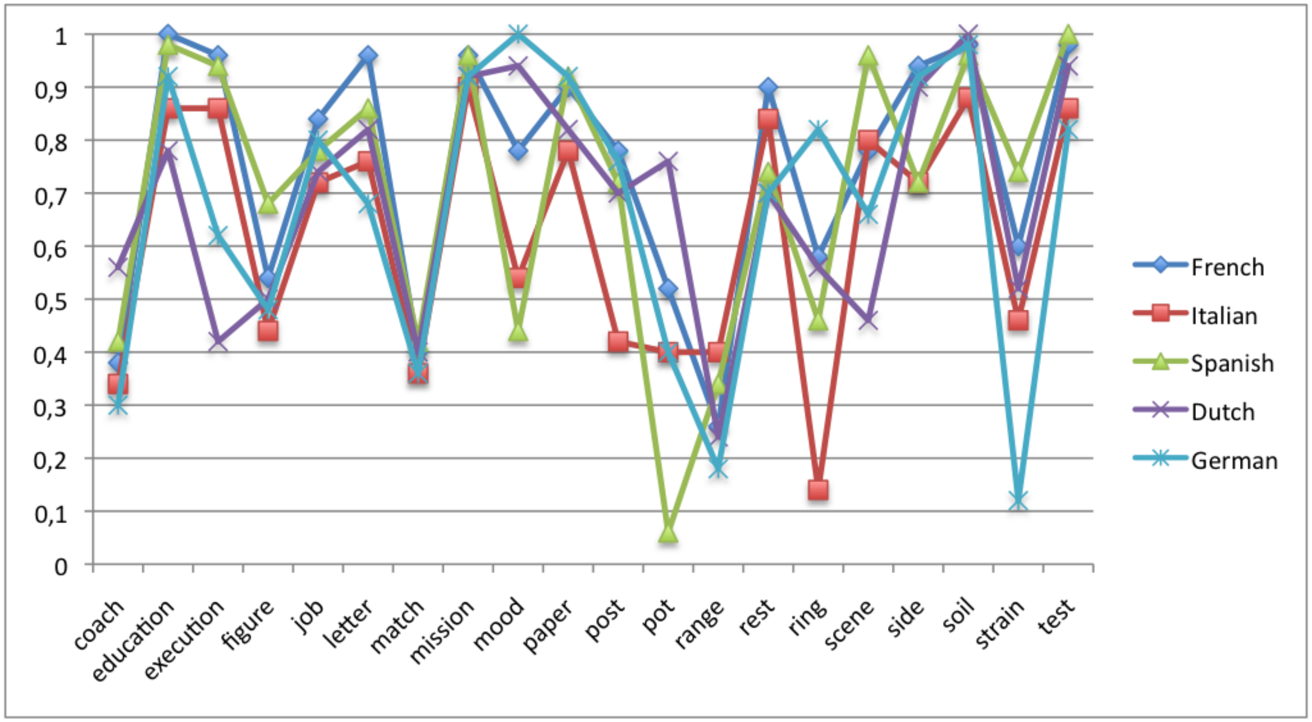


Figure 7.3: Accuracy figures for the ParaSense system incorporating the binary translation features for all twenty test words in all five supported languages.

7.6 Impact of word alignment errors

One major advantage of the presented cross-lingual WSD approach is that all steps to create the ParaSense system can be run automatically. As a consequence, the approach is very sensitive to error percolation between the different steps that are run automatically for the creation of the training and test feature vectors. As mentioned in Section 2.3.2, especially automatic word alignment is not error-prone yet, which leads to erroneous translation labels in the training corpus. We managed, however, as illustrated in Section 5.1.1, to achieve reasonable word alignment performance on the training data containing our set of ambiguous focus words (around 80% on average). This strengthened us in our belief that the automatic extraction of the translation labels was feasible for our corpus. Nevertheless, we believe it is important to measure the performance decrease caused by those errors that were introduced by the statistical word alignment procedure. In order to do so, we built a version of the ParaSense system which contains manually-validated translation labels, and compared it to the system which contains the automatically-generated translation labels. The construction of both sets of translation labels was described in more detail in Section 3.3.

Figure 7.4 shows the performance differences when using corrected or automatically generated translation labels for the ParaSense system containing binary translation features, whereas Figure 7.5 visualizes the performance differences for the ParaSense system using latent semantic translation features.

The figures clearly show that the classification scores decrease only slightly when the automatically-generated word alignments are used. In general, the ParaSense system using the latent semantic translation features benefits most from the manual validation of the translation labels, resulting in performance improvements ranging between 0.3% (Dutch) and 2% (French). The ParaSense system incorporating binary translation features hardly suffers from word alignment errors; only the German classifier considerably improves (+ 2.3%) when using the corrected translation labels. These results confirm the viability of our setup. Manual interventions in the data seem to result in very modest performance gains. As a consequence, our system can be developed fully automatically, which makes it very flexible and language-independent.

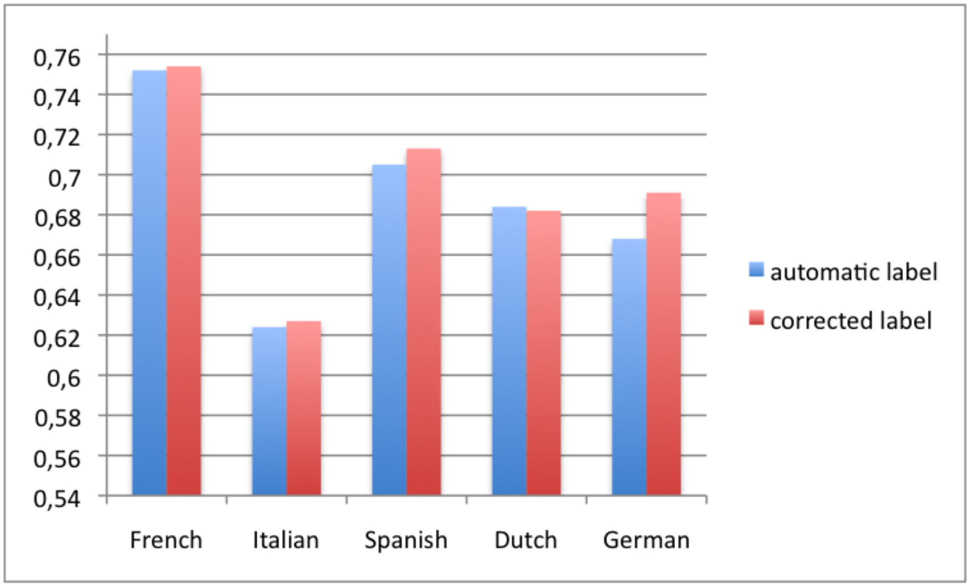


Figure 7.4: Accuracy scores for two flavors of the ParaSense system containing binary translation features; one flavor containing the automatically generated translation label and the other flavor containing the manually validated translation label.

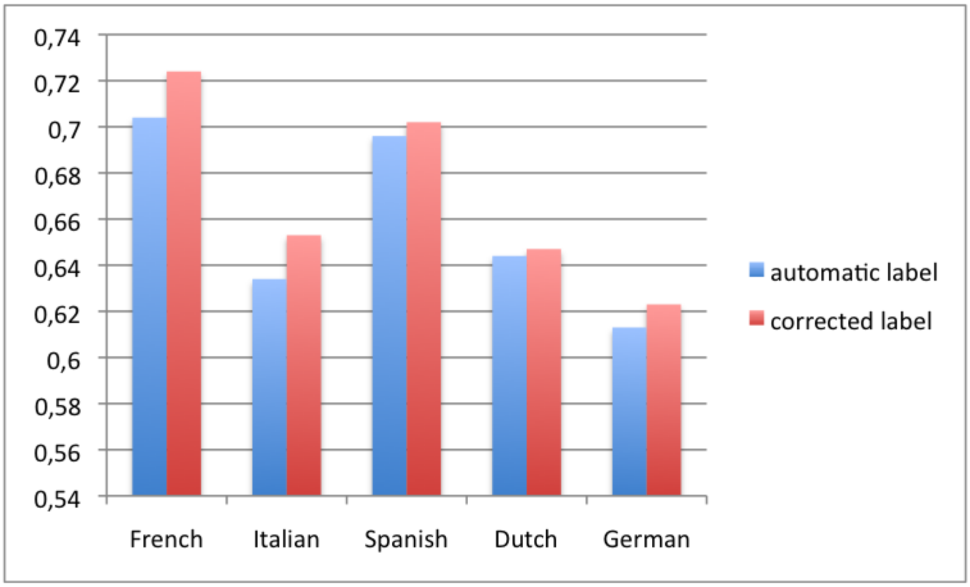


Figure 7.5: Accuracy scores for two flavors of the ParaSense system containing the latent semantic translation features; one flavor containing the automatically generated translation label and the other flavor containing the manually validated translation label.

7.7 Comparison with state-of-the-art systems

Finally, we also compared our results with all systems that participated in the SemEval-2 Cross-Lingual Word Sense Disambiguation task (Lefever and Hoste 2010). Five different teams participated to the CLWSD competition. All teams were allowed to submit up to 4 different flavors of their system. The winning systems were UvT-WSD (that only participated for Dutch and Spanish) and T3-COLEUR.

The UvT-WSD system (van Gompel 2010), which also uses a k Nearest Neighbor classifier and a variety of local and global context features, yielded the best scores for Spanish and Dutch in the SemEval CLWSD competition. Although we also used a memory-based learner, our method is different from this system in the way the feature vectors are constructed. Alongside similar local context features, we also included translational evidence from multiple languages in our feature vector.

For French, Italian and German, the T3-COLEUR system (Guo and Diab 2010) outperformed the other systems in the SemEval competition. This system adopts a different approach: during the training phase a monolingual WSD system processes the English input sentence and a word alignment module is used to extract the aligned translation. The English senses, together with their aligned translations (and probability scores), are then stored in a word sense translation table, in which look-ups are performed during the testing phase. This system also differs from the UvT-WSD and ParaSense systems in that the word senses are derived from WordNet, whereas the ParaSense and UvT-WSD systems do not use any external resources.

The OWNS system (Mahapatra et al. 2010) identifies the nearest neighbors of the test instances from the training data using a pairwise similarity measure that corresponds to the weighted sum of the word overlap and semantic overlap between two sentences. They also use WordNet similarity measures as an additional information source. The FCC-WSD system (Vilariño et al. 2010) uses a Naive Bayes classifier which is fed with the probabilities obtained from a bilingual translation table. The probability dictionary results from running Giza++ on the Europarl corpus. Finally, the UHD system (Silberer and Ponzetto 2010) builds for each focus word a multilingual co-occurrence graph based on the focus word's aligned contexts found in parallel corpora. The cross-lingual nodes are first linked by translation edges, that are labeled with the translations of the focus word in the corresponding contexts. The graph is transformed into a minimum spanning tree which is used to select the most relevant words in context to disambiguate a given test instance.

Figures 7.6 to 7.10 list the best precision scores for the five languages averaged over all twenty test words for the baseline, the participating SemEval systems and two flavors of the ParaSense system: *ParaSense_binary* contains the binary translation features while *ParaSense_LSA* contains the latent semantic translation features.

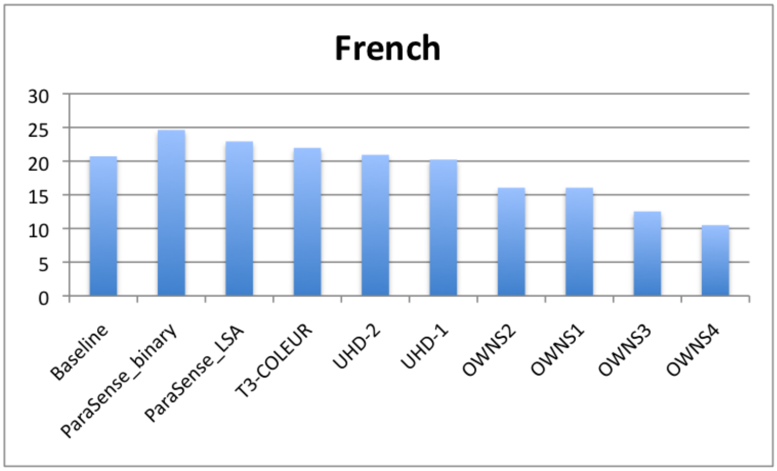


Figure 7.6: BEST precision scores for the baseline, the participating SemEval systems and two flavors of the ParaSense system for French.

The BEST precision scores show that both flavors of the ParaSense system as well as the two winning SemEval systems beat the most frequent translation baseline, except for Dutch, where the T3-COLEUR system performs below the baseline. In addition, both flavors of the ParaSense system clearly outperform all participating SemEval systems for French, Italian and German. For Spanish and Dutch, the scores are very similar for the two ParaSense flavors and the wining SemEval system, UvT-WSD. The ParaSense flavor containing the binary translation features, however, outperforms all systems in all five languages. These results confirm the potential advantages of using a multilingual approach to solving the cross-lingual WSD task.

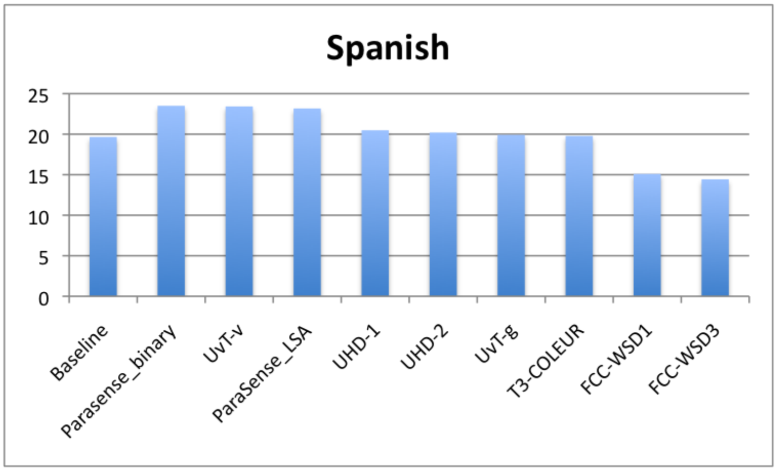


Figure 7.7: best precision scores for the baseline, the participating SemEval systems and two flavors of the ParaSense system for Spanish.

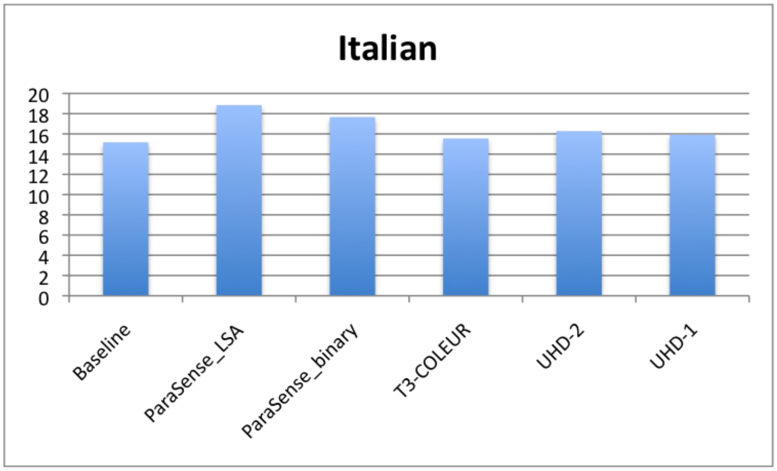


Figure 7.8: best precision scores for the baseline, the participating SemEval systems and two flavors of the ParaSense system for Italian.

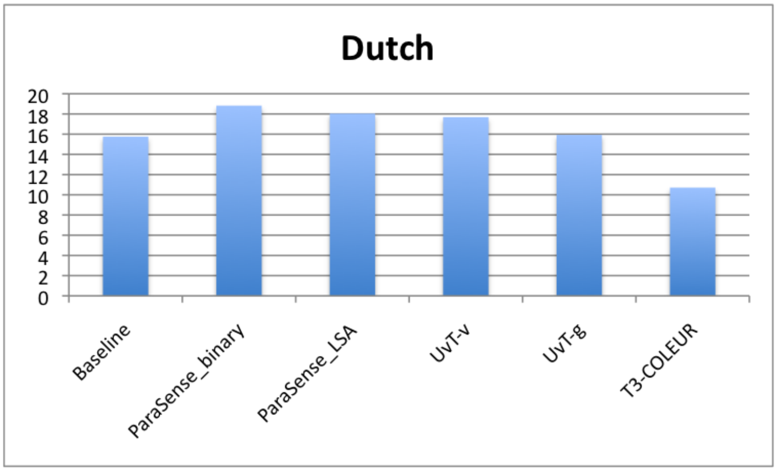


Figure 7.9: BEST precision scores for the baseline, the participating SemEval systems and two flavors of the ParaSense system for Dutch.

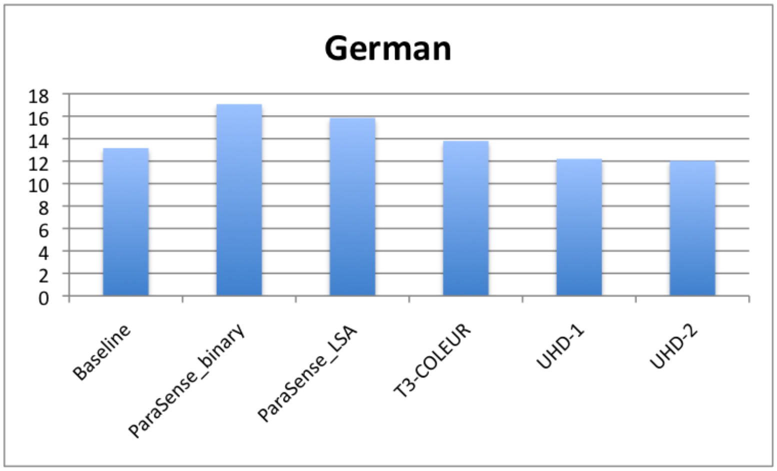


Figure 7.10: BEST precision scores for the baseline, the participating SemEval systems and two flavors of the ParaSense system for German.

We presented a detailed analysis of the classification results of our ParaSense system on the dedicated cross-lingual WSD benchmark data set and compared the results with state-of-the-art systems on the same task. In the next chapter, we will evaluate our results in a more practical machine translation setting. In order to do so, we will compare the output of the ParaSense system with the output of two statistical machine translation systems for our set of ambiguous focus words.

CHAPTER 8

Evaluating the translation quality of ambiguous words in an SMT framework.

In this chapter, we examine the potential benefits of a dedicated cross-lingual WSD system in a statistical machine translation framework. We compared the performance of our ParaSense system with the output of two state-of-the-art statistical machine translation (SMT) systems, Moses and Google Translate, on our lexical sample set of 20 ambiguous nouns.

8.1 Word Sense Disambiguation for Machine Translation

An important line of WSD research consists in the development of dedicated WSD modules for machine translation (MT). Instead of assigning a sense label from a monolingual sense-inventory to the ambiguous word, the WSD system has to predict a correct translation for the ambiguous word in a given context. A parallel research track investigates the improvements for statistical machine translation when integrating source context modeling directly into the SMT framework. The source language context can be modeled by using a wide range of contextual features, ranging from lexical local context features (Giménez and Marquez 2007, Stroppa, van den Bosch and Way 2007) or features extracted from the full sentence (Carpuat and Wu 2007) to shallow and deep syntactic features (Haque et al. 2011). For a detailed overview of related research on integrating source language context into statistical machine translation, we refer to Haque et al. (2011).

In our related research overview, we focus on studies that integrate source language context information into the SMT framework in order to filter the list of candidate translations by learning context-dependent translation probabilities, as opposed to studies that are more focussed on creating improved word alignment and translation lexicons.

The very first related studies tried to integrate context information into word-based SMT models. Brown et al. (1991) developed a dedicated WSD model to generate a correct French translation for English ambiguous words. New instances of the ambiguous focus word receive a sense label based on mutual information with the translation of the focus word in the corpus. In Vickrey et al. (2005), the problem was defined as a word translation task, where the correct translation of an ambiguous word is predicted based on the context of the word. The translation choices of ambiguous words are gathered from a parallel corpus by means of word alignment. The authors reported improvements on two simplified translation tasks: word translation and blank filling. The evaluation was done on an English-French parallel corpus but was faced with the important limitation of having only one valid translation (the aligned translation in the parallel corpus) as a gold standard translation. Carpuat and Wu (2005) used a Chinese WSD model to post-process the Chinese-English SMT output: translation candidates are directly replaced by the output that is generated by the WSD module. The authors report that the system that uses the WSD output does not yield significantly better translation quality than the default SMT system.

Gradually, the focus shifted from word-based to phrase-based SMT systems. Specia, Nunes and Stevenson (2006) used an inductive logic programming-based WSD system which was tested on ten ambiguous verbs in English-Portuguese translation. The system incorporates co-occurrence information from the context that refers to words which have been previously translated. The latter systems already present promising results for the use of WSD in MT, but really significant improvements in terms of general machine translation quality were for the first time obtained by Carpuat and Wu (2007) and Chan et al. (2007). Both papers describe the integration of a dedicated WSD module in a Chinese-English statistical machine translation framework and report statistically significant improvements in terms of standard MT evaluation metrics. Specia,

Sankaran and Nunes (2008) use n-best reranking to integrate a dedicated WSD module for English-Portuguese within a SMT system. The authors report significant improvements in BLEU scores.
Stroppa et al. (2007) directly introduce context-information features that exploit source similarity, in addition to target similarity that is modeled by the language model, in an SMT framework. For the estimation of these features, which are very similar to the typical WSD local context features (left and right context words, Part-of-Speech of the focus phrase and context words), they use a memory-based classification framework. Haque et al. (2011) combine a set of lexical features with semantic roles and dependency information. They observe that including contextual features of the source language in general produces improvements for the SMT output.

