Automatic Identification of Cognates, False Friends, and Partial Cognates

by

Oana Magdalena Frunză

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the M.Sc. degree in
Computer Science

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

© Oana Magdalena Frunză, Ottawa, Canada, 2006

Abstract

Cognates are words in different languages that have similar spelling and meaning. They can help second-language learners with vocabulary expansion and reading comprehension tasks. Special attention needs to be paid to pairs of words that appear similar but are in fact false friends: they have different meanings in all contexts.

Partial cognates are pairs of words in two languages that have the same meaning in some, but not all, contexts. Detecting the actual meaning of a partial cognate in context can be useful for Machine Translation and Computer-Assisted Language Learning tools.

Our research on cognate and false-friend words between two pair of languages (French and English in our case) consists in automatically classifying a pair of words from two languages as cognates or false friends. We use Machine Learning techniques with several measures of orthographic similarity as features for classification. We study the impact of selecting different features, averaging them, and combining them through Machine Learning techniques. The methods work on different pair of languages as long as a small amount of annotated pairs of words is provided as training data.

In addition to the work done on cognate and false-friend identification we propose a supervised and a semi-supervised method that uses bootstrapping for disambiguating partial cognates between French and English. The proposed methods use only automatically-labeled data and therefore they can be applied to other pairs of languages as well. The data that we use is automatically collected from parallel corpora. The impact of data collected from different domains is also taken into account in our research.

To complement the studies that we did on cognates, false friends and partial cognate pairs of words, we developed an annotation tool for this special type of words. The tool can automatically annotate cognates, false friends and partial cognates for any French text. The tool uses UIMA (Unstructured Information Management Architecture) from IBM and BaLIE (an open-source Java project designed to extract information from free text).

Acknowledgements

Dedic această teză părinţilor mei Elena şi Dorin, şi fratelui meu Codrin Frunză.

Mulţumesc pentru ajutorul şi sprijinul acordat.

I would like to thank my supervisor Dr. Diana Inkpen for all the support and guidance, Dr. Lise Duquette from the Second Language Institute, University of Ottawa, for giving us the idea to work on Cognates and False Friends between French and English, and Dr. Gregorz Kondrak, from University of Alberta, for providing us the orthographic similarity measures package.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Goals

The main reasons why we chose to do research in Natural Language Processing, Cross-Language Word Sense Disambiguation, Data Mining, and Machine Learning is because we believe that computers can help humans in different tasks.

Our research is focused on cross-language word pair identification and disambiguation in French and English, but our methods can be applied to other pairs of languages as well. The type of word pairs that we are working with are: *cognates* — pairs of words that reflect similarities between two languages, *false friends* — pairs of words that reflect dissimilarities between languages, and *partial cognates* — pairs of words that in some contexts have a cognate behavior and in others have a false friend behavior.

Nowadays learning a new language is not only a fashionable thing to do, but it sometimes becomes a necessity. One of the main reasons that make learning a new language attractive is the relations that exist between countries. Culture, economics and politics make the relations between countries grow and become stronger each day. Besides the global aspect of learning a new language, we can easily add our own motivation of becoming multicultural persons. These are only a few reasons that make us start learning a new language. Stronger motivation can also be found. The need for

multilingualism, especially in Europe. The European Union has been formulated very clearly and succinctly by the French linguist Claude Hagge when he said: "L'Europe sera multilingue ou elle ne sera pas" *(Europe will be multilingual or it will not be).*

The best way to learn a new language is to have a human tutor, but that is not always handy due to different reasons, mostly time and money. There are a lot of available online tools — some free, some commercial — that can be used in language learning. The main problem with most of the tools that exist is how well they impersonate a human tutor. The major problem that arises is that they do not provide feedback in language learning, and more than in any other domain we can follow the saying "we learn from mistakes". A tool capable to provide students visual explanations that helps noticing similarities and differences across languages can be helpful in the task of second language learning.

It is true that most of the CALL (Computer Assisted Language Learning) tools are not as accurate as a teacher or tutor, and this is the one important drawback of these systems, but with a little help of human work and knowledge that can be fixed. A teacher can eliminate errors produced by automatically designed systems in order to provide the students more accurate information. The human effort is less demanding than it would be for a human to do the whole process of detecting and analyzing cross language differences and similarities in a text.

Machine Translation (MT) tools can also benefit from being aware of cross-language differences and similarities, in order to improve their results when translating a certain word in context. Similarities (cognates) and dissimilarities (false friends) are used with success in this area of Natural Language Processing (NLP). Word alignment is another research area that showed an improvement when using cognates (Marcu, Kondrak, & Knight, 2003).

Information retrieval and most likely cross-language information retrieval systems can use cross-language word sense disambiguation in order to retrieve desired documents in a target language. Information extraction systems are also NLP application that can

produce better results when using word sense disambiguation and cross-language word sense disambiguation.

Data Mining and Machine Learning (ML) allows us to transform data into information and information into knowledge. These are the main things that a computer needs to be able to get as close as possible to a rational process, the main goal of the computed world.

All these areas of NLP along with their final applications made us define our aim — automatically identify cognates and false friends, and automatically disambiguate partial cognates in French and English, for this thesis. The hypothesis we prove in our work is that we can sucessfully accomplish our aim.

1.2 Cognates, False Friends, and Partial Cognates

Although French and English belong to different branches of the Indo-European family of languages, their vocabularies share a great number of similarities. Some are words of Latin and Greek origin e.g., *education* and *theory*. A small number of very old, genetic cognates go back all the way to Proto-Indo-European, e.g., *mére* - *mother* and *pied* - *foot*. The majority of these pairs of words penetrated the French and English language due to the geographical, historical, and cultural contact between the two countries over many centuries — and here we talk about borrowings. Other cognates can be traced to the conquest of Gaul by Germanic tribes after the collapse of the Roman Empire, and by the period of French domination of England after the Norman conquest.

Nowadays, new terms related to modern technology are often adopted in a similar form across completely unrelated languages. Even if languages are written in distinct scripts, approximate phonetic transcription of orthographic data is relatively straightforward in most cases. For example, after transcribing the Japanese word for *sprint* from the Katakana script into semi-phonetic *supurinto*, it is possible to detect its similarity to the French word *sprinter*, which has the same meaning.

Most of the borrowings have changed their orthography, following different orthographic rules, and most likely their meaning as well. Some of the adopted words replaced the original word in the language, while others were used together but with slightly or completely different meanings.

1.2.1 Definitions

The definitions that we adopt and that we are going to follow in our work are language-independent, but the examples that we give are for French and English, the focus of our work.

Cognates, or True Friends (Vrais Amis), are pairs of words that are perceived as similar and are mutual translations. The spelling can be identical or not, e.g., *nature* - *nature*, *reconnaissance* - *recognition*. Some researchers refer to cognates as being pairs of words that are orthographically identical and to near-cognates as the ones that have slightly different spelling. In our work, we adopt the cognate definition for both.

False Friends (Faux Amis) are pairs of words in two languages that are perceived as similar but have different meanings, e.g., *main* (= *hand*) - *main* (meaning *principal* or *essential*), *blesser* (= *to injure*) - *bless* (that is translated as *bénir* in French).

Partial Cognates are pairs of words that have the same meaning in both languages in some but not all contexts. They behave as cognates or as false friends, depending on the sense that is used in each context. For example, in French, *facteur* means not only *factor*, but also *mailman*, while *étiquette* can also mean label or sticker, in addition to the cognate sense.

Genetic Cognates are word pairs in related languages that derive directly from the same word in the ancestor (proto)-language. Because of gradual phonetic and semantic

changes over long periods of time, genetic cognates often differ in form and/or meaning, e.g., *père - father*, *chef - head*. This category excludes lexical borrowings, i.e., words transferred from one language to another such as *concierge*.

1.2.2 History and Related Facts

English, a Germanic language, often seems to resemble a Romance language, French in particular. English has been borrowing words from French since the Middle Ages. The process has greatly increased in intensity after 1066 when Norman French became the official language of government, church, and English aristocracy. It is suggested that by the end of 13^{th} century more than 10,000 French words have entered the English language and that more than 75% of these are still in use today (Crystal, 1995).

It seems that the English language has borrowed many words from French and the other way around. This is easy to spot if we look at the similarities between the vocabularies of these two languages.

English Loans from French

The process of borrowing words from a language is not a straight-forward one, even though some of the words are taken into the language without any change and with their originality intact, e.g., *sang-froid*, *cause célébre*, *par excellence*, and *déjà vu*. Other words are so easily used in English that we do not even think about where they come from, e.g., *boutique*, *detour*, *nuance* and *amateur*. But most of the words that are adapted to English underwent an orthographic change process. LeBlanc & Séguin (1996) have shown that between French and English at least 2,529 orthographic rules were applied to the French words that penetrated the English vocabulary. They have also shown that more than 38% of the vocabulary of these two languages has an overlap. From a list of 60,000 pairs of French-English dictionary entries, 23,000 were found to be cognates. 6,500 were exactly identical words (homographs) and 16,500 with slightly spelling differences.

Besides orthographic changes, semantic changes have also affected the words that entered the vocabulary of a language. The words that have shifted their original meaning partially or totally are the ones that need special attention from students.

Spelling errors are frequent between learners of French and English and it seems that English learners make them more often. *Comfortable* with an *n* instead of *m*, *literature* with two *t*'s, *baggage* with an indefinite article (a/an) or in plural form are some of the most common mistakes made by French learners of English.

Areas like administration, law, religion, gastronomy, fashion, literature, arts and science are the ones that have a strong French flavour. The adopted words that conserved their meanings either replaced the existing words or the two words lived side by side in the language. Sometimes they developed slightly different meanings or nuances. The Old English word *pig* is used for the live animal and the French-origin word *pork* for the edible meat. Words like *begin - commence*, *help - aid*, *wedding - marriage*, *hide - conceal* are some examples of words that are both in use in English. In these examples, the second word has a French origin and is borrowed in English without any changes.

The English process of borrowing words from French continues today as well. According to the *Oxford English Dictionary*[1] one of the latest French words that entered the English language is *pisteur* - the person that prepare the snow on a skiing track.

With this continuous process English manages to keep its French flavor.

French Loans from English

French is one of many languages that borrow words from English. Due to this continuous process, a considerable part of its vocabulary is taken from English. Resistance to this phenomenon has been initiated by *L'Académie Française* but it seems that is not so easy.

One of the first words that was borrowed from English is *le weekend* in 1926. The French integration of this word has been well documented. In 1964 it seems that the amount of English words that are spoken transformed the French language into a mixture

[1] http://www.oed.com/

of French and English that was called *Franglais*. *Parlez-vous Franglais* by René Étiemble (Étiemble, 1991) is a book that was published in that period and that describes the mixture of language that was spoken in that time.

New measures were taken in 1994 in order to decrease the infiltration of English words. A law that was aiming to replace the English words with French ones, e.g., *ordinateur* instead of *computer* was released in 1996. Economics and finance were some of the fields targeted by this new cultural measure. A list of replacement words was given for the most used English words e.g., *arbitrage* for *trade-off*; *la vente agressive* for *hard selling*; *jeune pousse* for *start-up* and *achat sur simulation électronique* for *virtual shopping*.

It seems that life is not always waiting for new French words to be invented and used. This was the case of an *Le Monde* reporter that was sent in Iraq. One of his headlines for the paper was *Un jour dans la vie d'un 'embedded'*.

Young persons being attracted by the American culture is a process that is happening not only in the French culture. Music, film, sports are areas that are consequently adjusted with English words.

As we could see from this section, languages are under continuous processes of transformation and therefore new studies will always bring new and interesting discoveries.

1.3 Thesis Outline

This section presents the organization of the thesis and the content of each chapter.

1.3.1 Related Work

In the second chapter of the thesis, *Related Work* we will present the literature survey for all the NLP domains that we have touched in our research.

A section will be reserved for cognate and false-friend identification. The section will present methods and work that has been done so far to identify the existence of cognates

and false friends between different languages.

A brief survey of the main areas that have been using cognates and false friends with success includes work in the following domains: Second Language Learning, Machine Translation, Word Alignment, and Information Retrieval.

A separate section will provide an overview of the Word Sense Disambiguation (WSD) task, monolingual in particular. Some references for what has been done in cross-language WSD, because this is closely related to one of our main topics of research from Chapter 4, will also be mentioned. In the WSD part we will also describe some ML techniques that are used to determine the sense of a polysemous word in context.

1.3.2 Identification of Cognates and False Friends

In the third chapter, *Identification Cognates and False Friends*, we will present our work on cognate and false friend identification. As we mention in Section 1.1, our experiments are performed for pairs of words in French and English. Our approach to identifying cognate and false friends is based on orthographic matching. We experiment with different combinations of orthographic measures through Machine Learning techniques. In addition to the methods used for identifying cognate and false friend pairs of words, our experiments also describe an automatic way to determine a threshold for each of the orthographic measures that we use. The resulting thresholds can be later used in different experiments for new pairs of words.

The method that we use does not have a semantic dimension the cognate and false friend pairs are orthographically similar. To discriminate cognate pairs from false-friends pair additional information is needed. If the pairs are translation of each other, then they are cognates; otherwise they are false friends. Experiments when the semantic aspect is taken into account are also conducted in order to create complete lists of cognates and false friends between French and English. The semantic dimension can be added to the method in two ways: adjusting the features used in the ML techniques to include a translation feature; or splitting the pairs after they are classified as cognates/false friends

into cognates or false friends depending whether they are translation of each other in a French-English dictionary.

We also took entries of a French-English dictionary and we determined how many of the entries are cognates using the thresholds determined by our proposed method. To determine the false friends from the entry list we paired words with each other, except their translation; this way we created pairs of words in French and English that are not translation of each other. From this list of pairs of words the ones that have an orthographic similarity value above the threshold are false friends.

1.3.3 Partial Cognate Disambiguation

The fourth chapter, *Partial Cognate Disambiguation*, presents our proposed techniques, based on Machine Learning, the Weka tool (Witten & Frank, 2005), to disambiguate partial cognates. As explained in Section 1.2.1, partial cognates are pairs of words in two languages that share the same meaning in some but not all contexts. We use a semi-supervised method based on monolingual and bilingual bootstrapping. We use parallel corpora to automatically create and tag our training seeds for the bootstrapping techniques. To be able to use our methods to disambiguate partial cognates in different domains, we combined and used corpora of different domains. Evaluation experiments with the semi-supervised method using 10 pairs of French-English partial cognates are presented and discussed in this chapter.

1.3.4 A Tool for Cross-Language Pair Annotations

A tool that implements and uses our theoretical research on cognates and false friends is described in the chapter named *A Tool for Cross-Language Pair Annotations*.

One of our main research goals is to use our methods to identify and disambiguate cross-language pairs of words in Computer-Assisted Language Learning tools (CALL). Using UIMA, an SDK product from IBM, we developed a tool that is capable of

annotating cognates and false friends in French texts. The tool is designed to assist a second-language learner of French in the reading comprehension task. A special attention is given to false friends. Not only are they annotated but additional information (definitions and brief examples) is also provided. The reason why we treated the false friend pairs differently is because they are the ones that cause most problems in language learning (Carroll 1992).

1.4 Research Contributions

This thesis brings research contributions to NLP by defining and solving novel tasks. The two tasks are: automatic identification of cognates and false friends between two pairs of languages and automatic disambiguation of partial cognates.

For the cognate and false friend identification task, we propose a new method that uses ML techniques and does not require a lot of human effort. The training data required are pairs of words judged as cognates/false friends and unrelated. The method uses different combinations of several orthographic measures that can be automatically applied to a pair of words in two languages. The values of the measures represent the feature value space for the ML algorithms. Besides the process of identifying cognates and false friends from a list of pair of words, our method can be extended and used to determine complete lists of cognates and false friends between two languages.

The results that we obtained in this task support our claim that this method of identifying cognates and false friends between two languages is a research contribution.

Automatic partial cognate disambiguation is a task that in our knowledge was not of research interest before, at least in the computational comunity. The contributions that we bring to the NLP community are: the task itself — we will show later that it is useful for many other NLP tasks; the methods proposed to disambiguate a partial cognate: a supervised method and a semi-supervised method based on a bootstrapping technique, both using ML algorithms; and the idea of using a combination of corpora

from different domains. We show that even though we started with small corpora from a parliamentary domain, the knowledge of the methods can be boosted using corpora from different domains, more general; this way the method can disambiguate partial cognates in a larger variety of contexts.

Our practical aim for the research studies that we have done for this thesis is to have a Computer-Assisted Language Learning tool that will use the knowledge of cognates and false friends between languages. Besides the research contributions that we bring with our studies, we also come with a practical contribution consisting in a CALL tool that can help second language learners of French in the reading comprehension task. The tool capabilities can also be integrated in other existing CALL tools.

More details of the proposed methods and experiments that we have done are presented mostly in the third and fourth chapter of the thesis.

Chapter 2

Related Work

2.1 Identification of Cognates and False Friends

Previous work on automatic cognate identification is mostly related to bilingual corpora and translation lexicons. Simard, Foster, & Isabelle (1992) use cognates to align sentences in bitexts. They employ a very simple test: French-English word pairs are assumed to be cognates if their first four characters are identical.

Brew & McKelvie (1996) extract French-English cognates and false friends from aligned bitexts using a variety of orthographic similarity measures based on DICE's coefficient measure. They look only at pairs of verbs in French and English, pairs that were automatically extracted from the aligned corpus. They conclude that XXDICE, a variation of DICE measure performed best for a threshold selected by hand. Mann & Yarowsky (2001) automatically induce translation lexicons on the basis of cognate pairs. They find that edit distance with variable weights outperforms both Hidden Markov Models and stochastic transducers.

Guy (1994) identifies letter correspondence between words and estimates the likelihood of relatedness. No semantic component is present in the system, the words are assumed to be already matched by their meanings. Hewson (1993),
Lowe & Mauzaudon (1994) use systematic sound correspondences to determine proto-

projections for identifying cognate sets.

One of the most active researchers in identifying cognates between pairs of languages is Kondrak (2001, 2004). His work is more related to the phonetic aspect of cognate identification, especially genetic cognates. He uses in his work algorithms that combine different orthographic and phonetic measures, recurrent sound correspondences, and some semantic similarity based on gloss overlap. In (Kondrak, 2004) looks directly at the vocabularies of related languages to determine cognates between languages. Kondrak & Dorr (2004) report that a simple average of several orthographic similarity measures outperforms all individual measures on the task of the identification of drug names.

Mackay & Kondrak (2005) identify cognates using Pair Hidden Markov Models, a variation on Hidden Markov Models that has been used successfully for the alignment of biological sequences. The parameters of the model are automatically learned from training data that consists of word pairs known to be similar. The results show that the system outperforms previously proposed techniques for the task of identifying cognates.

Complex sound correspondence was also used by Kondrak (2003) to help the process of identifying cognates between languages. The algorithm was initially designed for extracting non-compositional compounds from bitexts, and was shown to be capable of determining complex sound correspondences in bilingual word lists. He reports 90% results for precision and recall for cognate identification.

For French and English, substantial work on cognate detection was done manually. LeBlanc & Séguin (1996) collected 23,160 French-English cognate pairs from two general-purpose dictionaries: Robert-Collins (Robert-Collins, 1987) and
Larousse-Saturne. 6,447 of these cognates had identical spelling, disregarding diacritics. Since the two dictionaries contain approximately 70,000 entries, cognates appear to make up over 30% of the vocabulary.

There is a considerable amount of work done on cognate identification, but not as much for false-friend identification. In fact, we could not point out work that

has been focusing on false-friend identification between English and French from free lists of pairs of words. However work on cognate and false friend identification in German and English was done by Friel & Kennison (2001). They were looking at a set of 563 German-English pairs of nouns for the purpose of identifying cognates, false cognates and non-cognates. Two techniques for identifying cognates were used and compared: (i) (Groot & Nas, 1991) similarity-rating technique - using human knowledge to determine the similarity in sound and spelling of a pair of translated words and (ii) a translation-elicitation task similar to that of Kroll & Stewart (1994) who used native English speakers with no knowledge of Dutch or German to translate a list of Dutch words. Words that were correctly translated by more than 50% of the participants were treated as cognates. The results obtained by Friel & Kennison (2001) with English-speaking participants produced 112 cognates, 94 false cognates, and 357 non-cognates and indicated that the two techniques yielded similar findings.

Barker & Sutcliffe (2000) proposed a semi-automatic method to identify false cognates between English and Polish. The method uses a set of morphological transformation rules to convert an English word into a number of candidate Polish "words". The resulted Polish words are classified as being: false cognates, true cognates, unrelated and nonexistent based on human judgement.

Identifying cognates and false friends has been an attractive research area not only for people interested in NLP but also for linguists and psycholinguists.

2.2 Related Research Areas

In this section we present areas that have been successfully using cognates and false friends.

2.2.1 Cognates and False Friends in Language Learning

Linguists have studied the impact of false friends and cognates in second language learning for a long period of time. They suggest that cognate use and recognition bring improvement in vocabulary acquisition and reading comprehension, and provide a head start in language learning (LeBlanc, 1989).

