# A tool for detecting French-English cognates and false friends

Oana FRUNZA, Diana INKPEN
School of Information Technology and Engineering
University of Ottawa, Ottawa, ON, K1N 6N5, Canada
{ofrunza,diana}@site.uottawa.ca

**Résumé.** Les congénères sont des mots qui ont au moins un sens en commun entre deux langues en plus d'avoir une orthographie semblable. La reconnaissance de ce type de mots permet aux apprenants de langue seconde ou étrangère d'enrichir plus rapidement leur vocabulaire et d'améliorer leur compréhension écrite. Toutefois, les faux amis sont des paires de mots qui à l'écrit ont des similarités, mais ils ont des significations différentes. Pour leur part, les congénères partiels sont des mots qui ont la même signification dans certains contextes dans chacune des deux langues. Cet article présente une méthode pour la classification automatique des paires des mots classées en congénères ou faux amis, en utilisant des mesures de similarité orthographiques et des méthodes d'apprentissage automatique. Ainsi, nous construisons des listes complètes des congénères et des faux amis entre les deux langues. Nous désambiguisons les congénères partiels dans des contextes spécifiques. Nos méthodes sont évaluées pour le français et l'anglais, mais elles seraient applicables à d'autres paires des langues. Nous avons construit un outil qui prend ces listes et marque dans un texte français les mots qui ont des congénères ou des faux amis en anglais, dans le but d'aider les apprenants en français langue seconde ou étrangère à améliorer leur compréhension écrite et à développer une meilleure rétention.

**Abstract.** Cognates are pairs of words in different languages similar in spelling and meaning. They can help a second-language learner on the tasks of vocabulary expansion and reading comprehension. False friends are pairs of words that have similar spelling but different meanings. Partial cognates are pairs of words in two languages that have the same meaning in some, but not all contexts. In this article we present a method to automatically classify a pair of words as cognates or false friends, by using several measures of orthographic similarity as features for classification. We use this method to create complete lists of cognates and false friends between two languages. We also disambiguate partial cognates in context. We applied all our methods to French and English, but they can be applied to other pairs of languages as well. We built a tool that takes the produced lists and annotates a French text with equivalent English cognates or false friends, in order to help second-language learners improve their reading comprehension skills and retention rate.

**Mots-clés :** congénères, faux amis, congénères partiels, mesures de similarité orthographiques, apprentissage automatique, apprentissage des langues assisté par ordinateur.

**Keywords:** cognates, false friends, partial cognates, orthographic similarity measures, machine learning (ML), computer-assisted language learning (CALL).

# 1 Introduction

When learning a second language, a student can benefit from knowledge in his/her first language (Gass, 1987), (LeBlanc *et al.*, 1989). Cognate words can accelerate vocabulary acquisition and facilitate the reading comprehension task. On the other hand, a student has to pay attention to pair of words that are false friends and partial cognates. The following definitions are language-independent, but the examples that we give are for French and English, the focusses 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.

Although French and English belong to different branches of the Indo-European family of languages, their vocabularies share a great number of similarities. Most of these similar 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. Most of the borrowings have changed their orthography, following different orthographic rules (LeBlanc & Séguin, 1996) and most likely their meaning as well.

Cognates have been employed in Natural Language Processing (NLP) for different tasks. The applications include sentence alignment (Simard *et al.*, 1992), (Melamed, 1999), and improving statistical machine translation models (Marcu *et al.*, 2003). Machine Translation (MT) systems can benefit from extra information when translating a certain word in context. Knowing if a French word is a cognate or a false friend with an English word can improve the translation results. 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 focus on the automatic identification of cognates and false friends. Our approach is based on several orthographic similarity measures that we use as features for classification. We test each feature separately, we average the values of all features, and we also explore various ways to combine the measures, by applying several Machine Learning techniques from the Weka package[1].

The task of disambiguating partial cognates can be seen as a coarse grain cross-language Word-Sense Discrimination. The results of this process can be useful for different NLP tasks and applications. Our proposed methods can be applied to any pair of languages for which a parallel corpus is available, and two monolingual collections of text. One of our main focus is to be able to disambiguate a French partial cognate looking at its English cognate and false friend senses.