We strongly believe that our ParaSense WSD system, which presents a real multilingual approach to WSD (thus also integrating information from languages apart from the source and target language), can contribute to the performance of SMT, especially because it can easily be trained for different language pairs on exactly the same corpus that is used to train the SMT system, which should make integration much easier. In order to verify the potential of our ParaSense system in an SMT context, we evaluated the performance of two state-of-the-art SMT systems and ParaSense on the translation of ambiguous words. Although it is crucial to measure the general translation quality of the MT output after integrating a dedicated WSD module in the SMT system, we think it is equally interesting to conduct a dedicated evaluation of the translation quality on ambiguous nouns. Standard SMT evaluation metrics such as BLEU (Papineni et al. 2002) or edit-distance metrics (e.g. Word Error Rate) measure the global overlap of the translation with a reference, and are thus not very sensitive to WSD errors. The mistranslation of an ambiguous word might be a subtle change compared to the reference sentence, but it often drastically affects the global understanding of the sentence.

Example (23) illustrates the importance of correctly translated ambiguous nouns for the general understanding of the sentence. The English input sentence reports on a car accident, but two important words are translated wrongly: *wreckage* is translated as *schipbreuk*, which means “shipwreck” in English, whereas *neerstorting* refers to a plane *crash*[1].

(23) *ENGLISH*: Two elderly casualties in the car had to be cut free from the **wreckage** following the **crash** on the A75 near Gretna.

DUTCH: Twee bejaarde slachtoffers in de auto moesten vrij van de

[1] The example is taken from a student assignment from 2010, where the automatic translation was generated by the Babelfish system.

> **schipbreuk** na de **neerstorting** worden gesneden op A75 dichtbij Gretna.

Section 8.2 introduces the two machine translation systems we evaluated, while section 8.3 gives an overview of the experimental set-up and results.

8.2 Statistical Machine Translation Systems

For our experiments, we analyzed the behavior of two phrase-based statistical machine translation (SMT) systems on the translation of ambiguous nouns. SMT generates translations on the basis of statistical models whose parameters are derived from the analysis of sentence-aligned parallel text corpora. Phrase-based SMT is considered as the dominant paradigm in MT research today. It combines a phrase translation model (which is based on the noisy channel model) and a phrase-based decoder in order to find the most probable translation e of a foreign sentence f (Koehn, Och and Marcu 2003). Usually, Bayes' rule is used to reformulate this translation probability:

$$argmax_e p(e|f) = argmax_e p(f|e)p(e)$$

This allows for a language model $p(e)$ that guarantees the fluency and grammatical correctness of the translation, and a separate translation model $p(f|e)$ that focuses on the quality of the translation. Training of both the language model (on monolingual data) as well as the translation model (on bilingual text corpora) requires large amounts of text data.

Research has shown that adding more training data, both for the translation and for the language models, results in better translation quality (Callison-Burch et al. 2009). Therefore, it is important to notice that our comparison of the two SMT systems is somewhat unfair, as we compared the Moses research system (that was trained on the Europarl corpus) with the Google commercial system that is trained on a much larger data set. It remains an interesting exercise though, as we consider the commercial system as the upper bound of how far current SMT can go in case it has virtually unlimited access to text corpora and computational resources.

8.2.1 Moses

The first statistical machine translation system we used is the off-the-shelf Moses toolkit (Koehn et al. 2007). As the Moses system is open-source, well documented, supported by a very lively users forum and reaches state-of-the-art performance, it has quickly been adopted by the community and has significantly stimulated development in the SMT field. It also features factored translation models, which enable the integration of linguistic and other information at the word level. This makes Moses a good candidate to experiment with a dedicated WSD module, that requires more enhanced linguistic information (such as lemmas and Part-of-Speech tags).

We trained Moses for English–French and English–Dutch on the large subsection of the Europarl corpus that was introduced in Section 3.1.2, and performed some standard cleaning:

- Empty lines were removed.
- Redundant space characters were removed.
- Sentences (and their aligned counterpart) that were too short (violating the 9-1 sentence ratio limit of GIZA++) or too long (containing more than 80 words) were removed.

Table 8.1 lists the number of aligned sentences after cleaning the bilingual corpus, and the number of uni-, bi- and trigrams that are comprised in the language model.

	French	Dutch
Number of bilingual sentence pairs		
	872.689	873.390
Number of ngrams		
unigrams	103.027	173.700
bigrams	1.940.925	2.544.554
trigrams	2.054.906	1.951.992

Table 8.1: Statistics resulting from the Moses training phase

8.2.2 Google

In order to gain insights into the upper bounds for current SMT, we also analyzed the output of the Google Translate API[2] for our set of ambiguous nouns. Google Translate currently supports 64 languages.

Since both the volume of parallel and monolingual training data as well as computer power are crucial for statistical MT, Google – that disposes of large computing clusters and a network of data centers for Web search – has very valuable assets at its disposal for this task. We can only speculate about the number of resources that Google uses to train its translation engine. Part of the training data comes from transcripts of United Nations meetings (in six official languages) and those of the European Parliament (Europarl corpus). Google research papers report on a distributed infrastructure that is used to train on up to two trillion tokens, which result in language models containing up to 300 billion ngrams (Brants et al. 2007). Given that these figures were published in 2007, they are probably already outdated.

8.3 Evaluation

To evaluate the two machine translation systems as well as the ParaSense system on their performance on the lexical sample of twenty ambiguous words, we used the manually constructed gold-standard and test set that was presented in Chapter 5. The experiment was carried out for two language pairs, viz. English–French and English–Dutch. As evaluation metric, we used the straightforward accuracy measure that divides the number of correct answers by the total amount of test instances. As a baseline, we list again the most frequent lemmatized translation that resulted from the automated word alignment (GIZA++).

The output of the ParaSense WSD module consists of a lemmatized translation of the ambiguous focus word in the target language. The output of the two statistical machine translation systems, however, is a translation of the full English input sentence. Therefore, we manually selected the translation of the ambiguous focus word from the full translation generated by both SMT systems, and ensured the translation was rendered in its base form (masculine singular form for nouns and adjectives, infinitive form for verbs). Since the gold standard potentially contains nine valid translation labels for each test instance, we believe there is only a small chance that the SMT systems generated other valid synonyms for the

[2] `http://code.google.com/apis/language/translate/overview.html`

gold standard translations. We therefore decided to not further analyze the Moses and Google output.

Figure 8.1 lists the accuracy figures for the baseline, two flavors of the ParaSense system (*ParaSense_LSA*: local context features combined with latent semantic translation features, *ParaSense_binary*: local context features combined with binary translation features), Moses and Google for English–French and English–Dutch.

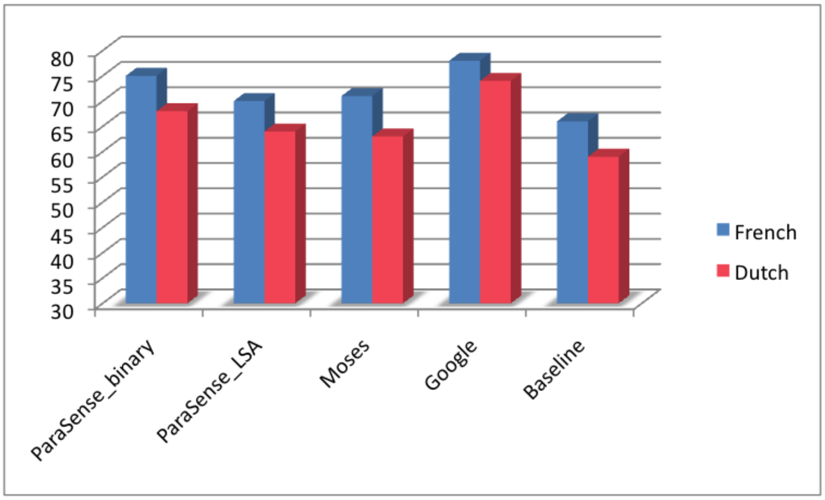


Figure 8.1: French and Dutch accuracy figures per system averaged over all 20 test words.

A first conclusion is that all systems beat the most frequent sense baseline. As expected, the Google system (which has no limitations on the amount of training data) achieves the best results, but for French the considerable difference in training size only leads to modest performance gains when we compare Google with the ParaSense system that incorporates the binary translation features (78% versus 75%). The ParaSense system using binary translation features outperforms Moses (75% versus 71% for French, 68% versus 63% for Dutch), whereas the ParaSense system built with latent semantic translation features obtains very similar results to Moses.

We carried out a statistical test for the equality of proportions to measure statistical significance of the performance differences. As we compare 5 systems in total (MFT baseline, ParaSense_binary, ParaSense_LSA, Moses

FRENCH	ParaSense	ParaSense	Moses	Google
	Binary	LSA		
ParaSense_LSA	0.1819	–	–	–
Moses	0.4896	1.0000	–	–
Google	1.0000	**0.0016**	**0.0070**	–
Baseline	**5.1e-05**	0.3086	0.1073	**2.4e-08**

Table 8.2: Statistic significance test results for French.

DUTCH	ParaSense	ParaSense	Moses	Google
	Binary	LSA		
ParaSense_LSA	0.64853	–	–	–
Moses	.20815	1.00000	–	–
Google	0.11964	**0.00011**	**1.2e-05**	–
Baseline	**0.00034**	0.24060	0.73243	**2.4e-10**

Table 8.3: Statistic significance test results for Dutch.

and Google), we corrected the p-values by means of the Bonferroni correction[3] for multiple comparisons (Miller 1991). Table 8.2 lists the statistical significance test results for French, whereas Table 8.3 presents the results for Dutch.

The conclusions of the significance test were the following:

- The results for French and Dutch are similar, whereas the absolute scores show larger differences between the various systems in the two languages.
- Only Google and the ParaSense_binary system are significantly better than the most frequent translation baseline.
- The ParaSense system outperforms Moses in absolute scores, but these performance differences are not statistically significant. The same holds for the differences between Google and the ParaSense_binary system.

To conclude, more test data will be needed to confirm or nuance the obtained test results.

We also compared the performance of the ParaSense system with the two machine translation systems for all individual test words. Figure 8.2

[3] The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously.

illustrates the accuracy figures for French for the ParaSense system that incorporates the binary translation features and for both MT systems.

In general, the three curves follow a similar pattern. For some words, however, Google clearly outperforms the other systems (*figure, job, match, range*), whereas for other words the ParaSense system shows the best results (*education, ring, strain, test*).

To summarize, the best results are obtained by Google, the SMT system that is built with less constraints on data size or computational resources. Furthermore, the results show that the ParaSense system incorporating binary translation features clearly outperforms Moses that is built with the same training corpus. Although the MT system has access to the same context information, the additional translational evidence that is used by the ParaSense system seems to further improve the translation quality of the ambiguous focus words. As a consequence, we believe that adding a dedicated multilingual WSD module to a statistical machine translation system could improve the translation adequacy on ambiguous words of the SMT system. Further research, however, is needed to examine the performances of the different systems on other word categories such as adjectives, verbs and adverbs.

We carried out a basic test to examine the potential benefits of adding a dedicated cross-lingual WSD module to an SMT framework. In future research, we would like to properly embed our ParaSense system into an SMT framework. For this purpose, we could for instance rerank the n-best translations based on the WSD output by assigning higher scores to translation candidates that contain the translation generated by the WSD system.

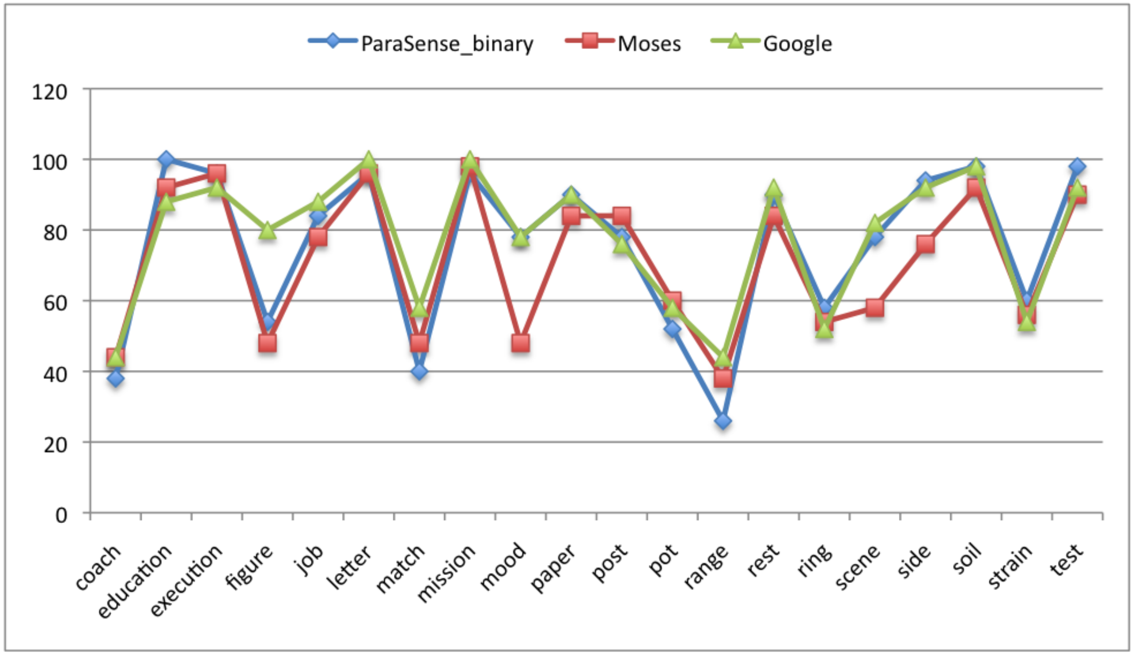


Figure 8.2: French accuracy figures per system for all 20 test words.

CHAPTER 9

Conclusions

In this thesis, a multilingual classification-based machine learning approach to Cross-lingual Word Sense Disambiguation was presented. Unlike other multilingual approaches to WSD, the ParaSense system does not use any manually created lexical resources, but relies only on parallel corpora for the automatic creation of the sense inventory and the construction of the training data set. In this way, it tackles the data acquisition bottleneck which remains an issue for many low-density languages. Although the collection of very large and varied parallel corpora might also be problematic for underresourced languages at this moment, multilingual public and private text resources are becoming increasingly available. Alongside the numerous projects aiming at the creation of parallel corpora for specific languages (Eg. Macken, De Clercq and Paulussen (2011)), there is a large interest in the research community to automatically build parallel corpora from the web (Resnik and Smith 2003, Mohler and Mihalcea 2008).

Using translations instead of a fixed predefined sense inventory such as WordNet offers a number of additional advantages. Firstly, using translations makes it easier to integrate a dedicated WSD module into real multilingual applications such as Machine Translation or Information Retrieval. In Chapter 8, we showed the potential benefits of integrating the ParaSense system in a statistical machine translation system. Secondly, our approach deals with the sense granularity issue since finer sense distinctions are only relevant insofar as they get lexicalized by different translations of the word. Since we build our sense inventory in an automatic way, we do not have to concern ourselves with subjective decisions lexicographers are supposed to take. In addition, the use of parallel corpora allows the automatic creation of dedicated sense inventories for specialized domains.

In order to test the viability of multilingual WSD, we created a lexical sample data set for 25 ambiguous English nouns that was divided into a trial data set of five words and a test data set of 20 words. In Chapter 5 a detailed overview was presented of the different steps that were taken to construct this dedicated CLWSD data set. For the creation of the gold standard, we first applied word alignment on the parallel corpus in order to retrieve the set of possible translations for the ambiguous focus words. In a second step, these translations were manually clustered by meaning. In a final step, the resulting sense inventory was used by three annotators to manually annotate all trial and test instances. The resulting benchmark data set was also used for the SemEval-2010 Cross-Lingual Word Sense Disambiguation evaluation task.

Section 9.1 briefly reviews the ParaSense system architecture, while Section 9.2 summarizes the classification results and conclusions we could draw from all experimental work. Section 9.3 gives some prospects for future research.

9.1 ParaSense System Architecture

The ParaSense system takes a classification-based approach to WSD, where the possible translations of an ambiguous word are the class labels and new occurrences of the word are assigned a correct translation based on disambiguating information. We elaborated on the information sources that were used to build the feature vectors for all training and test instances in Chapter 3. A combination of both local context information and translational evidence was used to discriminate between different senses of the word, the underlying hypothesis being that using multilingual information would be more informative than having access to monolingual or bilingual features. Two different flavors of the bag-of-words translation features were constructed. The first flavor contains binary translation features that simply indicate the presence or absence of a specific content word in the context of the ambiguous focus word. For the second flavor, we applied singular value decomposition on the document-term matrices in order to find more abstract *latent semantic* relations between the bag-of-words features. In this way, we were able to reduce our very sparse bag-of-words features and find hidden associations between synonyms of different instances. With respect to the translation labels, we generated them in an automated way by running statistical word alignment on the sentence-aligned parallel corpus.

Chapter 4 gives an overview of the machine learning methods that were applied in the ParaSense system: a Memory-based Learning (MBL) Method and a Support Vector Machine (SVM) method as implemented in SVMlight. For the MBL learner, we opted for the k Nearest neighbor method as implemented in timbl. As most classifiers can be initialized with a wide range of parameters, we decided to use a genetic algorithm to optimize the parameter settings for our classification task. The focus in the GA experiments was on the optimization of the timbl parameters. With respect to the resulting selected parameters, general tendencies could be observed across the different words and languages (Section 7.3.1): *Gain ratio* and the *shared variance weighting* were the optimal feature weighting techniques, *majority voting*, and to a minor extent the *inverse distance weighting* were the best types of class voting weights and the *IB1 algorithm* was shown to give the best classification results. Considering the number of nearest distances taken into account, we could observe a general use of high k values, that were in average ranging between 8 and 11. In general, we noticed large performance differences on the training data when combining different parameter settings for the selection of the best GA individual. This observation also holds for the test data, where performance increases of 18.9% for Dutch, 11.6% for French and 16.4% for German can be noticed for the classifier applied to the latent semantic translation features. The SVMlight classification results were optimized by applying the linear kernel with trade-off parameter $C = 1.0$, which has shown to perform well for the WSD task before. Here as well, performance improvements of 7.8% (German) and 7% (French) can be observed. We are well aware of the fact that we only performed a small fraction of all possible optimization experiments. We could, for instance, have applied the GA on the binary feature sets as well. Through the optimization experiments, we wanted to determine whether performance gains could be obtained by changing the algorithm parameters of the algorithm under consideration. The results confirm that such optimization is indeed crucial for obtaining reasonable classification performance.

9.2 Classification results: main observations

A diverse set of experiments was conducted to validate the research hypotheses of our multilingual classification-based approach to Word Sense Disambiguation:

(a) it is possible to rely on parallel corpora to generate the translation labels and disambiguating features in an automated way, without using external resources.

(b) adding multilingual information to the classifier improves the disambiguation of polysemous nouns.

Chapter 7 gives a detailed overview of all experimental results, which lead to the following conclusions.

Validity of the multilingual classification-based approach In general, the experimental results clearly confirm the validity of our multilingual approach to word sense disambiguation: the classifiers that employ translational evidence outperform the classifiers that merely use English local context features for four out of five target languages, viz. French, Spanish, German and Dutch. For Italian, however, only the classifier that utilizes the Spanish translations achieves better results than the classifier that does not use any translation features at all. For French, Spanish and Italian, the translations from the other romance languages seem to contribute most to the classification results. For Dutch and German, the classifiers integrating translations from the other germanic language also achieve very good classification scores. In order to draw final conclusions on the contribution of the different languages, one could test the approach with more distant languages as well.

Binary bag-of-words translation features versus latent semantic translation features At first sight, the ParaSense system that combines the English local context features with a set of binary translation features clearly outperforms the system using latent semantic translation features. As a consequence, the classifier does not seem to take advantage of the dimensionality reduction that is performed on the very sparse bag-of-words translation features. Nevertheless, a qualitative analysis revealed that the ParaSense system using the latent semantic translation features is less biased to generate the most frequent translation label, and predicts a more diverse set of more distinctive translation class labels.

Comparison with state-of-the-art systems A detailed comparison with all systems that participated in the SemEval cross-lingual word sense disambiguation task reveals that the ParaSense system obtains state-of-the-art results for the task in hand. The ParaSense flavor combining English local context features with the binary bag-of-words translation features outperforms all other systems for all five target languages. As most other systems used external resources combined with bilingual translation information, the results confirm the potential advantages of using a multilingual approach that extracts all disambiguating information from sentence-aligned parallel corpora.

To conclude, we are convinced that our multilingual classification-based approach offers a very flexible, efficient and language-independent solution for the word sense disambiguation task. As all steps are performed automatically, and we only use a parallel corpus, we believe our approach proposes a valid answer to the knowledge-acquisition bottleneck.

9.3 Future research goals

Integration of Cross-lingual WSD in practical applications. One major advantage of taking a multilingual approach to Word Sense Disambiguation is that it enables working directly with translations instead of more abstract sense labels that need to be mapped in their corresponding translations. This should facilitate the integration of a dedicated WSD module into real applications.