Studies undertaken for French (Tréville, 1990) and Spanish (Nagy, 1992), (Hancin-Bhatt & Nagy, 1993) show the importance of cognate recognition in reading comprehension and more importantly, the awareness of cognate relationships in reading strategies. Researchers have concluded that explicit instruction of cognate pairing will increase learner's utilization of cognate knowledge.

When learning a second language, a student can benefit from knowledge in his/her first language (Gass, 1987) (Ringbom, 1987). Kroll *et al.* (2002) look at the way students use their knowledge of the first language (L1) to transfer it to the second language (L2).

Morphological rules of conversion between English and French also proved helpful in cognate identification in language learning. LeBlanc (1989) proposed 2,529 such rules. An example is: *cal* $\rightarrow$ *que* in pairs such as *logical* - *logique*, *political* - *politique*.

Awareness of the morphological relationship among words creates a better metalinguistic and metacognitive knowledge and the more similarity in the structure of morphological rules between language pairs, the broader the possibility for cognate recognition in L2.

The morphological rules seem to be helpful in language learning when there has been an exposure to the language for a few years at high level of discourse. Studies done by Hancin-Bhatt & Nagy (1993) for French and Spanish with students of different agegroups support the claim. Second language learners of German that are native English speakers were studied by Dollenmayer & Hansen (2003) to show that students themselves attempt to guess the meaning of cognates rather than just point out the phonemic relationship resulting from historical sound shifts.

In 1988 Palmberg (1988) conducted experiments with Swedish-speaking students of

English to show that orthographic processing of words is a better facilitator of cognate recognition than oral input.

As we have seen by now, cognates have an important role in language learning but, on the other hand, a student has to pay attention to the pairs of words that look and sound similar but have different meanings — false-friends pair, and especially to pairs of words that share meaning in some but not all contexts — partial cognates.

It is good news for second-language learners that in general the number of false friends and partial cognates between languages are not as high as the number of cognates, especially for language that are etymologically closely related. Hammer (1976) draws our attention to the fact that in most related languages the number of cognates is much greater than the number of false friends. He compared English and French and concluded that the ratio of cognates to false friends was approximately eleven to one. On the other hand, Friel & Kennison (2001) have shown in a study that the number of false friends between German and English is greater than the number of cognate pairs.

Claims that false friends can be a hindrance in second language learning are supported by the studies of (Carroll 1992). She suggests that a cognate pairing process between two words that look alike happens faster in the learner's mind than a false-friend pairing. Experiments with second language learners of different stages conducted by Heuven, Dijkstra, & Grainger (1998) suggest that missing false-friend recognition can be corrected when cross-language activation is used — sounds, pictures, additional explanation, feedback.

DIDALECT (Balcom, Copeck, & Szpakowicz, 2006) is a project dedicated to second language learners of French developed by the Second Language Institute, University of Ottawa. It is a project for which we have collaborated with false-friend annotation of French texts.

Partial cognates, words that in some context behave like cognates and in others like false friends, can also be useful in Computer-Assisted Language Learning (CALL) tools.

Search engines for E-Learning can find a partial cognate annotator useful. A teacher

who prepares a test to be integrated into a CALL tool can save time by using our methods of automatically disambiguating partial cognates, even though the automatic classifications need to be checked by the teacher.

As pointed in this subsection, the use and usefulness of cognates and the awareness of false friends is strongly integrated with language learning.

2.2.2 Cognates and False Friends in various NLP applications

Machine Translation (MT) systems can benefit from additional information when translating a certain word in context. Knowing if a word in the source language is a cognate or a false friend with a word in the target language can improve translation results.

MT and Word Alignment have been using with success cognate pairs between two languages. (Marcu, Kondrak, & Knight, 2003) report results of experiments aimed to improve translation quality by incorporating the cognate information into translation models. The results confirm that the cognate identification approach can improve the quality of word alignment in bitexts without the need for extra resources.

MT benefits from cognate knowledge and Word Sense Disambiguation (WSD). (Vickrey *et al.*, 2005) have proposed an algorithm to determine the correct translation of a word in context. They have evaluated the proposed technique for French and English and reported a 95% recall for translating English ambiguous words into the correct French translation. They are looking at WSD as a task of finding the correct translation word in a target language for an ambiguous word in the source language.

Carpuat & Wu (2005) empirically argue that WSD does not help Statistical Machine Translation to produce better translations. Their method of integrating WSD into MT systems did not bring an improvement in the translation. Vickrey *et al.* (2005) argue that the method is not as flexible as theirs; this is why the improvement of the results did not appear.

Machine Translation is related to word alignment. The knowledge about cognate

existence between languages is used for word alignment as well, and implicitly for MT. Simard et al. (1992) have shown that this additional knowledge of cognates helps word alignment and MT systems. Isabelle (1993) proposed a method of identifying *deceptive cognates* false friends in our work — in bitexs. Their work was included in a translation checking tool called *TransCheck*[1].

Tufis et al. (2005) have used cognates between Romanian and English for a word alignment tool named **COWAL** that performed best on the shared task on word alignment, which was organized as part of the ACL 2005 Workshop on Building and Using Parallel Texts.

Cross-Language Information Retrieval systems can use the knowledge of the sense of certain words in a query in order to retrieve desired documents in the target language. We are not aware of work that has been done on cross-language information retrieval system that use cognate knowledge, so far.

As we could see in this section, cognates are useful not only in language learning but also in MT, one of the most interesting and challenging areas of NLP.

2.2.3 Linguistic Distances

In the Computational Linguistic (CL) research, linguistic distances and linguistic similarity are notions that are frequently used. They are present in almost all semantic tasks in NLP (e.g., Information Retrieval, Word Sense Disambiguation, Information Extraction, Question Answering etc).

A brief survey for some semantic measures that are frequently used for analyzing text is done by Lebart & Rajman (2000). Mesures used in Information Retrieval (IR) and Text Mining are also presented.

Besides the CL world, Linguistics also uses notions of similarity. Areas like historical linguistics, second-language learning (for learners' proficency), psycholinguistics, are just a few domains that use this notion.

[1] http://rali.iro.umontreal.ca/Traduction/TransCheck.en.html

We can see the distance between two words, two texts, two languages, etc., from any of the following aspects:

a. Phonetic

b. Syntactic

c. Semantic

a. Phonetic

Albright & Hayes (2003) have done research on the phonetic similarity looking at a model of phonological learning from the "minimal generalization" point of view. The minimal generalization refers to minimal distance in pronunciation. They show that children learn on the basis of slight generalization. They give as an example the formation of the past tense of verbs ending in 'ing' (e.g., sing, sting, string). These verbs have the past tenses ending in 'ung'.

Kessler (1995) work shows how can edit distance be used to automate pronunciation differences to better analyze the dialectology aspect of a language.

Kondrak & Sherif (2006) present and compare several phonetic similarity algorithms for the cognate identification task. The results show that Machine Learning techniques perform well for this task.

b. Syntactic

Syntactic Typology is an area of linguistic theory that tries to identify syntactic features that are associated in languages. The goal of this reserach is to show that some languages are more similar to one another than they would appear (Croft, 2001).

Thomason & Kaufmann (1988) looked at the syntactic level of the language contact and influence between two languages that are used in the same community. Languages change and borrow words from each other if they are in contact(e.g., political, cultural, economical). The same studies are done in second-language learning. As we cited in the previous sections, research on how the first language knowledge is projected in a second language is of great interest for psycholinguists.

c. Semantic

One of the directions that are followed on lexical semantics is to identify verb classes that have similar syntactic and semantic behavior. Levin (1993) studied some of the English verb classes. Context similarity is always a good measure to use to determine if two words are used with the same meaning.

From our point of view we measure the linguistic distance between two languages using the cognates and false friends that exist in the two languages. If the languages are strongly related (come from the same branch of languages e.g., latin, slavic) the number of cognates that exist between the two languages is large.

One of the methods that we propose to identify cognates and false friends between languages uses Machine Learning algorithms. The algorithms are trained on lists of cognates and false friends. If two languages are not related (distant) might have few cognates in common, therefore there will not be a lot of training data that can be used by our method. This method, is presented in Chapter 3.

We also propose a method to disambiguate partial cognates in Chapter 4. The method determines if a partial cognate is used with a cognate or a false friend sense. Our automatic way to disambiguate partial cognates uses parallel text in two languages and Machine Learning techniques. Therefore it can be applied to other languages (regardless their degree of relatedness) as long as parallel text is available.

2.2.4 Word Sense Disambiguation

Word Sense Disambiguation is considered one of the most interesting and longest-standing problems in NLP. It got the researcher's attention since the beginning of automated treatment of the language, in 1950's, and never stopped ever since. As Ide & Veronis (1998) suggest, the task is not necessary a standalone one, but it is needed in almost any other task of language processing.

Many words have more than one possible meaning, they are either *homonyms* — words that are written and pronounced the same way, but have different meanings (e.g.

lie can be a verb meaning to tell something that is not true or to be in a horizontal position) or *polysemous* — words that have multiple related meanings (e.g. *bank* can be a *a river bank*, *money bank* or a *memory bank*).

WSD is the task that tries to determine what sense of a homonym or polysemous word is used in a certain context.

Definitions

Word Sense Disambiguation is the problem of selecting a sense for a word from a set of predefined possibilities. The sense inventory usually comes from a dictionary or thesaurus. The problem can be resolved with knowledge-intensive methods, supervised learning, and (sometimes) through bootstrapping approaches.
Word Sense Discrimination is the problem of dividing the usages of a word into different meanings, without regard to any particular existing sense inventory. Word Sense Discrimination uses unsupervised techniques for sense clustering.

Research on WSD

Word Sense Disambiguation refers to the resolution of lexical semantic ambiguity and its goal is to attribute the correct senses to words in a given context. The task has been described as an Artificial Intelligence (AI)-complete problem, which usually refers to problems that require a human-level intelligence to be solved. If we look at the following newspaper headlines, we can understand that the task of disambiguating a sense of polysemous words is hard. (e.g. *RESIDENTS CAN DROP OFF TREES; INCLUDE CHILDREN WHEN BAKING COOKIES; MINERS REFUSE TO WORK AFTER DEATH.*)

WSD Approaches and Solutions WSD was noticed as a problem in Machine Translation by Weaver (1949), but after that it became a task that is required in almost any NLP application.

Some of the most known approaches for WSD that are used my researchers are based on the way they acquire the information. Two of the most known ones are:

i. Corpus-Based Approaches

ii. Knowledge-Based Approaches

i. Corpus-Based Approaches In the corpus-based approaches, the information needed to disambiguate a polysemous word is extracted from a collection of data, a corpus. The corpus will provide sets of samples for each sense of the word that has to be disambiguated. The knowledge gained from the corpus is obtained through a training process.

The corpus-based method can be further classified into two subclasses, based on the type of training corpus:

a. Supervised Disambiguation

b. Unsupervised Disambiguation

In the supervised approach the training data will be sense-tagged, each example from the training corpus has the sense attached. In unsupervised methods the training data is raw corpora, it is not semantically disambiguated.

a. Supervised Disambiguation Supervised disambiguation is closely related to supervised learning techniques that create classifiers that are later used for disambiguation.

Machine Learning (ML) techniques are the ones that are commonly used to build WSD classifiers. Machine learning is an automatic process of constructing classifiers from a large collection of instances (Mitchell 1997). In order to use machine learning techniques, each instance — sample from the given corpus — will first be transformed into a feature representation, usually a feature vector $f_v = ((f_1, v_1), (f_2, v_2), , (f_n, v_n))$, where f_i

is a feature and v_i is its corresponding value. Appropriate feature representations should capture features with high discrimination power, while the number of different features should be kept as small as possible in order to have classifiers with good generalization capabilities. The feature representation of the data corresponds to the first step required to use some ML techniques.

The second step consists in training the classifier on the disambiguated corpus, where each occurrence of an ambiguous word is annotated with the appropriate sense used in that context. The aim of supervised disambiguation is to build a model based on what was seen in the training step, and use it to classify new cases where an ambiguous word appears. The sense that will be assigned to a new target word use is decided by taking into account the context. The context of the word to be disambiguated includes information contained within the text or discourse in which the word appears together with extra-linguistic information about the text (e.g., syntactic relations), if available.

One of the major problems that appear with the supervised approaches is the need for large sense-tagged training corpora. Despite the availability of large corpora, manually sense-tagged corpora are very few mostly because the high cost of time and human effort.

Some of sense-tagged corpora available for use are: the SemCor corpus
(Landes, Leacock, & Tengi, 1998) and the SENSEVAL-1 corpus
(Kilgarriff & Rosenzweig, 2000). The SemCor corpus, created by the Princeton University, is a subset of the English Brown corpus[2] containing almost 700,000 content words. In SemCor, all the words are part-of-speech tagged and more than 200,000 content words are also lemmatized and sense-tagged according to Princeton WordNet 1.6 (mappings for later versions of WordNet are also available). SENSEVAL-1[3] corpus is derived from the HECTOR corpus (Atkins 1993) and dictionary project. It is a joint Oxford University Press and Digital project which took place in the early 1990s. In the course of the project a 20-million word corpus which also served as a pilot for the British National Corpus

[2]http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
[3]http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/index.html

(BNC)[4] was developed.

SENSEVAL-2[5] and SENSEVAL-3[6] competitions produced large sets of annotated data and for different languages as well. Open Mind Word Expert project (Chklovski & Mihalcea, 2002) is designed to be an active learning system that is capable of collecting word sense tagged corpora using the Web.

In order to avoid the lack of large sense-annotated corpora, several methods have been proposed to automatically sense-tag a training corpus using **bootstrapping** methods. Bootstrapping relies on a small number of instances of each sense for each word of interest. These sense-tagged instances are used as seeds to train an initial classifier. This initial classifier is then used to extract a larger training set from untagged data. The process is usually done in several steps, each time an increase of the training corpus is obtained.

Hearst (1991) proposed an algorithm, *CatchWord*, which used a small set of initial training data for a bootstrapping algorithm. From the initial manually annotated sets for a set of nouns, statistical information is extracted in order to be used later to automatically tag other occurrences. If another occurrence of an ambiguous noun was disambiguated with a certain level of confidence then the instance would be used for the next step of training and new additional statistical information would be added to the method. Based on all performed experiments some conclusions were suggested: an initial set of at least 10 occurrences is necessary for the procedure; 20 or 30 occurrences are necessary for high precision.

Yarowsky (1995) has used a few seeds and a set of untagged sentences in a bootstrapping algorithm based on decision lists. He added two constrains — words tend to have one sense per discourse and one sense per collocation. *One sense per collocation* argues that nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship. *One sense per discourse* constraint argues that the sense of a target word is highly consistent within

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

[5]http://193.133.140.102/senseval2/

[6]http://www.senseval.org/senseval3

any given document. He reported high accuracy scores for a set of 10 words.

Mihalcea & Moldovan (2001) and Mihalcea (2002) have also used bootstrapping techniques. Bootstrapping technique initializes a set of ambiguous words with all the nouns and verbs in the text, and then applies various disambiguation procedures and builds a set of disambiguated words. The sense tags for new words are determined based on their relation to the already disambiguated words and then added to the set. With this method, 55% of the verbs and nouns from the text are disambiguated with an accuracy of 92%.

Bootstrapping is not the only technique that was suggested in order to avoid the need of hand-tagged training data. Available **parallel corpora** are used with success in WSD.

Brown, Lai, & Mercer (1991), Gale, Church, & Yarowsky (1992), and
Gale & Church (1993) used parallel corpora to avoid hand-tagging corpora with the premise that differently senses will be lexicalized different in another language (for example, *pen* in English is *stylo* in French for its *writing implement* sense, and *enclos* for its *enclosure sense*). With parallel aligned corpus, the sense of a target word can be determined looking at the translation of each occurrence of the word. Some problems can be encountered because ambiguities can be preserved in a target language as well and because of the domain of the parallel corpus (*Hansard Corpus* of Canadian Parliamentary debates) that can influence the sense distribution.

Diab & Resnik (2002) has shown that WSD systems that use parallel corpora can achieve good results. They used parallel corpora in French, English, and Spanish,
automatically-produced with MT tools, to determine cross-language lexicalization sets of target words. The major goal of the work was to perform monolingual English WSD. Evaluation was performed on the nouns from the English all-words data in Senseval2. Additional knowledge was put in the system from WordNet in order to improve the results.

Li & Li (2004) have shown that word translation and bilingual bootstrapping is a

good combination for disambiguation. They used a set of 7 pairs of Chinese and English words. The two senses of the words were highly distinctive: e.g. *bass* as *fish* or *music*; *palm* as *tree* or *hand*.

b. Unsupervised Disambiguation In unsupervised disambiguation, the information needed for later use, is gathered from raw corpora which have not been semantically disambiguated. Unsupervised methods are closely related to clustering tasks rather than sense-tagging tasks. A completely unsupervised disambiguation will not be possible for word senses since sense tagging requires a sense inventory. Sense discrimination is a task that can be resolved fully unsupervised, while sense disambiguation needs an additional source of knowledge necessary to define the senses. If we see WSD as a combination of sense discrimination and sense labeling then sense discrimination is the part that can be determined using an unsupervised manner.

Based on the results that were obtained by different methods and techniques the accuracy of unsupervised WSD systems are in general with 5 to 10% lower than for the supervised approaches.

ii. Knowledge-Based Approaches The average number of lexical items that are treated with supervised WSD methods is not very high (4 to 12), except for the Senseval competitions. This is due to the amount of human effort needed to create a corpus that can be used with corpus-based methods. Large-scale WSD can be obtained when using large-scale lexical resources: dictionaries and thesauri.

Methods that use this kind of resources are presented in the following sections.

Machine-Readable Dictionaries Machine-readable dictionaries (MRD) provide a ready-made information source of word senses. The first attempt to use MRDs came from Lesk (1986). The main idea presented in his work is that a word's dictionary definitions are likely to be good indicators of its senses. By using Oxford Advanced Learner's Dictionary (OALD), he counted overlapping content words in the sense

definitions of the ambiguous word and the context words of the target lexeme in a certain sentence. The sense that achieves the maximum number of overlaps is selected. The accuracy of the method varies between 50-70%. The results were encouraging since a fine set of sense distinctions has been used.

In general, dictionaries are created for human use not for computers this is why some inconsistencies can appear (Véronis & Ide, 1991), (Ide & Véronis, 1993a, 1993b). The main drawback when using a dictionary is the lack of pragmatic information used for sense determination. As suggested by Ide & Veronis (1998) the relation between *ash* and *tobacco*, *cigarette* or *tray* is very indirect in a dictionary whereas the word *ash* co-occurs very frequently with these words in a corpus.

Thesauri One of the most used thesauri in WSD is Roget's International Thesaurus which was computationally available from 1950s. Thesauri are semantic resources that provide information about relationships that exist among words. A WSD system will use the semantic categorization that a thesaurus provides. The semantic categories of each context word can determine the semantic category of the whole context category; this can determine the correct sense of a polysemous word.

The main reason why thesauri are not that often used in WSD systems is because they do not provide enough information about word relations. Jarmasz & Szpakowicz (2001) have been looking at transforming Roget's thesaurus in a relational lexical semantic resource. They were looking at the similarities and differences between Roget's thesaurus and WordNet and also at a possible way to connect these useful resources.

Lexical relations are needed in linguistics and psycholinguistics. Research that was focused on creating large electronic databases with lexical relations has and it is still done. If in the early years the main interest was the English language, now relational resources are available or in progress for other languages as well: French, Spanish, Romanian, Polish, etc.

One of the most commonly used computational lexicon for English is **WordNet**.

Wordnets became available for other languages too. Tufis, Radu, & Ide (2004) used cross-lingual lexicalization, wordnets alignment for several languages, and a clustering algorithm to perform WSD on a set of polysemous English words. They report an accuracy of 74%. Ide, Erjavec, & Tufis (2001) used both parallel corpora in six languages and knowledge from WordNet. Mihalcea & Faruque (2004) proposed a minimally supervised sense tagger that attempts to disambiguate all content words in a text using the senses from WordNet. The algorithm called SENSELEARNER obtained an average of 64.6% accuracy on SENSEVAL-3 English all-words task.

As we have seen in this section WSD is an NLP task not only that is challenging and attractive, but also very useful.

Chapter 3

Identification of Cognates and False Friends

As we have seen in the previous chapter, cognates and false friends are cross-language pairs of words that are useful in many NLP tasks. Our main goal for using cognates and false friends is second language learning. Linguistic studies for several languages: French, English, German, and Dutch etc., have shown that the cognate knowledge and use are extremely beneficial for second language learning students.

Grosjean (1997, 1998) has pointed out that there is still confusion regarding the definition of a cognate despite the growing interest in cognates and the growing empirical literature related to the storage and processing of cognates in human memory. Although most would agree that cognates are translations similar in sound and appearance, researchers have been using several definitions. For example, Sanchez-Casas & Garcia-Albea (1992) defined their Spanish-English cognates as translation pairs having a common original root word. Similarly, Gollan & Frost (1997) defined their cognates as translation pairs in which Hebrew words were borrowed from English.

Groot & Nas (1991) asked Dutch-English bilinguals to rate Dutch-English translations on a scale from 1 to 7 in terms of similarity in sound and spelling to corroborate the translation pairs they labeled cognates in a previous experiment. In all but one case,

the translation pairs they labeled cognates were rated higher than the pairs they labeled non-cognates.

