Automatic Identification of Cognates and False Friends in French and English

Diana Inkpen and **Oana Frunza**
School of Information Technology and Eng.
University of Ottawa
Ottawa, ON, K1N 6N5, Canada
{diana,ofrunza}@site.uottawa.ca

Grzegorz Kondrak
Department of Computing Science
University of Alberta
Edmonton, AB, T6G 2E8, Canada
kondrak@cs.ualberta.ca

Abstract

Cognates are words in different languages that have similar spelling and meaning. They can help a second-language learner on the tasks of vocabulary expansion and reading comprehension. The learner also needs to pay attention to pairs of words that appear similar but are in fact *false friends*: they have different meaning in some contexts or in all contexts. In this paper we propose a method to automatically classify a pair of words as cognates or false friends. We focus on French and English, but the methods are applicable to other language pairs. We use 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.

Keywords: similarity measures, machine learning, cognates and false friends, second-language learning, machine translation.

1 Introduction

When learning a second language, a student can benefit from knowledge in his/her first language (Gass 87) (Ringbom 87). Cognates – words that have similar spelling and meaning in the two languages – help with vocabulary expansion and with reading comprehension. On the other hand, there are also pairs of words that appear similar, but have different meaning in some or all contexts: false friends. Dictionaries often include information about false friends, and there are dictionaries devoted exclusively to them – for example (Prado 96).

Cognates have also been employed in natural language processing. The applications include sentence alignment (Simard *et al.* 92; Melamed 99), inducing translation lexicons (Mann & Yarowsky 01; Tufis 02), improving statistical machine translation models (Al-Onaizan *et al.* 99), and identification of confusable drug names (Kondrak & Dorr 04). All those applications depend on an effective method of identifying cognates by computing a numerical score that reflects the likelihood that the two words are cognates.

In this paper we propose a method to automatically classify pairs of words as cognates or false friends. Our approach to the identification of cognates is based on several 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 00). 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.

Although French and English belong to different branches of the Indoeuropean family of languages, they share an extraordinary high number of cognates. The cognates derive from several distinct sources. The majority are words of Latin and Greek origin that permeate the vocabularies of European languages, e. g., *éducation - education* and *théorie - theory*. A small number of very old, "genetic" cognates go back all the way to Proto-Indoeuropean, e. g., *mère - mother* and *pied - foot*. 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.

While our focus is on French and English, the methods that we describe are also applicable to other language pairs. Nowadays, new terms related to modern technology are often adopted in 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 a French word *sprinter*, which has the same meaning.

2 Related Work

Previous work on automatic cognate identification is mostly related to bilingual corpora and translation lexicons. Simard et al. (Simard *et al.* 92) 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 and McKelvie (Brew & McKelvie 96) extract French-English cognates and false friends from bitexts using a variety of orthographic similarity measures. Mann and Yarowsky (Mann & Yarowsky 01) automatically induce translation lexicons on the basis of cognate pairs. They found that edit distance with variable weights outperformed both hidden Markov models and stochastic transducers. Kondrak (Kondrak 04) identifies genetic cognates directly in the vocabularies of related languages by combining the phonetic similarity of lexemes with the semantic similarity of glosses. Kondrak & Dorr 04) report that a simple average of several orthographic similarity measures outperforms all individual measures on the task of the identification of drug names.

For French and English, substantial work on cognate detection was done manually. LeBlanc and Seguin (LeBlanc & Séguin 96) collected 23,160 French-English cognate pairs from two general-purpose dictionaries: Robert-Collins (Robert-Collins 87) and Larousse-Saturne (Dubois 81). 6,447 of the 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.

The use of cognates in second language teaching was shown to accelerate vocabulary acquisition and to facilitate reading comprehension tasks (LeBlanc *et al.* 89). Morphological rules for conversion from English to French were also proved to help. Tréville (Tréville 90) proposed 25 such rules. An example is: *cal* → *que* in pairs such as *logical* - *logique*, *political* - *politique*.

3 Background

3.1 Definitions

We adopt the following definitions. The definitions are language-independent, but the examples are pairs of French and English words, respectively.

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*, *recognition* - *reconnaissance*.

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*, *blesser* "to injure" - *bless*.

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".

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 at some point of time, such as *concierge*.

Unrelated pairs are words that exhibit no orthographic similarity. They can be translations of each other, e. g., *glace* - *ice*, but not necessarily, e. g., *glace* - *chair*.