In Chapter 8, we investigated the potential benefits of adding a dedicated WSD module to a *statistical machine translation* framework. Our experiments showed that the ParaSense system outperforms state-of-the-art SMT systems that are trained on the same amount of parallel data when it comes to translation quality of ambiguous nouns. Therefore, it would be interesting to integrate the ParaSense system in a real MT framework and measure the quantitative and qualitative impact of adding a WSD module to the SMT system.

Another direction for future research is to integrate the cross-lingual WSD architecture in a *Cross-lingual link discovery system.* Wikipedia pages typically contain inter-language links to the corresponding pages in other languages, allowing users to consult the relevant information in their mother tongue. These links, however, are often incomplete; sometimes the corresponding pages in other languages are missing, or, when they do exist, no human contributor has established the appropriate inter-language link yet. The final goal of the link discovery system is to provide a human editor with a list of possible missing Wikipedia inter-language links that should be manually verified.

We performed a set of initial experiments to investigate the viability of discovering missing inter-language links between Wikipedia pages for ambiguous nouns (Lefever, Hoste and De Cock 2012). The input for the system was a set of Dutch pages for a given ambiguous noun and the output of the system was a set of links to the corresponding pages in three target languages (viz. French, Spanish and Italian). The system contains two sub-modules. In a first step, all pages are retrieved that contain a translation (in the three target languages) of the ambiguous word in the page title, whereas in a second step all corresponding pages are linked in the focus language (Dutch) and the three target languages. The linking of two web pages is recast as a classification problem: for every pair of documents, the classifier determines whether they should be linked or not. The framework of the classification approach is adopted from the ParaSense framework: the training feature vectors contain bag-of-words translation features that are extracted from the Dutch Europarl sentences containing the ambiguous Wikipedia concept and their aligned translations in five other languages (viz. English, French, Spanish, Italian and German). In order to construct the same set of translation features for the test vectors, the Dutch Wikipedia pages are translated by means of the Google Translate API. For the evaluation, all possible links were manually validated between the source and target Wikipedia documents for four ambiguous Wikipedia concepts: *muis*, *graad*, *stam* and *operatie*. The experimental results showed that although it is a very challenging task, the system succeeds to detect missing inter-language links between Wikipedia documents. We detected for instance a link between the French *Souris* page and the Dutch *Muis_van de hand* ("ball of the thumb" sense of the Dutch word *muis*) page that is not present in Wikipedia. We detected even more important missing links for more frequent usages of the noun, such as the analogies between the Dutch *muis_animal* and the Spanish and Italian corresponding pages. There are a number of remaining issues, though, that require further research: the bag-of-words translation features need to be enhanced as they appeared not to be informative enough; sometimes there is overlap on the content words of two different meanings of the word, which prevents the classifier from distinguishing between the different meanings of the concept. As an example, we can refer to the word *muis*. For both the "ball of the thumb" and "computer mouse" senses, the content word *hand* appears to be equally important. In addition, we need to expand the training corpus that currently only contains Europarl material, which is very different in nature and vocabulary from the Wikipedia pages. Finally, it would also be interesting to compare the cross-lingual WSD approach with an unsupervised clustering approach that only takes term-document matrices as input.

System optimization In addition to further enhancing the ParaSense system through the optimization of the parameter settings and feature spaces (Daelemans et al. 2003), it would be interesting to apply *instance selection* to the training set. As we work with an automatically constructed set of training feature vectors and classification labels, instance selection could help to remove the noise that is introduced by erroneous word alignments. We envision two paths to perform instance selection: the first path uses a Genetic Algorithm to optimize the content of the training set, while an alternative research path investigates the use of fuzzy-rough instance selection (Jensen and Cornelis 2010, Verbiest, Cornelis and Herrera 2012) to remove non-informative or bad instances from the training base.

Further research on the viability of a multilingual classification-based approach to Word Sense Disambiguation In this dissertation, we showed that adding multilingual evidence helps the classifier to predict a contextually correct translation for a set of polysemous target nouns. As the approach does not depend on manually annotated training corpora or predefined sense-inventories, we are convinced that the system could be extended to a more generic one that covers all ambiguous words.

Many interesting research opportunities accompany this extension of the system, such as

(a) **adding more distant languages** to the feature vector in order to measure whether languages from other language families contribute more/less to the classification results. As different languages tend to lexicalize different meaning of the word, one would expect that more distant languages will enable the system to distinguish between sense distinctions that are possibly not made by languages from the same language family.

(b) testing the approach for **other Part-of-Speech categories**. We have validated the ParaSense approach for ambiguous nouns, but the scope could be expanded to other grammatical categories of words, such as verbs or adjectives. We expect more noise due to erroneous word alignment for both PoS categories, but our system revealed to be quite robust against word alignment mistakes. In addition, one could decide to only keep those training instances where word alignment was performed with a high level of confidence.

(c) applying a **decompounding module** to the training data and test the impact on the word alignment performance and classification accuracy. Further research is required to measure the effect of error percolation (in case of false decompounding) on the correct disambiguation of the given focus words.

(d) adding **more test data** in order to confirm the obtained evaluation results. A new shared task will be proposed within the SemEval-2013 framework, that will provide the WSD community with another 1000 sense-tagged instances for the same set of ambiguous test words.

To conclude, we strongly believe that our multilingual corpus-based approach to Word Sense Disambiguation offers a very flexible framework that could be extended to cover more languages as well as a broader set of ambiguous focus words.

Bibliography

Agirre, E., Aldezabal, I., Etxeberria, J., Iruskieta, M., Izagirre, E., Mendizabal, K. and Pociello, E.: 2006, Improving the Basque WordNet by corpus annotation., *Proceedings of the Third International WordNet Conference*, Jeju Island, Korea, pp. 287–290.

Agirre, E., Ansa, O., Martínez, D. and Hovy, E.: 2001, Enriching WordNet concepts with topic signatures, *Proceedings of the NAACL Workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations*, Pittsburg, PA, pp. 23–28.

Agirre, E. and Edmonds, P.: 2006, *Word Sense Disambiguation. Algorithms and Applications*, Text, Speech and Language Technology, Springer.

Agirre, E., Lopez de Lacalle, O. and Martínez, D.: 2005, Exploring feature spaces with svd and unlabeled data for Word Sense Disambiguation., *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'05)*, Borovets, Bulgaria.

Agirre, E. and Martínez, D.: 2000, Exploring automatic word sense disambiguation with decision lists and the Web, *Proceedings of the Semantic Annotation and Intelligent Annotation Workshop, organized by COLING*, Luxembourg, pp. 11–19.

Agirre, E. and Martínez, D.: 2004a, Smoothing and Word Sense Disambiguation, *Proceedings of EsTAL - España for Natural Language Processing*, Alicante, Spain.

Agirre, E. and Martínez, D.: 2004b, Unsupervised WSD based on automatically retrieved examples: the importance of bias, *Proceedings of the 10th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, pp. 25–32.

Agirre, E., Otegi, A. and Zaragoza, H.: 2010, Using Semantic Relatedness and Word Sense Disambiguation for (CL)IR, *Lecture Notes in Computer Science* **6241**, 166–173.

Agirre, E. and Soroa, A.: 2009, Personalizing pagerank for word sense disambiguation, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, Athens, Greece.

Aha, D., Kibler, D. and Albert, M.: 1991, Instance-based learning algorithms, *Machine Learning* **6**, 37–66.

Apidianaki, M.: 2009, Data-driven semantic analysis for multilingual WSD and lexical selection in translation, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Athens, Greece, pp. 77–85.

Banerjee, S. and Pedersen, T.: 2002, An adapted lesk algorithm for word sense disambiguation using wordnet, *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLING-02*, pp. 136–145.

Bellman, R.: 1957, *Dynamic Programming.*, Princeton University Press, Princeton.

Berger, A., Caruana, R., Cohn, D., Freitag, D. and Mittal, V.: 2000, Bridging the Lexical Chasm: Statistical Approaches to Answer Finding, *Proc. Int. Conf. Research and Development in Information Retrieval*, pp. 192–199.

Boden, B.: 2010, *Word Sense Disambiguation based on parallel corpora: A comparison of the meanings and translations of five ambiguous words found in the European corpus with those found in the dictionary.*, Master's thesis, Ghent University College.

Brants, T., Popat, A., Xu, P., Och, F. and Dean, J.: 2007, Large Language Models in Machine Translation, *Proceedings of the 2007 Joint Conference on Empirical methods in Natural Language Processing and Computational Natural Language Learning*, pp. 858–867.

Brockmann, C. and Lapata, M.: 2003, Evaluating and Combining Approaches to Selectional Preference Acquisition, *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 27–34.

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A. and Mercer, R. L.: 1993, The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics* **19**(2), 263–311.

Brown, P., Pietra, S., Pietra, V. and Mercer, R.: 1991, Word-sense disambiguation using statistical methods, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, pp. 264–270.

Budanitsky, A. and Hirst, G.: 2001, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA, pp. 29–34.

Buitelaar, P., Magnini, B., Strapparava, C. and Vossen, P.: 2006, *Word Sense Disambiguation: Algorithms, Applications, and Trends*, Kluwer, chapter Domain-specific WSD, pp. 275–298.

Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J.: 2009, Findings of the 2009 Workshop on Statistical Machine Translation, *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, Athens, Greece, pp. 1–28.

Carpuat, M. and Wu, D.: 2005, Word sense disambiguation vs. statistical machine translation, *Annual Meeting of the Association for Computational Linguistics (ACL-05)*, Ann Arbor, Michigan.

Carpuat, M. and Wu, D.: 2007, Improving statistical machine translation using word sense disambiguation, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, pp. 61–72.

Chan, Y. and Ng, H.: 2005, Scaling Up Word Sense Disambiguation via Parallel Texts, *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, Pittsburgh, Pennsylvania, USA, pp. 1037–1042.

Chan, Y., Ng, H. and Chiang, D.: 2007, Word sense disambiguation improves statistical machine translation, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 33–40.

Chklovski, T. and Mihalcea, R.: 2002, Building a sense tagged corpus with open mind word expert, *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, USA, pp. 116–122.

Chklovski, T. and Pantel, P.: 2004, Verbocean: Mining the web for fine-grained semantic verb relations, *Proceedings of the 2004 Conference* *on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Claus, E.: 2011, *L'utilisation de textes parallèles pour la désambiguisation sémantique: une étude contrastive des traductions de mots polysémiques tires d'Europarl et proposées par des dictionnaires bilingues*, Master's thesis, Ghent University College.

Clough, P. and Stevenson, M.: 2004, Cross-language information retrieval using eurowordnet and word sense disambiguation., *Advances in Information Retrieval, 26th European Conference on IR Research (ECIR)*, Sunderland, UK, pp. 327–337.

Cover, T. and Hart, P.: 1967, Nearest neighbor pattern classification, *Institute of Electrical and Electronics Engineers Transactions on Information Theory* **13**, 21–27.

Cowie, J., Guthrie, J. A. and Guthrie, L.: 1992, Lexical disambiguation using simulated annealing, *Proceedings of the International Conference on Computational Linguistics (COLING)*, Nantes, France, pp. 157–161.

Cristianini, N. and Shawe-Taylor, J.: 2000, *Support Vector Machines and other kernel-based learning methods*, Cambridge University Press.

Crow, J. and Kimura, M.: 1970, *An Introduction to Population Genetics Theory*, New York: Harper and Row.

Daelemans, W., Buchholz, S. and Veenstra, J.: 1999, Memory-based shallow parsing, *CoNLL-99*, Bergen, Norway.

Daelemans, W., Hoste, V., De Meulder, F. and Naudts, B.: 2003, Combined optimization of feature selection and algorithm parameters in machine learning of language, *Machine Learning* pp. 84–95.

Daelemans, W. and van den Bosch, A.: 2005, *Memory-based Language Processing*, Cambridge University Press.

Daelemans, W., van den Bosch, A. and Zavrel, J.: 1999, Forgetting Exceptions is Harmful in Language Learning, *Machine Learning* **34**(1-3), 11–41.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A.: 2002, TiMBL: Tilburg Memory-Based Learner, version 4.3, Reference Guide, *Technical Report ILK Technical Report - ILK 02-10*, Tilburg University.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A.: 2009, TiMBL: Tilburg Memory Based Learner, version 6.2, Reference Guide, *Technical Report 09-01*, ILK Research Group.

Dagan, I., Itai, A. and Schwall, U.: 1991, Two Languages are more Informative than One, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 130–137.

DeConinck, M.: 2010, *Word Sense Disambiguation using a parallel corpus: A contrastive study of the translations of polysemous nouns as found in the Europarl corpus and in dictionaries.*, Master's thesis, Ghent University College.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R.: 1990, Indexing by latent semantic analysis, *Journal of the American Society for Information Science* **41**, 391–407.

Diab, M.: 2004, *Word Sense Disambiguation within a Multilingual Framework*, Phd, University of Maryland, USA.

Diab, M. and Resnik, P.: 2002, An Unsupervised Method for Word Sense Tagging Using Parallel Corpora, *Proceedings of ACL*, pp. 255–262.

Dudani, S.: 1976, The Distance-Weighted k-Nearest-Neighbor Rule, *IEEE Transactions on Systems, Man and Cybernetics* **6**(4), 325–327.

Dyvik, H.: 2004, Translations as semantic mirrors: from parallel corpus to wordnet, *Language and Computers* **49**(1), 311–326.

Eckart, C. and Young, G.: 1936, The approximation of one matrix by another of lower rank, *Psychometrika* **1**, 211–218.

Edmonds, P. and Kilgarriff, A.: 2002, Introduction to the special issue on evaluating word sense disambiguation systems., *Journal of Natural Language Engineering* **8**(4), 279–291.

Escudero, G., Màrquez, L. and Rigau, G.: 2000a, Boosting Applied to Word Sense Disambiguation, *European Conference on Machine Learning*, pp. 129–141.

Escudero, G., Màrquez, L. and Rigau, G.: 2000b, Naive bayes and exemplar-based approaches to word sense disambiguation revisited, *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI)*, Berlin, Germany, pp. 421–425.

Escudero, G., Màrquez, L. and Rigau, G.: 2000c, On the portability and tuning of supervised word sense disambiguation systems, *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, Hong Kong, China, pp. 172–180.

Fellbaum, C.: 1998, *WordNet: An Electronic Lexical Database*, MIT Press.

Florian, R., Cucerzan, S., Schafer, C. and Yarowsky, D.: 1998, Combining classifiers for word sense disambiguation, *Journal of Natural Language Engineering* **8**(4), 1–14.

Fujii, A., Inui, K., Tokunaga, T. and Tanaka, H.: 2001, Selective sampling for example-based word sense disambiguation, *Computational Linguistics* **24**(4), 573–598.

Gale, W. A. and Church, K. W.: 1991a, A program for aligning sentences in bilingual corpora, *Proceeding of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, USA, pp. 177–184.

Gale, W. and Church, K.: 1991b, Identifying Word Correspondences in Parallel Text, *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 152–157.

Gale, W. and Church, K.: 1993, A program for aligning sentences in bilingual corpora, *Computational Linguistics* **19**(1), 75–102.

Gale, W., Church, K. and Yarowsky, D.: 1992a, A method for disambiguating word senses in a large corpus, *Computers and the Humanities* **26**, 415–439.

Gale, W., Church, K. and Yarowsky, D.: 1992b, One sense per discourse, *Proceedings of the DARPA Speech and Natural Language Workshop*, New York, USA, pp. 233–237.

Gentner, D.: 1981, Some interesting differences between verbs and nouns, *Cognition and brain theory* **4**(2), 161–178.

Gentner, D.: 1982, Why nouns are learned before verbs: Linguistic relativity versus natural partitioning., *Language development: Vol. 2. Language, thought and culture* pp. 301–334.

Giménez, J. and Marquez, L.: 2007, Context-aware discriminative phrase selection for statistical machine translation, *Workshop on Statistical Machine Translation*, Prague.

Gliozzo, A., Magnini, B. and Strapparava, C.: 2004, Unsupervised domain relevance estimation for word sense disambiguation, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, pp. 380–387.

Goldberg, D.: 1989, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley.

Goldberg, D. and Deb, K.: 1991, A Comparative Analysis of Selection Schemes Used in Genetic Algorithms, *Foundations of Genetic Algorithms*, Morgan Kaufmann Publishers, San Mateo, California, USA, pp. 69–93.

Guo, W. and Diab, M.: 2010, COLEPL and COLSLM: An Unsupervised WSD Approach to Multilingual Lexical Substitution, Tasks 2 and 3 SemEval 2010, *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, Sweden, pp. 129–133.

Guo, Y., Che, W., Hu, Y., Zhang, W. and Liu, T.: 2007, HIT-IR-WSD: A WSD System for English Lexical Sample Task, *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Association for Computational Linguistics, Prague, pp. 165–168.

Haque, R., Naskar, S., van den Bosch, A. and Way, A.: 2011, Integrating source-language context into phrase-based statistical machine translation, *Machine Translation* **23**(3), 239–285.

Harris, Z. S.: 1968, *Mathematical structures of language*, Wiley.

Holland, J.: 1975, *Adaptation in natural and artificial Systems*, MIT Press.

Hoste, V., Hendrickx, I., Daelemans, W. and van den Bosch, A.: 2002, Parameter Optimization for Machine-Learning of Word Sense Disambiguation, *Natural Language Engineering, Special Issue on Word Sense Disambiguation Systems* **8**, 311–325.

Ide, N., Erjavec, T. and Tufiş, D.: 2002, Sense discrimination with parallel corpora. , *ACL-2002 Workhop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, pp. 54–60.

Ide, N. and Véronis, J.: 1998, Word sense disambiguation: The state of the art., *Computational Linguistics* **24**(1), 1–40.

Jensen, R. and Cornelis, C.: 2010, Fuzzy-rough instance selection, *Proceedings of the 19th International Conference on Fuzzy Systems (FUZZ-IEEE 2010)*, pp. 1776–1782.

Joachims, T.: 1998, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *Proceedings of the European Conference on Machine Learning*, Springer.

Kaplan, A.: 1955, An experimental study of ambiguity and context, *Mechanical Translation* **2**(2), 39–46.

Kay, M. and Röscheisen, M.: 1993, Text-Translation Alignment, *Computational Linguistics* **19**(1), 121–142.

Kilgarriff, A.: 1997, I don't believe in word senses., *Computers and the Humanities* **31**, 91–113.

Kilgarriff, A. and Palmer, M.: 2000, Special issue on SENSEVAL, *Computers and the Humanities* **34**(1–2), 91–113.

Klein, D., Toutanova, K., Ilhan, T., Kamvar, S. and Manning, C.: 2002, Combining heterogeneous classifiers for word-sense disambiguation, *Proceedings of the ACL workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, PA, pp. 74–80.

Koehn, P.: 2005, Europarl: a parallel corpus for statistical machine translation, *Tenth Machine Translation Summit*, Phuket, Thailand, pp. 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: 2007, Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180.

Koehn, P., Och, F. and Marcu, D.: 2003, Statistical Phrase-based translation, *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 48–54.

Landauer, T. and Dumais, S.: 1997, A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge., *Psychology Review* **104**, 211–240.

Landauer, T., Foltz, P. and Laham, D.: 1998, An introduction to latent semantic analysis, *Discourse processes* **25**, 259–284.

Landes, S., Leacock, C. and Tengi, R.: 1998, *Building Semantic Concordances*, MIT Press, Cambridge, MA, chapter Chapter 8, pp. 199–216.

Leacock, C., Chodorow, M. and Miller, G.: 1998, Using corpus statistics and WordNet relations for sense identification, *Computational Linguistics* **24**(1), 147–165.

Lefever, E. and Hoste, V.: 2010, SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation, *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, Uppsala, Sweden, pp. 15–20.

Lefever, E. and Hoste, V.: 2011, Examining the Validity of Cross-Lingual Word Sense Disambiguation, *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*, Tokyo, Japan.

Lefever, E., Hoste, V. and De Cock, M.: 2012, Discovering missing wikipedia inter-language links by means of cross-lingual word sense disambiguation, *in* N. C. C. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis (eds), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey.

Lesk, M.: 1986, Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone, *1986 ACM SIGDOC Conference*, Toronto, Canada, pp. 24–26.

Lopez de Lacalle, O.: 2009, *Domain-Specific Word Sense Disambiguation*, Phd, Lengoiaia eta Sistema Informatikoak Saila (UPV-EHU). Donostia 2009ko Abenduaren 14ean.

Lund, K. and Burgess, C.: 1996, Producing high-dimensional semantic spaces from lexical co-occurrence, *Behavior Research Methods, Instruments, and Computers* **28**, 203–208.

Macken, L., De Clercq, O. and Paulussen, H.: 2011, Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus, *Meta* **56**(2).

Mahapatra, L., Mohan, M., Khapra, M. and Bhattacharyya, P.: 2010, OWNS: Cross-lingual Word Sense Disambiguation Using Weighted Overlap Counts and Wordnet Based Similarity Measures, *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, Association for Computational Linguistics, Uppsala, Sweden, pp. 138–141–989.

Màrquez, L., Escudero, G., Martìnez, D. and Rigau, G.: 2006, Supervised corpus-based methods for WSD, *in* E. Agirre and P. Edmonds (eds), *Word Sense Disambiguation: Algorithms and Applications*, Eds Springer, New York, NY, pp. 167–216.

Martin, J., Mihalcea, R. and Pedersen, T.: 2005, Word alignment for languages with scarce resources, *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Ann Arbor, MI.

Martínez, D. and Agirre, E.: 2000, One sense per collocation and genre/topic variations, *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, Hong Kong, China, pp. 207–215.

Martínez, D., Agirre, E. and Màrquez, L.: 2002, Syntactic features for high precision word sense disambiguation, *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan.