Kroll & Stewart (1994) developed another method of identifying cognates. Native English speakers with no knowledge of Dutch or German were asked to translate a list of Dutch words. Words that were correctly translated by more than 50% of the participants were treated as cognates.

The definitions that we adopt in our work are the ones that have been presented in the first chapter. Based on these definitions, our method of cognate and false friend identification differs from the previous ones not only in approach that we follow (combine different orthographic similarity measures with ML techniques), but also by the usefulness of the results that we obtain: automatically determined thresholds for several orthographic similarity measures. We work with 13 orthographic similarity measures, separately and by combining them through ML techniques. All 13 orthographic similarity measures are implemented in a Perl package by Kondrak (2005).

A cognate pair can be recognized for its acoustic match or for its orthographic resemblance. Both the acoustic match and the orthographic one can sometime introduce errors. In our work we look only at orthographic matching of cognate pairs between two languages.

Our focus is to automatically identify cognates and false friends for the purpose of preparing lists of them to be included in dictionaries and other learning aid tools. Cognate lists exist only for a few language pairs. Moreover, they are expensive to create because they require large human effort. Automatic ways to create such lists will save lot of time and human effort — but will be less accurate.

Our approach is based on 13 orthographic similarity measures that we use as features for classification. We test each feature separately; we also test for each pair of words the average value of all the features. Then we explore various ways to combine the features, by applying several machine learning techniques from the Weka package (Witten & Frank, 2005). The two classes for the automatic classification are:

Cognates/False-Friends and Unrelated. Cognates and False-Friends can be distinguished on the basis of an additional "translation" feature: if the two words are translations of each other in a bilingual dictionary, they are classified as Cognates; otherwise, they are assumed to be False Friends.

In the following sections of this chapter we will describe the orthographic measures that we used in our method, our proposed method, and the data that we used. Evaluation results for all conducted experiments and conclusions obtained from our work are also presented in this chapter.

3.1 Orthographic Similarity Measures

In this section, we will describe the measures that we use as features for the cognate and false friend classification task. Many different orthographic similarity measures have been proposed and their goal is to quantify human perception of similarity, which is often quite subjective. Each measure that we used returns a real value (between 0 and 1, inclusive) that describes the similarity between two words. In the following part of the section we explain each method followed by some examples.

- IDENT is a measure that we use as a baseline. The measure returns 1 if the words are identical, and 0 otherwise.

- PREFIX is a simple measure that returns the length of the common prefix divided by the length of the longer string[1]. E. g., the common prefix for *factory* and *fabrique* has length 2 (the first two letters) which, divided by the length string 8, yields 0.25.

- DICE (Adamson & Boreham 1974) is calculated by dividing twice the number of shared letter bigrams by the total number of bigrams in both words:

$$\text{DICE}(x, y) = \frac{2|bigrams_{(x)} \cap bigrams_{(y)}|}{|bigrams_{(x)}| + |bigrams_{(y)}|}$$

[1]The PREFIX measure can be seen as a generalization of the approach of Simard, Foster, & Isabelle (1992).

where *bigrams(x)* is a multi-set of character bigrams in word x. E. g., DICE(*colour*, *couleur*) = 6/11 = 0.55 (the shared bigrams are *co*, *ou*, *ur*).

- TRIGRAM is defined in the same way as DICE, but employs trigrams instead of bigrams.

- XDICE (Brew & McKelvie 1996) is also defined in the same way as DICE, but employs "extended bigrams", which are trigrams without the middle letter.

- XXDICE (Brew & McKelvie 1996) is an extension of the XDICE measure that takes into account the positions of bigrams. Each pair of shared bigrams is weighted by the factor:

$$\frac{1}{1+(pos(a)-pos(b))^2}$$

where $pos(a)$ is the string position of the bigram a[2].

- LCSR (Melamed 1999) stands for the Longest Common Subsequence Ratio, and is computed by dividing the length of the longest common subsequence by the length of the longer string. E. g., LCSR(*colour*, *couleur*) = 5/7 = 0.71

- NED is a normalized edit distance. The edit distance (Wagner & Fischer 1974) is calculated by counts up the minimum number of edit operations necessary to transform one word into another. In the standard definition, the edit operations are substitutions, insertions, and deletions, all with the cost of 1. A normalized edit distance is obtained by dividing the total edit cost by the length of the longer string.

- SOUNDEX (Hall & Dowling 1980) is an approximation to phonetic name matching. SOUNDEX transforms all but the first letter to numeric codes and after

[2]The original definition of XXDICE does not specify which bigrams should be matched if they are not unique within a word. In our implementation, we match non-unique bigrams in the order of decreasing positions, starting from the end of the word.

removing zeros truncates the resulting string to 4 characters. For the purposes of comparison, our implementation of SOUNDEX returns the edit distance between the corresponding codes.

- BI-SIM, TRI-SIM, BI-DIST, and TRI-DIST belong to a family of n-gram measures (Kondrak & Dorr 2004) that generalize LCSR and NED measures. The difference lies in considering letter bigrams or trigrams instead of single letter (i. e., unigrams). For example, BI-SIM finds the longest common subsequence of bigrams, while TRI-DIST calculates the edit distance between sequences of trigrams. n-gram similarity is calculated by the formula:

  $$s(x_1 \ldots x_n, y_1 \ldots y_n) = \frac{1}{n} \sum_{i=1}^{n} id(x_i, y_i)$$

  where $id(a, b)$ returns 1 if a and b are identical, and 0 otherwise.

In Table 3.1 we give an example of all orthographic similarity measures for the fallowing pair of words: *acompte account*.

3.2 Method

Our contribution to the task of identifying cognates and false friends between languages is the method itself, the way we approach the identification task by using ML techniques. Other methods that have been proposed for cognate and false friend identification require intensive human knowledge (Barker & Sutcliffe, 2000), (Friel & Kennison, 2001).

As we described in the second chapter, ML techniques require the data to be in a certain format. Each instance that is going to be used by the technique has to be transformed into a feature value representation. We can associate the representation with a flat relational data base where each row is an instance and the columns are the features used to represent the data.

From the Weka package (Witten & Frank 2005) we used different supervised ML algorithms to best discriminate between the two classes that we have chosen: Cog-

Measure	Value
DICE	0.3333
EDIT	0.4286
IDENT	0.0000
LCSR	0.5714
SIMARD	0.2857
SOUNDEX	0.7500
TRI	0.0000
XDICE	0.1818
XXDICE	0.1364
BI-DIST	0.4286
BI-SIM	0.5714
TRI-DIST	0.4286
TRI-SIM	0.5714

Table 3.1: Result of all orthographic measures for the pair of words: *acompte account*.

nates/False Friends — are orthographically similar, and Unrelated — are not orthographically similar. Our classes are similar to those of Friel & Kennison (2001) and Barker & Sutcliffe (2000), described in chapter 2, in their classification using human judges. Our method is based on different ML algorithms that we try with our data and with the chosen feature representation.

Appendix 1 will present a small part of our data set along with their feature value representation, representation that is used for the ML classifiers.

3.2.1 Instances

Instances are small parts of the whole data that are used in a ML technique, that have a label attached. The label is the class to which the instance belongs. In ML techniques we want to create classifiers that are able to discriminate between different classes.

What an instance is and how it is represented are interesting aspects of ML techniques. A big role here is taken by the human knowledge and intuition. The choices of data representation differ from method to method and from task to task.

In our method an instance is a pair of words containing a French word and an English word. The data that we use are different lists of pairs of words that will be described in detail in the next section.

3.2.2 Features and Feature Values

The features that we have chosen to use in our method are the 13 orthographic similarity measures that we described in Section 3.1. In different experiments we used different features: each orthographic measure separately as a feature — the ML algorithm will use data represented by 13 features, features that correspond to one of the measures; the average of the results from the whole 13 orthographic measures as a single feature — the ML algorithm will use data represented by one feature that corresponds to the average value of all measures.

When we want to determine the threshold for each measure we use only one feature. When we want to see if the average of all measures performs better than each measure in part or all put together, we used again only one feature that has as value the average of all measures.

No matter what are the features that the method uses, the values of the features are real numbers between 0 and 1 (inclusively) that reflect the orthographic similarity between two words that belong to one instance, a pair of French-English words, in our experiments.

In addition to other research that has been done on cognate and false friend identification, we look at different orthographic measures combined by ML techniques to classify pairs of words as being cognates, false friends or unrelated pairs. In Section 3.4 we will present the results that we obtained using our method.

3.2.3 Semantic Dimension of the Method

The main focus of the method is to determine how well different combinations of orthographic similarity measures perform in order to discriminate cognates and false friends from other pair of words that we call Unrelated.

In our work, when we refer to the semantic dimension we do not refer to a deep aspect of it. When we refer to the semantic dimension we look at the meaning of words contained in a pair, more exactly if they share the same meaning. The words are translations of each other if and only if they have the same meaning. Throughout all the sections that follow, when we refer to the semantic dimension of the method we refer to the words having the same meaning looking only if they are translations of each other or not.

We added the semantic dimension to the method in two ways. Both ways are using a bilingual dictionary or prior knowledge that the pairs are translation of each other. The methods are described in the following two paragraphs.

a. Additional Feature with Three-Class Classification — adding a boolean feature to the orthographic features. This feature will tell us if the words are translations of each other. In these experiments the classes in which a pair will be classified are three: Cognates, False Friends and Unrelated. Here we split the class Cognates/FalseFriends in two based on the translation feature. Instead of having two classes: Cognates/FalseFriends — orthographically similar and Unrelated — not orthographically similar, we will have three classes that will separate the first class taking into account the fact that they are translation of each other.

b. Data Collection and Dictionary Use - if the pairs on which we evaluate the method are translations of each other and if the method determines that they are orthographically similar then we have obtained lists of cognates, and otherwise we have obtained false friend lists. When we have pairs of words of which we do not know if they are mutual translation, we use an additional semantic resource: a French-English bilingual dictionary.

As we will see in subsection 3.4.4 and Section 3.5, we experimented with both ways

of adding the semantic dimension.

In the following sections we present the data that we used, how we obtained it, and the evaluation results for the collected data.

3.3 The Data

This section will present the data used to evaluate our proposed method to identify cognates and false friends.

3.3.1 Training and Testing Data Set

The training dataset that we used consists of 1454 pairs of French and English words (see Table 3.2). None of the the pairs that we worked with contain multi-word expressions. They were extracted from the following sources:

1. An on-line[3] bilingual list of 1047 basic words and expressions. (After excluding multi-word expressions, we manually classified 203 pairs as Cognates and 527 pairs as Unrelated.)
2. A manually word-aligned bitext (Melamed 1998). (We manually identified 258 Cognate pairs among the aligned word pairs.)
3. A set of exercises for Anglophone learners of French (Tréville 1990) (152 Cognate pairs).
4. An on-line[4] list of "French-English False Cognates" (314 False-Friends).

A separate test set is composed of 1040 pairs (see Table 3.2), extracted from the following sources:

1. A random sample of 1000 word pairs from an automatically generated translation lexicon. We manually classified 603 pairs as Cognates and 343 pairs as Unrelated.

[3]http://mypage.bluewin.ch/a−z/cusipage/basicfrench.html

[4]http://french.about.com/library/fauxamis/blfauxam.htm

2. The above-mentioned on-line list of "French-English False Cognates" , 94 additional False-Friends not used for training.

	Training set	Test set
Cognates	613 (73)	603 (178)
False Friends	314 (135)	94 (46)
Unrelated	527 (0)	343 (0)
Total	1454	1040

Table 3.2: The composition of data sets. The numbers in brackets are counts of word pairs that are identical (ignoring accents).

In order to avoid any overlap between the two sets, we removed from the test set all pairs that happened to be already included in the training set.

The dataset has a 2:1 imbalance in favor of the class Cognates/False-Friends; this is not a problem for the classification algorithms (the precision and recall values are similar for both classes in the experiments presented in Evaluation section 3.4). All the Unrelated pairs in our datasets are translation pairs. It would have been easy to add more pairs that are not translations, but we wanted to preserve the natural proportion of cognates in the sample translation lexicons.

From the whole data set 73 cognates and 135 false friends in the training dataset have identical spelling in both languages. When counting identical words we ignore the accents in the French words. The number of identical pairs without ignoring diacritics is: 58 cognates and 121 false friends.

This is the data on which we evaluated our method and on which we determined the threshold for each measure. On the same data we evaluated our method when adding the extra boolean feature that gave us the semantic dimension.

3.3.2 Genetic Cognates Data Set

We also experiment with our method on a small set of genetic cognates, pairs of words derived from the same word in the proto-language; many of which underwent a major orthographic change. Greenberg (1987) gives a list of "most of the cognates from French and English". The list serves as an illustration how difficult it would be to demonstrate that French and English are genetically related by examining only the genetic cognates between these two languages. Inkpen, Frunza, & Kondrak (2005) transcribed the list of 82 cognate pairs from International Phonetic Alphabet (IPA) to standard orthography. They augmented the list with 14 pairs from the Comparative Indo-European Data Corpus[5] and 17 pairs that we identified ourselves. The final lists[6] contains 113 true genetic cognates from the Proto-Indo-European language.

3.3.3 Data Sets Collected for the Semantic Dimension of the Method

The experiments when we add semantics to the method (presented in subsection 3.2.3) are done using the pairs of words described in the next two paragraphs. The data that we are going to present in this subsection is not labeled. We do not know if the pairs are cognates or false friends. For the data sets that were presented in subsection 3.3.1 and subsection 3.3.2 we knew what was the label of the pairs. We knew which lists contained cognates and which lists contained false friends, they were created this way. Our proposed method uses supervised ML techniques and one of the conditions that have to be satisfied in order to use and evaluate these techniques is to have data that is labeled with the corresponding class. In this section, we will present an extension of the method and how we can determine cognates and false friend word pairs.

The data that we will present below was collected in order to be able to create

[5]http://www.ntu.edu.au/education/langs/ielex/

[6]http://www.cs.ualberta.ca/~kondrak/cognatesEF.html

complete lists of cognates and false friends for later use, in different NLP tasks. We show how we can collect and how we can create lists using our trained ML classifiers.

Bilingual Dictionary Entries To collect pairs of words that are translation of each other we used the dictionary entries from the *Internet Dictionary Project* (IDP)[7]. From the 3,246 dictionary entries we extracted 2,591 entries that were not multi-word expressions. The dictionary is not very big but it is one of the few that has its entries available for free download. We wanted to perform experiments with dictionary entries to see what percentage of the entries is recognized as cognates by our method.

To determine pairs of words that are not translations of each other, possible false friends, we paired each entry word with all the others except its translation. Using this approach we obtained a list of 5,619,270 pairs of words that are not translations of each other.

In section 3.5 we will show how many of the pairs that we obtained from the dictionary entries are cognates and how many are false friends.

Bilingual List of Words As mentioned at the beginning of the chapter one of our goals is to be able to produce complete list of cognates and false friends to be used in CALL tools. The pairs that we determine are not 100% accurate — they are produced automatically, they could be if validated by a human judge. This would require significant less effort than manually building the lists from scratch. If we look at the way we determine the cognate and false friends we see that we are very close to 100% recall, we might miss the genetic cognates, those that have a common origin and that have changed their spelling significantly.

In order to produce complete lists of cognates and false friends, we used the English entries from the LDOCE[8] dictionary, which is a dictionary intended for adult learners of English. We extracted 38,768 entries, and paired each entry with a list of 65,000

[7]http://www.june29.com/IDP/IDPfiles.html

[8]http://www.longman.com/ldoce/

lemmas of French content words (Nouns, Adjectives and Adverbs) from the *Analyse et Traitement Informatique de la Langue Française* (ATILF[9]) project.

After we paired each English word with each French word we obtained a list of pairs of words that we try to classify in cognates and false friends using an on-line French-English Dictionary[10] of approximately 75,000 terms.

3.4 Evaluation on Training and Testing Data Sets

We present evaluation experiments using the two datasets described in Section 3.3: a training/development set, a test set, and the genetic cognates set. We classify the word pairs on the basis of similarity into two classes: Cognates/False-Friends and Unrelated. Cognates are later distinguished from False-Friends by virtue of being mutual translations. We report the accuracy values for the classification task (the precision and recall values for the two classes are similar to the accuracy values). We test various feature combinations for our classification task. We test each orthographic similarity measure individually, and we also average the values returned by all the 13 measures. Then, in order to combine the measures, we run several machine learning classifiers from the Weka package.

3.4.1 Results on the Training Data Set

Individual Orthographic Measures

Table 3.3 presents the results of testing each of the 13 orthographic measures individually. For each measure, we need to choose a specific similarity threshold for separating Cognates/False-Friends from the Unrelated pairs. The separation has to be made such that all the pairs with similarity above or equal to the threshold are classified

[9]http://actarus.atilf.fr/morphalou/

[10]http://humanities.uchicago.edu/orgs/ARTFL/forms_unrest/FRENG.html

as Cognates/False-Friends, and all the pairs with similarity below the threshold are classified as Unrelated.

For the IDENT measure, the threshold was set to 1 (identical in spelling ignoring accents). This threshold leads to 49.4% accuracy, since the number of pairs with identical spelling in the training data is small (208 pairs out of 1454 that is 14.3% identical pairs, ignoring accents). We could also use the value 0 for the threshold; in this case all the pairs would be classified as Cognates/False-Friends since all scores are greater or equal to zero. This achieves 63.75% accuracy, the same as the baseline obtained by always choosing the class that is the most frequent in the training set (reported in Table 3.4).

For the rest of the measures, we determined the best thresholds by running Decision Stump classifiers with a single feature. Decision Stumps are Decision Trees that have a single node containing the feature value that produces the best split. When we run the decision stump classifier for one feature (each measure in part), we obtained the best thresholds. An example, for the XXDICE measure Decision Stump tree, is presented in Fig3.1. The values of the thresholds obtained in this way are also included in Table 3.3.

```
XXDICE <= 0.21710000000000002 : UNREL
XXDICE > 0.21710000000000002  : CG_FF
```

Figure 3.1: Example of Decision Decision Stump classifier.

Combining the Measures

The training dataset representation for machine learning experiments consists of 13 features for each pair of words: the values of the 13 orthographic similarity measures. We trained several machine learning classifiers from the Weka package: OneRule (a shallow Decision Rule that considers only the best feature and several values for it), Naïve Bayes, Decision Trees, Instance-based Learning (IBK), Ada Boost, Multi-layered Perceptron, and a light version of Support Vector Machine (SMO).

Orthographic similarity measure	Threshold	Accuracy
IDENT	1	43.90%
PREFIX	0.03845	92.70%
DICE	0.29669	89.40%
LCSR	0.45800	92.91%
NED	0.34845	93.39%
SOUNDEX	0.62500	85.28%
TRI	0.0476	88.30%
XDICE	0.21825	92.84%
XXDICE	0.12915	91.74%
BI-SIM	0.37980	94.84%
BI-DIST	0.34165	94.84%
TRI-SIM	0.34845	95.66%
TRI-DIST	0.34845	95.11%
Average Measure	0.14770	93.83%

Table 3.3: Results of each orthographic similarity measure individually, on the training dataset. The last line presents a new measure which is the average of all measures for each pair of words.

Unlike some other machine learning algorithms, Decision Tree classifier has the advantage of being relatively transparent. Figure 3.2 shows the Decision Tree obtained with the default Weka parameter settings. For each node, the numbers in round brackets show how many training examples were in each class.

Some of the nodes in the decision tree contain counter-intuitive decisions. For example, one of the leaves classifies an instance as Unrelated if the BI-SIM value is *greater* than 0.3. Since all measures attempt to assign high values to similar pairs and low values to dissimilar pairs, the presence of such a node suggests overfitting. One possible remedy to this problem is more aggressive pruning. We kept lowering the *confidence level* threshold from the default $CF = 0.25$ until we obtained a tree without counter-intuitive decisions, at $CF = 0.16$ (Figure 3.3). Our hypothesis was that the latter tree would perform better on a test set.

The results presented in the rightmost column of Table 3.4 are obtained by 10-fold cross-validation on training dataset (the data is randomly split in 10 parts, a classifier is trained on 9 parts and tested on the tenth part; the process is repeated for all 10 splits). We also report, in the middle column, the results of testing on the training set: they are artificially high, due to overfitting. The baseline algorithm in the Table 3.4 always chooses the most frequent class in the dataset, which happened to be Cognates/False-Friends. The best classification accuracy (for cross-validation) is achieved by Decision Trees, OneRule, and Ada Boost (95.66%). The performance equals the one achieved by the TRI-SIM measure alone in Table 3.3.

