Integrating Language Knowledge Resources to Extend the English Inclusion Classifier to a New Language

Integrating Language Knowledge Resources to Extend the English Inclusion Classifier to a New Language

Beatrice Alex

School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW, UK
balex@inf.ed.ac.uk

Abstract

This paper presents an unsupervised system that classifies English inclusions in written text. It will demonstrate that extending this English inclusion classifier, which was originally designed for German, requires minimal time and effort to adapt to a new language, in this case French. The analysis of several evaluation experiments carried out on French and German data shows that the system performs well for both languages and on unseen data from the same domain and language.

1. Introduction

With increasing globalisation and a rapidly expanding digital society, the influence of English as an international language is growing constantly. As the influx of English vocabulary into other languages is becoming increasingly prevalent, natural language processing systems must be able to deal with this language mixing phenomenon. This paper demonstrates that extending an existing unsupervised system, which detects English inclusions in German text, to a new language requires little time and effort.

The existing German system yields considerably high precision, recall and F-scores for identifying English inclusions (Alex, 2005). In an attempt to carry out similar experiments for a new base language and ascertain the performance for a different language scenario, the system was updated to process French input text as well. The extension of the system, which is described in this paper, facilitates token level identification of English inclusions in either French or German text. By means of English inclusion identification experiments on specially prepared French and German corpora, I illustrate the appeal of this system derived from its ease of portability to new languages.

Section 2 briefly examines the issue of anglicisms appearing in French and provides an overview of research efforts in the field of automatic language identification. An indication as to the time necessary to convert each component of the existing system is given in Section 3. The French development and test datasets created for evaluating the English inclusion classifier are described in Section 4. The individual components of the French system are presented in Section 5. Section 6 provides a detailed overview of a series of evaluation experiments and discusses their results.

2. English Inclusions in French

The occurrence of anglicisms and pseudo-anglicisms in French is not a new phenomenon. One well known anglicism in French is the word *weekend* which was borrowed from English at the beginning of the 20th Century. However, with growing internationalisation and the development of the internet, the influx of English expressions into the French language has taken on a different dimension in recent years. Despite serious efforts from the French government in the 1990s, which tried to restrict this trend by introducing new French words to replace already prevalent anglicisms, the French media does not often object to the use of anglicisms. This is particularly the case when a French term has not yet been invented or when a specific English term is more modern and therefore more popular than its French equivalent (Rollason, 2005; Nicholls, 2003). The following sentence, taken from an online article published by ZDNet France (Dumont, 2005), contains some examples of English inclusions in French.

(1) Tous les **e-mails** entrants, qui ne seront pas dûment authentifiés par **Sender ID**, seront considérés automatiquement comme du **spam**.
All incoming emails which will not be duly authenticated by Sender ID, will be automatically considered as spam.

As such mixed-lingual documents are becoming more frequent, particularly on the Web, it is desirable to identify individual language portions for appropriate further text processing. This additional language knowledge could prove beneficial, for example, for rendering the correct pronunciation in text-to-speech (TTS) synthesis.

The majority of existing state-of-the-art language identification systems rely on word-level information such as diacritics and special characters (Newman, 1987; Beesley, 1988), common short words (Johnson, 1993), characteristic letter sequences (Dunning, 1994) or character n-gram statistics (Cavnar and Trenkle, 1994). A comparison of different techniques (Grefenstette, 1995) demonstrates that there is no one best language identification method and results largely depend on the type and number of languages involved as well as the number of input words. This means that language identification accuracy increases with the length of the test sentence and is not satisfactory for individual words. Most systems are successful in identifying the base language of a document. They are however not designed to deal with mixed-lingual text to identify the origin of foreign words within a given sentence. Pfister

Data	Development Set					Test Set				
Domain: IT	Tokens	%	Types	%	TTR	Tokens	%	Types	%	TTR
French										
Total Tokens	16188		3233		0.20	16125		3437		0.21
English Tokens	986	6.1	339	10.5	0.34	1089	6.8	367	10.7	0.34
German										
Total Tokens	15919		4152		0.26	16219		4404		0.27
English Tokens	963	6.0	283	6.8	0.29	1034	6.4	258	5.9	0.25

Table 1: Corpus statistics including type-token-ratios (TTRs) of the French development and test sets compared to the German data.

and Romsdorfer (2003) developed a morpho-syntactic analyser to identify foreign inclusions in German text. Their analyser functions by means of language-specific lexicons, word and sentence grammars as well as relevant inclusion grammars. As the system is not evaluated, it is unclear how well the analyser performs on real mixed-lingual data. Although Pfister and Romsdorfer have taken an interesting approach to dealing with mixed-lingual documents, a system working with large grammars is costly given that linguistic experts have to write the necessary grammars for each language scenario. This paper presents an unsupervised English inclusion classifier and demonstrates its ease of portability to a new language.