McCarthy, D., Koeling, R., Weeds, J. and Carroll, J.: 2004, Finding predominant senses in untagged text, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp. 577–583.

McCarthy, D. and Navigli, R.: 2007, SemEval-2007 Task 10: English Lexical Substitution Task, *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, pp. 48–53.

Mihalcea, R.: 2002, Word sense disambiguation with pattern learning and automatic feature selection, *Natural Language Engineering, Special Issue on Word Sense Disambiguation Systems* **8**, 343–358.

Mihalcea, R.: 2004, Co-training and self-training for word sense disambiguation, *Proceedings of the Conference on Natural Language Learning (CoNLL)*, Boston, USA, pp. 33–40.

Mihalcea, R.: 2006, *Word Sense Disambiguation: Algorithms, Applications, and Trends*, Kluwer, chapter Knowledge Based Methods for Word Sense Disambiguation, pp. 107–131.

Mihalcea, R. and Moldovan, D.: 1999, An automatic method for generating sense tagged corpora, *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI)*, Orlando, USA, pp. 461–466.

Miller, R.: 1991, *Simultaneous Statistical Inference.*, Springer-Verlag.

Mitchell, M.: 1996, *An Introduction to Genetic Algorithms*, MIT Press.

Mohler, M. and Mihalcea, R.: 2008, Babylon parallel text builder: Gathering parallel texts for low-density languages, *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis and D. Tapias (eds), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco.

Navigli, R.: 2005, Semi-automatic extension of large-scale linguistic knowledge bases, *Proceedings of the 18th Florida Artificial Intelligence Research Society Conference (FLAIRS)*, Clearwater Beach, FL, pp. 548–553.

Navigli, R.: 2009, Word Sense Disambiguation: a Survey, *ACM Computing Surveys* **41**(2), 1–69.

Navigli, R. and Lapata, M.: 2007, Graph connectivity measures for unsupervised word sense disambiguation, *Proceedings of IJCAI*.

Navigli, R. and Ponzetto, S.: 2010, BabelNet: Building a very large multilingual semantic network, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 216–225.

Ng, H. and Lee, H.: 1996, Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics (ACL-96)*, pp. 40–47.

Ng, H., Wang, B. and Chan, Y.: 2003, Exploiting parallel texts for word sense disambiguation: An empirical study, *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, pp. 455–462.

Och, F. J. and Ney, H.: 2003, A systematic comparison of various statistical alignment models, *Computational Linguistics* **29**(1), 19–51.

Palmer, M.: 1998, Are wordnet sense distinctions appropriate for computational lexicons?, *SIGLEX-98, SENSEVAL*, Herstmonceux, Sussex, UK.

Papineni, K., Roukos, S., Ward, T. and W.-J., Z.: 2002, BLEU: a method for automatic evaluation of machine translation., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Pedersen, T. and Bruce, R.: 1997, Distinguishing word sense in untagged text, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, USA, pp. 197–207.

Pham, T., Ng, H. and Lee, W.: 2005, Word sense disambiguation with semi-supervised learning, *Proceedings of the 20th national Conference on Artificial Intelligence (AAAI)*, Pittsburgh, USA, pp. 1093–1098.

Pierce, J., Carroll, J., Hamp, E., Hays, D., Hockett, C., Oettinger, A. and Perlis, A.: 1966, Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, *Technical report*, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.

Procter, P.: 1978, *Longman Dictionary of Contemporary English*, Longman Group, London.

Quinlan, J.: 1993, *C4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA.

Rada, R., Mili, H., Bicknell, E. and Blettner, M.: 1989, Development and application of a metric on semantic nets, *IEEE Transactions on Systems, Man, and Cybernetics* **19**(1), 17–30.

Rehůřek, R. and Sojka, P.: 2010, Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50. http://is.muni.cz/publication/884893/en.

Resnik, P.: 1999, *Natural Language Processing Using Very Large Corpora*, Dordrecht: Kluwer Academic Publishers, chapter Disambiguating noun groupings with respect to WordNet senses, pp. 77–98.

Resnik, P. and Smith, N.: 2003, The web as a parallel corpus, *Computational Linguistics* **29**(3), 349–380.

Resnik, P. and Yarowsky, D.: 2000, Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation, *Natural Language Engineering* **5**(3), 113–133.

Rieger, C. and Small, S.: 1979, Word expert parsing, *Proceedings of the 6th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 723–728.

Roy, D. and Reiter, E.: 2005, Connecting Language to the World, *Artificial Intelligence* **167**, 1–12.

Salton, G.: 1989, *Automatic text processing: the transformation, analysis and retrieval of information by computer*, Addison Wesley.

Santy, J.: 2008, *L'utilisation de textes parallèles pour la désambiguisation sémantique: une étude contrastive des traductions tires d'Europarl et des traductions proposées par des dictionnaires bilingues*, Master's thesis, Ghent University College.

Schmid, H.: 1994, Probabilistic part-of-speech tagging using decision trees, *Proceedings of the International Conference on new methods in Language Processing*, Manchester, UK.

Schütze, H.: 1998, Automatic word sense discrimination, *Computational Linguistics* **24**(1), 97–123.

Shah, R., Lin, B., Gershman, A. and Frederking, R.: 2010, SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation, *Proceedings of the Second Workshop on African Language Technology (AFLAT 2010)*, Valletta, Malt.

Silberer, C. and Ponzetto, S.: 2010, UHD: Cross-Lingual Word Sense Disambiguation Using Multilingual Co-Occurrence Graphs, *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, Association for Computational Linguistics, Uppsala, Sweden, pp. 134–137.

Sinha, R. D., McCarthy, D. and Mihalcea, R.: 2009, SemEval-2010 Task 2: Cross-lingual Lexical Substitution, *Proceedings of the NAACL-HLT 2009 Workshop: SEW-2009 - Semantic Evaluations*, Boulder, Colorado.

Sinha, R. and Mihalcea, R.: 2007, Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity., *Proceedings of the IEEE International Conference on Semantic Computing (ICSC2007*, Irvine, CA, USA.

Specia, L., Das Graças, M., Nunes, V. and Stevenson, M.: 2005, Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation, *Proceedings of RANLP-05*, Borovets, pp. 525–531.

Specia, L., Nunes, M., Ribeiro, G. and Stevenson, M.: 2006, Multilingual versus Monolingual WSD., *Workshop Making Sense of Sense:*

Bringing Psycholinguistics and Computational Linguistics Together (EACL-2006), Trento, Italy, pp. 33–40.

Specia, L., Nunes, M. and Stevenson, M.: 2006, Translation Context Sensitive WSD, *Annual COnference of the European Association for Machine Translation (EAMT-2006)*, Oslo, Norway, pp. 227–232.

Specia, L., Nunes, M. and Stevenson, M.: 2007, Learning Expressive Models for Word Sense Disambiguation, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 41–48.

Specia, L., Sankaran, B. and Nunes, M.: 2008, N-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation., *Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, LNCS 4919, pp. 399–410.

Strapparava, C., Gliozzo, A. and Giuliano, C.: 2004, Pattern abstraction and term similarity for word sense disambiguation: IRST at Senseval-3, *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, pp. 229–234.

Stroppa, N., van den Bosch, A. and Way, A.: 2007, Exploiting source similarity for smt using context-informed features, *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*.

Suárez, A. and Palomar, M.: 2002, A maximum entropy-based word sense disambiguation system, *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, pp. 960–966.

Tufiş, D., Cristea, D. and Stamou, S.: 2004, BalkaNet: Aims, methods, results, and perspectives. A general overview, *Romanian Journal of Information Science and Technology* **7**(1–2), 9–43.

Tufiş, D., Ion, R. and Ide, N.: 2004, Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets, *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Association for Computational Linguistics, Geneva, Switzerland, pp. 1312–1318.

van Gompel, M.: 2010, UvT-WSD1: A Cross-Lingual Word Sense Disambiguation System, *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, Association for Computational Linguistics, Uppsala, Sweden, pp. 238–241.

Vapnik, V.: 1998, *Statistical Learning Theory*, John Wiley and Sons, New York.

Vasilescu, F., Langlais, P. and Lapalme, G.: 2004, Evaluating variants of the Lesk approach for disambiguating words, *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 633–636.

Verbiest, N., Cornelis, C. and Herrera, F.: 2012, Selección de Prototipos Basada en Conjuntos Rugosos Difusos., *Proceedings of XVI Congreso español sobre Tecnologías y Lógica (ESTYLF2012)*, pp. 638–643.

Vereeken, K.: 2012, *GA GRID. User Manual. Not published.*

Vickrey, D., Biewald, L., Teyssier, M. and Koller, D.: 2005, Word-sense disambiguation for machine translation, *Proceedings of EMNLP05*, pp. 771–778.

Vilariño Ayala, D., Balderas Posada, C., Pinto Avendaño, D., Rodríguez Hernández, M. and León Silverio, S.: 2010, FCC: Modeling Probabilities with GIZA++ for Task 2 and 3 of SemEval-2, *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010*, Association for Computational Linguistics, Uppsala, Sweden, pp. 112–116.

Vossen, P. (ed.): 1998, *EuroWordNet: a multilingual database with lexical semantic networks*, Kluwer Academic Publishers, Norwell, MA, USA.

Vossen, P., Görög, A., Izquierdo, R. and van den Bosch, A.: 2012, DutchSemCor: Targeting the ideal sense-tagged corpus, *in* N. Calzolari, K. Choukri, T. Declerck, M. Ugur Dogan, B. Maegaard, J. Mariani, J. Odijk and P. E. L. R. A. E. Piperidis, S. (eds), *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC2012)*, Istanbul, Turkey, pp. 584–589.

Vossen, P., Görög, A., Laan, F. and Van Gompel, M.: 2011, DutchSemCor: building a semantically annotated corpus for Dutch, *Proceedings of lectronic Lexicography in the 21st century: New Applications for new users (eLEX2011)*, Bled, Slovenia.

Vossen, P., Rigau, G., Alegira, I., Farwell, D. and Fuentes, M.: 2006, Meaningful results for information retrieval in the meaning project, *Proceedings of the 3rd Global WOrdnet COnference*, Jeju Islands, Korea.

Wasow, T., Perfors, A. and Beaver, D.: 2005, *Morphology and The Web of Grammar: Essays in Memory of Steven G. Lapointe*, CSLI Publications, chapter The Puzzle of Ambiguity.

Weaver, W.: 1949, *Machine translation of languages: fourteen essays*, Technology Press of the Massachusetts Institute of Technology, chapter Translation, pp. 15–23.

Wettschereck, D., Aha, D. W. and Mohri, T.: 1997, A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms, *Artificial Intelligence Review* **11**(1-5), 273–314.

White, A. and Liu, W.: 1994, Bias in information-based measures in decision tree induction, *Machine Learning* **15**(3), 321–329.

Wilks, Y.: 1975, Preference semantics, *in* E. Keenan (ed.), *Formal Semantics of Natural Language*, Cambridge University Press, Cambridge, U.K., pp. 329–348.

Wilks, Y. and Stevenson, M.: 1996, The grammar of sense: Is word sense tagging much more than part-of-speech tagging?, *Technical Report CS-96-05*, University of Sheffield, Sheffield, UK.

Yarowsky, D.: 1993, One sense per collocation, *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, USA, pp. 265–271.

Yarowsky, D.: 1995, Unsupervised word sense disambiguation rivaling supervised methods, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Cambridge, USA, pp. 189–196.

Yarowsky, D.: 2000, Hierarchical decision lists for word sense disambiguation., *Computers and the Humanities* **34**, 179–186.

Zipf, G. K.: 1949, *Human Behavior and the Principle of Least Effort*, Addison-Wesley (Reading MA).

APPENDIX A

Publications

This Appendix contains a list of all peer-reviewed journal and conference proceedings publications from the period 2007-2012.

- 2012
  - Macken, L., Lefever, E. and Hoste, V. *TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment.* Terminology. Accepted for publication.
  - Lefever, E., Hoste, V. and De Cock, M. *Discovering Missing Wikipedia Inter-language Links by means of Cross-lingual Word Sense Disambiguation.* Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey.
  - Mourisse, D., Lefever, E., Verbiest, N., Saeys, Y., De Cock, M. and Cornelis, C. *SBFC: An Efficient Feature Frequency-Based Approach to Tackle Cross-Lingual Word Sense Disambiguation.* Proceedings of the 15th International Conference on Text, Speech and Dialogue (TSD 2012), Brno, Czech Republic.
  - Lefever, E., Hoste, V. and De Cock, M. *Parallel corpora make sense: bypassing the knowledge acquisition bottleneck for WSD.* submitted to an A1 Journal.

- 2011
  - Lefever, E. and Hoste, V. *Computational Linguistics in the Netherlands Journal.*
  - Lefever, E., Macken, L. and Hoste, V. *Taal- en spraaktechnologie: een stand van zaken.* Over Taal, 50 (1), 20-22. UGA.
  - Lefever, E. and Hoste, V. *Examining the Validity of Cross-Lingual Word Sense Disambiguation.* Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011). Tokyo, Japan.
  - Lefever, E., Hoste, V. and De Cock, M. *ParaSense or how to use Parallel Corpora for Word Sense Disambiguation.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA.
  - Lefever, E. and Hoste, V. *An Evaluation and Possible Improvement Path for Current SMT Behavior on Ambiguous Nouns.* Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, ACL HLT 2011. Portland, Oregon, USA.
  - Lefever, E., Hoste, V. and De Cock, M. *Using Parallel Corpora for Word Sense Disambiguation.* Proceedings of the 23rd Benelux Conference on Artificial Intelligence. Gent, Belgium.
- 2010
  - Lefever, E., Fayruzov, T., Hoste, V. and De Cock, M. *Clustering Web People Search Results using Fuzzy Ants.* Information Sciences, 180 (17), 3192-3209, SCI2011:2.833. Elsevier Science Inc., New York, USA.
  - Hoste, V., Vanopstal, K., Lefever, E. and Delaere, I. *Classification-based scientific term detection in patient information.* Terminology, 16 (1), 1-29, SCI2010:0.4. John Benjamins Publishing Company, Amsterdam, Netherlands.
  - Lefever, E. and Hoste, V. *SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation.* Proceedings of the ACL Workshop on Semantic Evaluations (SemEval- 2010). Uppsala, Sweden.
  - Lefever, E. and Hoste, V. *Construction of a Benchmark Data Set for Cross-Lingual Word Sense Disambiguation.* In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds.), Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association, Valletta, Malta.

- 2009
  - Lefever, E., Macken, L. and Hoste, V. *Language-independent bilingual terminology extraction from a multilingual parallel corpus.* Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. The Association for Computational Linguistics, Athens, Greece.
  - Lefever, E. and Hoste, V. *SemEval-2 Task 3: Cross-lingual Word Sense Disambiguation.* Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009). Boulder, Colorado.
  - Lefever, E., Fayruzov, T., Hoste, V. and De Cock, M. *Fuzzy Ants Clustering for Web People Search.* Proceedings of WePS2 (2nd Web People Search Evaluation Workshop), workshop at WWW2009. Madrid, Spain.
- 2008
  - Macken, L. and Lefever, E. *Translational Equivalence in Statistical Machine Translation or Meaning as co-occurrence.* In S. Vandepitte (ed.), Looking for meaning: Methodological issues in translation studies (7): Linguistica Antverpiensia, New Series, 193-208. University Press Antwerp, Antwerp, Belgium.
  - Macken, L., Lefever, E. and Hoste, V. *Linguistically-based subsentential alignment for terminology extraction from a bilingual automotive corpus.* Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, UK.
  - Hoste, V., Lefever, E., Vanopstal, K. and Delaere, I. *Learning-based Detection of Scientific Terms in Patient Information.* In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias (eds.), Proceedings of the Sixth Conference on International Language Resources and Evaluation (LREC'08). European Language Resources Association, Marrakech, Morocco.
- 2007
  - Lefever, E., Fayruzov, T. and Hoste, V. *A combined classification and clustering approach for web people disambiguation.* Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic.
  - Hoste, V., Vanopstal, K. and Lefever, E. *The Automatic Detection of Scientific Terms in Patient Information.* Proceedings of the Workshop on Acquisition and Management of Multilingual Lexicons. Borovetz, Bulgaria.

APPENDIX B

Test set for the ambiguous focus word *post*

This Appendix contains the list of 50 English test sentences selected for the ambiguous focus word *post*.

1. And there'd be nothing easier than popping a cheque in the post to them, or sending them some cast-off woollies, or ladling some soup into them, or offering them advice.
2. A take away and caff at the South end of the main street near sign posts in miles and furlongs, also unique in our experience it serves the finest battered cod we can ever recall eating.
3. Almost all freeholders would have taken strong exception to the suggestion that their support had been purchased, and this was as true of those gentlemen who had recently obtained posts for themselves or their sons as it was for those who had been less fortunate.
4. American diplomats were so appalled by Mr Zappala's nomination that they leaked to the Spanish press a copy of the misnamed competence certificate, which is sent to the Senate, which has the final say on ambassadorial posts.
5. And normally they have had not less than three years' experience in a junior or first line management post directly concerned with the provision of food or beverages or with the provision and servicing of accommodation.

6. As she walked from the post office empty handed, she pretended to be an American tourist, one who could leave in a few days, one for whom the trip would soon be nothing more than a few travel stories and postcards.
7. At that time the settlement of Cape Town, which had been just a small staging post on the voyage between the Netherlands and the East Indies, was beginning to grow, but although the English had occupied the Cape in 1806, the white population remained mainly Dutch until the arrival of many new settlers from Britain in 1820.
8. Attacks on Baltic republics' border posts by Soviet troops continued during June, although OMON (the special Interior Ministry troops) officers on June 26 denied charges of involvement in more than 20 such attacks.
9. During one of his many civil disobedience campaigns, which ranged from non-payment of taxes to blocking border posts into the US, he led a hunger strike in protest at alleged vote rigging in the presidential race.
10. Finally, after the third month passed with no rent received, Debbie phoned the tenants, asked them to pay her directly, and sent Landlords a recorded-delivery letter, which was returned, and finally one by ordinary post stating that Landlords were no longer employed as the managing agents.
11. For many journals the referee receives a typescript in the post, sends comments off, and receives no other feedback than to see the article appearing in print later.
12. Government security laws, passed to tackle terrorism, caused controversy especially when left-wing activists were excluded from civil service posts.
13. He was also, from 1444 to 1455, the first holder of the newly created post of master (or warden) of the children of the chapel royal, the household chapel of the English kings.
14. He was nearing 60, with warnings of ill health, but immediately undertook with his departmental colleagues an imaginative restructuring of anatomical teaching, introducing new techniques and persuading the university to fund new posts.
15. If there is anyone in your parish whom you think might be interested and suitable, perhaps you would draw their attention to these posts and encourage them to write to me for application forms and job description.

16. In fact when you enter a large mental hospital you do enter a special kind of world, a village with streets and sign posts, and usually, for in such places, everyone has plenty of time on their hands, no shortage of people willing to direct you to where you want to go.

17. In it, 1,500 individuals were asked: "If you were to buy life assurance, which of the following methods, if any, would you choose - from a salesman in your own home, from a broker in his office, by post (either from a newspaper advertisement or from a mailshot), from a bank, or from a building society."

18. In total, in the "no training officer" group there was less evidence of team decisions, less evidence of decision-making in the frame-work of existing plans / priorities, more mention of need to have local authority approval, and more mention of relevancy to post as a criterion.

19. It may be that "post" will need progressive reinterpretation to include telex, facsimile transmission and other forms of "electronic mail" but international conventions appear not to have explored these possibilities thus far.

20. Local electricity, gas, telephone and post office counter services all satisfied more than 70% of their customers, with the electricity board coming out on top with an 85% approval rating.

21. Meantime the British were trying to evaluate the seriousness of American proposals for decolonization or for international trusteeships as staging posts on the road from colony to self-government and independence.

22. Next morning Sophie searched through her post anxiously and was beginning to despair when, at the bottom of the pile, she found an unstamped envelope marked "personal" and underlined.

23. Not surprisingly the variations in salary made transfers almost as common an object of solicitation as first appointments and promotions, but requests for a change of post might, however, be occasioned by more significant matters than the possibility of attracting a few additional pounds in salary.

24. On the contrary, immense efforts were made to secure a military treaty with independent India, the Chiefs of Staff having given it as their opinion that India's manpower resources and location as a staging post made a military understanding with her "essential from the aspect of imperial strategy".

25. Selection for senior posts is therefore more important than training; if people who display the appropriate qualities are placed in the correct context then they will flourish.

26. Senior Tories forecast there would be radical ministerial changes, with a "50-50" chance that Mr Lamont, the Chancellor, could be switched to another senior post, possibly Defence or the Home Office.
27. Some Afghanis - their thinnish ranks supplemented by eager non-veteran Algerians who copy their dress and swagger - proclaim themselves a paramilitary wing of Algeria's now-banned Islamic Salvation Front; they attacked a military post near the Libyan border in late 1991.
28. Some of the culture-based arguments clearly have political ingredients, but one strictly political argument, or rather political-system argument, is equally applicable to any new political movement (whether left, right or centrist) seeking long-term viability - that the British first-past-the-post electoral system has features that make it extremely difficult for a new party to "break through".
29. That is likely to be British-style first-past-the-post rules for both houses of parliament - the system that 82% of voters backed in the referendum, though their only alternative was the existing system.
30. The alternative would be to ensure that top civil service posts were held by political supporters of the government: such people would have to lose their jobs if ministers lost office.
31. The amount of such mail handled by the Post Office increased threefold from 1975 to 1987, reaching 1,626 million items and making up 7.7 percent of total national expenditure on advertising.
32. The offer coincided with a move to separate the Cheltenham Schools of Architecture and Art, with which he disagreed, so he applied for the post and succeeded despite his now apparent and debilitating illness.
33. The post has to be sorted, letters attached to previous correspondence and any mail requiring attention dealt with or distributed to the appropriate department.
34. The post is, meanwhile, losing out to electronic "mail" systems and a lengthy postal strike last year served only to hammer another nail in the coffin.
35. The raids had been going on for months, but had become increasingly violent: by mid-May they involved border posts being strafed with bullets or set on fire.
36. The train has been painstakingly restored to all its former glory, even down to its traditional wood-burning stoves and original post box.
37. There are conditions you must meet in all services, such as using proper packaging, addressing things correctly and not sending certain things in the post.