Error Analysis: We examined the misclassified pairs for the classifiers built on the training data. There were many shared pairs among the 60–70 pairs misclassified by several of the best classifiers. Here are some examples, from the decision tree classifier, of false negatives (Cognates/False-Friends classified as Unrelated): *égale* - *equal*, *boisson* - *beverage*, *huit* - *eight*, *cinquante* - *fifty*, *cinq* - *five*, *fourchette* - *fork*, *quarante* - *forty*, *quatre* - *four*, *plein* - *full*, *coeur* - *heart*, *droit* - *right*, *jeune* - *young*, *faire* - *do*, *oreille* - *ear*, *oeuf* - *egg*, *chaud* - *hot*. Most of the false negatives were genetic cognates that have different

Classifier	Accuracy on training set	Accuracy cross-val
Baseline	63.75%	63.75%
OneRule	95.94%	95.66%
Naive Bayes	94.91%	94.84%
Decision Trees	97.45%	95.66%
DecTree (pruned)	96.28%	95.66%
IBK	99.10%	93.81%
Ada Boost	95.66%	95.66%
Perceptron	95.73%	95.11%
SVM (SMO)	95.66%	95.46%

Table 3.4: Results of several classifiers for the task of detecting Cognates/False-Friends versus Unrelated pairs on the training data (cross-validation).

orthographic form due to changes of language over time (13 out of the 16 examples above). False positives, on the other hand, were mostly caused by accidental similarity: *arrêt* - *arm*, *habiter* - *waiter*, *peine* - *pear*. Several of the measures are particularly sensitive to the initial letter of the word, which is a strong clue of cognation. Also, the presence of an identical prefix made some pairs look similar, but they are not cognates unless the word roots are related.

3.4.2 Results on the Test Set

The rightmost column of Table 3.5 shows the results obtained on the test set described in Section 3.3. The accuracy values are given for all orthographic similarity measures and for the machine learning classifiers that use all the orthographic measures as features. The classifiers are the ones built on the training set.

The ranking of measures on the test set differs from the ranking obtained on the training set; this may be caused by the absence of genetic cognates in the test

set. Surprisingly, only the Naïve Bayes classifier outperforms the simple average of orthographic measures. The pruned Decision Tree shown in Figure 3.3 achieves higher accuracy than the overtrained Decision Tree, from Figure 3.2, but still below the simple average. Among the individual orthographic measures, XXDICE performs the best, supporting the results on French-English cognates reported in (Brew & McKelvie, 1996). Overall, the measures that performed best on the training set achieve more than 93% on the test set. We conclude that our classifiers are generic enough: they perform very well on the test set.

3.4.3 Results on the Genetic Cognates Dataset

We decided to also test the classifier trained in Section 3.3.1 on the genetic cognate set. The results on the genetic cognates set for the classifiers built on the training set (individual measures and machine learning combinations) are shown in the middle column of Table 3.5. Among the individual measures, the best accuracy is achieved by SOUNDEX, because it is designed for semi-phonetic comparison. Most of the simple orthographic measures perform poorly. One exception is PREFIX, which can be attributed to the fact that the initial segments are the most stable diachronically. TRI-SIM and TRI-DIST also did relatively well, thanks to their robust design based on approximate matching of trigrams. The IDENT measure is almost useless here because there are only two identical pairs (*long* - *long*, *six* - *six*) among the 113 pairs. Since the set contains only cognates, our baseline algorithm would achieve 100% accuracy by always choosing the Cognates/False Friends class.

Error Analysis The misclassifications are due to radical changes in spelling, such as: *frère* - *brother*, *chaud* - *hot*, *chien* - *hound*, *faire* - *do*, *fendre* - *bite*. The majority of errors are made on the genetic cognates.

Disscusion of Results The results for individual measures and their combinations on the datasets vary: the best accuracy is achieved by the instance-based learning. We note that there happened to be an overlap of 23 pairs between the training set and the

genetic cognates set. We did not remove them from the training set (because the test on the genetic cognates was not the main focus of this work), but we can say that most classifiers misclassified these pairs in the tests on the training set reported in section 3.3.1. Nonetheless, some of machine learning algorithms, especially instance-based learning may have performed better because of this.

The results on genetic cognates suggest that a different approach may be more appropriate when dealing with closely related languages (e.g., Dutch and German), which share a large number of genetic cognates. For such languages, recurrent sound and/or letter correspondences should also be considered. Methods for detecting recurrent correspondences exist (Tiedemann, 1999), (Kondrak, 2004) and could be used to improve the accuracy on genetic cognates. However, for languages that are unrelated or only remotely related, the identification of genetic cognates is of little importance. For example, in our lexicon sample of 1000 words used for testing, only 4 out of 603 French-English cognate pairs were genetic cognates.

3.4.4 Results for Three-Class Classification

Since all the examples of pairs of the class Unrelated in our training set were mutual translations, we had to add Unrelated pairs that are not translations. (Otherwise all pairs with the translation feature equal to 0 would have been classified as False Friends by the machine learning algorithms.) We generated these extra pairs automatically, by taking French and English words from the existing Unrelated pairs, and pairing them with words other than their pairs. We manually checked to ensure that all these generated pairs were not translations of each other by chance.

Table 3.6 presents a summary of the data that we used in this experiment.

As expected, this experiment achieved similar but slightly lower results than the ones from Table 3.3 when running on the same dataset (cross-validation). Most of the machine learning algorithms (except the Decision Tree) did not perfectly separate the Cognate/False-Friends class. We conclude that it is better to do the two-way

classification that we presented above (into Cognates/False-Friends and Unrelated), and then split the first class into Cognates and False-Friends on the basis on the value of the translation feature. Nevertheless, the three-way classification could still be useful provided that the translation feature is assigned a meaningful score, such as the probability that the two words occur as mutual translations in a bitext.

Results for the three-class classification are presented in Table 3.7.

3.5 Results for Building Large List of Cognates and False Friends

In this section we will present results obtained when we tried to create complete list of cognates and false friends. The experiments presented also follows the method by which we add semantics to our technique to classify cognates and false friends.

Experiments when we used pairs of words that we knew that are/are not translation of each other, were performed with data described in subsection 3.3.3. These experiments were conducted in order to create large lists of cognates and false friends.

Languages change all the time, new words and new meanings are added to or changed in a language vocabulary. It is not correct to say that once we manually create lists of cognates and false friends they will be complete. In our work, we design an automatic way that will decrease the time and human effort required to build such lists that showed to be really helpful. Our method is not 100% accurate, but the lists could be validated with some humman effort.

Results for Bilingual Dictionary Entries

From the list of dictionary entries of the IDP project, subsection 3.3.3, first paragraph, we concluded that 55% are cognates. To determine if a pair of words is a possible cognate pair, we used the threshold of one of the orthographic measures. The threshold was

the one automatically determined by our method on the training set, as presented in subsection 3.3.1.

For this experiment we used the XXDICE measure with a threshold of 0.14. The reason why we have chosen to use this measure is that it was the one that performed best on the test set that was described in 3.3.1.

The threshold automatically determined by the method was 0.12. We increased a little bit the threshold used in our experiment because we wanted to obtain pairs that are classified with a better confidence. The pair of words that are classified as cognates are not evaluated by the human judges, and since the end use for this pairs are in CALL tools, we have chosen the strategy: fewer pairs but with greater confidence in classification.

The same threshold was used for the pairs of words that were not translation of each other and were obtained from the same dictionary entry. We took each French entry word and paired it with all the English words except its translation. From the list of pairs obtained this way, only 2% were determined to be false friends.

A summary of all results for this experiment is presented in Table 3.8.

Results for Bilingual Lists of Words

For our task to determine complete lists of cognates and false friends we used the data set presented in subsection 3.3.3, second paragraph: data was obtained from the monolingual English dictionary LDOCE and a list of French content words.

From all pairs of words that were created with the algorithm described in subsection 3.3.3, we selected only the ones that have an XXDICE orthographic similarity value greater than 0.14. We are looking only at the pairs that are orthographically similar to determine cognates and false friends. The number of pairs that are selected as similar is 11,469,662.

Cognates

In order to determine cognate pairs we took each pair from the list that we created and

checked it with the dictionary referred in subsection 3.3.3, second paragraph to see if they are translation of each other. This bilingual dictionary was one of the few that allowed us to perform a huge number of queries on-line and was free to use. It is possible that we missed some cognate pairs since the dictionary that we used is not complete. The reason why we could have missed some pair of cognates is because the French word or the English word from the pair was not an entry in the respective dictionary.

The method that we propose to identify complete lists of cognates between two languages is simple and language-independent. We do not require aligned corpora (Brew & McKelvie, 1996) or any human expertise. We need list of words in each of the two languages, a threshold that can be automatically determined from a small sample of annotated data by the method described in section 3.2, and a bilingual dictionary. Using this method from the total list of 11,469,662 pairs of words we found that 3,496 pairs are cognates.

False Friends

The false-friends pairs are selected as pairs that are not translation of each other but are orthographically similar. We select only pairs for which the French word of the pair is an entry word in the dictionary and all the possible English translations do not contain the English word of the pair. Following this rule we obtained that 3,767,435 from the total number of pairs are false friends. The false friend identification can introduce some errors. The English word of a certain pair can be the translation for the French word but the dictionary does not contain it. Pairs of words that might be false friends but were not detected by our method can also be missed, if the French word or the English word from a pair is not an entry in the corresponding bilingual dictionary.

The number of false friends is higher than the number of cognates because the threshold was not very high (we wanted to have a good recall for both cognates and false friends) and because the dictionary did not contain all the posible translations for an entry. Some of the wrongly classified false friends can be eliminated by using a higher threshold as an additional filter for this class.

The pairs of words that could not be classified as cognates or as false friends in the manner described earlier might belong to one of these two classes, but since the bilingual dictionary was not complete, we could not make any decision about them.

In this section we presented experiments and results for our proposed method to determine complete lists of cognates and false friends between two languages. We have shown that a simple method that uses free tools can perform a task that otherwise would require lot of time and human effort to be created from the scratch.

Discussion of Results A problem that has to be taken into consideration for this type of experiments is the availability of a bilingual dictionary that can be queried/used. The number of entries of the dictionary is as well an issue. The more complete the dictionary, the better the recall and accuracy of the results.

A summary of all the results for this experiment are presented in Table 3.9.

3.6 Conclusion and Future Work

We presented several ways to automatically identify cognates and false friends. We tested a number of orthographic similarity measures individually, and then combined them using several different machine learning classifiers. We evaluated the methods on a training set, on a test set, and on a list of genetic cognates. We also use bilingual dictionary entries list and bilingual lists of words to automatically create complete list of cognates and false friends. The results show that for French and English it is possible to achieve very good accuracy even without the training data by employing orthographic measures of word similarity.

In future work we plan to apply our method to other pairs of languages that use a latin alphabet (since the orthographic similarity measures are not language-dependent) and increase the accuracy of the automatically generated lists of cognates and false friends.

```
TRI-SIM <= 0.3333
| TRI-SIM <= 0.2083: UNREL (447/17)
| TRI-SIM > 0.2083
| | XDICE <= 0.2
| | | PREFIX <= 0: UNREL (74/11)
| | | PREFIX > 0
| | | | SOUNDEX <= 0.5
| | | | | BI-SIM <= 0.3
| | | | | | SOUNDEX <= 0.25: CG_FF (6/2)
| | | | | | SOUNDEX > 0.25
| | | | | | | LCSR <= 0.1818: UNREL (3)
| | | | | | | LCSR > 0.1818
| | | | | | | | TRI-DIST <= 0.29: CG_FF (2)
| | | | | | | | TRI-DIST > 0.29: UNREL (2)
| | | | | BI-SIM > 0.3: UNREL (7)
| | | | SOUNDEX > 0.5: CG_FF (3)
| | XDICE > 0.2
| | | BI-SIM <= 0.3: UNREL (3)
| | | BI-SIM > 0.3: CG_FF (9)
TRI-SIM > 0.3333
| BI-SIM <= 0.4545
| | LCSR <= 0.25: UNREL (5/1)
| | LCSR > 0.25
| | | BI-DIST <= 0.4091
| | | | TRI-DIST <= 0.3333
| | | | | XXDICE <= 0.1222: CG_FF (7)
| | | | | XXDICE > 0.1222: UNREL (2)
| | | | TRI-DIST > 0.3333: CG_FF (26)
| | | BI-DIST > 0.4091
| | | | TRI-DIST <= 0.4286
| | | | | XXDICE <= 0.2273: UNREL (7/1)
| | | | | XXDICE > 0.2273: CG_FF (4/1)
| | | | TRI-DIST > 0.4286: CG_FF (11/1)
| BI-SIM > 0.4545: CG_FF (836/3)
```

Figure 3.2: Example of Decision Tree classifier (default Weka parameters, CF=25%). The two classes are Cognates/False-Friends (CG_FF) and Unrelated (UNREL). Decisions are based on values of the orthographic similarity measures. The numbers in parentheses show how many examples were classified under each leaf node.

```
TRI-SIM <= 0.3333
|   TRI-SIM <= 0.2083: UNREL (447.0/17.0)
|   TRI-SIM > 0.2083
|   |   XDICE <= 0.2: UNREL (97.0/20.0)
|   |   XDICE > 0.2
|   |   |   BI-SIM <= 0.3: UNREL (3.0)
|   |   |   BI-SIM > 0.3: CG_FF (9.0)
TRI-SIM > 0.3333: CG_FF (898.0/17.0)
```

Figure 3.3: Example of Decision Tree classifier, heavily pruned (confidence threshold for pruning CF=16%).

Classifier (measure or combination)	Accuracy on genetic cognates set	Accuracy on test set
IDENT	1.76%	55.00%
PREFIX	36.28%	90.97%
DICE	13.27%	93.37%
LCSR	24.77%	94.24%
NED	23.89%	93.57%
SOUNDEX	39.82%	84.54%
TRI	4.42%	92.13%
XDICE	15.92%	94.52%
XXDICE	13.27%	95.39%
BI-SIM	29.20%	93.95%
BI-DIST	29.20%	94.04%
TRI-SIM	35.39%	93.28%
TRI-DIST	34.51%	93.85%
Average measure	36.28%	94.14%
Baseline	—	66.98%
OneRule	35.39%	92.89%
Naive Bayes	29.20%	94.62%
Decision Trees	35.39%	92.08%
DecTree (pruned)	38.05%	93.18%
IBK	43.36%	92.80%
Ada Boost	35.39%	93.47%
Perceptron	42.47%	91.55%
SVM (SMO)	35.39%	93.76%

Table 3.5: Results of testing the classifiers built on the training set (individual measures and machine learning combinations). The middle column tests on the set of 113 genetic cognate pairs. The rightmost column tests on the test set of 1040 pairs.

Type of Pairs	No. of Pairs
Cognates Pairs	484
False Friends	326
Unrelated Translation	258
Unrelated Non translation	157

Table 3.6: Summary of the data set used with three class classification.

Classifier	Accuracy on cross-validation
Baseline	39.51%
OneRule	71.18%
Naive Bayes	92.08%
Decision Trees	96%
DecTree (pruned)	96%
IBK	95.18%
Ada Boost	96%
Perceptron	95.75%
SVM (SMO)	95.4%

Table 3.7: Results of several classifiers for the task of detecting Cognates, False-Friends and Unrelated pairs using cross-validation.

Translation Pairs	Cognates
2,591	1,438
NonTranslation Pairs	False Friends
5,619,270	133,178

Table 3.8: Number of cognates and false friends collected from IDP dictionary.

Pair of Words	Cognates	False Friends
11,469,662	3,496	3,767,435

Table 3.9: Number of cognates and false friends collected from bilingual word lists.

Chapter 4

Partial Cognate Disambiguation

Partial cognates are pairs of words in two languages that have the same meaning in some, but not all contexts. Detecting the actual meaning of a partial cognate in context can be useful for Machine Translation tools and for Computer-Assisted Language Learning tools.

In this chapter we describe and propose a supervised and a semi-supervised method of disambiguating partial cognates between two languages: French and English. The methods use only automatically-labeled data; therefore they can be applied to various pairs of languages as well. We also show that our methods perform well when using corpora from different domains.

The goal is to disambiguate a French partial cognate to help second language learners of French in a reading comprehension task. The same approach that we propose in this chapter can be followed to help second language learners of English in a reading comprehension task. The only difference is the partial cognate pairs that will be used, the methods will be similar.

Our task, disambiguating partial cognates, is in a way equivalent to coarse-grain cross-language Word-Sense Disambiguation. Our focus is disambiguating French partial cognates in context: deciding if they are used with the sense corresponding to cognate English words, or if they are used as false friends.

A French second-language learner has to be able to distinguish if the French partial cognate word is used with the same meaning as the English word (cognate word, orthographically similar and with similar meaning) or with a different meaning (a false friend). For example in the sentence *L'avocat a accepté et lui a conseillé de ne rien dire à la police.* the French partial cognate *police* has the same meaning as the English word *police*, but in the following sentence *Il s'agit ni plus ni moins d'une police d'assurance.* the same French partial cognate has a different meaning than the English word *police*. The aim of our work is to automatically detect the meaning of an French partial cognate in a specific context.

Related Research

There is a lot of work done on monolingual Word Sense Disambiguation systems that use supervised and unsupervised methods and report good results on Senseval data, but there is less work done to disambiguate cross-language words. The results of this process can be useful in many NLP tasks. As far as we know there is no work done to disambiguate partial cognates between two languages. There is work done for cross-language lexicalization. Ide (2000) has shown on a small scale that cross-lingual lexicalization can be used to define and structure sense distinctions. Tufis, Radu, & Ide (2004) used cross-lingual lexicalization, wordnet alignment for several languages, and a clustering algorithm to perform WSD on a set of polysemous English words.

Diab & Resnik (2002) used cross-language lexicalization for an English monolingual unsupervised WSD system. Besides the parallel data and MT tools they also used additional knowledge from WordNet in order to improve the results. Their task and technique are different from our task and our methods. The difference is that our technique uses the whole sentence from the parallel text, while Diab & Resnik (2002) are using only the target words (the translation of certain English words.)

Vickrey *et al.* (2005) propose a method that determines the correct translation of a word in context, a task that they consider as a different formulation of the word-

sense disambiguation task. They used the European Parliament English French parallel corpus as a training data for the logistic regression model in order to determine the correct translation in context for a set of ambiguous words. A combination of context words and part-of-speech of the context words was used as feature for the model. They were also interested in improving the MT results by using the correct translation for a word that has multiple senses and multiple translation possibilities. The paper reports an increase of accuracy over a baseline that always chooses the most common translation of a word.

Our task, disambiguating partial cognates between two languages, is a new task. We will show later on in the chapter that the method that we propose is different than all the methods used before. Our method is based on a supervised and also a semi-supervised technique that uses bootstrapping, to discriminate the senses of a partial cognate between French and English. In addition to all the methods that use bootstrapping and parallel text, presented in Chapter 2, we also bootstrap our method with corpora from different domains. As Vickrey *et al.* (2005) mention, usually parallel texts represent only a certain domain. Hansard, the French-English parallel text, is one of the largest and well-known parallel corpora, but its disadvantage is that it contains only text from the parliamentary domain. Our method uses a small set of seeds from Hansard, but additional knowledge from different domains is added using a bootstrapping technique.

In the following sections we present the way we collected the data, the methods that we used, and evaluation experiments with results for both proposed methods. A shorter version of this chapter is published as (Frunza & Inkpen, 2006).

4.1 Data for Partial Cognate Disambiguation

In this section of the chapter, we present the data that we used in our task of disambiguating partial cognates.

We performed experiments using our proposed methods with ten pairs of partial

cognates. We list them in Table 4.1. For a French partial cognate we list its English cognate word and several false friends in English. Often the French partial cognate has two senses (one for cognate, one for false friend), but sometimes it has more than two senses: one for cognate and several for false friends (nonetheless, we treat the false friend senses together). For example, the false friend words for the French partial cognate *note* include one sense for grades and one sense for bills. In our experiments, the false friends contain both senses.

The partial cognate (PC), the cognate (COG) and false-friend (FF) words were collected from a Web[1] resource. The resource contains a list of 400 false friends including 64 partial cognates. All partial cognates are words frequently used in the language. We selected ten partial cognates (presented in the first column of Table 4.1) according to the number of extracted sentences (a balance between the two meanings — cognates and false friends), to experiment and evaluate with our proposed methods.

To show how frequent the ten pairs of partial cognates are, we ran some experiments on the LeMonde[2] corpus, a collection of French newspaper news from 1994 and 1995. We counted the absolute frequency of all content words that we found in the corpus. We did not use any lemmatization in the experiment. To filter out the stop words, we used a list of 463 stop French words from the web. The total number of content words from the corpus that remain after filtering out the stop words was 216,697. From all extracted content words we took into account only the ones that have a frequency greater or equal to 100, below 100 almost all words had the frequency 1 and very few had a frequency between 1 and 100, a total of 13,656. If we compute the average frequency for the chosen content words, the value is 695,52. In Table 4.2 we show that our chosen partial cognates frequency are above the average frequency of all content words from the corpus.

With the ten pairs of partial cognates collected, the human effort that we required for our methods was to add more false-friend English words than the ones we found in the

[1]http://french.about.com/library/fauxamis/blfauxam_a.htm

[2]http://www.lemonde.fr/

French partial cognate	English cognate	English false friends
blanc	blank	white, livid
circulation	circulation	traffic
client	client	customer, patron, patient spectator, user, shopper
corps	corps	body, corpse
détail	detail	retail
mode	mode	fashion, trend, style, vogue
note	note	mark, grade, bill check, account
police	police	policy, insurance, font, face
responsable	responsible	in charge, responsible party, official representative, person in charge, executive, officer
route	route	road, roadside

Table 4.1: The ten pairs of partial cognates.