3.2 Orthographic Similarity Measures

Many different orthographic similarity measures have been proposed. Their goal is to quantify human perception of similarity, which is often quite subjective. In this section, we briefly describe the measures that we use as features for the cognate classification task.

- IDENT is a baseline measure that 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 com-

[1] The PREFIX measure can be seen as a generalization of Simard et al. (Simard *et al.* 92) approach.

mon prefix for *factory* and *fabrique* has length 2 (the first two letters) which, divided by the length of 8, yields 0.25.

- DICE (Adamson & Boreham 74) is calculated by dividing twice the number of shared letter bigrams by the total number of bigrams in both words:

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

  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 96) is also defined in the same way as DICE, but employs "extended bigrams", which are trigrams without the middle letter.

- XXDICE (Brew & McKelvie 96) 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 99) 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 74) 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 80) is an approximation to phonetic name matching. SOUNDEX transforms all but the first letter to numeric codes and after removing zeroes 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 04) 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.

4 The Data

The training dataset that we used consists of 1454 pairs of French and English words (see Table 1). 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 98). (We manually identified 258 Cognate pairs among the aligned word pairs.)

3. A set of exercises for Anglophone learners of French (Tréville 90) (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 1), 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.)

[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.

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

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

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

Table 1: The composition of data sets. The numbers in brackets are counts of word pairs that are identical (ignoring accents).

2. The above-mentioned on-line list of "French-English False Cognates" (94 additional False-Friends).

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 favour 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 Section 5). 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.

5 Evaluation

We present evaluation experiments using the two datasets described in Section 4: a training/development set, and a test set. We classify the word pairs on the basis of similarity into two classes: Cognates/False-Friends and Unrelated. Cognates are distinguished from False-Friends by virtue of being mutual translations. 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.

5.1 Results on the Training Data Set

Table 2 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. For the IDENT measure, the threshold was set to 1 (identical spelling ignoring accents). 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. The values of the thresholds obtained in this way are also included in Table 2.

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 2: 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.

The training dataset 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), Naive Bayes, Decision Trees, Instance-based Learning (IBK), Ada Boost, Multi-layered Perceptron, and a light version of Support Vector Machine.

The Decision Tree classifier has the advantage of being relatively transparent. 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 suggest overtraining. 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

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

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

counter-intuitive decisions, at $CF = 0.16$ (Figure 1). Our hypothesis was that the latter tree would perform better on a test set.

The results presented in the rightmost column of Table 3 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 the possible splits). We also report, in the middle column, the results of testing on the training set: they are artificially high, due to overtraining. The baseline algorithm in the Table 3 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 2.

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. Most of the false negatives were genetic cognates that have different orthographic form due to changes of language over time. False positives, on the other hand, were mostly caused by accidental similarity. 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.

5.2 Results on the Test Set

The rightmost column of Table 4 shows the results obtained on the test set described in Section 4. 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, which may be caused by the absence of genetic cognates in the test set. Surprisingly, only the Naive Bayes classifier outperforms the simple average of orthographic measures. The pruned Decision Tree shown in Figure 1 achieves higher accuracy than the overtrained Decision Tree, 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 96). 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.

5.3 Results on the Genetic Cognates Dataset

Greenberg (Greenberg 87) gives a list of "most of the cognates from French and English". The list serves as an illustration how difficult it would to demonstrate that French and English are genetically related by examining only the genetic cognates between those two languages. We transcribed the list of 82 cognate pairs from IPA to standard orthography. We augmented the list with 14 pairs from the Comparative Indoeuropean Data Corpus[5]and 17 pairs that we identified ourselves. The final list contains 113 true genetic cognates that go back to Proto-Indoeuropean[6].

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

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

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 4: 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.

We decided to also test the classifier trained in Section 5.1 on this genetic cognates set. The results are shown in the middle column of Table 4. 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. The misclassifications are due to radical changes in spelling, such as: *frère* - *brother*, *chaud* - *hot*, *chien* - *hound*, *faire* - *do*, *fendre* - *bite*. 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.

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 exist (Tiedemann 99; Kondrak 04) 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, only 4 out of 603 French-English cognate pairs were genetic cognates.

5.4 Three-way Classification

We also experimented with a three-way classification into Cognates, False-Friends and Unrelated. We used an extra feature in our machine learning experiments, which is set to 1 if the two words are translations of each other, and to 0 otherwise. Since all the examples of pairs of 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 then their pairs. We manually checked to insure that all these generated pairs were not translations of each other by chance.