38. There were not enough teachers, those there were harassed almost beyond bearing and driven from pillar to post, and no one ever seemed to know who Jasper was, still less remember his name.

39. There were probably a number of them in Britain, but they must have left very little archaeological trace, needing only a flat piece of ground, probably outside the towns, marked out with wood rails and probably a stand at the finishing post for the local dignitaries and wealthy citizens.

40. They also learn the position of the wetlands that provide them with crucial staging posts where they can feed and rest before starting on the next part of their journey.

41. To his professional duties he had added the role of Departmental Safety Officer, and it was this experience which took him in 1977 to Imperial College in the new post of College Safety Director.

42. Town's drugs are often made in Britain, flown to the Far East or some other convenient staging post and then brought back on the next night - to be sold more cheaply than if they had never left Britain.

43. We are committed to maintaining a nation-wide letter service with delivery to every address in the United Kingdom, within a uniform structure of prices, and with a nation-wide network of post offices.

44. Whenever he is stopped at military posts now, he produces his book and says, "Look, these are all my friends who will protest if I am arrested again".

45. Where the year-group was large, year leadership became a post of some importance, counterbalancing the cross-school role of curriculum leader, and introducing potential tensions over who was responsible for what.

46. With that background, John went to Peggy van Praagh, ballet mistress of the young company, and asked whether he could have the vacant post provided that he could manage those lifts.

47. A notification referred to in paragraph (a) shall be deemed to be received by the Community and the States seven days after the date of the transmission by registered post of the notification by the Agency to the Community and the States.

48. Each Member State shall communicate to the other Member States and the Commission the list of frontier posts to be used for the introduction of bovine animals and swine into its territory.

49. If you order goods as a private individual from mail order advertisements in this magazine and pay by post in advance of delivery, what

Personal Computer will consider you for compensation if the advertiser should become subject to bankruptcy or go into liquidation.

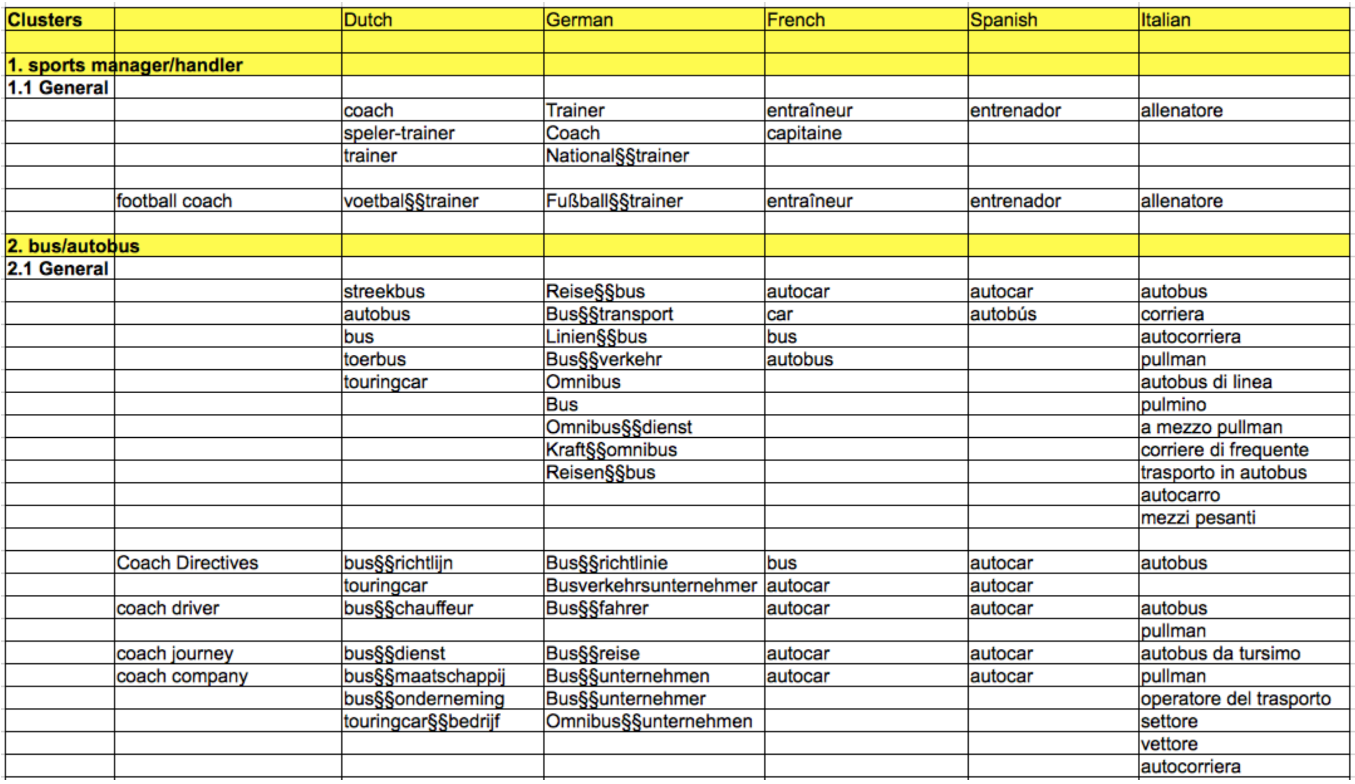
50. Management education can be broadly defined as being predominantly concerned with the training of men and women for the higher-level supervisory and decision-taking posts.

APPENDIX C

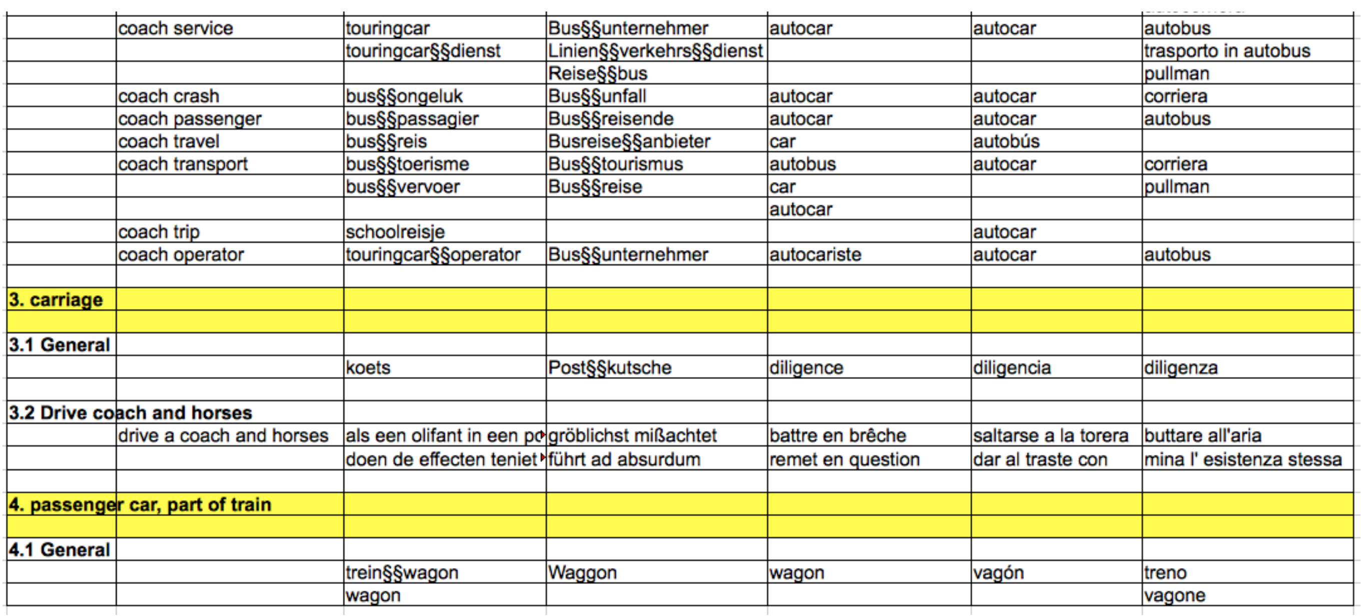
Clustering tables

This Appendix contains the clustering tables that were created for the annotation of the test sentences for three ambiguous words: (1) coach, (2) execution and (3) figure. A detailed description of the manual clustering process can be found in Section 5.1.2. The clusters are organized in different levels: the top level always reflects the main meanings of the word, whereas the lower levels correspond to finer sense distinctions. It is important to notice that there is no horizontal correspondence between the translations of the different languages: all translations in a given language that are listed next to (1) a cluster definition or (2) an English compound or multiword expression containing the ambiguous word, are to be considered as belonging to the set of translations that correspond to the English entry.

C.1 Coach



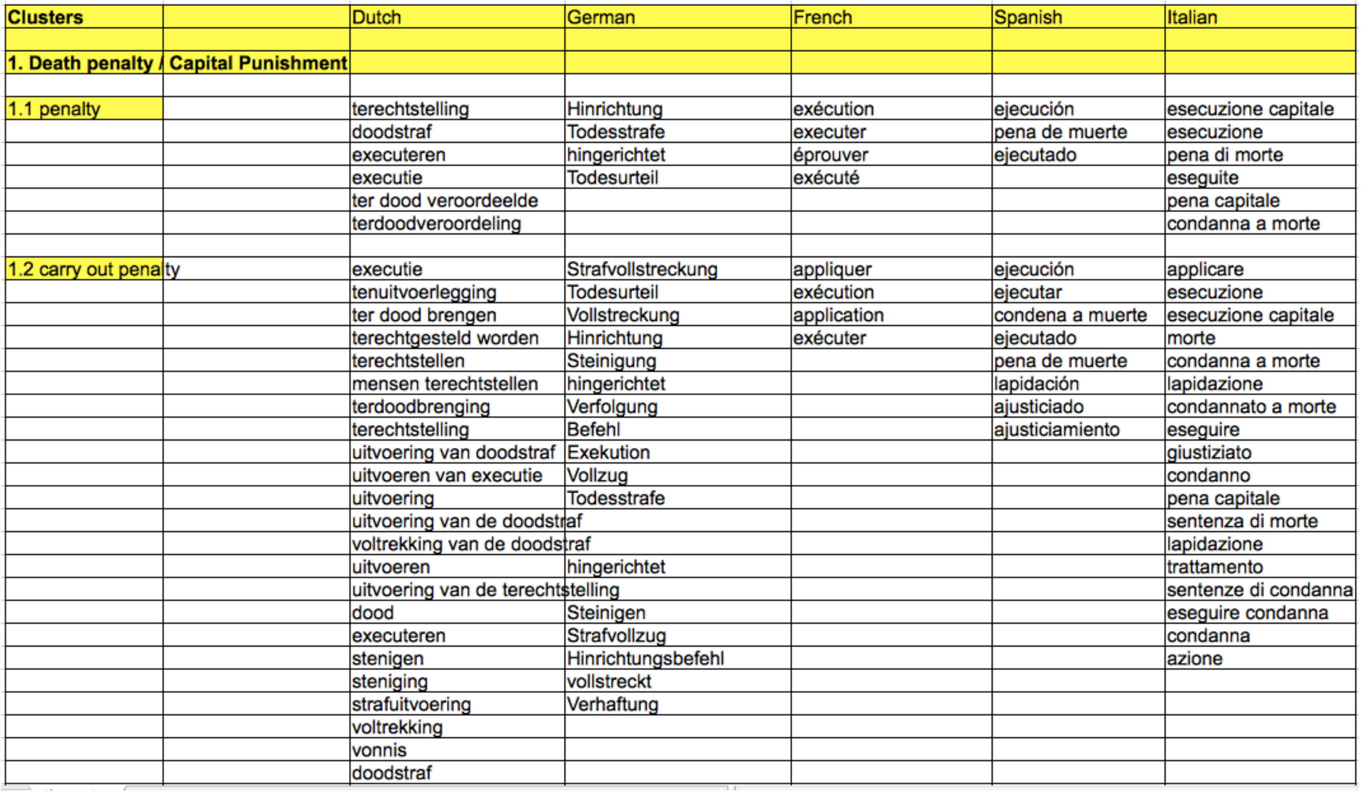
Clusters		Dutch	German	French	Spanish	Italian
1. sports manager/handler						
1.1 General						
		coach	Trainer	entraîneur	entrenador	allenatore
		speler-trainer	Coach	capitaine		
		trainer	National§§trainer			
	football coach	voetbal§§trainer	Fußball§§trainer	entraîneur	entrenador	allenatore
2. bus/autobus						
2.1 General						
		streekbus	Reise§§bus	autocar	autocar	autobus
		autobus	Bus§§transport	car	autobús	corriera
		bus	Linien§§bus	bus		autocorriera
		toerbus	Bus§§verkehr	autobus		pullman
		touringcar	Omnibus			autobus di linea
			Bus			pulmino
			Omnibus§§dienst			a mezzo pullman
			Kraft§§omnibus			corriere di frequente
			Reisen§§bus			trasporto in autobus
						autocarro
						mezzi pesanti
	Coach Directives	bus§§richtlijn	Bus§§richtlinie	bus	autocar	autobus
		touringcar	Busverkehrsunternehmer	autocar	autocar	
	coach driver	bus§§chauffeur	Bus§§fahrer	autocar	autocar	autobus
						pullman
	coach journey	bus§§dienst	Bus§§reise	autocar	autocar	autobus da tursimo
	coach company	bus§§maatschappij	Bus§§unternehmen	autocar	autocar	pullman
		bus§§onderneming	Bus§§unternehmer			operatore del trasporto
		touringcar§§bedrijf	Omnibus§§unternehmen			settore
						vettore
						autocorriera



	coach service	touringcar	Bus§§unternehmer	autocar	autocar	autobus
		touringcar§§dienst	Linien§§verkehrs§§dienst			trasporto in autobus
			Reise§§bus			pullman
	coach crash	bus§§ongeluk	Bus§§unfall	autocar	autocar	corriera
	coach passenger	bus§§passagier	Bus§§reisende	autocar	autocar	autobus
	coach travel	bus§§reis	Busreise§§anbieter	car	autobús	
	coach transport	bus§§toerisme	Bus§§tourismus	autobus	autocar	corriera
		bus§§vervoer	Bus§§reise	car		pullman
				autocar		
	coach trip	schoolreisje			autocar	
	coach operator	touringcar§§operator	Bus§§unternehmer	autocariste	autocar	autobus
3. carriage						
3.1 General						
		koets	Post§§kutsche	diligence	diligencia	diligenza
3.2 Drive coach and horses						
	drive a coach and horses	als een olifant in een po	gröblichst mißachtet	battre en brêche	saltarse a la torera	buttare all'aria
		doen de effecten teniet	führt ad absurdum	remet en question	dar al traste con	mina l' esistenza stessa
4. passenger car, part of train						
4.1 General						
		trein§§wagon	Waggon	wagon	vagón	treno
		wagon				vagone

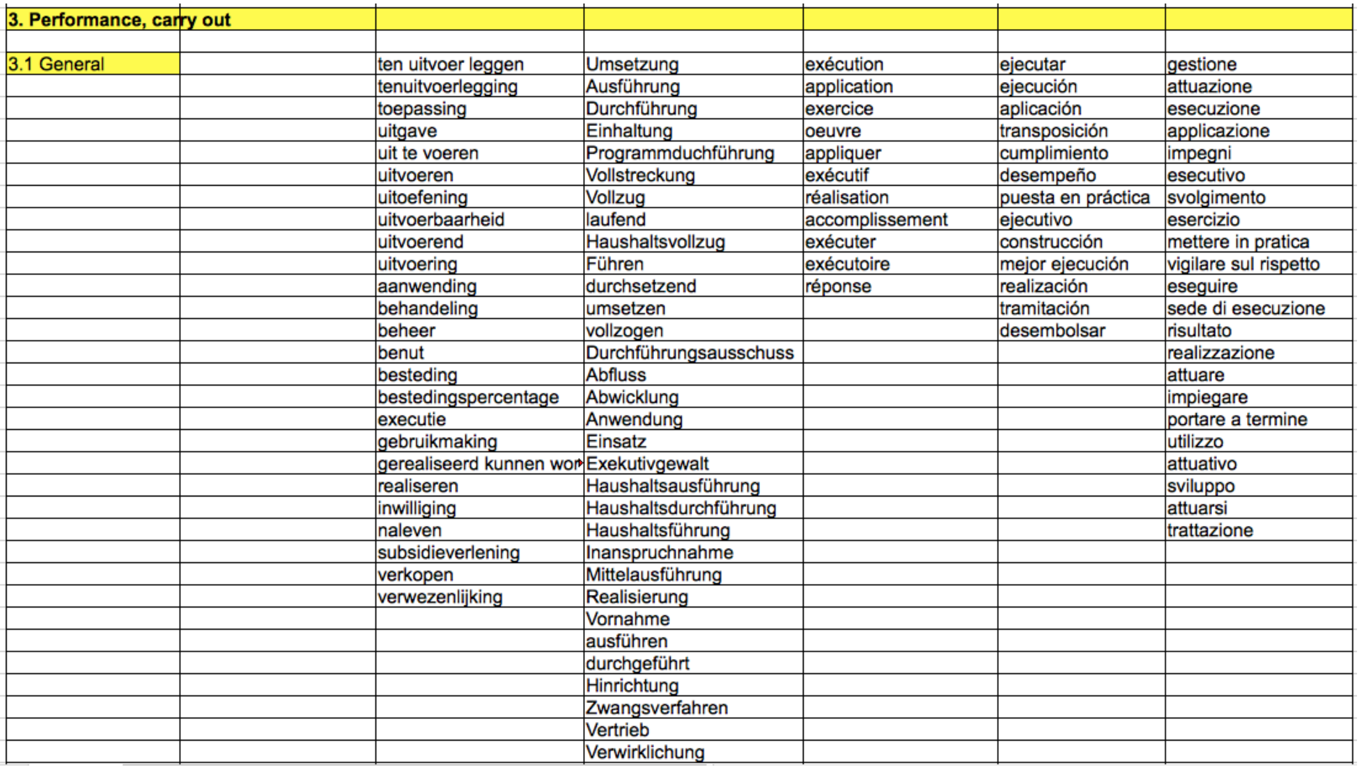
Figure C.1: Clustering file for the test word *coach*.