Web resource. We wanted to be able to distinguish the senses of cognate and false-friends for a wider variety of senses. This task was done using a bilingual dictionary[3]. After adding more false friend words, the final set of pairs for which we evaluate our methods is the one from Table 4.1.

4.1.1 Seed Set Collection

Both the supervised and the semi-supervised method that we will describe in the next section use a set of seeds. The seeds are parallel sentences, French and English, which contain the partial cognate. For each partial-cognate word, a part of the set contains the cognate sense and another part the false-friend sense.

The seed sentences that we use are not hand-tagged with the sense (the cognate sense

[3]http://www.wordreference.com

Partial Cognate	LeMonde_Frequency
Blanc	2,986
Circulation	1,134
Client	745
Corps	3,689
Détail	779
Mode	2,422
Note	1,979
Police	5,506
Responsable	3,409
Route	2,251

Table 4.2: The partial cognate absolute frequency in the LeMonde corpus.

or the false-friend sense); they are automatically annotated by the way we collect them. To collect the set of seed sentences we use parallel corpora from Hansard[4], and EuroParl[5] and the manually aligned BAF[6] corpus from University of Montreal.

The cognate sense sentences were created by extracting parallel sentences that had on the French side the French cognate and on the English side the English cognate. See the upper part of Table 4.3 for an example.

The same approach was used to extract sentences with the false-friend sense of the partial cognate, only this time we used the false-friend English words. See the lower the part of Table 4.3.

To keep the methods simple and language-independent, no lemmatization was used. We took only sentences that had the exact form of the French and English word as described in Table 4.1. Some improvement might be achieved when using lemmatization. We wanted to see how well we can do by using sentences as they are extracted from the

[4]http://www.isi.edu/natural-language/download/hansard/ and http://www.tsrali.com/

[5]http://people.csail.mit.edu/koehn/publications/europarl/

[6]http://rali.iro.umontreal.ca/Ressources/BAF/

Fr (PC:COG)	Je *note*, par exemple, que l'accusé a fait une autre déclaration très incriminante à Hall environ deux mois plus tard.
En (COG)	I *note*, for instance, that he made another highly incriminating statement to Hall two months later.
Fr (PC:FF)	S'il gèle les gens ne sont pas capables de régler leur *note* de chauffage.
En (FF)	If there is a hard frost, people are unable to pay their *bills*.

Table 4.3: Example sentences from parallel corpus.

parallel corpus, with no additional preprocessing and without removing any noise that might be introduced during the collection process.

From the extracted sentences, we used 2/3 of the sentences for training (seeds) and 1/3 for testing when applying both the supervised and semi-supervised approach. In Table 4.4 we present the number of seeds used for training and testing as well as the number of features selected from the training seed sets for each partial cognate.

We will show later on in the chapter that even though we started with a small set of seeds in a certain domain — the nature of the parallel corpus that we had, an improvement was obtained in discriminating the senses of partial cognates using free text from other domains.

In Appendix B we present the distribution of each false friend English word in the training set, testing set and the unlabeled data that we add in the unsupervised method that we use in our experiments.

With all the data presented in this section, and some more data, which will be presented later in the chapter we experimented with and evaluated the proposed partial cognate disambiguation methods.

Partial Cognates	Train COG	Train FF	Test COG	Test FF	French Features	English Features
Blanc	54	78	28	39	83	76
Circulation	213	75	107	38	363	328
Client	105	88	53	45	229	187
Corps	88	82	44	42	198	163
Détail	120	80	60	41	195	178
Mode	76	104	126	53	163	156
Note	250	138	126	68	377	326
Police	154	94	78	48	373	329
Responsable	200	162	100	81	484	409
Route	69	90	35	46	150	127
AVERAGE	132.9	99.1	66.9	50.1	261.5	227.9

Table 4.4: The number of parallel sentences used as seeds.

4.2 Methods

In this section we describe the supervised and the semi-supervised methods that we use in our experiments. We will also describe the data sets that were used for the monolingual and bilingual bootstrapping techniques.

For both methods, we have the same goal: to determine which of the two senses (the cognate or the false-friend sense) of a partial-cognate word is present in a test sentence. The classes in which we classify a sentence that contains a partial cognate are: COG (cognate) and FF (false friend). Our goal is to determine the sense of a partial cognate in a French sentence, to determine if the partial cognate is used with a cognate sense with the corresponding English word or with a false friend sense. Both the cognate and false friend English words are translated as the same French word.

4.2.1 Supervised Method

For both the supervised and semi-supervised method we used the bag-of-words (BOW) approach of modeling context with binary values for the features. The features were words from the training corpus that appeared at least 3 times in the training sentences. We removed the stop words from the features. A list of stop words for French was used on the French sentences and one for English was used on the English parallel sentences. We ran some additional experiments when we kept the stop words as features but the results did not improve. In Table 4.4, the last two columns present the number of features extracted from the training seed sets.

Since we wanted to learn the contexts in which a partial cognate has a cognate sense and the contexts in which it has a false-friend sense, the cognate and false friend words themselves were not taken into account as features. Leaving them in would mean to indicate the classes when applying the methods for the English sentences since all the sentences with the cognate sense contain the cognate word and all the false-friend sentences do not contain it. For the French side all collected sentences contain the partial cognate word, the same for both senses. For the French features the information gain that the partial cognate has is 0, since it is present in sentences from both classes, COG and FF.

As a baseline for the experiments that we present we used the ZeroR classifier from WEKA; it predicts the class that is the most frequent in the training corpus. The classifiers for which we report results are: Naïve Bayes with a kernel estimator, Decision Trees — J48, and a Support Vector Machine implementation — SMO. All the classifiers can be found in the WEKA package. We used these classifiers because we wanted to have a probabilistic, a decision-based and a functional classifier. The decision tree classifier allows us to see which features are most discriminative.

Experiments were also performed with other classifiers and with different levels of tuning, on a 10-fold cross validation approach as well; see Table 4.6 results for the French training seed data and Table 4.7 results for the English training seed data; the

classifiers we mentioned above were consistently the ones that gave the best accuracy results.

The supervised method used in our experiments consists in training the chosen classifiers on the automatically-collected training seed sentences, separately for French and for English, for each partial cognate, and then testing their performance on the test set. Results for this method are presented later, in section 4.3.

4.2.2 Semi-Supervised Methods

Besides the supervised method that we propose to disambiguate a partial cognate, we look at a semi-supervised method as well. For the semi-supervised method we add unlabeled examples, sentences that contain the partial cognate with one of the two senses: cognate or false friend, from monolingual corpora: the French newspaper LeMonde 1994, 1995 (LM), and the BNC[7] corpus — different domain corpora than the seeds. The procedure of adding and using this unlabeled data is described in the Monolingual Bootstrapping (MB) and Bilingual Bootstrapping (BB) algorithms.

Monolingual Bootstrapping

The monolingual bootstrapping algorithm that we used for experiments on French sentences (MB-F) or on English sentences (MB-E) is:

For *each pair of partial cognates (PC)*

1. Train a classifier on the training seeds — using the BOW approach and a NB-K classifier with attribute selection on the features.

2. Apply the monolingual classifier on unlabeled data — sentences that contain the PC word, extracted from LeMonde (MB-F) or from BNC (MB-E)

3. Take the first k newly classified sentences, both from the COG and FF class and add them to the training seeds (the most confident ones — the prediction accuracy greater or equal than a threshold =0.85)

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

4. Rerun the experiments training on the new training set

5. Repeat steps 2 and 3 for **t** *times*

endFor

For the first step of the algorithm we used the NB-K classifier because it was the classifier that consistently performed better. We chose to perform attribute selection on the features after we tried the method without attribute selection. We obtained better results when using attribute selection. This sub-step was performed with the WEKA tool, the Chi-Square attribute selection was chosen. The attribute selection was performed on the feature only when we trained the classifiers to be used to label the unlabeled data from the additional resources.

In the second step of the MB algorithm the classifier that was trained on the training seeds was then used to classify the unlabeled data that was collected from the two additional resources, separately. For the MB algorithm on the French side we trained the classifier on the French side of the training seeds and then we applied the classifier to classify the sentences that were extracted from LeMonde and contained the partial cognate, as belonging to the COG class or the FF class. The same approach was used for the MB on the English side only this time we were using the English side of the training seeds to train the classifier and the BNC corpus to extract new examples. In fact, the MB-E step is needed only for the BB method.

Only the sentences that were classified with a probability greater than 0.85 were selected for later use in the bootstrapping algorithm. This value of the parameter is a heuristic value for our experiments. All results that will be described in Section 4.3 use the threshold for the probability distribution 0.85.

The number of sentences that were selected from the new corpora and used in the MB and BB algorithms is presented in Table 4.5.

For the partial-cognate *blanc* with the cognate sense, the number of sentences that had a probability distribution greater or equal with the threshold was low. For the rest of partial cognates the number of selected sentences was limited by the value of the

PC	LM COG	LM FF	BNC COG	BNC FF
Blanc	45	250	0	241
Circulation	250	250	70	180
Client	250	250	77	250
Corps	250	250	131	188
Détail	250	163	158	136
Mode	151	250	176	262
Note	250	250	178	281
Police	250	250	186	200
Responsable	250	250	177	225
Route	250	250	217	118

Table 4.5: Number of sentences selected from the LeMonde and BNC corpus.

parameter **k** that was 250, in the algorithm.

Bilingual Bootstrapping

The algorithm for bilingual bootstrapping that we have proposed and tried in our experiments is:

1. *Translate the English sentences that were collected in the MB-E step into French using an online MT tool*[8] *and add them to the French seed training data.*
2. *Repeat the MB-F and MB-E steps* **T** *times.*

For both monolingual and bilingual bootstrapping techniques the value of the parameters **t** and **T** is 1 in our experiments.

In the bilingual bootstrapping algorithm we take the English sentences that were extracted from the BNC corpus, as described in the MB-E algorithm, translate them into English using an on-line MT tool and then we add them to the French training corpus.

[8]http://www.freetranslation.com/free/web.asp

Our proposed methods are bilingual, they can be applied to any pair of languages for which a parallel corpus is available, and two monolingual collections of text. Our main focus for our methods is to be able to disambiguate a French partial cognate looking at its English cognate and false friend senses.

4.3 Evaluation and Results

In this section we present the results that we obtained with the supervised and semi-supervised methods that we applied to disambiguate partial cognates.

In Table 4.6 and Table 4.7 we present the results that we obtained with the supervised method when using the 10-fold cross validation technique on the training data for French and respectively for the English data. The results obtained are reported for several classifiers and with different levels of tuning. For the Decision Tree classifier we tuned the Confidence Factor (CF) value. A lower value of the CF will trigger more pruning, the algorithm will discharge nodes that contain a small number of instances that cause the overfitting problem.

The *Exp* value for the SMO algorithm describes the exponent for the polynomial kernel of the functional classifier. The K parameter for Instance Based Learning classifier describes the number of neighbors to be used.

The best accuracy results are obtained using the Naïve Bayes classifier with a Kernel distribution (NB-K). The highest values of accuracy are 89% for the French training data set and 87% for the English training data set. We experimented all these classifiers because we wanted to see which one performs best. As expected, the Naïve Bayes classifier was the one that outperformed the others.

4.3.1 Evaluation Results for the Supervised Method

In this subsection we present the results that we obtained with our supervised method to disambiguate partial cognates. As we mentioned in Section 4.2 subsection 4.2.1 we

PC	ZeroR	NB	NB-K	Dec. Tree CF=0.25	Dec.Tree CF=0.16	SMO Exp=2	SMO Exp=1	LBK K=1	IBK K=5	AdaBoost
Blanc	58.00%	96.96%	94.69%	95.45%	95.45%	94.69%	93.18%	93.18%	92.42%	95.45%
Circulation	74.00%	90.62%	92.36%	79.51%	79.86%	92.01%	90.27%	73.95%	80.20%	73.95%
Client	54.08%	72.53%	82.38%	68.39%	68.39%	79.79%	79.27%	75.12%	67.35%	61.13%
Corps	51.16%	94.11%	96.47%	86.47%	86.47%	90%	78.82%	73.52%	53.52%	81.17%
Détail	59.40%	91.50%	93.00%	90.00%	90.00%	93.50%	85%	81.00%	70.00%	89.00%
Mode	58.24%	81.11%	81.11%	74.44%	74.44%	83.00%	77.22%	72.77%	73.33%	70.00%
Note	64.94%	90.93%	92.00%	84.45%	83.16%	89.37%	89.11%	81.60%	79.27%	82.12%
Police	61.41%	89.87%	91.49%	87.00%	87.00%	89.47%	91.09%	89.87%	73.68%	89%
Responsable	55.24%	85.91%	89.22%	76.51%	76.51%	84.80%	81.76%	61.60%	67.67%	75.69%
Route	56.79%	71.06%	75.47%	56.60%	56.60%	72.32%	69.81%	68.55%	57.23%	61.63%
AVERAGE	59.00%	86.00%	89.00%	80.00%	80.00%	87.00%	84.00%	77.00%	71.00%	78.00%

Table 4.6: Results on the French training seeds using 10-fold cross validation.

report results only for three classifiers, selected based on the results obtained by 10-fold cross-validation evaluation on the training data and based on the chosen classifiers: probabilistic, decisional and functional.

Table 4.8 presents the results obtained on the French data using our supervised technique. We used for training the 2/3 of the seed sets and for testing the other 1/3 of the seeds.

In Table 4.9 we present the results using the same supervised method only for the English set of seeds. The results obtained are close to the ones obtained on the French seed sets.

4.3.2 Results for Semi-Supervised Methods

We want to disambiguate partial cognates not only in a parliamentary domain, the domain of our collected seeds but in different domains. To vary the domain of the training data, and improve the classification results, we proposed to algorithms MB and

PC	ZeroR	NB	NB-K	Dec. Tree CF=0.25	Dec.Tree CF=0.16	SMO Exp=2	SMO Exp=1	IBK K=1	LBK K=5	AdaBoost
Blanc	58.00%	95.45%	93.18%	96.21%	96.21%	94.69%	93.93%	93.93%	92.42%	96.21%
Circulation	74.00%	86.11%	90.27%	75.00%	73.95%	85.41%	86.45%	82.63%	80.55%	73.95%
Client	54.08%	76.68%	79.27%	66.83%	63.73%	79.79%	79.27%	76.16%	53.88%	64.76%
Corps	51.16%	94.11%	92.94%	88.82%	89.41%	92.94%	85.88%	81.76%	62.35%	86.47%
Détail	59.40%	90.50%	94.00%	81.00%	81.00%	86.50%	85.00%	80.00%	75.00%	77.00%
Mode	58.24%	75.00%	80.00%	66.67%	66.67%	76.00%	70.55%	69.44%	70.55%	67.77%
Note	64.94%	90.69%	89.00%	79.84%	79.06%	89.92%	87.33%	82.42%	73.38%	77.26%
Police	61.41%	93.14%	93.54%	76.00%	76.00%	90.72%	77.01%	60.88%	48.38%	74.00%
Responsable	55.24%	83.7%	87.01%	76.24%	76.79%	85.63%	78.45%	58.83%	46.68%	75.13%
Route	56.79%	70.25%	75.94%	56.32%	56.32%	73.41%	70.88%	64.55%	66.45%	56.96%
AVERAGE	59.00%	86.00%	87.00%	76.00%	76.00%	85.00%	81.00%	75.00%	67.00%	75.00%

Table 4.7: Results on the English training seeds using 10-fold cross validation.

BB presented in Section 4.2.

Results that we obtained with these two algorithms are presented in Table 4.11 for the French MB (MB-F), Table 4.13 for the English MB (MB-E), and Table 4.15 for BB. For the MB experiments the training examples (training seeds) both for the French side of the parallel corpus and the English one, are complemented with sentences extracted from LeMonde corpus for the French experiments, and sentences extracted from BNC corpus for the English experiments. The training data and the number of features extracted after we added the new training data for the French MB experiments are presented in Table 4.10. Similar information is presented in Table 4.12 for the English MB experiments.

The next table, 4.14 presents the data and results obtained with the BB algorithm on the French side. To the French training seeds we added the translated sentences extracted from the BNC corpus and trained the classifier on them. The classifier performance is tested on the seed testing set, 1/3 of the collected seed sets. Results of this experiment are presented in Table 4.15.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.00%	95.52%	98.5%	98.5%
Circulation	74.00%	91.03%	80.00%	89.65%
Client	54.08%	67.34%	66.32%	61.22%
Corps	51.16%	62.00%	61.62%	69.76%
Détail	59.40%	85.14%	85.14%	87.12%
Mode	58.24%	89.01%	89.01%	90.00%
Note	64.94%	89.17%	77.83%	85.05%
Police	61.41%	79.52%	93.70%	94.40%
Responsable	55.24%	85.08%	70.71%	75.69%
Route	56.79%	54.32%	56.79%	56.79%
AVERAGE	59.33%	**80.17%**	77.96%	80.59%

Table 4.8: Results for the Supervised Method on the French test set data.

We also combined MB and BB and evaluated the classifiers for this combination. We trained the classifiers on the training seed sentences plus the sentences from LeMonde plus the sentences from BNC. The result of this experiment is presented in Table 4.16.

For all the results reported until now we tested the classifiers on the test set of the automatically collected seeds. These results are discussed in Section 4.4.

Results for additional experiments with different data sets.

Besides the experiments that we did with the semi-supervised method using unlabeled corpus from LeMonde and BNC, we run additional experiments with another set of automatically collected sentences from a multi-domain parallel corpus.

The set of new sentences (multi-domain) was extracted in the same manner as the seeds from Hansard and EuroParl. The new parallel corpus is a small one, approximately 1.5 million words, but contains texts from different domains: magazine articles, modern fiction, texts from international organizations and academic textbooks. The corpus was

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.00%	97.01%	97.01%	97.01%
Circulation	74.00%	95.86%	75.17%	84.13%
Client	54.08%	77.55%	55.10%	70.40%
Corps	51.16%	77.90%	58.13%	75.58%
Détail	59.4%	80.19%	70.29%	80.19%
Mode	58.24%	85.71%	90.10%	80.00%
Note	64.94%	88.65%	77.31%	78.86%
Police	61.41%	69.04%	66.04%	66.67%
Responsable	55.24%	86.18%	75.13%	82.87%
Route	56.79%	59.25%	55.55%	59.25%
AVERAGE	59.33%	**81.73%**	72.28%	77.52%

Table 4.9: Results for the Supervised Method on the English test set data.

provided to us by Prof. Raphael Salkie, Brighton University, UK. We use this set of sentences in our experiments to show that our methods perform well on multi-domain corpora, and also because our aim is to be able to disambiguate PC in different domains. From this parallel corpus we were able to extract the number of sentences shown in Table 4.17.

With this new set of sentences we performed different experiments both for the MB and BB the algorithms. All the results are described in Table 4.18. The results are reported for the average of the accuracies for the ten pairs of partial cognates.

The symbols that we use in Table 4.18 represent: S — the seed training corpus, TS — the seed test set, BNC and LM — sentences extracted from LeMonde and BNC (Table 4.5), and NC — the sentences that were extracted from the multi-domain new corpus. When we use the + symbol we put together all the sentences extracted from the respective corpora.

Figure 4.1 presents in a graphical way the results obtained with the four methods,

PC	Train COG	Train FF	No. Features
Blanc	99	328	369
Circulation	463	325	1052
Client	355	338	817
Corps	338	332	876
Détail	370	243	669
Mode	227	354	667
Note	500	388	981
Police	404	344	1018
Responsable	450	412	1109
Route	319	340	725
AVERAGE	352.5	370.7	828.3

Table 4.10: Amount of data and number of features for Monolingual Bootstrapping on the French side.

No Bootstrapping, Monolingual French Bootstrapping, Bilingual Bootstrapping and Monolingual plus Bilingual Bootstrapping on different sets of French sentences for the average over all 10 pairs of partial cognates. The sets of French sentences set that the method uses are shown on the X axis of the chart. The set used initially for training, no bootstrapping, is presented before the underscore line, and the set used for testing is presented after the underscore line.

All the results for each pair of partial cognate that are averaged in Table 4.18, for the different training and testing sets, are presented in Appendix C.

Error Analysis Most of the errors that the classifiers made were on the hard to disambiguate words, but still improvement was obtained even for these partial cognates. The errors that appear can also be caused by the noise that was introduced in the seed set collection process and in the bootstrapping process.