As expected, this experiment achieved slightly lower results than the ones from Table 2 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.

6 Conclusion and Future Work

We presented several methods 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. 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 automatically identify partial cognates, which have senses that behave as cognates and senses that behave as false friends. Word sense disambiguation would make it possible to place the partial cognates in their context. We plan to use translation probabilities from a word-aligned parallel corpus. Another direction of future work is to produce complete lists of cognates and false friends, given two vocabulary lists for the two languages. We would also like to apply the methods presented in this paper to other pairs of languages.

Acknowledgments

We thank to Lise Duquette for giving us the idea to work on this project. Our research is supported by the Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada (SHERC), the University of Ottawa, and the University of Alberta.

References

(Adamson & Boreham 74) George W. Adamson and Jillian Boreham. 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, 1974.

(Al-Onaizan *et al.* 99) Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. Statistical machine translation. Technical report, Johns Hopkins University, 1999.

(Brew & McKelvie 96) Chris Brew and David McKelvie. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55, Ankara, Turkey, 1996.

(Dubois 81) Marguerite M. Dubois. *Saturn Larousse French-English, English-French Dictionary: Dictionnaire Larousse Saturne Français-Anglais-Français.* French & European Publications, 1981.

(Gass 87) S.M. Gass, editor. *The use and acquisition of the second language lexicon (Special issue).* Studies in Second Language Acquisition 9(2), 1987.

(Greenberg 87) Joseph H. Greenberg. *Language in the Americas.* Stanford University Press, Stanford, CA, USA, 1987.

(Hall & Dowling 80) Patrick A. V. Hall and Geoff R. Dowling. Approximate string matching. *Computing Surveys*, 12(4):381–402, 1980.

(Kondrak & Dorr 04) Grzegorz Kondrak and Bonnie Dorr. Identification of confusable drug names: A new approach and evaluation methodology. In *Proceedings of COLING 2004: 20th International Conference on Computational Linguistics*, pages 952–958, 2004.

(Kondrak 04) Grzegorz Kondrak. Combining evidence in cognate identification. In *Proceedings of Canadian AI 2004: 17th Conference of the Canadian Society for Computational Studies of Intelligence*, pages 44–59, 2004.

(LeBlanc & Séguin 96) Raymond LeBlanc and Hubert Séguin. Les congénères homographes et parographes anglais-français. In *Twenty-Five Years of Second Language Teaching at the University of Ottawa*, pages 69–91. University of Ottawa Press, 1996.

(LeBlanc *et al.* 89) Raymond LeBlanc, Jean Compain, Lise Duquette, and Hubert Séguin, editors. *L'enseignement des langues secondes aux adultes: recherches et pratiques.* Les Presses de l'Université d'Ottawa, 1989.

(Mann & Yarowsky 01) Gideon S. Mann and David Yarowsky. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 151–158, 2001.

(Melamed 98) I. Dan Melamed. Manual annotation of translational equivalence: The Blinker project. Technical Report IRCS #98-07, University of Pennsylvania, 1998.

(Melamed 99) I. Dan Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130, 1999.

(Prado 96) Marcial Prado. *NTC's Dictionary of Spanish False Cognates.* McGraw-Hill, 1996.

(Ringbom 87) H. Ringbom. *The Role of the First Language in Foreign Language Learning.* Multilingual Matters Ltd., Clevedon, England, 1987.

(Robert-Collins 87) Robert-Collins. *Robert-Collins French-English English-French Dictionary.* Collins, London, 1987.

(Simard *et al.* 92) Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montreal, Canada, 1992.

(Tiedemann 99) Jörg Tiedemann. 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*, College Park, Maryland, USA, 1999.

(Tréville 90) Marie-Claude Tréville. *Rôle des congénères interlinguaux dans le dévelopment du vocabulair réceptif.* Unpublished PhD thesis, Université de Montreal, 1990.

(Tufis 02) Dan Tufis. A cheap and fast way to build useful translation lexicons. In *Proceedings of COLING 2002: 19th International Conference on Computational Linguistics*, pages 1030–1036, 2002.

(Wagner & Fischer 74) Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.

(Witten & Frank 00) Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools with Java implementations.* Morgan Kaufmann, San Francisco, USA, 2000.