C.2 Execution

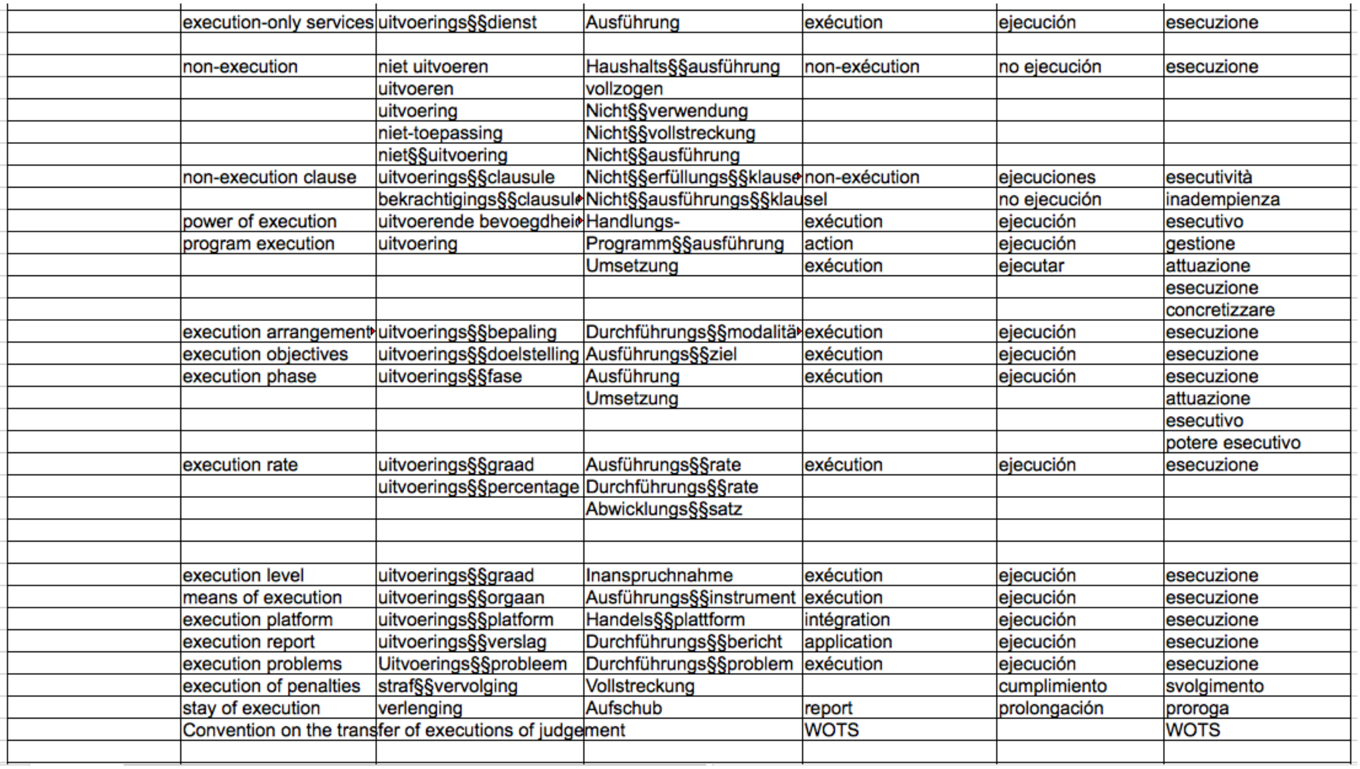


Clusters		Dutch	German	French	Spanish	Italian
1. Death penalty / Capital Punishment						
1.1 penalty		terechtstelling	Hinrichtung	exécution	ejecución	esecuzione capitale
		doodstraf	Todesstrafe	executer	pena de muerte	esecuzione
		executeren	hingerichtet	éprouver	ejecutado	pena di morte
		executie	Todesurteil	exécuté		eseguite
		ter dood veroordeelde				pena capitale
		terdoodveroordeling				condanna a morte
1.2 carry out penalty		executie	Strafvollstreckung	appliquer	ejecución	applicare
		tenuitvoerlegging	Todesurteil	exécution	ejecutar	esecuzione
		ter dood brengen	Vollstreckung	application	condena a muerte	esecuzione capitale
		terechtgesteld worden	Hinrichtung	exécuter	ejecutado	morte
		terechtstellen	Steinigung		pena de muerte	condanna a morte
		mensen terechtstellen	hingerichtet		lapidación	lapidazione
		terdoodbrenging	Verfolgung		ajusticiado	condannato a morte
		terechtstelling	Befehl		ajusticiamiento	eseguire
		uitvoering van doodstraf	Exekution			giustiziato
		uitvoeren van executie	Vollzug			condanno
		uitvoering	Todesstrafe			pena capitale
		uitvoering van de doodstraf				sentenza di morte
		voltrekking van de doodstraf				lapidazione
		uitvoeren	hingerichtet			trattamento
		uitvoering van de terechtstelling				sentenze di condanna
		dood	Steinigen			eseguire condanna
		executeren	Strafvollzug			condanna
		stenigen	Hinrichtungsbefehl			azione
		steniging	vollstreckt			
		strafuitvoering	Verhaftung			
		voltrekking				
		vonnis				
		doodstraf				

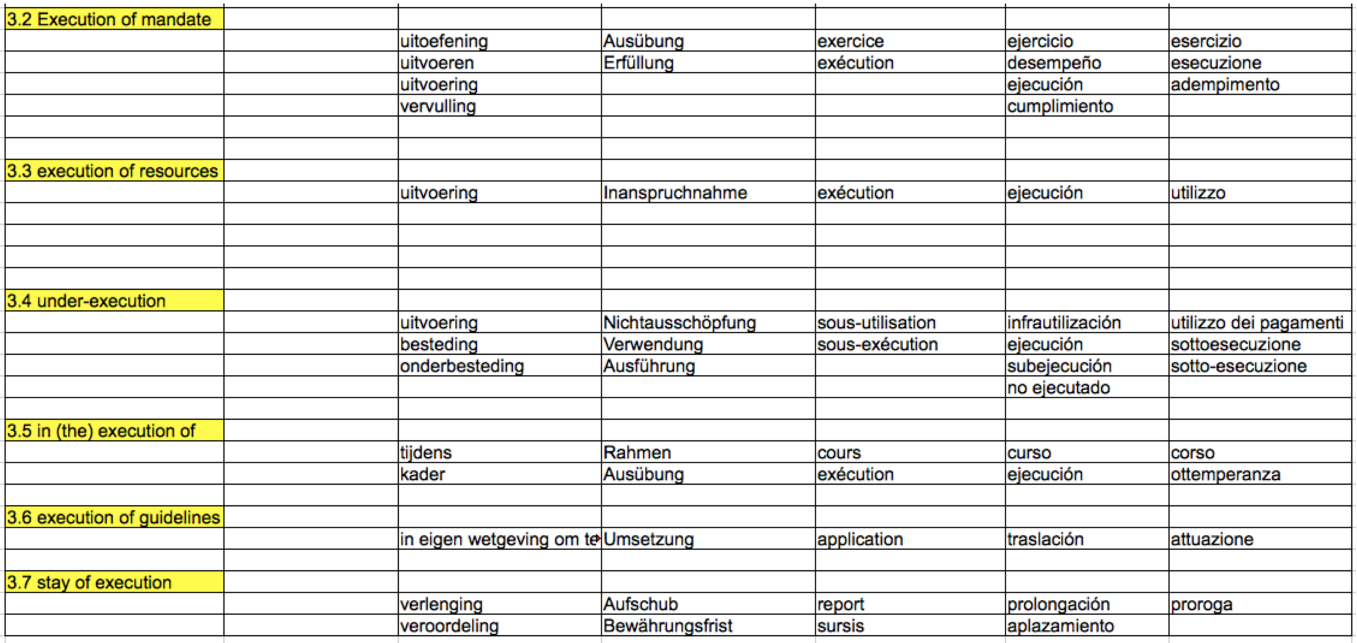
		doodsvonnissen				
		doodvonnissen				
		doodvonnissen voltrekken				
	stay of execution	tenuitvoerlegging	Todesurteil	exécution	ejecución	esecuzione
		executie	Vollstreckung	sursis	suspensión	sentenza
		uitstel	Aufschub	grâce	prórroga	rinvio
		vervallenverklaring	Ausführung	prise	aplazamiento	sospensiva
			Galgenfrist		suspender	applicazione
			Vollstreckungsaufschub		suspensión	indulto
			Aussetzung			
	execution order	executiebevel	Hinrichtungsbefehl	exécution	ejecución	esecuzione
	executions capital	terechtstelling	Hinrichtungshauptstadt	exécution	ejecución	esecuzione
	place of execution		Schaffot	exécution	ejecución	patibolo
2. Murder						
		executie	Hinrichtung	exécution	ejecución	esecuzione
		terechtstelling	Tötung	exécuter	matar	passato alle armi
		executeren	Ermordung	liquidation	fusilamiento	uccidere
		moord	hingerichtet			omicidio
		vermoorden	ermordet			
		doodschieten	Exekution			
		afrekening	tötet			
		doden	Liquidierung			
		fusillade				
		liquidatie				
	mass execution	massa-executie	Massen§§erschießung	exécution	fusilamiento	esecuzione
		massa§§executie	Massen§§exekution	exécuter	ejecución	sterminato
		executie	Massen§§hinrichtung			
	executions of hostages	executie	Geiseler§§schießung	exécution	ejecución	esecuzione



3. Performance, carry out						
3.1 General		ten uitvoer leggen	Umsetzung	exécution	ejecutar	gestione
		tenuitvoerlegging	Ausführung	application	ejecución	attuazione
		toepassing	Durchführung	exercice	aplicación	esecuzione
		uitgave	Einhaltung	oeuvre	transposición	applicazione
		uit te voeren	Programmduchführung	appliquer	cumplimiento	impegni
		uitvoeren	Vollstreckung	exécutif	desempeño	esecutivo
		uitoefening	Vollzug	réalisation	puesta en práctica	svolgimento
		uitvoerbaarheid	laufend	accomplissement	ejecutivo	esercizio
		uitvoerend	Haushaltsvollzug	exécuter	construcción	mettere in pratica
		uitvoering	Führen	exécutoire	mejor ejecución	vigilare sul rispetto
		aanwending	durchsetzend	réponse	realización	eseguire
		behandeling	umsetzen		tramitación	sede di esecuzione
		beheer	vollzogen		desembolsar	risultato
		benut	Durchführungsausschuss			realizzazione
		besteding	Abfluss			attuare
		bestedingspercentage	Abwicklung			impiegare
		executie	Anwendung			portare a termine
		gebruikmaking	Einsatz			utilizzo
		gerealiseerd kunnen wor	Exekutivgewalt			attuativo
		realiseren	Haushaltsausführung			sviluppo
		inwilliging	Haushaltsdurchführung			attuarsi
		naleven	Haushaltsführung			trattazione
		subsidieverlening	Inanspruchnahme			
		verkopen	Mittelausführung			
		verwezenlijking	Realisierung			
			Vornahme			
			ausführen			
			durchgeführt			
			Hinrichtung			
			Zwangsverfahren			
			Vertrieb			
			Verwirklichung			



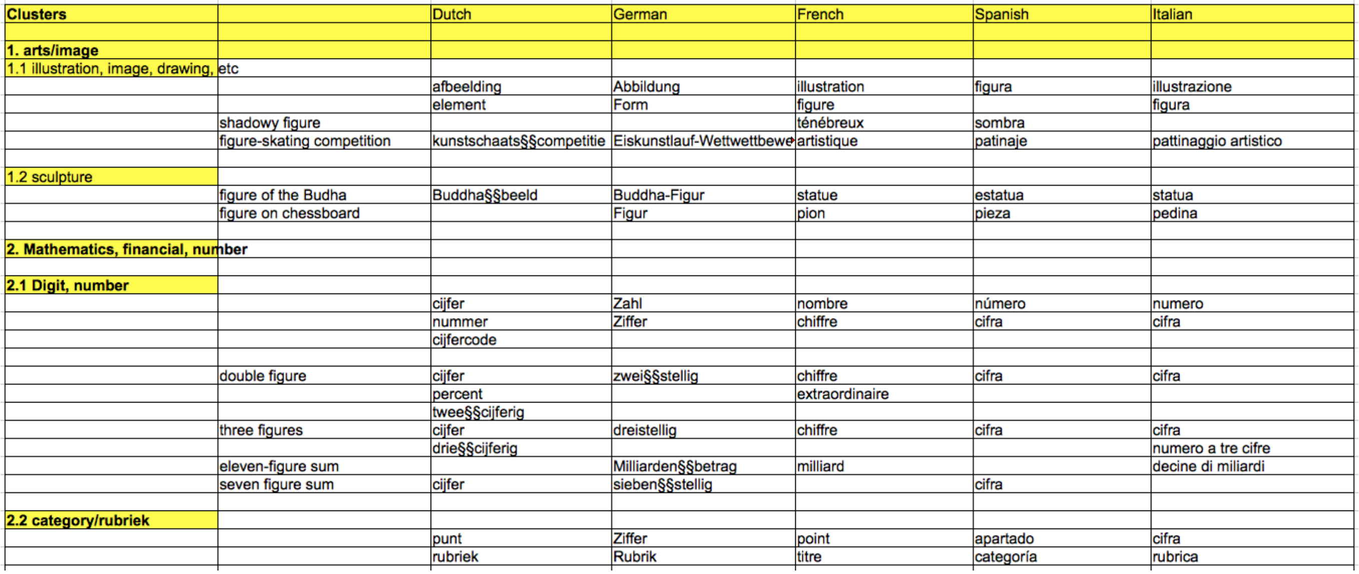
	execution-only services	uitvoerings§§dienst	Ausführung	exécution	ejecución	esecuzione
	non-execution	niet uitvoeren	Haushalts§§ausführung	non-exécution	no ejecución	esecuzione
		uitvoeren	vollzogen			
		uitvoering	Nicht§§verwendung			
		niet-toepassing	Nicht§§vollstreckung			
		niet§§uitvoering	Nicht§§ausführung			
	non-execution clause	uitvoerings§§clausule	Nicht§§erfüllungs§§klause	non-exécution	ejecuciones	esecutività
		bekrachtigings§§clausul	Nicht§§ausführungs§§klausel		no ejecución	inadempienza
	power of execution	uitvoerende bevoegdhei	Handlungs-	exécution	ejecución	esecutivo
	program execution	uitvoering	Programm§§ausführung	action	ejecución	gestione
			Umsetzung	exécution	ejecutar	attuazione
						esecuzione
						concretizzare
	execution arrangement	uitvoerings§§bepaling	Durchführungs§§modalitä	exécution	ejecución	esecuzione
	execution objectives	uitvoerings§§doelstelling	Ausführungs§§ziel	exécution	ejecución	esecuzione
	execution phase	uitvoerings§§fase	Ausführung	exécution	ejecución	esecuzione
			Umsetzung			attuazione
						esecutivo
						potere esecutivo
	execution rate	uitvoerings§§graad	Ausführungs§§rate	exécution	ejecución	esecuzione
		uitvoerings§§percentage	Durchführungs§§rate			
			Abwicklungs§§satz			
	execution level	uitvoerings§§graad	Inanspruchnahme	exécution	ejecución	esecuzione
	means of execution	uitvoerings§§orgaan	Ausführungs§§instrument	exécution	ejecución	esecuzione
	execution platform	uitvoerings§§platform	Handels§§plattform	intégration	ejecución	esecuzione
	execution report	uitvoerings§§verslag	Durchführungs§§bericht	application	ejecución	esecuzione
	execution problems	Uitvoerings§§probleem	Durchführungs§§problem	exécution	ejecución	esecuzione
	execution of penalties	straf§§vervolging	Vollstreckung		cumplimiento	svolgimento
	stay of execution	verlenging	Aufschub	report	prolongación	proroga
	Convention on the transfer of executions of judgement			WOTS		WOTS



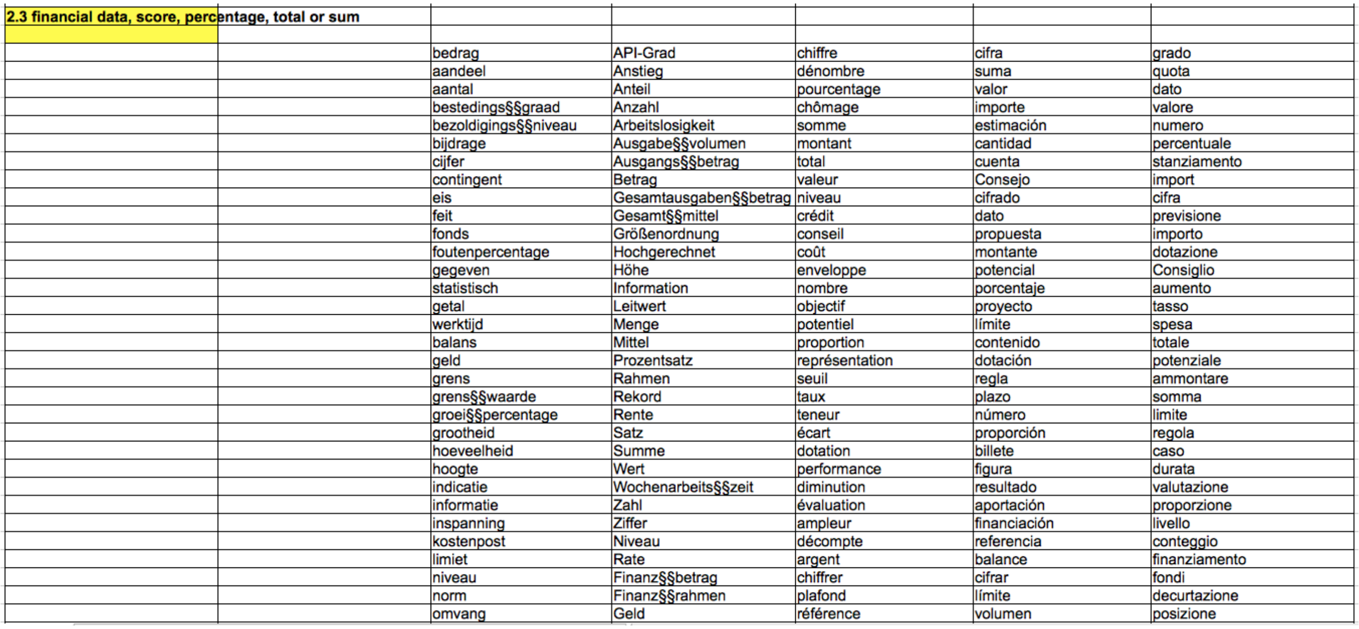
3.2 Execution of mandate						
		uitoefening	Ausübung	exercice	ejercicio	esercizio
		uitvoeren	Erfüllung	exécution	desempeño	esecuzione
		uitvoering			ejecución	adempimento
		vervulling			cumplimiento	
3.3 execution of resources						
		uitvoering	Inanspruchnahme	exécution	ejecución	utilizzo
3.4 under-execution						
		uitvoering	Nichtausschöpfung	sous-utilisation	infrautilización	utilizzo dei pagamenti
		besteding	Verwendung	sous-exécution	ejecución	sottoesecuzione
		onderbesteding	Ausführung		subejecución	sotto-esecuzione
					no ejecutado	
3.5 in (the) execution of						
		tijdens	Rahmen	cours	curso	corso
		kader	Ausübung	exécution	ejecución	ottemperanza
3.6 execution of guidelines						
		in eigen wetgeving om te	Umsetzung	application	traslación	attuazione
3.7 stay of execution						
		verlenging	Aufschub	report	prolongación	proroga
		veroordeling	Bewährungsfrist	sursis	aplazamiento	

Figure C.2: Clustering file for the test word *execution*.