For example for the partial cognate *circulation*, on the seed testing set, from a total

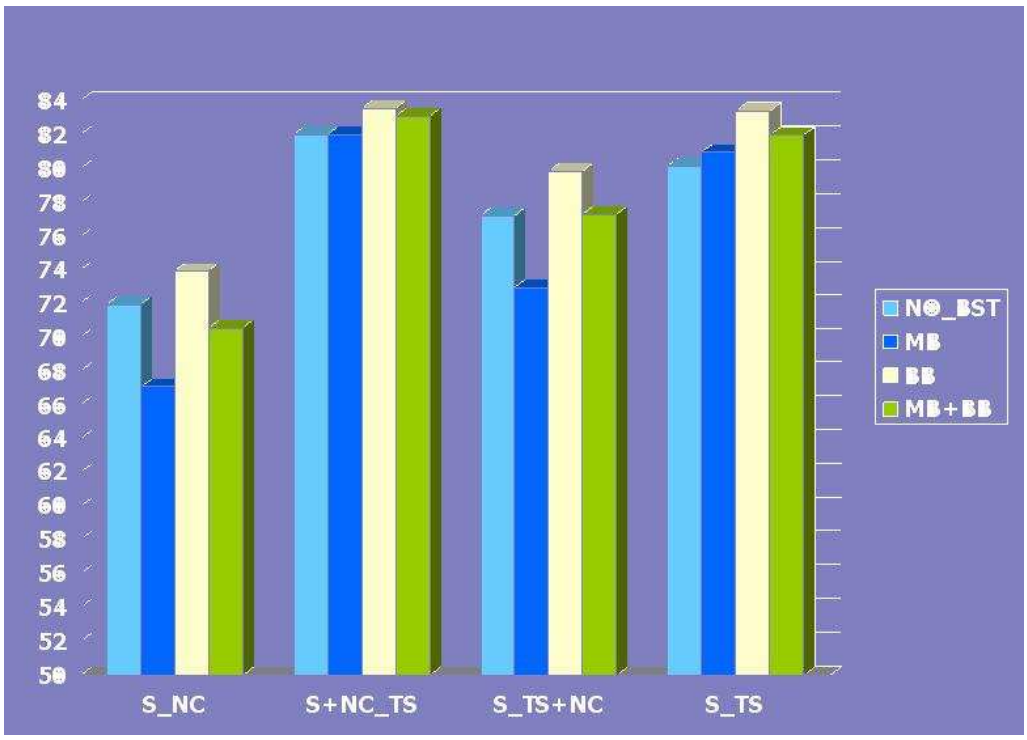

Figure 4.1: Results for the average of the PC set with different methods and data sets.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	97.01%	97.01%	98.5%
Circulation	73.79%	90.34%	70.34%	84.13%
Client	54.08%	77.55%	55.10%	70.40%
Corps	51.16%	78%	56.97%	69.76%
Détail	59.4%	88.11%	85.14%	82.17%
Mode	58.24%	89.01%	90.10%	85.00%
Note	64.94%	85.05%	71.64%	80.41%
Police	61.41%	71.65%	92.91%	71.65%
Responsable	55.24%	87.29%	77.34%	81.76%
Route	56.79%	51.85%	56.79%	56.79%
AVERAGE	59.33%	**80.96%**	75.23%	77.41%

Table 4.11: Monolingual Bootstrapping results (accuracies) on the French side.

of 145 testing instances, the Naïve Bayes classifier, trained on a number of 288 instances, made 13 mistakes. 6 mistakes were on the Cognate class, the actual class was the Cognate class but the classifier predicted the False Friend class, and the rest of 7 mistakes were made on the False Friend class.

The first two sentences are errors made on the Cognate class and the last two are errors made by the classifier on the False Friend class.

Examples of errors on the Cognate Class

Nous ne pouvons, donc, que partager les justes manifestations de protestation des oléiculteurs du Sud de l'Italie, dont le revenu s'est réduit à cause, aussi, de l'absence de tout contrôle sur la provenance et la qualité de l'huile en circulation. (We can only agree with the fair demonstrations of civil protest made by the olive growers of southern Italy, who have seen their income fall, as a result of a lack of control over the origin and quality of the oil in circulation.)

Ces deux conventions, comme vous le savez, Mesdames et Messieurs, portent sur le

PC	Train COG	Train FF	No. Features
Blanc	54	319	279
Circulation	283	254	637
Client	182	337	550
Corps	218	269	503
Détail	278	215	473
Mode	252	365	626
Note	428	419	788
Police	340	293	812
Responsable	377	386	943
Route	286	207	476
AVERAGE	269.8	306.4	608.7

Table 4.12: Amount of data and number of features for Monolingual Bootstrapping on the English side.

renforcement de la sécurité du transport desdits équipements et sur la garantie de leur libre circulation au sein du marché commun communautaire. (Such agreements, as you know, ladies and gentlemen , are intended to strengthen safety in the transport of such equipment and ensure free circulation of the same in the common Community market.)

Examples of errors on the False Friend Class

Des représentants des services policiers ont dit à des comités de la Chambre et, apparemment, du Sénat, que nous ne contrôlons pas la circulation des conteneurs, et c'est un fait. (It is a fact that police have told us at committees of the House and apparently of the Senate that we do not control container traffic.)

Nous nous sommes entendus sur des façons d'accélérer la circulation des biens et des personnes entre nos deux pays. (We agreed on ways to speed up cross-border traffic.)

The errors occur because the contexts of the two classes for some partial cognates are

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	94.02%	95.52%	97.01%
Circulation	73.79%	92.41%	62.06%	82.75%
Client	45.91%	70.40%	45.91%	62.24%
Corps	48.83%	91.86%	50.00%	86.04%
Détail	59.40%	86.13%	63.36%	83.16%
Mode	58.24%	81.31%	58.24%	88.00%
Note	64.94%	85.05%	51.54%	80.41%
Police	61.90%	70.63%	69.04%	72.22%
Responsable	44.75%	85.63%	80.66%	80.11%
Route	43.20%	65.43%	58.02%	67.90%
AVERAGE	55.92%	**82.29%**	63.44%	79.98%

Table 4.13: Monolingual Bootstrapping results (accuracies) on the English side.

close. This was the case with the missclassified sentences presented above. The tested sentences had the majority of content words belonging to the training data of the class they do not belong to.

4.4 Discussion of the Results

The results of the experiments and the methods that we propose show that we can use with success unlabeled data to learn from, and that the noise introduced due to the seed set collection is tolerable by the ML techniques that we use. The noise that can be introduced is due to the fact that we do not use a word-aligned corpus. For example a French sentence can contain the partial cognate and the English parallel sentence can contain the cognate English word, but the meaning of the English sentence could be the false friend one, and the cognate word could appear in another part of the sentence.

Some results of the experiments we present in Table 4.18 are not as good as others.

PC	Train COG	Train FF	No. Features
Blanc	54	319	331
Circulation	283	255	686
Client	182	337	636
Corps	219	269	582
Détail	278	215	548
Mode	252	365	714
Note	428	419	922
Police	340	293	915
Responsable	377	386	1028
Route	286	207	533
AVERAGE	269.8	306.4	689.5

Table 4.14: Data sets for Bilingual Bootstrapping on the French side.

What is important to notice is that every time we used MB or BB or both, there was an improvement. For some experiments MB did better, for others BB was the method that improved the performance; nonetheless for some combinations MB together with BB was the method that worked best.

The supervised method results, Table 4.8 and Table 4.9 were outperformed by the semi-supervised methods both on the French set, Table 4.11 and respectively English test set Table 4.13, showing that unlabeled data can be used with succes to boost the resuts of the classification task.

In Tables 4.8 and 4.15 we show that BB improved the results on the NB-K classifier with 3.24%, compared with the supervised method (no bootstrapping), when we tested only on the test set (TS), the one that represents 1/3 of the initially-collected parallel sentences. This improvement is not statistically significant, according to a t-test.

In Table 4.18 we show that our proposed methods bring improvements for different combinations of training and testing sets. Table 4.18, lines 1 and 2 show that BB with

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	95.52%	97.01%	98.50%
Circulation	73.79%	92.41%	63.44%	87.58%
Client	45.91%	70.40%	45.91%	63.26%
Corps	48.83%	83.00%	67.44%	82.55%
Détail	59.00%	91.08%	85.14%	86.13%
Mode	58.24%	87.91%	90.10%	87.00%
Note	64.94%	85.56%	77.31%	79.38%
Police	61.41%	80.31%	96.06%	96.06%
Responsable	44.75%	87.84%	74.03%	79.55%
Route	43.20%	60.49%	45.67%	64.19%
AVERAGE	55.87%	**83.41%**	74.21%	82.40%

Table 4.15: Accuracies results for Bilingual Bootstrapping on the French side.

NB-K brought an improvement of 1.95% from no bootstrapping, when we tested on the multi-domain corpus NC. For the same setting, there was an improvement of 1.55% when we tested on TS (Table 4.18, lines 6 and 8). When we tested on the combination TS+NC, again BB brought an improvement of 2.63% from no bootstrapping (Table 4.18, lines 10 and 12). The difference between MB and BB with this setting is 6.86% (Table 4.18, lines 11 and 12). According to a t-test the 1.95% and 6.86% improvements are statistically significant.

The results presented in these tables are only performed with French data, since our goal was to be able to disambiguate a French partial cognate in a certain context. The improved results obtained with the semi-supervied method on the English data suggest that similar experiments to those presented in Table 4.18 can improve results for the English side.

Unlike previous work with monolingual or bilingual bootstrapping Diab & Resnik (2002), Li & Li (2004), we tried to disambiguate not only words that have senses that are very

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	98.50%	97.01%	98.50%
Circulation	73.79%	88.96%	69.65%	86.20%
Client	45.91%	71.42%	54.08%	70.40%
Corps	48.83%	80.00%	60.46%	76.74%
Détail	59.40%	90.09%	85.14%	87.12%
Mode	58.24%	86.81%	90.10%	82.00%
Note	64.94%	85.05%	70.61%	78.35%
Police	61.41%	74.80%	95.27%	70.86%
Responsable	44.75%	88.39%	75.69%	79.00%
Route	43.20%	55.55%	45.67%	58.02%
AVERAGE	55.87%	**81.98%**	74.37%	78.76%

Table 4.16: Accuracies for Monolingual Bootstrapping plus Bilingual Bootstrapping on the French side.

different, e.g., *plant* — with a sense of biological plant or with the sense of factory. In our set of partial cognates the French word *route* is a difficult word to disambiguate even for humans: it has a cognate sense when it refers to a maritime or trade route and a false-friend sense when it is used as road. The same observation applies to *client* (the cognate sense is client, and the false friend sense is customer, patron, or patient) and to *circulation* (cognate in air, blood, etc. circulation and false friend for street traffic).

We also showed that our method is able to bootstrap the initial knowledge of the chosen classifiers, parliamentary domain knowledge, with information from different domains that was obtained in the monolingual and bilingual bootstrapping steps. Appendix D presents an example of a Decision Tree for a partial cognate extracted from the training seeds, then one Decision Tree obtained from the MB experiments, and then one from the BB experiments. The value of the feature shows how the knowledge of the classifier changes and also increases, from a specific domain to other more varied. The

PC	COG	FF
Blanc	18	222
Circulation	26	10
Client	70	44
Corps	4	288
Détail	50	0
Mode	166	12
Note	214	20
Police	216	6
Responsable	104	66
Route	6	100

Table 4.17: Number of sentences collected from the New Corpus (NC).

number of features that was extracted at each semi-supervised step more than doubled compared with the number that was initially extracted from the seeds.

4.5 Conclusions and Future Work

In this chapter we have shown that the task of partial cognate word disambiguation can be done with success using a supervised and more likely with a semi-supervised method that uses a bootstrapping technique. We proposed two semi-supervised algorithms that use unlabeled data from different languages, French and English, which can improve the accuracy results of a simple supervised method. We have also shown that classifiers, computer algorithms, are able to capture knowledge in an incremental process like humans, and that are sensitive to knowledge from different domains. The unlabeled data that was extracted and added to the training data for the different algorithms was collected from different domains than the initial training seed data.

Simple methods and available tools have proved to be resources of great value to

Train	Test	ZeroR	NB-K	Dec. Trees	SMO
S (no bootstrapping)	NC	67%	**71.97%**	73.75%	76.75%
S+BNC (BB)	NC	64%	**73.92%**	60.49%	74.8%
S+LM (MB)	NC	67.85%	67.03%	64.65%	65.57%
S+LM+BNC (MB+BB)	NC	64.19%	70.57%	57.03%	66.84%
S+NC (no bootstrapping)	TS	57.44%	**82.03%**	76.91%	80.71%
S+NC+LM (MB)	TS	57.44%	82.02%	73.78%	77.03%
S+NC+BNC (BB)	TS	56.63%	**83.58%**	68.36%	82.34%
S+NC+LM+BNC (MB+BB)	TS	58%	83.10%	75.61%	79.05%
S (no bootstrapping)	TS+NC	62.70%	**77.20%**	77.23%	79.26%
S+LM (MB)	TS+NC	62.7%	**72.97%**	70.33%	71.97%
S+BNC (BB)	TS+NC	61.27%	**79.83%**	67.06%	78.8%
S+LM+BNC (MB+BB)	TS+NC	61.27%	77.28%	65.75%	73.87%

Table 4.18: Results for different experiments with Monolingual and Bilingual Bootstrapping (MB and BB), when the New Corpus (NC) is used either in training or in testing.

achieve good results in the task of partial cognate disambiguation.

The accuracy results might be increased by using dependency relations, lemmatization, part-of-speech tagging — extract only sentences where the partial cognate has the same POS, and other types of data representation combined with different semantic tools (e.g. decision lists, rule based systems). In our experiments we use a machine language representation — binary feature values, and we show that nonetheless machines are capable of learning from new information. New information was collected and extracted by classifiers when additional corpora were used for training.

In future work we plan to try different representations of the data, to use knowledge of the relations that exists between the partial cognate and the context words, and to run experiments when we iterate the MB and BB steps more than once.

Chapter 5

A Tool for Cross-Language Pair Annotations

In this chapter, we will describe our tool called Cross-Language Pair Annotator (CLPA) that is capable of automatically annotating cognates and false friends in a French text. The tool uses the Unstructured Information Management Architecture (UIMA)[1] Software Development Kit (SDK) from IBM and Baseline Information Extraction (BaLIE)[2], an open source Java project capable of extracting information from raw texts.

5.1 Tool Description

CLPA is a tool that has a Graphical User Interface (GUI) capability that makes it easy for the user to distinguish between different annotations of the text. We designed the tool as a Java open source downloadable kit that contains all the additional projects (Balie and UIMA) that are needed. It can be downloaded from the following address: CLPA[3].

The tool is a practical follow up to the research that we did on cognates and false

[1]http://www.research.ibm.com/UIMA/

[2]http://balie.sourceforge.net/

[3]www.site.uottawa.ca/ ofrunza/CLPA.html

friends between French and English. Since one of our main goals is to be able to use the research that we did in a CALL tool can help second-language learners of French, CLPA is intended to be the first version of such a tool. At this point, the tool uses as knowledge a list of 1,766 cognates and a list of 428 false friends. The list of false friends contains a French definition for the French word and an English definition for the English word of the pair. Both lists contain the cognates and false friend pairs that were used in the Machine Learning experiments for the cognate and false friend identification task described in Chapter 3.

The tool offers an easy management of the resources. If the user would like to adjust/use other lists of cognates and/or false friends he/she needs to add the new source files to the resource directory of the project. The directory can be found in the home project directory.

UIMA is an open platform for creating, integrating, and deploying unstructured information management solutions from a combination of semantic analysis and search components. It also has various GUI document analyzers that make it easy for the user to visualize the text annotations.

UIMA offers CLPA the GUI interface and an efficient management of the annotations that are done for a certain text. The user can select/deselect the cognate or false friend annotations. By default, both type of cross language pairs are annotated.

BaLIE is a trainable Java open source project that can perform: Language Identification, Sentence Boundary Detection, Tokenization, Part of Speech Tagging and Name Entity Recognition for English, French, German, Spanish and Romanian.

BaLIE is the project that provided the tokenization and part-of-speech tagging tasks for the French texts. The tokenization is done using a rule based method and the part-of-speech by using a probabilistic part-of-speech tagger, QTag[4].

In a single run, the tool can annotate not only one document, but a directory that contains more than a single text document. UTF-8 is the character encoding chosen

[4]http://www.english.bham.ac.uk/staff/omason/software/qtag.html

to represent a document. The reason why we chose this format is due to the French characters and also for a consistency with the other projects that are used by CLPA the BaLIE project.

The following figures provide snapshots of the interface that the user will see after the annotation process is completed.

The user has to click on one of the text annotations to obtain additional information about the chosen annotation, (e.g. at what position in the text does the chosen word starts, what position does it end, the French definition of the French false friend word, the English definition of the English false friend word, etc.).

5.2 Tool Capabilities

In its early stage of existence, first version of CLPA, can annotate cognates and false friends between French and English in a French text. The cognate and false friend knowledge that the tool has is provided by lists of pairs of cognates and false friends. For now we intended high accurate lists instead of automatically produced lists. Instead of the lists that we use in this first version, we can use the lists that we automatically produced and described in Chapter 3.

In addition to the colored annotations (cognates are annotated with one color and false friends with another color), the tool provides other useful information about the annotations. For the cognate words, it provides the position in the text and the English cognate word. For the false friend words it provides: the position in the text, the English false friend word/words, the definition of the French word, and the definition of the English words. The definitions were collected from the same resource[5] as the false friend word pairs.

The lists that the CLPA uses to annotate the French texts are free to download and can be used for future research. They are contained in the same package as the tool.

[5]http://french.about.com/library/fauxamis/blfauxam_a.htm

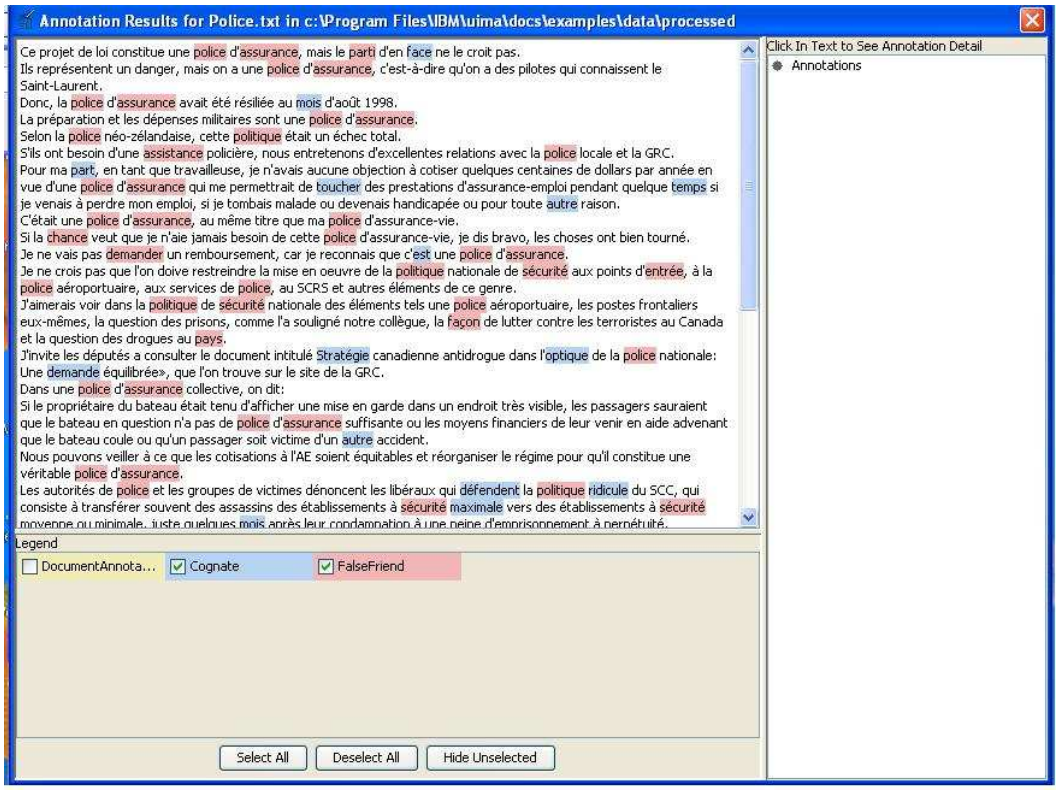

Figure 5.1: Cognate and False Friend annotations.

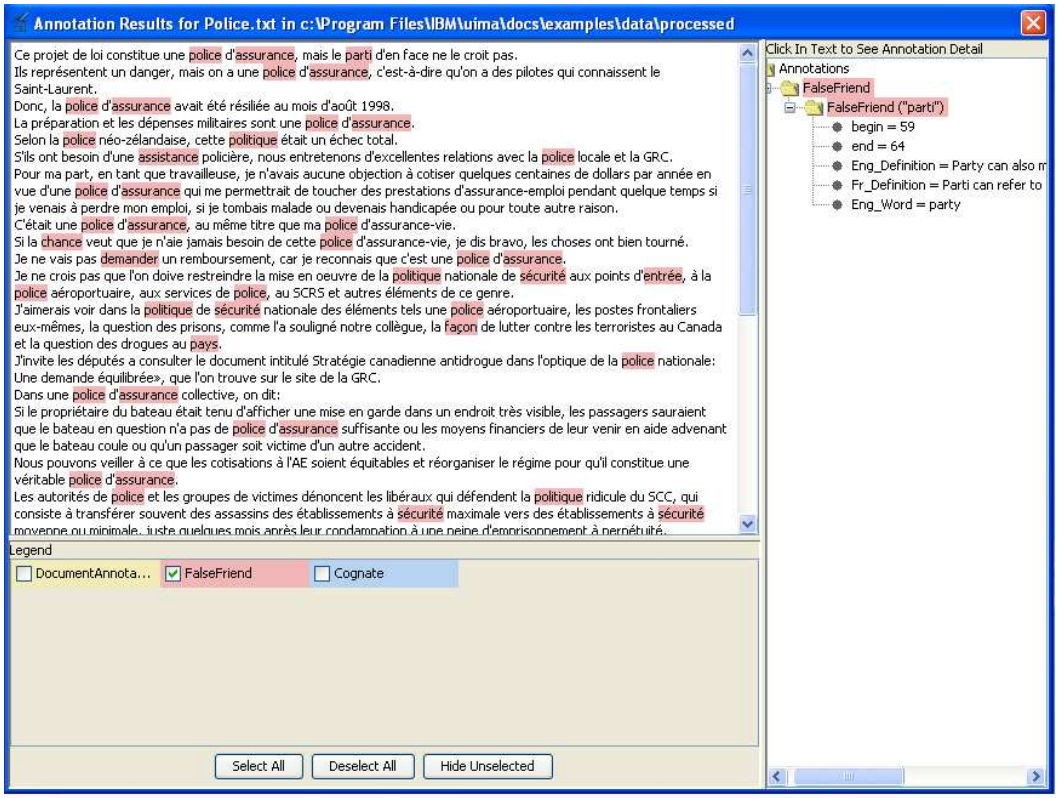

Figure 5.2: False Friend annotations.