[1] http ://www.cs.waikato.ac.nz/ml/weka/

We implemented a tool, a Computer-Assisted Language Learning (CALL) tool that is capable to annotate cognates and false friends in French texts, in order to help second language learners of French (native English speakers) in a reading comprehension and vocabulary retention task.

In the following sections we present related work, followed by our methods for cognate and false friend identification and its evaluation ; then we briefly describe our partial cognate disambiguation method and its evaluation, and at the end we present the CALL tool.

## 2 Related work

Previous work on automatic cognate identification is mostly related to bilingual corpora and translation lexicons. (Simard *et al.*, 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.

One of the most active researchers in automatic identification of cognates between various pairs of languages is (Kondrak, 2001), (Kondrak, 2004). His work is related to the phonetic aspect of cognate identification, especially **genetic cognates** – pairs in related languages that derive directly from the same word in the ancestor (proto)-language. He uses algorithms that combine different orthographic and phonetic measures, recurrent sound correspondences, and semantic similarity based on gloss overlap.

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 and Larousse-Saturne). They concluded that cognates appear to make up over 30% of the vocabulary.

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 *et al.*, 1998) suggest that missing false-friend recognition can be corrected when cross-language activation is used.

Word Sense Disambiguation (WSD) is an NLP task that has attracted researchers since 1950 and it is still a topic of high interest. We define the partial cognate disambiguation task as a cross-language WSD task. Determining the sense of an ambiguous word, using bootstrapping and texts from a different language was done by (Yarowsky, 1995), (Hearst, 1991), (Diab & Resnik, 2002), and (Li & Li, 2004). We follow a similar approach for our partial cognate disambiguation task. The difference between our approach and the ones mentioned above, is that our technique uses the whole sentence from the parallel text, not only the target words (the translation of certain English words), unlike (Diab & Resnik, 2002) ; we do not impose any constrains like (Yarowsky, 1995), our focus is not only on nouns as in (Hearst, 1991) ; and we look at words that are difficult to disambiguate even for humans, not only at very different words as in (Li & Li, 2004).

# 3 Cognates and false friends identification

## 3.1 Data sets for cognate and false friend identification

The data sets that we used to perform experiments for the task of cognates and false friends identification consist in a training and a testing list of pairs of words that are manually annotated as being cognates or false friends. The training dataset that we used contains 1454 pairs of French and English words.

They were extracted from different resources[2]. A separate test set composed of 1040 pairs were also extracted. A summary of the data that we have used is presented in Table 1.

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

TAB. 1 – The composition of data sets. The numbers in brackets are counts of word pairs that are identical, ignoring accents.

## 3.2 Method for cognate and false friend identification

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). We used different supervised ML algorithms to best discriminate between the two classes that we have chosen : Cognates/False Friends — are orthographically similar, and Unrelated — are not orthographically similar.

In our method an instance is a pair of words containing a French word and an English word. The features that we have chosen to use in our method are the 13 orthographic similarity measures. We performed experiments when we use different feature combinations, each pair of words is represented by all 13 orthographic similarity measures, by one of the measures (we want to determine a threshold value for each measure) or by the average result 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 from a French-English pair (see (Inkpen *et al.*, 2005) for a detailed description of the measures).

## 3.3 Evaluation results for cognate and false friend identification

We present evaluation experiments using the two datasets described in Section 3.1. We classify a pair of words 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).

[2] See (Inkpen *et al.*, 2005) for a description of these resources.

**Results on the test data set.** Table 2 presents the results that we obtained for the cognate and false friend identification task on the test set. We report results for each measure separately, the average of all measures and when we used all 13 measures as features for ML algorithms.

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

We also 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) to experiment our method with all 13 measures. Surprisingly, only the Naïve Bayes classifier outperforms the simple average of orthographic measures. Among the individual orthographic measures, XXDICE performs the best, supporting the results on French-English cognates reported in (Brew & McKelvie, 1996).

We run similar experiments on the training data using 10-fold cross validation and we obtained similar results (better with 1-2%) as the ones that we obtained on the test set. 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.

**Results for three-class classification.** We also experiment our method by adding one more feature and increasing the number of classes to three. The new feature which is set to 1 if the two words are translations of each other, and to 0 otherwise. The three classes are : Cognates, False-Friends and Unrelated.