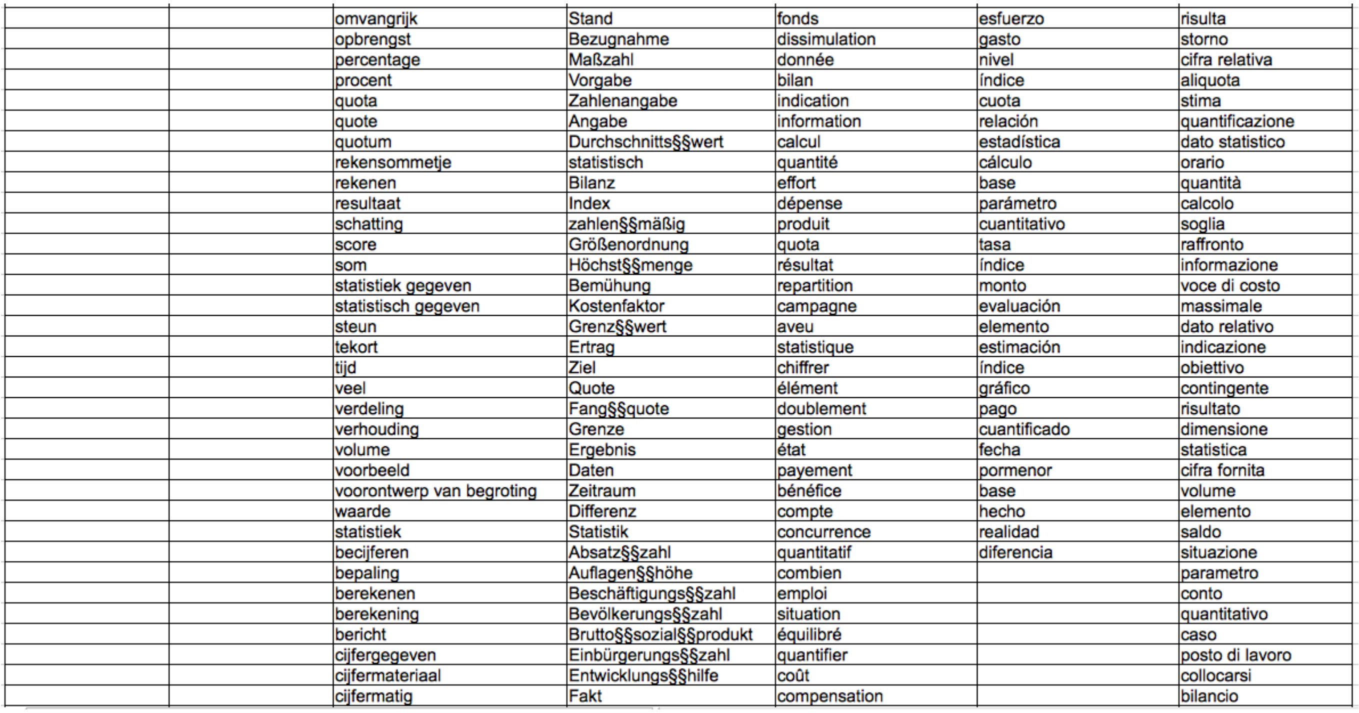
C.3 Figure



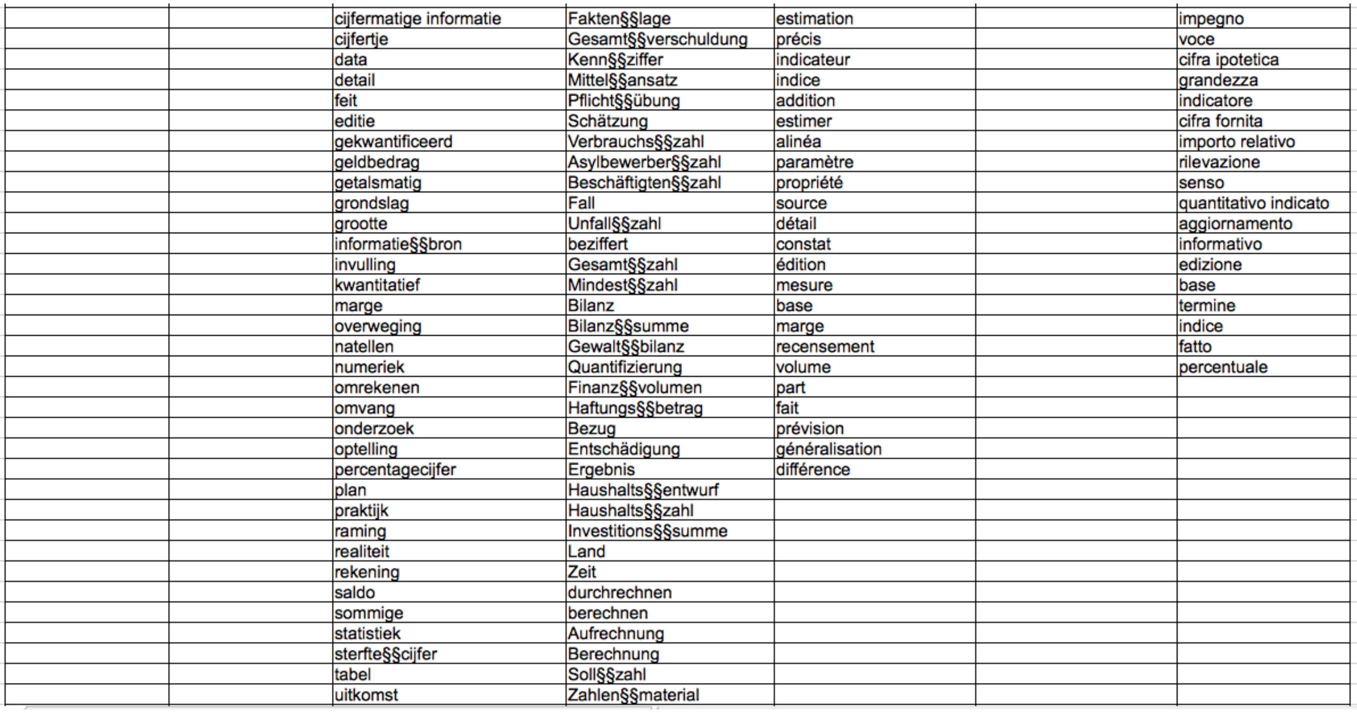
Clusters		Dutch	German	French	Spanish	Italian
1. arts/image						
1.1 illustration, image, drawing, etc						
		afbeelding	Abbildung	illustration	figura	illustrazione
		element	Form	figure		figura
	shadowy figure			ténébreux	sombra	
	figure-skating competition	kunstschaats§§competitie	Eiskunstlauf-Wettwettbewe	artistique	patinaje	pattinaggio artistico
1.2 sculpture						
	figure of the Budha	Buddha§§beeld	Buddha-Figur	statue	estatua	statua
	figure on chessboard		Figur	pion	pieza	pedina
2. Mathematics, financial, number						
2.1 Digit, number						
		cijfer	Zahl	nombre	número	numero
		nummer	Ziffer	chiffre	cifra	cifra
		cijfercode				
	double figure	cijfer	zwei§§stellig	chiffre	cifra	cifra
		percent		extraordinaire		
		twee§§cijferig				
	three figures	cijfer	dreistellig	chiffre	cifra	cifra
		drie§§cijferig				numero a tre cifre
	eleven-figure sum		Milliarden§§betrag	milliard		decine di miliardi
	seven figure sum	cijfer	sieben§§stellig		cifra	
2.2 category/rubriek						
		punt	Ziffer	point	apartado	cifra
		rubriek	Rubrik	titre	categoría	rubrica



2.3 financial data, score, percentage, total or sum						
		bedrag	API-Grad	chiffre	cifra	grado
		aandeel	Anstieg	dénombre	suma	quota
		aantal	Anteil	pourcentage	valor	dato
		bestedings§§graad	Anzahl	chômage	importe	valore
		bezoldigings§§niveau	Arbeitslosigkeit	somme	estimación	numero
		bijdrage	Ausgabe§§volumen	montant	cantidad	percentuale
		cijfer	Ausgangs§§betrag	total	cuenta	stanziamento
		contingent	Betrag	valeur	Consejo	import
		eis	Gesamtausgaben§§betrag	niveau	cifrado	cifra
		feit	Gesamt§§mittel	crédit	dato	previsione
		fonds	Größenordnung	conseil	propuesta	importo
		foutenpercentage	Hochgerechnet	coût	montante	dotazione
		gegeven	Höhe	enveloppe	potencial	Consiglio
		statistisch	Information	nombre	porcentaje	aumento
		getal	Leitwert	objectif	proyecto	tasso
		werktijd	Menge	potentiel	límite	spesa
		balans	Mittel	proportion	contenido	totale
		geld	Prozentsatz	représentation	dotación	potenziale
		grens	Rahmen	seuil	regla	ammontare
		grens§§waarde	Rekord	taux	plazo	somma
		groei§§percentage	Rente	teneur	número	limite
		grootheid	Satz	écart	proporción	regola
		hoeveelheid	Summe	dotation	billete	caso
		hoogte	Wert	performance	figura	durata
		indicatie	Wochenarbeits§§zeit	diminution	resultado	valutazione
		informatie	Zahl	évaluation	aportación	proporzione
		inspanning	Ziffer	ampleur	financiación	livello
		kostenpost	Niveau	décompte	referencia	conteggio
		limiet	Rate	argent	balance	finanziamento
		niveau	Finanz§§betrag	chiffrer	cifrar	fondi
		norm	Finanz§§rahmen	plafond	límite	decurtazione
		omvang	Geld	référence	volumen	posizione

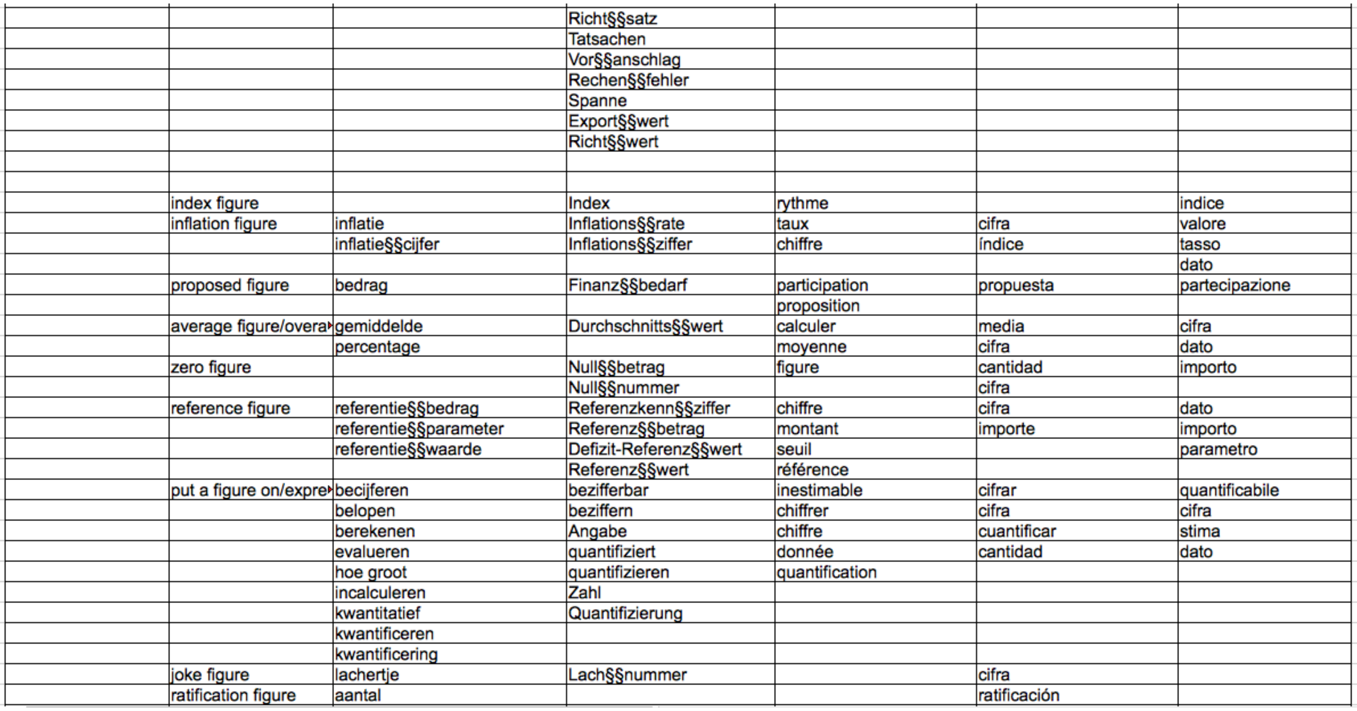


		omvangrijk	Stand	fonds	esfuerzo	risulta
		opbrengst	Bezugnahme	dissimulation	gasto	storno
		percentage	Maßzahl	donnée	nivel	cifra relativa
		procent	Vorgabe	bilan	índice	aliquota
		quota	Zahlenangabe	indication	cuota	stima
		quote	Angabe	information	relación	quantificazione
		quotum	Durchschnitts§§wert	calcul	estadística	dato statistico
		rekensommetje	statistisch	quantité	cálculo	orario
		rekenen	Bilanz	effort	base	quantità
		resultaat	Index	dépense	parámetro	calcolo
		schatting	zahlen§§mäßig	produit	cuantitativo	soglia
		score	Größenordnung	quota	tasa	raffronto
		som	Höchst§§menge	résultat	índice	informazione
		statistiek gegeven	Bemühung	repartition	monto	voce di costo
		statistisch gegeven	Kostenfaktor	campagne	evaluación	massimale
		steun	Grenz§§wert	aveu	elemento	dato relativo
		tekort	Ertrag	statistique	estimación	indicazione
		tijd	Ziel	chiffrer	índice	obiettivo
		veel	Quote	élément	gráfico	contingente
		verdeling	Fang§§quote	doublement	pago	risultato
		verhouding	Grenze	gestion	cuantificado	dimensione
		volume	Ergebnis	état	fecha	statistica
		voorbeeld	Daten	payement	pormenor	cifra fornita
		voorontwerp van begroting	Zeitraum	bénéfice	base	volume
		waarde	Differenz	compte	hecho	elemento
		statistiek	Statistik	concurrence	realidad	saldo
		becijferen	Absatz§§zahl	quantitatif	diferencia	situazione
		bepaling	Auflagen§§höhe	combien		parametro
		berekenen	Beschäftigungs§§zahl	emploi		conto
		berekening	Bevölkerungs§§zahl	situation		quantitativo
		bericht	Brutto§§sozial§§produkt	équilibré		caso
		cijfergegeven	Einbürgerungs§§zahl	quantifier		posto di lavoro
		cijfermateriaal	Entwicklungs§§hilfe	coût		collocarsi
		cijfermatig	Fakt	compensation		bilancio

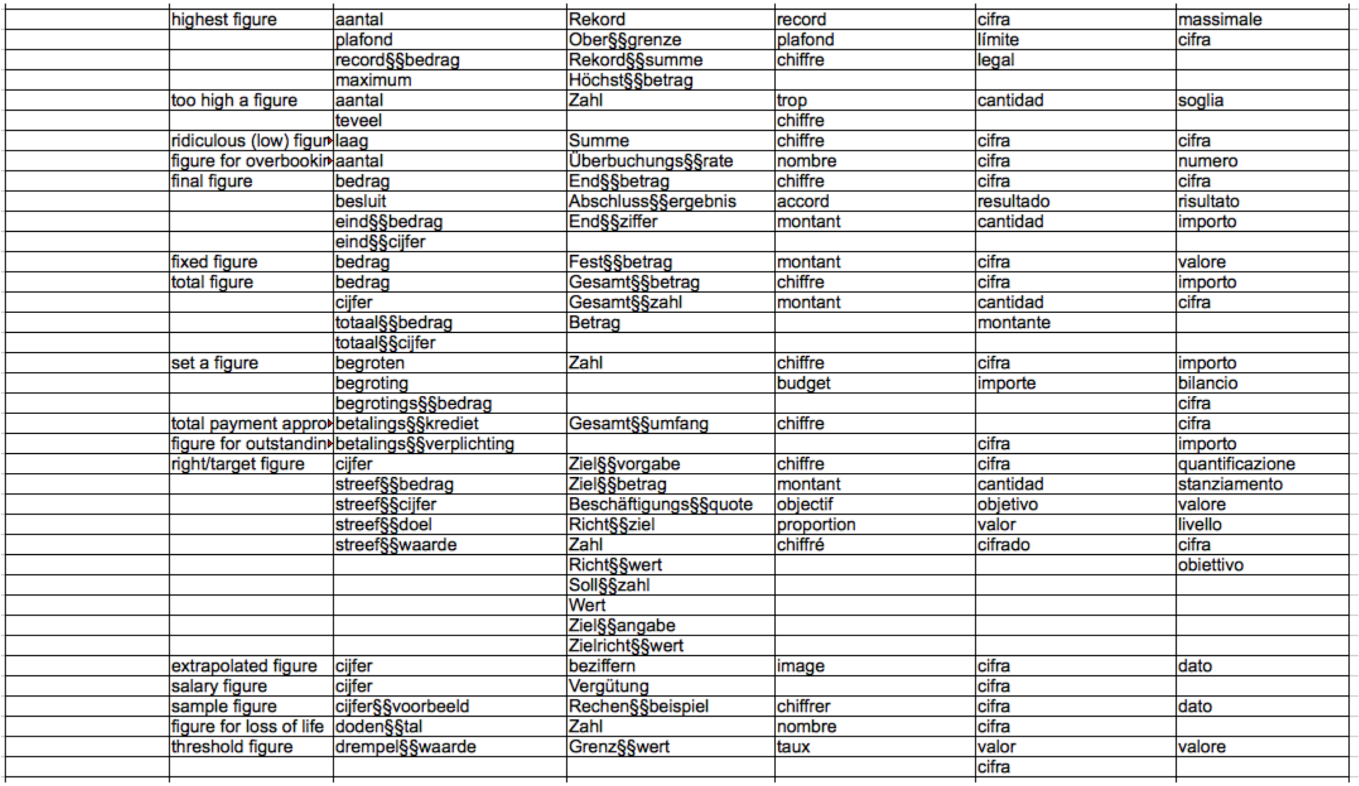


		cijfermatige informatie	Fakten§§lage	estimation		impegno
		cijfertje	Gesamt§§verschuldung	précis		voce
		data	Kenn§§ziffer	indicateur		cifra ipotetica
		detail	Mittel§§ansatz	indice		grandezza
		feit	Pflicht§§übung	addition		indicatore
		editie	Schätzung	estimer		cifra fornita
		gekwantificeerd	Verbrauchs§§zahl	alinéa		importo relativo
		geldbedrag	Asylbewerber§§zahl	paramètre		rilevazione
		getalsmatig	Beschäftigten§§zahl	propriété		senso
		grondslag	Fall	source		quantitativo indicato
		grootte	Unfall§§zahl	détail		aggiornamento
		informatie§§bron	beziffert	constat		informativo
		invulling	Gesamt§§zahl	édition		edizione
		kwantitatief	Mindest§§zahl	mesure		base
		marge	Bilanz	base		termine
		overweging	Bilanz§§summe	marge		indice
		natellen	Gewalt§§bilanz	recensement		fatto
		numeriek	Quantifizierung	volume		percentuale
		omrekenen	Finanz§§volumen	part		
		omvang	Haftungs§§betrag	fait		
		onderzoek	Bezug	prévision		
		optelling	Entschädigung	généralisation		
		percentagecijfer	Ergebnis	différence		
		plan	Haushalts§§entwurf			
		praktijk	Haushalts§§zahl			
		raming	Investitions§§summe			
		realiteit	Land			
		rekening	Zeit			
		saldo	durchrechnen			
		sommige	berechnen			
		statistiek	Aufrechnung			
		sterfte§§cijfer	Berechnung			
		tabel	Soll§§zahl			
		uitkomst	Zahlen§§material			

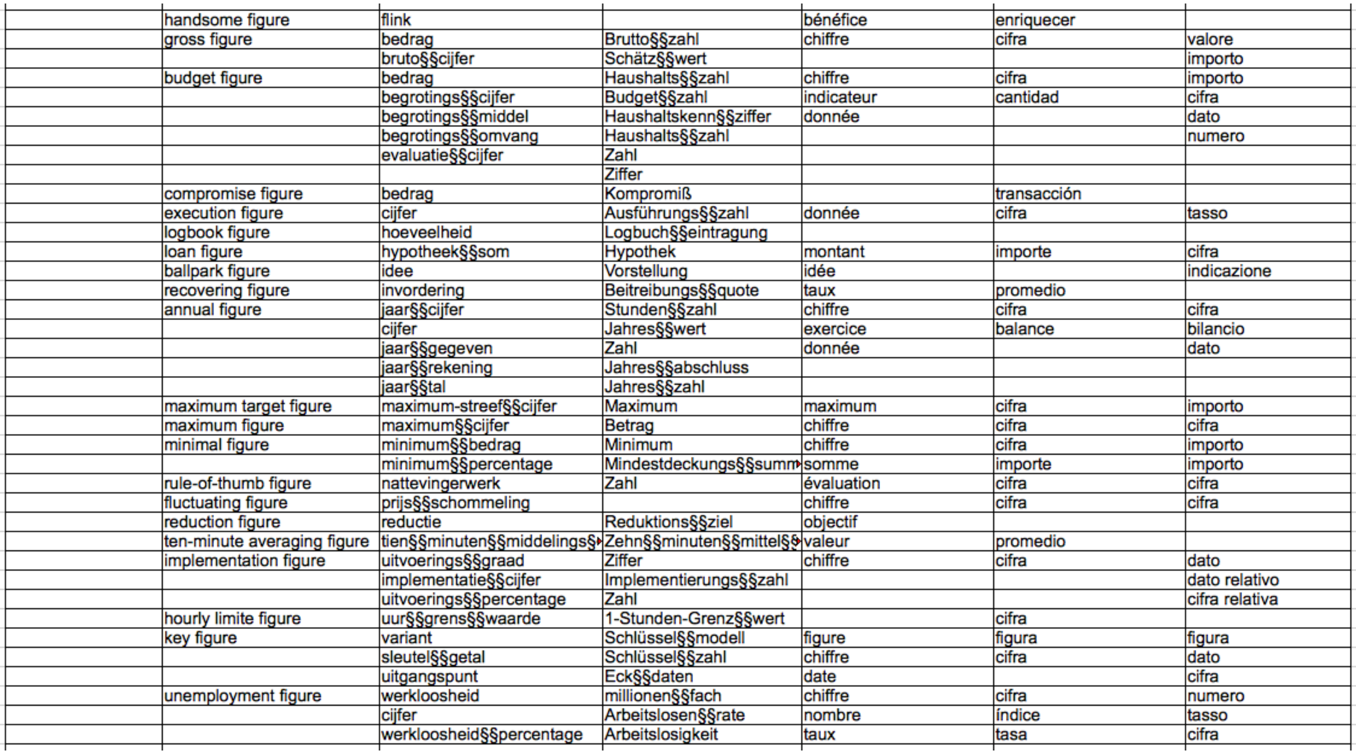
		praktijk§§gegeven	quantifiziert			
		verschil	Zahlenbeleg			
		voorlichtings§§percentage	Zahlenwerk			
			Auskunft			
			Auslands§§investition			
			Finanz§§zahl			
			Gesamt§§zahl			
			Indikator			
			Index			
			Parameter			
			Prozentzahl			
			Rauschgift§§bilanz			
			Schätzung			
			Stimmen§§zahl			
			Wirklichkeit			
			Zahlen§§friedhof			
			Zahlen§§kolonne			
			Zahlen§§kosmetik			
			Zahlen§§schlacht			
			Zahlen§§vergleich			
			Zahlen§§werk			
			beziffern			
			finanziell			
			Termin			
			Einzelheit			
			Abgas§§wert			
			Beschäftigungs§§daten			
			Element			
			Grundlage			
			Produktions§§zahl			
			Faktor			
			nachzahlen			
			Umfang			
			beziffern			