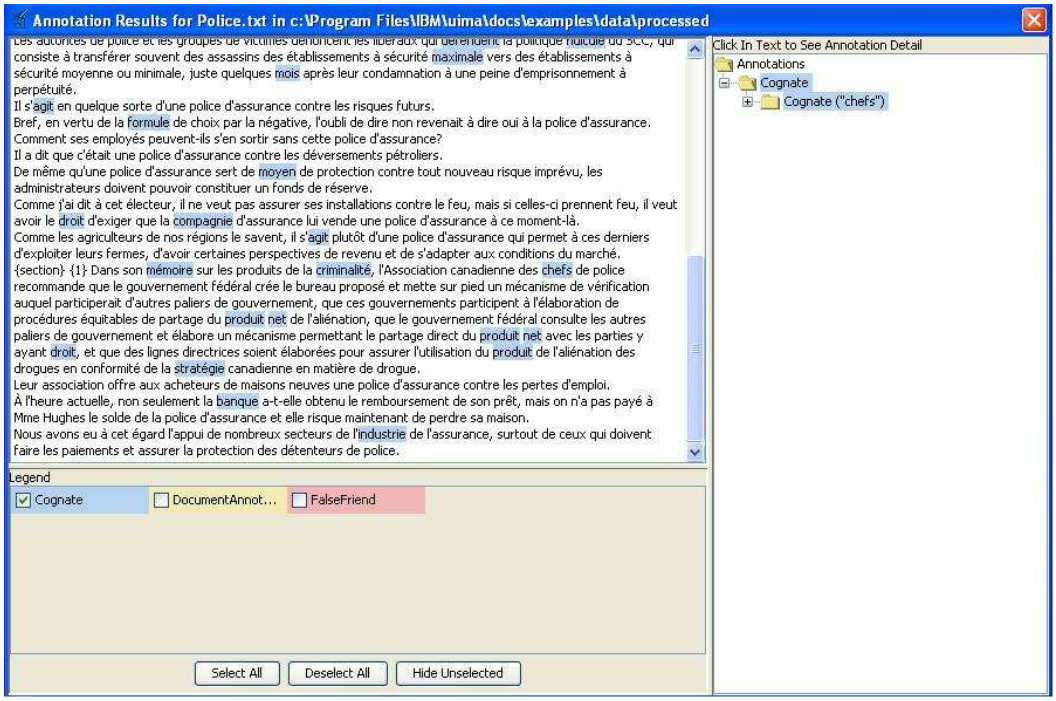

Figure 5.3: Cognate annotations.

The annotations that the tool makes are only for French content words: nouns, adjectives, adverbs and verbs. We have chosen to annotate only the content words not to introduce some false alarms (e.g. the French word *pour* can be either adverb (*pro*), or preposition (*for; to*), and it is a false friend with the English word *pour* that is a verb), and also because they are of more interest for second language learners.

Since BaLIE can provide information regarding the part-of-speech tag for each token in the text, it was easy for us to make the distinction between the content and close class French words.

The tool does not lemmatize the text, and for this reasons some words might not be annotated or some errors might be introduced. Some annotations might be missed because the words are not in the base form and some errors might be introduced because the inflected form corresponds to the base form for another word (e.g. the verb *être* has the singular third person *est* form that corresponds to the base form of the cardinal point *est* that is cognate with the English word *east*).

The annotation will be done only for the tokens in the text that have the same form as the pair of words in the lists, the base form of the French and English words. For the next version of the tool, we will have the lemmatization step performed on the text before we do the annotation step.

Both UIMA and BaLIE are Java projects that can be easily downloaded and used with Eclipse[6] SDK. In fact, UIMA has some of the features to be easily used with Eclipse. For both projects, documentation on how to install the projects is available from the corresponding web pages. For CLPA, the web page will provide instructions on how to install and put all the resources together so they can be ready to run for French text annotations.

[6]http://www.eclipse.org/

Chapter 6

Conclusions and Future Work

This chapter contains a brief overview of the research contributions that we bring in this thesis. It also contains ideas for future development of the main research topics that we have researched on.

All the research presented and evaluated in Chapters 3 and 4 is focused on special types of cross-language pair of words between French and English. The pairs of words that we looked at are: Cognates, False Friends, and Partial Cognates.

6.1 Conclusions

Identification of Cognates and False Friends In Chapter 3, we presented and evaluated a new method of identifying Cognates and False Friends between French and English. The method uses 13 orthographic similarity measures that are combined through different ML techniques. For each measure separately, we also determined a threshold of orthographic similarity that can be used to identify new pairs of cognates and false friends. The novelty that we bring to this task is the way we use and combine different orthographic similarity measures by ML techniques. The results show that the method can be used with success.

In addition to the ML technique that identifies cognates and false friends, we proposed

a method that uses a bilingual dictionary to create complete lists of cognates and false friends between two languages. For highly accurate results, the human effort that is needed is significantly lower than in the case of using only human knowledge, as done in previous work on creating list of these types of words. Moreover manually created lists of cognates and false friends are not very large because of the amount human effort required. Our method can create complete lists of cognates and false friends with a relatively high accuracy without much human effort. The human effort will be needed to prepare initial lists of cognates and false friends that will be used as training data to automatically determine the thresholds. These lists usually can be already found as resource for languages that are frequently used in research.

Partial Cognate Disambiguation Chapter 4 presents our proposed methods (a supervised and a semi-supervised method) for a task that, to our knowledge, was not in focus of research before, at least not in the NLP community.

Our research contribution consists in defining the task itself and in the methods that we propose to solve the task. We try to disambiguate French partial cognates — French words that in some contexts share the same meaning with a similar English word, and in others contexts have totally different meanings. For this task, we perform cross-language words sense disambiguation — we have French words that we want to disambiguate looking at an English word sense inventory.

For our task, partial cognate disambiguation, we propose two methods that were subject of experiments and evaluated on a list of 10 French partial cognates. The first method is a pure ML supervised approach that uses data automatically labeled; the second one is a semi-supervised method that contains two algorithms: Monolingual Bootstrapping and Bilingual Bootstrapping, which use free unlabeled texts. Our results show that simple methods and freely available tools lead to good results and cope well with the noise that might be present in the data, in a task that is hard to solve even for humans.

Another contribution is the fact that we showed that even though we started with data from the parliamentary domain we successfully bootstrap the knowledge of the methods with knowledge from different domains. The number of features that were extracted from the initial labeled data more than doubled at each MB and BB experiment, showing that even though we started with seeds from a restricted domain, the method can capture knowledge from different domains as well. Besides the change in the number of features, the domain of the features has also changed from the parliamentary one to others, more general, showing that the method will be able to disambiguate sentences where the partial cognates cover different types of context.

We focused not only on partial cognates that have completely distinct senses, we tried to look at those that have closely related senses as well. We evaluated our proposed methods with different data sets for training and testing, to better support our claims: we can solve a hard task using minimal human effort, we can bootstrap the knowledge of the classifiers with unlabeled data from different domains, and simple methods provide good results.

In our experiments we use the machine language representation — binary feature values, and we show that nonetheless machines are capable of learning from new information.New information was collected and extracted by classifiers when additional corpora were used for training.

A Tool for Cross-Language Pair Annotations In the Chapter 5 we present and describe a CALL tool, CLPA, which can annotate cognates and false friends in a French text. In its first version, the tool uses the list of cognates and false friends that were used to experiment with the cognate and false friend identification technique in Chapter 3 but any other list can be easily integrated. CLPA has an easy to use GUI that allows users to choose between annotations — only cognate annotation, only false friend annotation, or both, and also provide additional information to the users. This information can be useful to a second language learner similar to the feedback from a tutor.

6.2 Future Work

Identification of Cognates and False Friends As future work we want to apply the cognate and false friend identification task to other pairs of languages that lack this kind of resource (since the orthographic similarity measures are not language-dependent).

We want to increase the accuracy of the automatically generated lists of cognates and false friends by increasing the threshold used — we could obtain better precission but less recall for both classes. We could eliminate some falsely determined false friends by using other orthographic measures or the same measure with a higher threshold on the initial list determined with the same threshold for both cognates and false friends.

We also want to create complete lists of cognates and false friends for other languages. For this task, we will need initial lists of cognates and false friends, manually judged by humans, to determine the best threshold automatically. We also need a bilingual dictionary that will be used to make the distinction between cognates and false friends.

Partial Cognate Disambiguation In future work we could try to increase the accuracy of the results by using dependency relations, lemmatization, part-of-speech tagging — extract only sentences where the partial cognate has the same POS, and other types of data representation combined with different semantic tools (e.g., decision lists, rule based systems).

In future work for the disambiguation task, we want to look at different data representations, use lemmatization and POS tagging, and execute more steps for the Monolingual and Bilingual Bootstrapping algorithms.

To apply our method to new pairs of languages, all we need is parallel corpora for these languages, and a list of partial cognate words.

A Tool for Cross-Language Pair Annotations For the future we want to continue to develop the tool, add other features, perform the lemmatization step, and also annotate partial cognates with the corresponding meaning in the texts.

We also want to use French second-language students to evaluate the usefulness of the tool. We want to see if the cognate and false friend annotations are helpful, and more likely if the additional information that we provide helps students in the learning process.

Trying to develop the tool for other languages is also one of our future aims. In order to do this, all we need is to plug in lists of cognates and false friends for the corresponding languages.

The overall contribution that we bring to the NLP research community is the new methods that we proposed, experimented with and evaluated, and the new directions that we followed for cognate, false friend, and partial cognate words between French and English.

References

Adamson, G. W., and Boreham, J. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval* 10:253–260.

Albright, A., and Hayes, B. 2003. Rules vs. analogy in english past tenses: Acomputational/experimental study. *Cognition* (90):119–161.

Atkins, S. 1993. Tools for computer-aided corpus lexicography: the Hector project. (41):5–72.

Balcom, P.; Copeck, T.; and Szpakowicz, S. 2006. *Didalect:conception, implantation et évaluation initiale*. chapter Les cahiers scientifiques de l'association francophone pour le savoir, Colloque C-521: Les nouvelles technologies et le traitement automatique des langues au coeur des dispositifs d'apprentissage, 123–141.

Barker, G., and Sutcliffe, R. F. E. 2000. An experiment in the semi-automatic identification of false-cognates between english and polish. Technical report, Department of Languages and Cultural Studies, University of Limerick, Ireland.

Brew, C., and McKelvie, D. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, 45–55.

Brown, P.; Lai, J. C.; and Mercer, R. 1991. Aligning sentences in parallel corpora. *Proceedings of The Association for Computational Linguistics (ACL-91)* 169–176.

Carpuat, M., and Wu, D. 2005. Word sense disambiguation vs statistical machine translation. In *Proceedings of The Association for Computational Linguistics (ACL-2005)*.

Carroll, S. 1992. On cognates. Technical report, Second Language Research.

Chklovski, T., and Mihalcea, R. 2002. Building a sense tagged corpus with open

mind word expert. In *Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, 116–122.

Croft, W. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

Crystal, D. 1995. *The Cambridge Encyclopedia of the English Language*. Cambridge University Press.

Diab, M., and Resnik, P. 2002. An unsupervised method for word sense tagging using parallel corpora. *Proceedings of The Association for Computational Linguistics (ACL 2002)* 255–262.

Dollenmayer, D. B., and Hansen, T. S. 2003. *Neue Horizonte*. 6th edition.

Étiemble, R. 1991. *Parlez-vous Franglais*. Editions Flammarion.

Friel, B. M., and Kennison, S. M. 2001. Identifying German-English cognates, false cognates, and non-cognates: Methodological issues and descriptive norms. *Bilingualism: Language and Cognition* 4:249–274.

Frunza, O., and Inkpen, D. 2006. Semi-supervised learning of partial cognates using bilingual bootstrapping. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, COLING-ACL 2006*, 433–440.

Gale, W. A., and Church, K. W. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities,* 26:415–439.

Gale, W.; Church, K.; and Yarowsky, D. 1992. Using bilingual materials to develop word sense disambiguation methods. *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation,* 101–112.

Gass, S. 1987. *The use and acquisition of the second language lexicon*, volume 9. Studies in Second Language Acquisition 9(2).

Gollan, T. H., and Frost, K. I. F. . R. 1997. Translation priming with different scripts: masked priming with cognates and non-cognates in hebrewenglish bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23:1122–1139.

Greenberg, J. H. 1987. *Language in the Americas*. Stanford, CA, USA: Stanford University Press.

Groot, A. M. B. D., and Nas, G. L. J. 1991. Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language* 30:90–123.

Grosjean, F. 1997. Processing mixed language: issues, findings, and models. In *Tutorials in bilingualism: psycholinguistics perspectives*, 225–254. A. M. B. De groot & J.F. Kroll (eds.).

Grosjean, F. 1998. Studying bilinguals: methodological and conceptual issues. *Bilingualism: Language and Cognition* 131–149.

Guy, J. B. M. 1994. The use and acquisition of the second language lexicon (special issue). 9(2):35–42.

Hall, P. A. V., and Dowling, G. R. 1980. Approximate string matching. *Computing Surveys* 12(4):381–402.

Hammer, P. 1976. English-French Cognate Dictionary. Technical report, University of Alberta.

Hancin-Bhatt, B., and Nagy, W. 1993. Bilingual students developing understanding of morphologically complex cognates. Technical report.

Hearst, M. 1991. Noun homograph disambiguation using local context in large corpora. *Proceedings of the 7th Annual Conf. of the University of Waterloo Centre for the New OED and Text Research* 1–19.

Heuven, W. V.; Dijkstra, A.; and Grainger, J. 1998. Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language* 39:458–483.

Hewson, J. 1993. A computer-generated dictionary of proto-algonquian. Technical report, Ottawa:Canadian Museum of Civilization.

Ide, N., and Véronis, J. 1993a. Knowledge extraction from machine-readable dictionaries: an evaluation. *Third International EAMT Workshop Machine Translation and the Lexicon* 19–34.

Ide, N., and Véronis, J. 1993b. Refining taxonomies extracted from machine-readable dictionaries. *In Hockey, Susan and Nancy Ide (Eds.) Research in Humanities Computing II* 149–159.

Ide, N., and Veronis, J. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Association of Computational Linguistics* 24:1–40.

Ide, N.; Erjavec, T.; and Tufis, D. 2001. Automatic sense tagging using parallel corpora. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium* 83–89.

Ide, N. 2000. Cross-lingual sense determination: Can it work? *Computers and the Humanities, Special Issue on the Proceedings of the SIGLEX SENSEVAL Workshop* 34(1-2):223–234.

Inkpen, D.; Frunza, O.; and Kondrak, G. 2005. Automatic identification of Cognates and False Friends in French and English. In *In RANLP-2005*, 251–257.

Isabelle, P. 1993. Translation analysis and translation automation. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, 201–217.

Jarmasz, M., and Szpakowicz, S. 2001. Roget's Thesaurus: a Lexical Resource to Treasure. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, 186–188.

Kessler, B. 1995. Computational dialectology in irish gaelic. In *In Proceedings of the European ACL*, 60–67. Dublin.

Kilgarriff, A., and Rosenzweig, J. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities,* 34:15–48.

Kondrak, G., and Dorr, B. J. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING 2004)*, 952–958.

Kondrak, G., and Sherif, T. 2006. Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. In *Proceedings of the Workshop on Linguistic Distances*, 43–50. Association for Computational Linguistics.

Kondrak, G. 2001. Identifying Cognates by Phonetic and Semantic Similarity. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, 103–110.

Kondrak, G. 2003. Identifying complex sound correspondences in bilingual wordlists. In *Proceedings of the Fourth International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2003)*, 432–443.

Kondrak, G. 2004. Combining evidence in cognate identification. In *Proceedings of Canadian AI 2004: 17th Conference of the Canadian Society for Computational Studies of Intelligence*, 44–59.

Kondrak, G. 2005. Cognates and word alignment in bitexts. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, 305–312.

Kroll, J. F., and Stewart, E. 1994. Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33:149–174.

Kroll, J.; Michael, E.; Tokowicz, N.; and Dufour, R. 2002. The development of lexical fluency in a second language. *Second Language Research* 18:137–171.

Landes, S.; Leacock, C.; and Tengi, R. 1998. Building semantic concordances. *WordNet: An Electronic Lexical Database* 199–216.

Lebart, L., and Rajman, M. 2000. Computing similarity. In *Handbook of Natural Language Processing*, 477–505.

LeBlanc, R., and Séguin, H. 1996. *Les congénères homographes et parographes anglais-français*, volume Twenty-Five Years of Second Language Teaching at the University of Ottawa.

LeBlanc, R. 1989. *L'enseignement des langues secondes aux adultes: recherches et pratiques*. Les Presses de l'Université d'Ottawa.

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

Levin, B. 1993. English Verb Classes and Alternations: a Preliminary Investigation. Technical report, University of Chicago Press, Chicago and London.

Li, H., and Li, C. 2004. Word translation disambiguation using bilingual bootstrap. *Computational Linguistics* 30(1):1–22.

Lowe, J. B., and Mauzaudon, M. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics* 20:381–417.

Mackay, W., and Kondrak, G. 2005. Computing Word Similarity and Identifying Cognates with Pair Hidden Markov Models. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL 2005)*, 40–47.

Mann, G. S., and Yarowsky, D. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, 151–158.

Marcu, D.; Kondrak, G.; and Knight, K. 2003. Cognates can improve statistical translation models. *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)* 46–48.

Melamed, I. D. 1998. Manual annotation of translational equivalence: The Blinker project. Technical report, University of Pennsylvania.

Melamed, I. D. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics* 25:107–130.

Mihalcea, R., and Faruque, E. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of ACL/SIGLEX Senseval-3*, 155–158.

Mihalcea, R., and Moldovan, D. 2001. A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal on Artificial Intelligence Tools* 10(1-2):5–21.

Mihalcea, R. 2002. Bootstrapping large sense tagged corpora. *Proceedings of the 3rd International Conference on Languages Resources and Evaluations (LREC 2002).*

Mitchell, T. 1997. *Machine Learning.* McGraw Hill.

Nagy, W. E. 1992. Cross-language transfer of lexical knowledge: Bilingual students' use of cognates. Technical report.

Palmberg, R. 1988. Five experiments of EFL vocabulary learning: A project report. *Annual Meeting of International Association of Applied Linguistics* 1988.

Ringbom, H. 1987. *The Role of the First Language in Foreign Language Learning.* Clevedon, England: Multilingual Matters Ltd.

Robert-Collins. 1987. *Robert-Collins French-English English-French Dictionary.* London: Collins.

Sanchez-Casas, R. M., and Garcia-Albea, C. W. D. . J. E. 1992. Bilingual lexical processing: exploring the cognate/non-cognate distinction. *European Journal of Cognitive Psychology* 4:293–310.

Simard, M.; Foster, G. F.; and Isabelle, P. 1992. Using cognates to align sentences

in bilingual corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, 67–81.

Thomason, S., and Kaufmann, T. 1988. Language contact, creolization, and genetic linguistics. Technical report, University of California.

Tiedemann, J. 1999. Automatic construction of weighted string similarity measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Tréville, M.-C. 1990. *Rôle des congénères interlinguaux dans le dévelopment du vocabulair réceptif*. Ph.D. Dissertation, Université de Montreal.

Tufis, D.; Radu, I.; 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)* 1312–1318.

Véronis, J., and Ide, N. 1991. An assessment of information automatically extracted from machine readable dictionaries. *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics (EACL 1991)* 227–232.

Vickrey, D.; Biewald, L.; Teyssier, M.; and Koller, D. 2005. Word-sense disambiguation for machine translation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* 779–786.

Wagner, R. A., and Fischer, M. J. 1974. The string-to-string correction problem. *JACM* 21:168–173.

Weaver, W. 1949. In machine translation of languages. Technical report, MIT Press,Cambridge.

Witten, I. H., and Frank, E. 2005. *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition.

Yarowsky, D. 1995. Unsupervised word sence disambiguation rivaling supervised methods. *In Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL-95)* 189–196.