As expected, this experiment achieved similar but slightly lower results than the ones from Table 2. 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 taken from a bilingual dictionary or a bilingual list of words.

**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. Several of the measures are particularly sensitive to the initial letter of the word, which is a strong clue of cognation, *arrêt* - *arm*, *peine* - *pear*. 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 Complete lists of cognates and false friends between two languages

We applied the methods of classifying pairs of words into cognates or false friends to the task of creating complete lists of cognates and false friends between two languages. As knowledge of which pairs are translations we use a bilingual dictionary and a bilingual list of words.

| Classifier (measure or combination) set | Accuracy on test set |
|---|---|
| IDENT | 55.00% |
| PREFIX | 90.97% |
| DICE | 93.37% |
| LCSR | 94.24% |
| NED | 93.57% |
| SOUNDEX | 84.54% |
| TRI | 92.13% |
| XDICE | 94.52% |
| XXDICE | 95.39% |
| BI-SIM | 93.95% |
| BI-DIST | 94.04% |
| TRI-SIM | 93.28% |
| TRI-DIST | 93.85% |
| Average measure | 94.14% |
| Baseline | 66.98% |
| OneRule | 92.89% |
| Naive Bayes | 94.62% |
| Decision Trees | 92.08% |
| DecTree (pruned) | 93.18% |
| IBK | 92.80% |
| Ada Boost | 93.47% |
| Perceptron | 91.55% |
| SVM (SMO) | 93.76% |

TAB. 2 – Results on the test set of the classifiers built on the training set (individual measures and machine learning combinations).

**Method description.** For each pair of words that have high ortographic similarity, if they are transation of each other we put the pair in the list of cognates, otherwise we put it in the list of false friends.

For these experiments 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 (see Table 2). The threshold automatically determined by the method in the previous section was 0.12. We increased it a little because we wanted to obtain pairs that are classified with a higher confidence.

**Results for building large lists of cognates and false friends from dictionary entry lists.** To collect pairs of words that are translations of each other we used the dictionary entries from the *Internet Dictionary Project* (IDP)[3]. 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. We concluded that 55% of the dictionary entries are classified as cognates.

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

To determine pairs of words that are not translations of each other — possible false friends, we paired each entry word with all 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. From the total number of pair of words that we created and are not translation of each other only 2% were determined to be orthographically similar enough to be false friends.

**Results for building large lists of cognates and false friends from monolingual lists of words.** In order to produce complete lists of words between two languages, we used the English entries from the LDOCE[4] 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 lemmas of French content words (nouns, adjectives, verbs, and adverbs) from the *Analyse et Traitement Informatique de la Langue Française* (ATILF[5]) project. After we paired each English word with each French word we obtained a list of pairs of words that we tried to classify in cognates and false friends using an on-line French-English Dictionary[6] of approximately 75,000 terms. From all pairs of words that were created we selected only the ones that have an XXDICE orthographic similarity value greater than 0.14. The number of pairs that are selected as similar is 11,469,662. From this number, only 3,496 pairs were identified as cognates and 3,767,435 as false friends.

As mentioned, 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 significantly 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 close to 100% recall ; we might miss the genetic cognates that have a common origin and but changed their spelling significantly.

# 4 Partial cognate disambiguation

This section presents our proposed techniques, based on Machine Learning, to disambiguate partial cognates. Partial cognates behave as cognates in some contexts, and as false friends in others. We use a semi-supervised method based on Monolingual and Bilingual Bootstrapping and parallel corpora to automatically create and tag our training seeds for the bootstrapping techniques. In addition to all the methods that use bootstrapping and parallel text, we also bootstrap our method with corpora from different domains. Our method uses a small set of seeds from Hansard, but additional knowledge from different domains is added using bootstrapping.

We use a supervised method to train classifiers on a part of automatically annotated data (collected from Hansard and EuroParl) and test their performance on the part of the data set aside for testing. We use 2/3 of the automatically tagged data as training and the 1/3 part for testing, an average of 130 sentences for the cognate class for training and 66 for testing an average of 100 sentences for the false friend class for training and 50 for testing. We used a set of 10 French partial cognates for which we had the corresponding English cognate and false friend words.