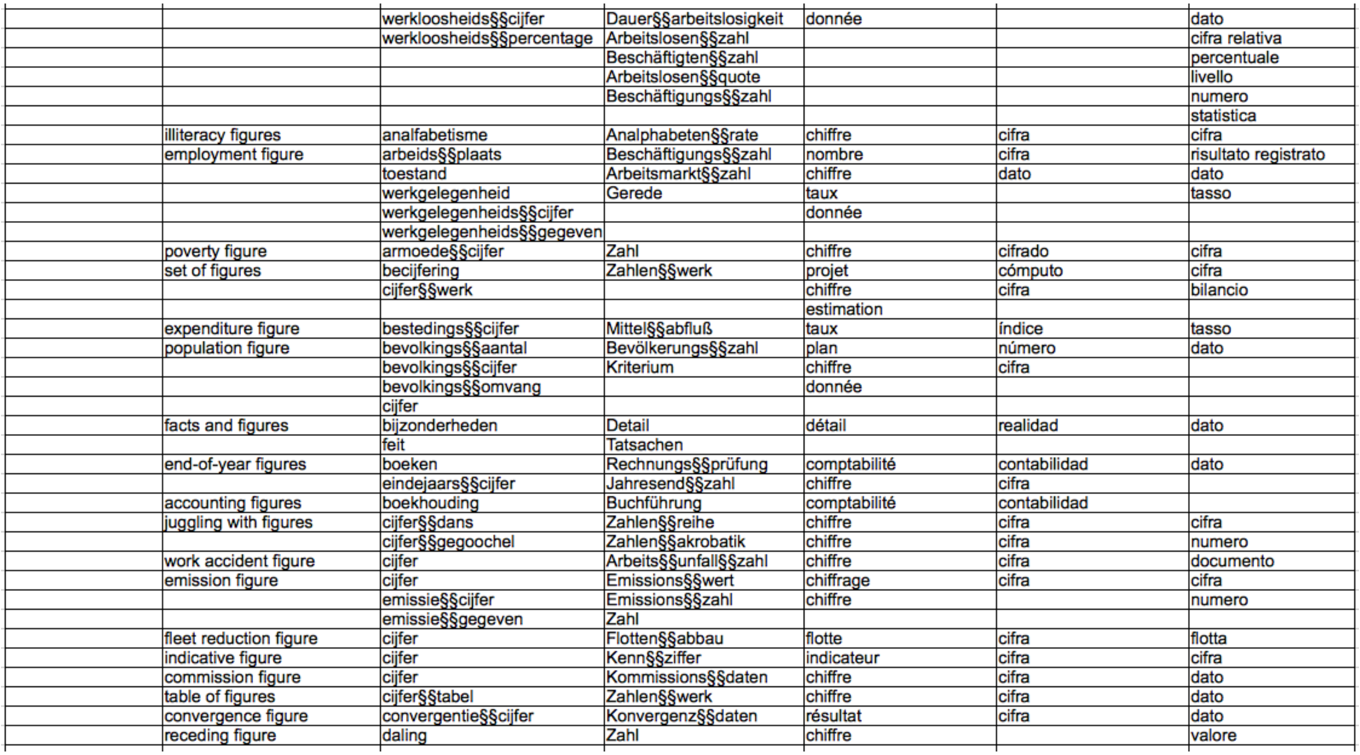
			Richt§§satz			
			Tatsachen			
			Vor§§anschlag			
			Rechen§§fehler			
			Spanne			
			Export§§wert			
			Richt§§wert			
	index figure		Index	rythme		indice
	inflation figure	inflatie	Inflations§§rate	taux	cifra	valore
		inflatie§§cijfer	Inflations§§ziffer	chiffre	índice	tasso
						dato
	proposed figure	bedrag	Finanz§§bedarf	participation	propuesta	partecipazione
				proposition		
	average figure/overa	gemiddelde	Durchschnitts§§wert	calculer	media	cifra
		percentage		moyenne	cifra	dato
	zero figure		Null§§betrag	figure	cantidad	importo
			Null§§nummer		cifra	
	reference figure	referentie§§bedrag	Referenzkenn§§ziffer	chiffre	cifra	dato
		referentie§§parameter	Referenz§§betrag	montant	importe	importo
		referentie§§waarde	Defizit-Referenz§§wert	seuil		parametro
			Referenz§§wert	référence		
	put a figure on/expre	becijferen	bezifferbar	inestimable	cifrar	quantificabile
		belopen	beziffern	chiffrer	cifra	cifra
		berekenen	Angabe	chiffre	cuantificar	stima
		evalueren	quantifiziert	donnée	cantidad	dato
		hoe groot	quantifizieren	quantification		
		incalculeren	Zahl			
		kwantitatief	Quantifizierung			
		kwantificeren				
		kwantificering				
	joke figure	lachertje	Lach§§nummer		cifra	
	ratification figure	aantal			ratificación	



	highest figure	aantal	Rekord	record	cifra	massimale
		plafond	Ober§§grenze	plafond	límite	cifra
		record§§bedrag	Rekord§§summe	chiffre	legal	
		maximum	Höchst§§betrag			
	too high a figure	aantal	Zahl	trop	cantidad	soglia
		teveel		chiffre		
	ridiculous (low) figur	laag	Summe	chiffre	cifra	cifra
	figure for overbookin	aantal	Überbuchungs§§rate	nombre	cifra	numero
	final figure	bedrag	End§§betrag	chiffre	cifra	cifra
		besluit	Abschluss§§ergebnis	accord	resultado	risultato
		eind§§bedrag	End§§ziffer	montant	cantidad	importo
		eind§§cijfer				
	fixed figure	bedrag	Fest§§betrag	montant	cifra	valore
	total figure	bedrag	Gesamt§§betrag	chiffre	cifra	importo
		cijfer	Gesamt§§zahl	montant	cantidad	cifra
		totaal§§bedrag	Betrag		montante	
		totaal§§cijfer				
	set a figure	begroten	Zahl	chiffre	cifra	importo
		begroting		budget	importe	bilancio
		begrotings§§bedrag				cifra
	total payment appro	betalings§§krediet	Gesamt§§umfang	chiffre		cifra
	figure for outstandin	betalings§§verplichting			cifra	importo
	right/target figure	cijfer	Ziel§§vorgabe	chiffre	cifra	quantificazione
		streef§§bedrag	Ziel§§betrag	montant	cantidad	stanziamento
		streef§§cijfer	Beschäftigungs§§quote	objectif	objetivo	valore
		streef§§doel	Richt§§ziel	proportion	valor	livello
		streef§§waarde	Zahl	chiffré	cifrado	cifra
			Richt§§wert			obiettivo
			Soll§§zahl			
			Wert			
			Ziel§§angabe			
			Zielricht§§wert			
	extrapolated figure	cijfer	beziffern	image	cifra	dato
	salary figure	cijfer	Vergütung		cifra	
	sample figure	cijfer§§voorbeeld	Rechen§§beispiel	chiffrer	cifra	dato
	figure for loss of life	doden§§tal	Zahl	nombre	cifra	
	threshold figure	drempel§§waarde	Grenz§§wert	taux	valor	valore
					cifra	



	handsome figure	flink		bénéfice	enriquecer	
	gross figure	bedrag	Brutto§§zahl	chiffre	cifra	valore
		bruto§§cijfer	Schätz§§wert			importo
	budget figure	bedrag	Haushalts§§zahl	chiffre	cifra	importo
		begrotings§§cijfer	Budget§§zahl	indicateur	cantidad	cifra
		begrotings§§middel	Haushaltskenn§§ziffer	donnée		dato
		begrotings§§omvang	Haushalts§§zahl			numero
		evaluatie§§cijfer	Zahl			
			Ziffer			
	compromise figure	bedrag	Kompromiß		transacción	
	execution figure	cijfer	Ausführungs§§zahl	donnée	cifra	tasso
	logbook figure	hoeveelheid	Logbuch§§eintragung			
	loan figure	hypotheek§§som	Hypothek	montant	importe	cifra
	ballpark figure	idee	Vorstellung	idée		indicazione
	recovering figure	invordering	Beitreibungs§§quote	taux	promedio	
	annual figure	jaar§§cijfer	Stunden§§zahl	chiffre	cifra	cifra
		cijfer	Jahres§§wert	exercice	balance	bilancio
		jaar§§gegeven	Zahl	donnée		dato
		jaar§§rekening	Jahres§§abschluss			
		jaar§§tal	Jahres§§zahl			
	maximum target figure	maximum-streef§§cijfer	Maximum	maximum	cifra	importo
	maximum figure	maximum§§cijfer	Betrag	chiffre	cifra	cifra
	minimal figure	minimum§§bedrag	Minimum	chiffre	cifra	importo
		minimum§§percentage	Mindestdeckungs§§summ	somme	importe	importo
	rule-of-thumb figure	nattevingerwerk	Zahl	évaluation	cifra	cifra
	fluctuating figure	prijs§§schommeling		chiffre	cifra	cifra
	reduction figure	reductie	Reduktions§§ziel	objectif		
	ten-minute averaging figure	tien§§minuten§§middelings§	Zehn§§minuten§§mittel§§	valeur	promedio	
	implementation figure	uitvoerings§§graad	Ziffer	chiffre	cifra	dato
		implementatie§§cijfer	Implementierungs§§zahl			dato relativo
		uitvoerings§§percentage	Zahl			cifra relativa
	hourly limite figure	uur§§grens§§waarde	1-Stunden-Grenz§§wert		cifra	
	key figure	variant	Schlüssel§§modell	figure	figura	figura
		sleutel§§getal	Schlüssel§§zahl	chiffre	cifra	dato
		uitgangspunt	Eck§§daten	date		cifra
	unemployment figure	werkloosheid	millionen§§fach	chiffre	cifra	numero
		cijfer	Arbeitslosen§§rate	nombre	índice	tasso
		werkloosheid§§percentage	Arbeitslosigkeit	taux	tasa	cifra



		werkloosheids§§cijfer	Dauer§§arbeitslosigkeit	donnée		dato
		werkloosheids§§percentage	Arbeitslosen§§zahl			cifra relativa
			Beschäftigten§§zahl			percentuale
			Arbeitslosen§§quote			livello
			Beschäftigungs§§zahl			numero
						statistica
	illiteracy figures	analfabetisme	Analphabeten§§rate	chiffre	cifra	cifra
	employment figure	arbeids§§plaats	Beschäftigungs§§zahl	nombre	cifra	risultato registrato
		toestand	Arbeitsmarkt§§zahl	chiffre	dato	dato
		werkgelegenheid	Gerede	taux		tasso
		werkgelegenheids§§cijfer		donnée		
		werkgelegenheids§§gegeven				
	poverty figure	armoede§§cijfer	Zahl	chiffre	cifrado	cifra
	set of figures	becijfering	Zahlen§§werk	projet	cómputo	cifra
		cijfer§§werk		chiffre	cifra	bilancio
				estimation		
	expenditure figure	bestedings§§cijfer	Mittel§§abfluß	taux	índice	tasso
	population figure	bevolkings§§aantal	Bevölkerungs§§zahl	plan	número	dato
		bevolkings§§cijfer	Kriterium	chiffre	cifra	
		bevolkings§§omvang		donnée		
		cijfer				
	facts and figures	bijzonderheden	Detail	détail	realidad	dato
		feit	Tatsachen			
	end-of-year figures	boeken	Rechnungs§§prüfung	comptabilité	contabilidad	dato
		eindejaars§§cijfer	Jahresend§§zahl	chiffre	cifra	
	accounting figures	boekhouding	Buchführung	comptabilité	contabilidad	
	juggling with figures	cijfer§§dans	Zahlen§§reihe	chiffre	cifra	cifra
		cijfer§§gegoochel	Zahlen§§akrobatik	chiffre	cifra	numero
	work accident figure	cijfer	Arbeits§§unfall§§zahl	chiffre	cifra	documento
	emission figure	cijfer	Emissions§§wert	chiffrage	cifra	cifra
		emissie§§cijfer	Emissions§§zahl	chiffre		numero
		emissie§§gegeven	Zahl			
	fleet reduction figure	cijfer	Flotten§§abbau	flotte	cifra	flotta
	indicative figure	cijfer	Kenn§§ziffer	indicateur	cifra	cifra
	commission figure	cijfer	Kommissions§§daten	chiffre	cifra	dato
	table of figures	cijfer§§tabel	Zahlen§§werk	chiffre	cifra	dato
	convergence figure	convergentie§§cijfer	Konvergenz§§daten	résultat	cifra	dato
	receding figure	daling	Zahl	chiffre		valore

	fraud figure	fraude§§bedrag	Betrugs§§zahl	chiffre	cifra	caso
	detailed figure	gedetailleerd	aufgezählt			termine
	comparative figure	getal	Vergleichs§§zahl	chiffre	cifra	
	growth figure	groei§§cijfer	Wachstums§§rate	croissance	índice	livello
		groei§§waarde	Wachstums§§zahl	chiffre	cifra	cifra
		stijging	Zahl	estimation	incremento	dato
			Zuschlag	taux		tasso
				augmentation		indicatore
				hausse		valore
	trade figure	handels§§cijfer	Zahl	chiffre	cifra	cifra
		handels§§volume	Handels§§volumen			volume
	benchmark figure	hoeksteen	Eck§§wert	élément	punto	punto
	higher figure	hoger	Zahl	chiffre	cifra	numero
	shock figure	horror§§cijfer	Horror§§zahl	chiffre	cifra	cifra
	import figure	import§§cijfer	Zahl	chiffre	cifra	cifra relativa
		invoer§§cijfer	Einfuhr§§zahl	caractère	volumen	dato
	investment figure	investerings§§cijfer	Investitions§§zahl	donnée	cifra	dato relativo
			Zahl	chiffre		
	viewer figure	kijk§§cijfer	Eischalt§§quoten	audimat	índice	audience
		kijk§§dichtheid			audiencia	
	figures in the billions	miljarden§§cijfer	Milliarden§§betrag	montant	cifra	importo
	crime figure	misdaad§§cijfer	Verbrechens§§zahl	courbe	cifra	dato
	estimated figure	model§§berekening	Model§§lrechnung	modélisation	modelación	calcolo
	different figure	mogelijkheid	Modell	forme	figura	configurazione
	accident figure	ongevallen§§risico		risque	número	dato
	development figure	ontwikkelings§§cijfer	Entwicklungs§§zahl	chiffre	cifra	dato
	education figure	opleidings§§cijfer	Ausbildungs§§zahl	nombre	cifra	dato
	popularity figure	populariteits§§cijfer	Beliebtheits§§quoten	cote	índice	popolarità
	production figure	productie§§cijfer	Produktions§§zahl	chiffre	cifra	dato
		tabaks§§productie	Tabak§§erzeugung			
	voting figure	stem§§uitslag	Abstimmungs§§ergebnis	chiffre	resultado	dato relativo
		stem§§verhouding				
	aid figure	steun§§bedrag	Unterstützung	subvention	cantidad	importo
		steun§§cijfer	Zahl	chiffre	cifra	dato
	spending figure	uitgave	Ausgabe	quantitatif	gasto	dato
		uitgaven§§maximum	Ausgabenhöchst§§betrag	chiffre	cifra	
		uitgaven§§niveau	Ausgaben§§budget			
	consumption figure	verbruiks§§gegeven	Verbrauchs§§wert	donnée	dato	dato

	allocation figure	verdeling	Aufteilungs§§zahl	chiffre	cifra	cifra
	sales figure	verkoop§§cijfer	Verkaufs§§zahl	chiffre	cifra	dato
					volumen	fatturato
	draft figure	voorontwerp	Zahlen§§werk	chiffre		bilancio
3. Person						
3.1 Person general						
		persoonlijkheid	Amt	figure	figura	figura
		lid	Engagierter	prêtre	miembro	individuo
		figuur	Figur	personne	personaje	protagonista
		functie	Person	personnalité	persona	rappresentante
		iemand	Persönlichkeit	stature	personalidad	personalità
		man	Beispiel	personnage	exponente	membro
		orgaan	Erscheinung	homme	funcionario	personaggio
		personage	Europäer	poste	organización	esponente
		persoon	Gestalt	institution	actor	uomo
		rechts§§figuur	Beamte	collaborateur	autoridad	dirigente
		rol	Gemeinschafts§§organ	membre		elemento
		structuur	Sonder§§beauftragten	rôle		persona
		verschijning	Mann	élément		funzionario
		zwaargewicht	Verantwortlicher	responsable		soggetto
		beiden	Diplomat	représentant		attore
		afgevaardigde	Gebilde	fonctionnaire		
		exponent	Mitglied	acteur		
		functionaris	Vertreter	dignitaire		
		instantie	Politiker			
		kracht	Redefigur			
		kring	Typ			
		mens	Gremium			
		personaliteit	Gruppierung			
		president	Akteur			
			Schlüssel§§figur			
	hate figure	boeman			símbolo	
	important figure	denker	Mann	géant	gigante	
	figure of the ombudsman	figuur	Bürgerbeauftragten		figura	figura
	solitary figure	figuur	Einzeler§§scheinung	figure	figura	figura
	leading figure	figuur	Führungs§§persönlichkeit	dirigeant	dirigente	dirigente

		hoofdrolspeler	Persönlichkeit	institution	protagonista	protagonista
		leider	Vertreter	personne	figura	figura
		opinieleider	Steuerungs§§figur	leader	líder	personaggio
		hoogwaardigheidsbekleder	Verantwortlicher	responsable	personalidad	responsabile
		kopstuk	Kopf	sommité	responsable	personalità
		leiding		figure	representante	rappresentante
		verantwoordelijke		entourage		
		vertegenwoordiger		personnalité		
				représentant		
	central figure	figuur	Mittelpunkt	acteur	figura	figura
		functionaris	Akteur	responsable	protagonista	protagonista
			Beamte		responsable	responsabile
	key figure	figuur	Schlüssel§§figur	personnage	figura	figura
		kopstuk	Protagonist	protagoniste	protagonista	protagonista
		speler	Vertreter	figure	actor	
			Akteur	acteur		
	symbolic figure	figuur	Symbol§§figur	figure	figura	figura
		symbool§§figuur	Figur			
	identifiable figure	identificatie§§figuur	Identifikations§§figur	caractère	figura	figura
	Islamic figure	islamiet	Figur	personnalité	figura	figura
	political figure	kopstuk	Persönlichkeit	personnalité	personalidad	personalità
		leider	Politiker	politicien	político	personaggio
		politicus	Person	intégrisme	representante	esponente
		gezagsdrager	Verantwortlicher	personnel	personaje	rappresentante
		persoonlijkheid	Akteur	acteur	responsable	responsabile
		piet	Jurist	dirigeant	potentado	
		vertegenwoordiger	Spitzen§§politiker	responsable		
				figure		
	opposition figure	oppositie§§leider	Oppositionel	figure	figura	esponente
		oppositie§§voerder	Oppositions§§politiker	opposant	oponente	rappresentante
		lid	Opposition	membre	figura	figura
		mens	Persönlichkeit	personnalité	personalidad	
		opposant	Oppositions§§führer	personne	opositor	
		oppositie§§lid			líder	
	father figure	pater	Mentor	pilier	figura	figura
	reference figure	referentie§§punt	Bezugs§§person	modèle	referencia	
	giant figure	staatsman	Persönlichkeit	géant	gigante	personalità
	world figure	wereld§§leider	Persönlichkeit	figure	personaje	figura

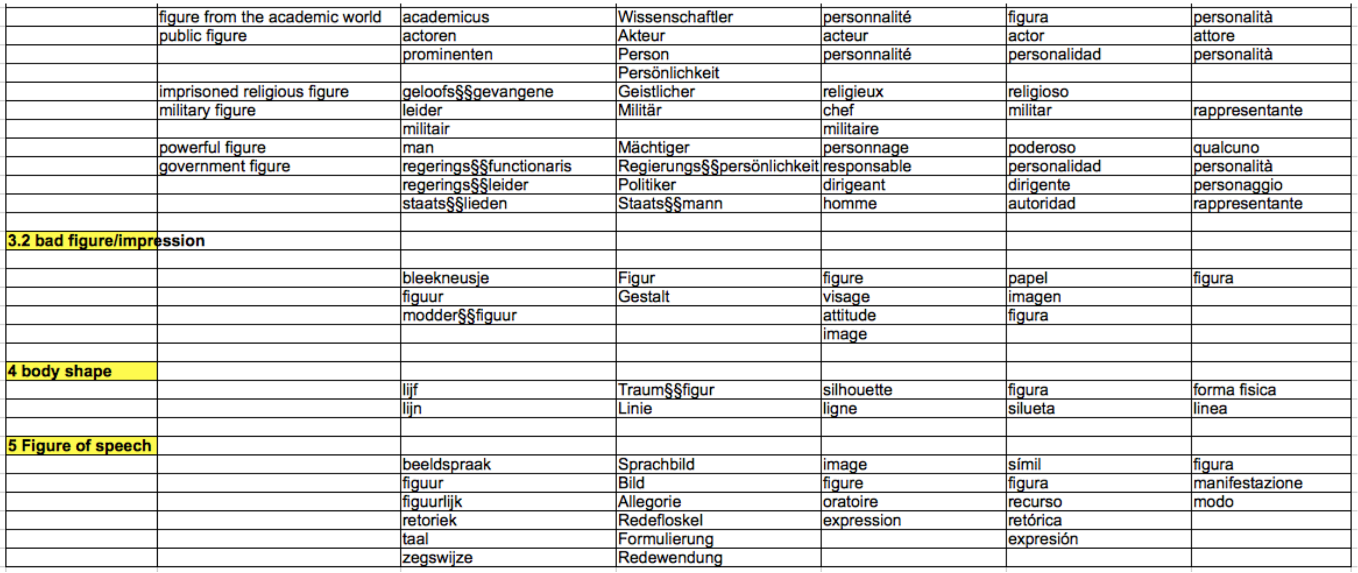


	figure from the academic world	academicus	Wissenschaftler	personnalité	figura	personalità
	public figure	actoren	Akteur	acteur	actor	attore
		prominenten	Person	personnalité	personalidad	personalità
			Persönlichkeit			
	imprisoned religious figure	geloofs§§gevangene	Geistlicher	religieux	religioso	
	military figure	leider	Militär	chef	militar	rappresentante
		militair		militaire		
	powerful figure	man	Mächtiger	personnage	poderoso	qualcuno
	government figure	regerings§§functionaris	Regierungs§§persönlichkeit	responsable	personalidad	personalità
		regerings§§leider	Politiker	dirigeant	dirigente	personaggio
		staats§§lieden	Staats§§mann	homme	autoridad	rappresentante
3.2 bad figure/impression						
		bleekneusje	Figur	figure	papel	figura
		figuur	Gestalt	visage	imagen	
		modder§§figuur		attitude	figura	
				image		
4 body shape						
		lijf	Traum§§figur	silhouette	figura	forma fisica
		lijn	Linie	ligne	silueta	linea
5 Figure of speech						
		beeldspraak	Sprachbild	image	símil	figura
		figuur	Bild	figure	figura	manifestazione
		figuurlijk	Allegorie	oratoire	recurso	modo
		retoriek	Redefloskel	expression	retórica	
		taal	Formulierung		expresión	
		zegswijze	Redewendung			

Figure C.3: Clustering file for the test word *figure*.