Appendix A

Feature-Value Representation for Word Pairs

For the following list of pairs of English-French words

abbé abbey

accidenté accidental

accuser accuse

achèvement achievement

acompte account

action action

actuel actual

actuellement actually

addition addition

adepte adept

abeille bee

aimer like

environ about

conseil advice

après afterwards

contre against

consentir agree

administré administered

ado ado

adresse address

the feature value representation that we used for ML algorithms is:

@relation CogFFriends_NonRelated

@attribute DICE numeric

@attribute EDIT numeric

@attribute IDENT numeric

@attribute LCSR numeric

@attribute SIMARD numeric

@attribute SOUNDEX numeric

@attribute TRI numeric

@attribute XDICE numeric

@attribute XXDICE numeric

@attribute BI-DIST numeric

@attribute BI-SIM numeric

@attribute TRI-DIST numeric

@attribute TRI-SIM numeric

@attribute Class CG_FF,NR

@data

0.5714, 0.6000, 0.0000, 0.6000, 0.6000, 0.7500, 0.2222, 0.5000, 0.5000, 0.7000, 0.7000, 0.7333, 0.7333, CG_FF

0.8235, 0.8000, 0.0000, 0.8000, 0.8000, 1.0000, 0.6316, 0.8125, 0.8125, 0.8500, 0.8500, 0.8667, 0.8667, CG_FF

0.9091, 0.8571, 0.0000, 0.8571, 0.8571, 0.7500, 0.6154, 0.9000, 0.9000, 0.8571, 0.8571, 0.8571, 0.8571, CG_FF

0.7368, 0.8182, 0.0000, 0.8182, 0.2727, 0.5000, 0.4762, 0.6667, 0.4167, 0.8182, 0.8182, 0.7879, 0.7879, CG_FF

0.3333, 0.4286, 0.0000, 0.5714, 0.2857, 0.7500, 0.0000, 0.1818, 0.1364, 0.4286, 0.5714, 0.4286, 0.5714, CG_FF

1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 0.6667, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, CG_FF

0.6000, 0.8333, 0.0000, 0.8333, 0.6667, 1.0000, 0.3333, 0.6667, 0.6667, 0.8333, 0.8333, 0.8889, 0.8889, CG_FF

0.4444, 0.5000, 0.0000, 0.5000, 0.3333, 1.0000, 0.2000, 0.4118, 0.4118, 0.5417, 0.5417, 0.5556, 0.5556, CG_FF

1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 0.7500, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, CG_FF

0.8889, 0.8333, 0.0000, 0.8333, 0.8333, 1.0000, 0.5455, 0.8750, 0.8750, 0.8333, 0.8333, 0.8333, 0.8333, CG_FF

0.2500, 0.4286, 0.0000, 0.4286, 0.0000, 0.2500, 0.0000, 0.1429, 0.0714, 0.2857, 0.2857, 0.2381, 0.2381, UNREL

0.0000, 0.4000, 0.0000, 0.4000, 0.0000, 0.2500, 0.0000, 0.1667, 0.1667, 0.3000, 0.3000, 0.2667, 0.2667, UNREL

0.0000, 0.0000, 0.0000, 0.1429, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.1429, 0.0000, 0.0952, UNREL

0.0000, 0.0000, 0.0000, 0.2857, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.1429, 0.0000, 0.0952, UNREL

0.0000, 0.3000, 0.0000, 0.3000, 0.1000, 0.5000, 0.0000, 0.1667, 0.0055, 0.3000, 0.3000, 0.3000, 0.3000, UNREL

0.0000, 0.0000, 0.0000, 0.2857, 0.0000, 0.0000, 0.0000, 0.1000, 0.0200, 0.0000, 0.1429, 0.0000, 0.1429, UNREL

0.0000, 0.1111, 0.0000, 0.1111, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.1111, 0.1111, 0.0741, 0.0741, UNREL

0.7000, 0.7500, 0.0000, 0.7500, 0.6667, 1.0000, 0.5455, 0.7368, 0.7368, 0.7500, 0.7500, 0.7500, 0.7500, CG_FF

1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 0.3333, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, 1.0000, CG_FF

0.8333, 0.7143, 0.0000, 0.8571, 0.2857, 1.0000, 0.4286, 0.7273, 0.4091, 0.7143, 0.8571, 0.6667, 0.8095, CG_FF

0.9091, 0.8571, 0.0000, 0.8571, 0.8571, 1.0000, 0.6154, 0.9000, 0.9000, 0.8571, 0.8571, 0.8571, 0.8571, CG_FF

Appendix B

False Friends Word Distribution

This appendix contains the distribution of the false friend words in the training set, testing set, and BNC sentences that we used in our experiments from Chapter 4. For each French partial cognate word, we counted the number of false friend words that appeared in the English sentences. For Example, the French partial cognate *Corps* has two English false friend words: *body* that appeared 79 times in the English training seeds and *corpse* that appeared 2 times.

The following tables contain the distribution for the set of partial cognates in all three English data sets.

PC	False Friend word	Frequency
Blanc	white	78
Circulation	traffic	75
Client	shopper	3
	patient	3
	customer	84
	user	1
Corps	body	79
	corpse	2
Détail	retail	80
Mode	fashion	35
	vogue	5
	style	61
	trend	7
Note	education	14
	account	18
	grade	2
	restaurant	2
	bill	99
	hotel	6
Police	face	24
	insurance	12
	policy	60
Responsible	in charge	61
	officer	81
	representative	1
	official	2
	person in charge	5
	executive	45
Route	roadside	2
	road	90

Table B.1: False Friend word distribution in the English training set.

PC	False Friend word	Frequency
Blanc	white	39
Circulation	traffic	38
Client	patient	3
	customer	41
	patron	1
Corps	body	42
Détail	retail	41
Mode	fashionable	33
	fashion	20
Note	mark	1
	check	14
	account	2
	grade	48
	restaurant	2
	bill	9
Police	insurance	4
	policy	48
Responsible	in charge	53
	officer	3
	representative	23
	official	6
	executive	1
Route	road	46

Table B.2: False Friend word distribution in the English testing set.

PC	False Friend word	Frequency
Blanc	white	163
	livid	16
Circulation	traffic	250
Client	shopper	4
	patient	22
	customer	152
	user	29
	spectator	13
	patron	35
Corps	body	147
	corpse	105
Détail	retail	250
Mode	fashionable	69
	fashion	88
	vogue	30
	style	37
	trend	36
Note	education	18
	account	24
	grade	53
	restaurant	13
	score	10
	hotel	10
	mark	67
	book out	3
	check out	2
	check	47
	bill	41
Police	insurance	69
	font	35
	face	59
	policy	89
Responsible	person in charge	1
	executive	61
	in charge	1
	responsible party	6
	officer	50
	representative	59
	official	77
Route	roadside	99
	road	156

Table B.3: False Friend word distribution in the BNC corpus.

Appendix C

Monolingual and Bilingual Experimental Results

The appendix presents the results for the partial cognate set using different training and testing data sets.

PC	ZeroR	NB-K	Trees	SMO
Blanc	92.50%	92.50%	92.50%	92.50%
Circulation	72.22%	66.66%	33.33%	66.66%
Client	61.40%	54.38%	47.36%	47.36%
Corps	1.36%	99.00%	97.26%	98.63%
Détail	100%	88.00%	100%	100%
Mode	6.74%	67.41%	46.06%	63.00%
Note	91.45%	41.88%	71.79%	70.94%
Police	97.29%	75.67%	93.69%	87.38%
Responsable	61.17%	55.29%	61.17%	60.00%
Route	94.33%	79.24%	94.33%	81.13%
AVERAGE	67.85%	71.97	**73.75%**	76.75%

Table C.1: Results for the partial cognate set when training on the training seed set (S) and testing on the new corpus (NC) set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	92.50%	91.66%	92.50%	92.50%
Circulation	72.22%	55.55%	33.33%	66.66%
Client	38.59%	57.89%	38.59%	59.64%
Corps	98.63%	97.00%	93.83%	90.41%
Détail	100%	92.00%	100%	100%
Mode	6.74%	82.02%	6.74%	79.00%
Note	91.45%	76.92%	70.94%	74.35%
Police	91.45%	81.08%	94.59%	93.69%
Responsable	38.82%	56.47%	61.17%	60.00%
Route	5.66%	49.05%	13.20%	32.07%
AVERAGE	63.61%	**73.92%**	60.49%	74.80%

Table C.2: Results for the partial cognate set when training on the training seed set (S) plus BNC and testing on the new corpus (NC) set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	92.50%	93.33%	91.66%	92.50%
Circulation	72.22%	50.00%	50.00%	33.33%
Client	61.40%	47.36%	43.85%	49.12%
Corps	1.36%	99.00%	96.57%	91.09%
Détail	100%	92.00%	100%	92.00%
Mode	6.74%	62.92%	6.74%	63%
Note	91.45%	39.31%	23.07%	34.18%
Police	97.29%	56.75%	97.29%	61.26%
Responsable	61.17%	50.58%	58.82%	48.23%
Route	94.33%	81.13%	94.33%	73.58%
AVERAGE	67.85%	67.03%	64.65%	65.57%

Table C.3: Results for the partial cognate set when training on the training seed set (S) plus Lemonde (LM) corpus and testing on the new corpus (NC) set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	92.50%	91.66%	92.50%	93.33%
Circulation	72.22%	55.55%	33.33%	55.55%
Client	38.59%	59.64%	38.59%	54.38%
Corps	98.63%	99.00%	97.26%	89.04%
Détail	100%	96.00%	100%	96.00%
Mode	6.74%	73.03%	14.60%	74.00%
Note	91.45%	41.02%	24.78%	31.62%
Police	97.29%	64.86%	93.69%	71.17%
Responsable	38.82%	52.94%	62.35%	54.11%
Route	5.66%	71.69%	13.20%	49.05%
AVERAGE	64.19%	70.57%	57.03%	66.84%

Table C.4: Results for the partial cognate set when training on the training seed set (S) plus Lemonde (LM) corpus plus BNC corpus and testing on the new corpus (NC) set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	97.01%	94.02%	97.01%
Circulation	73.79%	91.03%	84.13%	91.03%
Client	54.08%	72.44%	63.26%	63.26%
Corps	48.83%	73.00%	48.83%	77.9%
Détail	59.40%	89.10%	82.17%	84.15%
Mode	41.75%	83.51%	92.30%	86.00%
Note	64.94%	87.11%	76.80%	81.95%
Police	61.41%	83.46%	92.91%	93.70%
Responsable	55.24%	82.87%	77.90%	76.79%
Route	56.79%	60.49%	56.79%	55.55%
AVERAGE	57.44%	82.03%	76.91%	80.71%

Table C.5: Results for the partial cognate set when training on the training seed set (S) plus the new corpus (NC) set and testing on the testing set (TS).

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	98.50%	97.01%	98.50%
Circulation	73.79%	90.34%	70.34%	84.82%
Client	54.08%	74.48%	54.08%	66.32%
Corps	48.83%	81.00%	56.97%	70.93%
Détail	59.40%	90.00%	75.24%	82.17%
Mode	41.75%	87.91%	91.20%	87.00%
Note	64.94%	85.05%	74.22%	78.35%
Police	61.41%	70.86%	92.91%	68.5%
Responsable	55.24%	82.32%	69.06%	79.55%
Route	56.79%	59.25%	56.79%	54.32%
AVERAGE	57.44%	82.03%	76.91%	80.71%

Table C.6: Results for the partial cognate set when training on the training seed set (S) plus the new corpus (NC) plus Lemonde corpus (LM) set and testing on the test set (TS).

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.2%	94.02%	97.01%	98.50%
Circulation	73.79%	91.72%	63.44%	88.96%
Client	45.91%	68.36%	45.91%	62.24%
Corps	48.83%	83.00%	54.65%	82.55%
Détail	59.40%	91.08%	82.17%	86.13%
Mode	41.75%	87.91%	45.05%	86.00%
Note	64.94%	86.08%	75.77%	80.41%
Police	61.41%	78.74%	96.06%	96.85%
Responsable	55.24%	86.18%	77.90%	77.90%
Route	56.79%	69.13%	45.67%	64.19%
AVERAGE	56.63%	**83.58%**	68.36%	82.34%

Table C.7: Results for the partial cognate set when training on the training seed set (S) plus the new corpus (NC) plus the BNC corpus and testing on the test set (TS).

PC	ZeroR	NB-K	Trees	SMO
Blanc	58.20%	98.50%	97.01%	98.50%
Circulation	73.79%	88.27%	69.65%	85.51%
Client	45.91%	72.44%	54.08%	68.36%
Corps	43.83%	83.00%	60.46%	74.41%
Détail	59.40%	91.08%	85.14%	88.11%
Mode	58.24%	90.10%	92.30%	84.00%
Note	84.20%	78.35%	76.80%	64.94%
Police	61.41%	73.22%	96.06%	74.01%
Responsable	55.24%	87.84%	77.34%	79.55%
Route	56.79%	62.96%	45.67%	61.72%
AVERAGE	57.78%	**83.10%**	75.61%	79.05%

Table C.8: Results for the partial cognate set when training on the training seed set (S) plus the new corpus (NC) set plus LeMonde (LM) corpus plus BNC corpus and testing on the test set (TS).

PC	ZeroR	NB-K	Trees	SMO
Blanc	85.24%	93.44%	94.09%	94.09%
Circulation	73.74%	86.03%	70.94%	84.91%
Client	58.09%	59.52%	54.28%	53.80%
Corps	13.00%	90.15%	89.36%	92.28%
Détail	72.66%	88.66%	90.00%	91.33%
Mode	24.25%	75.00%	60.82%	72.00%
Note	79.57%	63.38%	74.64%	77.69%
Police	84.43%	77.52%	93.94%	90.20%
Responsable	58.16%	70.77%	65.90%	68.19%
Route	78.37%	67.56%	78.37%	69.72%
AVERAGE	62.70%	**77.20%**	77.23%	79.46%

Table C.9: Results for the partial cognate set when training on the training seed set (S) and testing on the test set (TS) plus new corpus (NC) set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	85.24%	93.44%	93.77%	94.75%
Circulation	73.79%	73.74%	63.12%	77.65%
Client	58.09%	53.33%	45.71%	53.33%
Corps	12.50%	86.00%	87.76%	86.17%
Détail	72.66%	89.33%	90.00%	85.33%
Mode	24.25%	72.01%	35.07%	54.69%
Note	79.57%	60.09%	45.07%	54.69%
Police	84.43%	62.82%	95.96%	65.12%
Responsable	58.16%	69.62%	68.48%	65.61%
Route	78.37%	69.18%	78.37%	66.48%
AVERAGE	62.70%	**72.97%**	70.33%	71.97%

Table C.10: Results for the partial cognate set when training on the training seed set (S) plus LeMonde (LM) corpus and testing on the test set (TS) plus new corpus (NC) set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	85.24%	92.78%	93.77%	94.09%
Circulation	73.74%	84.91%	57.54%	83.79%
Client	41.90%	63.33%	41.90%	60.95%
Corps	87.50%	93.00%	88.03%	88.56%
Détail	72.66%	91.33%	90.00%	90.66%
Mode	24.25%	84.32%	35.07%	82.00%
Note	79.57%	80.75%	73.94%	76.76%
Police	84.43%	81.26%	95.38%	94.81%
Responsable	41.83%	72.77%	67.90%	70.20%
Route	21.62%	53.51%	27.02%	46.48%
AVERAGE	61.27%	**79.83%**	67.06%	78.80%

Table C.11: Results for the partial cognate set when training on the training seed set (S) plus BNC corpus and testing on the test set (TS) plus new corpus (NC) set.

PC	ZeroR	NB-K	Trees	SMO
Blanc	85.24%	93.11%	93.77%	94.75%
Circulation	73.74%	82.12%	62.56%	80.44%
Client	41.90%	65.71%	45.71%	60.47%
Corps	87.50%	95.00%	89.09%	86.17%
Détail	72.66%	92.00%	90.00%	90.00%
Mode	24.25%	77.98%	40.00%	78.00%
Note	79.57%	61.16%	94.52%	71.18%
Police	84.43%	69.16%	94.52%	71.18%
Responsable	41.83%	71.34%	69.34%	67.04%
Route	21.62%	65.40%	27.02%	57.83%
AVERAGE	61.27%	**77.28%**	65.75%	73.87%

Table C.12: Results for the partial cognate set when training on the training seed set (S) plus LeMonde (LM) corpus plus BNC corpus and testing on the test set (TS) plus new corpus (NC) set.

Appendix D

Examples of Decision Trees

This appendix presents the Decision Tree classifiers for the *Corps* partial cognate when different training data sets are used. The number of features and the domain are changing with the additional data that we use for the bootstrapping technique.

```
europen <= 0
|   civil <= 0
|   |   diplomatique <= 0
|   |   |   arme <= 0
|   |   |   |   pays <= 0
|   |   |   |   |   amliorer <= 0
|   |   |   |   |   |   intervention <= 0: FF (93.0/14.0)
|   |   |   |   |   |   intervention > 0: CG (4.0)
|   |   |   |   |   amliorer > 0
|   |   |   |   |   |   sault <= 0: FF (3.0/1.0)
|   |   |   |   |   |   sault > 0: CG (2.0)
|   |   |   |   pays > 0: CG (5.0/1.0)
|   |   |   arme > 0: CG (9.0)
|   |   diplomatique > 0: CG (11.0)
|   civil > 0: CG (13.0)
europen > 0: CG (30.0)
```

Figure D.1: Decision Tree Classifier generated when using the training seed set. Example of Decision Tree classifier for the *Corps* partial cognate trained on the training seed set.

```
cadavre <= 0
|   arme <= 0
|   |   militaire <= 0
|   |   |   pays <= 0
|   |   |   |   crer <= 0
|   |   |   |   |   esprit <= 0
|   |   |   |   |   |   civil <= 0
|   |   |   |   |   |   |   organisation <= 0
|   |   |   |   |   |   |   |   paix <= 0: FF (211.0/82.0)
|   |   |   |   |   |   |   |   paix > 0: CG (12.0/5.0)
|   |   |   |   |   |   |   organisation > 0: CG (12.0/1.0)
|   |   |   |   |   |   civil > 0: CG (36.0/17.0)
|   |   |   |   |   esprit > 0: CG (15.0/6.0)
|   |   |   |   crer > 0: CG (16.0/7.0)
|   |   |   pays > 0: CG (16.0/7.0)
|   |   militaire > 0: CG (29.0/10.0)
|   arme > 0: CG (65.0/10.0)
cadavre > 0: FF (84.0)
```

Figure D.2: Decision Tree Classifier generated when using BB.

Example of Decision Tree classifier for the *Corps* partial cognate trained on the training seed set plus the BNC corpus (BB).

```
arme <= 0
|   guerre <= 0
|   |   grands <= 0
|   |   |   monde <= 0
|   |   |   |   cadavre <= 0
|   |   |   |   |   politique <= 0
|   |   |   |   |   |   militaire <= 0
|   |   |   |   |   |   |   commission <= 0
|   |   |   |   |   |   |   |   paix <= 0
|   |   |   |   |   |   |   |   |   sens <= 0: FF (833.0/190.0)
|   |   |   |   |   |   |   |   |   sens > 0
|   |   |   |   |   |   |   |   |   |   tte <= 0
|   |   |   |   |   |   |   |   |   |   |   semble <= 0
|   |   |   |   |   |   |   |   |   |   |   |   tide <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   machine <= 0: CG(18.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   machine > 0: FF (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   tide > 0: FF (2.0)
|   |   |   |   |   |   |   |   |   |   |   semble > 0: FF (2.0)
|   |   |   |   |   |   |   |   |   |   tte > 0: FF (4.0)
|   |   |   |   |   |   |   |   |   paix > 0: CG (40.0/16.0)
|   |   |   |   |   |   |   |   commission > 0: CG (25.0/9.0)
|   |   |   |   |   |   |   militaire > 0: CG (29.0/9.0)
|   |   |   |   |   |   politique > 0: CG (31.0/10.0)
|   |   |   |   |   cadavre > 0: FF (84.0)
|   |   |   |   monde > 0
|   |   |   |   |   propre <= 0
|   |   |   |   |   |   merveilleux <= 0
|   |   |   |   |   |   |   dcouverte <= 0
|   |   |   |   |   |   |   |   matire <= 0: CG (33.0/3.0)
|   |   |   |   |   |   |   |   matire > 0: FF (3.0/1.0)
|   |   |   |   |   |   |   dcouverte > 0: FF (2.0)
|   |   |   |   |   |   merveilleux > 0: FF (2.0)
|   |   |   |   |   propre > 0: FF (3.0/1.0)
|   |   |   grands > 0
|   |   |   |   pieds <= 0: CG (34.0/4.0)
|   |   |   |   pieds > 0: FF (2.0)
|   |   guerre > 0: CG (29.0/1.0)
arme > 0: CG (110.0/11.0)
```

Figure D.3: Decision Tree Classifier generated when using MB plus BB.

Example of Decision Tree classifier for the *Corps* partial cognate trained on the training seed set plus the BNC corpus plus the LeMonde (LM) corpus plus the new corpus (NC) (MB+BB).