For the supervised method we obtained an average of 80% accuracy in disambiguating sentences that contain a French partial cognate. The best classifier among several was Naïve Bayes.

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

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

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

Its results are much higher than the baseline of 59% when choosing the most frequent class.

For the semi-supervised methods, we use unlabeled data from the LeMonde and the Hansard corpus and we obtained an increase in accuracy of 2%. Statistically significant improvements were obtained when we performed experiments combining corpora from different domains. More detailed results for all these experiments can be found in (Frunza & Inkpen, 2006).

# 5 A tool for cross-language pair annotations

In this section, we describe our tool called Cross-Language Pair Annotator (CLPA) that is capable of automatically annotating cognates and false friends in French texts. The tool uses the Unstructured Information Management Architecture (UIMA)[7] Software Development Kit (SDK) from IBM and the Baseline Information Extraction (BaLIE)[8], an open source Java project capable of extracting information from raw texts. 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). Our tool is a practical follow up to the research that we did on cognates and false friends between French and English and it has the goal to help second-language learners of French. In its first version, the CLPA 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 Section 3.

UIMA is an open platform for creating, integrating, and deploying unstructured information management solutions from a combination of semantic analysis and search components. It 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. We use BaLIE for the tokenization and part-of-speech tagging capabilities.

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 to not 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.

The user can click on one of the text annotations in the GUI 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.) Snapshots of the tool along with the tool itself free to download can be found here[9].

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

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

[9] www.site.uottawa.ca/∼ofrunza/Pages/CLPA.html

# 6 Conclusions and future work

In Section 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 we 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 and 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.

For the task of partial cognate disambiguation (Section 4) we proposed a method that has a pure ML supervised approach and a semi-supervised method that contains two algorithms : Monolingual 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.

In the Section 5 we presented a CALL tool, CLPA, which is able to annotate cognates and false friends in a French text. 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.

In 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 precision 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 classes. For the disambiguation task, we want to look at different data representations, use lemmatization and POS tagging, and apply our method to new pairs of languages (all we need is a parallel corpus, and monolingual corpora). We also 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.

The overall contribution of this paper is the new methods that we proposed, experimented and evaluated, and the new directions that we followed for cognate, false friend, and partial cognate words between French and English.

# References

BARKER G. & SUTCLIFFE R. F. E. (2000). *An Experiment in the Semi-Automatic Identification of False-Cognates between English and Polish*. Rapport interne, Department of Languages and Cultural Studies, University of Limerick, Ireland.

BREW C. & MCKELVIE D. (1996). Word-pair extraction for lexicography. In *Proceedings of 2nd International Conf. on New Methods in Language Processing*, p. 45–55, Ankara, Turkey.

CARROLL S. (1992). *On Cognates*. Rapport interne, Second Language Research.

DIAB M. & RESNIK P. (2002). An unsupervised method for word sense tagging using parallel corpora. *In Proceedings of Association for Computational Linguistics (ACL '02)*, p. 255–262.

FRUNZA O. & 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*, p. 433–440, Sydney, Australia.

GASS S. (1987). The use and acquisition of the second language lexicon. *Studies in Second Language Acquisition 9(2)*, **9**, 128–262.

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*, p. 1–19.

HEUVEN W. V., DIJKSTRA A. & GRAINGER J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, **39**, 458–483.

INKPEN D., FRUNZA O. & KONDRAK G. (2005). Automatic identification of Cognates and False Friends in French and English. In *RANLP-2005*, p. 251–257, Bulgaria.

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*, p. 103–110.

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*, p. 44–59.

LEBLANC R., COMPAIN J., DUQUETTE L. & SÉGUIN H. (1989). *L'enseignement des langues secondes aux adultes : recherches et pratiques*. Les Presses de l'Université d'Ottawa.

LEBLANC R. & SÉGUIN H. (1996). Les congénères homographes et parographes anglais-français. In *Twenty-Five Years of Second Language Teaching at the Univ. of Ottawa*, p. 69–91.

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

MARCU D., KONDRAK G. & 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)*, p. 46–48.

MELAMED I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, **25**, 107–130.

SIMARD M., FOSTER G. F. & 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*, p. 67–81, Montreal, Canada.

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)*, p. 189–196.