3. Time Spent on System Extension

The initial English inclusion classifier was designed specifically for German text. The two main aims of extending the system to a new language are: (1) to prove that its underlying concept of English inclusion identification is not specific to one language scenario and (2) to determine the time to do so. It took approximately one person week to convert the core system to French, another Indo-European language with a Latin alphabet. This involved implementing a French tokeniser (1.5 days), incorporating the French TreeTagger (1 day), extending the lexicon module (1.5 days) and converting the search engine module to French (0.2 days).

A subsequent error analysis of the output was performed in order to generate post-processing rules. As the process of analysing errors is essentially difficult to time, a limit of one week was set for this task. This strategy proved beneficial in terms of fine-tuning the system to improve its overall performance (see Section 6). The actual evaluation of the system, requires French data that is manually annotated with English inclusions. Three working days were spent on collecting and annotating a French development and test set of approximately 16,000 tokens each which are described in more detail in the Section 4.

A further issue that must be considered for extending the system to a new language is the time required for identifying necessary resources and tools available and familiarising oneself with them. This is evidently dependent on the chosen language. In the case of French, I researched for approximately two working days and identified the POS tagger TreeTagger and the lexicon Lexique as appropriate resources. If a POS tagger and a lexicon are not available for a particular language scenario, more time and effort would need to be invested to create such resources.

As the English inclusion classifier is essentially unsupervised, i.e. it does not rely on manually annotated training data, it can be easily run on new data without any further cost. The search engine module then deals with any new vocabulary entering a language over time. This represents a serious advantage over a supervised system that relies on annotated training data. The latter is built on a snapshot of a particular language in use and would need to be adjusted by retraining on additional annotated data as this language evolves over time. It would therefore require much more time and effort to keep up-to-date.

4. French Development and Test Data

In order to evaluate the system performance of classifying English inclusions in French text, I collected a random selection of online articles published by ZDNet France[1], an online magazine reporting on the latest news in the high tech sector. These articles were published in the period between October 2003 and September 2005 in the domain of internet and telecoms (IT). All French articles were manually annotated for English inclusions using an annotation tool based on NXT (Carletta et al., 2003). As with the experiments on German, the data is split into a development set and a test set of approximately 16,000 tokens each.

Table 1 lists the total number of tokens and types plus the number of English inclusions both in the French and German development and test sets. The French datasets have similar characteristics, particularly regarding their type-token-ratios (TTRs) for each entire set (0.20 versus 0.21) and for the English inclusions alone (0.34 each). The French test set contains slightly more English inclusions (+0.7%) than the development set. Comparing these figures with those of the previously annotated German IT datasets shows that the proportion of English tokens in this domain is extremely similarly at approximately 6%. However, the percentage of English types varies to some extent both for the development and test sets. They only amount to 6.8% and 5.9% in the German data, compared to 10.5% and 10.7% in the French data. Moreover, the TTRs of English inclusions are 0.5 and 0.9 points higher in the French datasets, signalling that they are less repetitive than those contained in the German articles. However, overall TTRs are 0.6 points lower for French than for German which means that the remaining vocabulary in the French articles is somewhat less heterogeneous than in the German data.

[1] http://www.zdnet.fr/

5. System Module Conversion to French

The overall system architecture of the English inclusion classifier consists of several pre-processing steps, followed by a lexicon module, a search engine module and a post-processing module. Converting the search engine module to a new language required little computational cost and time. Conversely, the pre- and post-processing as well as the lexicon modules necessitated some language knowledge resources or tools and therefore demanded more time and effort to be customised for French. The core system was adapted in approximately one person week in total (see Section 3).

5.1. Pre-processing Module

Firstly, I developed a French tokeniser and implemented a French part-of-speech (POS) tagger into the system. The French tokeniser consists of two rule-based tokenisation grammars. It not only identifies tokens surrounded by white space and punctuation but also resolves typical abbreviations, numerals and URLs. Both grammars are applied by means of improved upgrades of the XML tools described in Thompson et al. (1997) and Grover et al. (2000). These tools process an XML input stream and rewrite it on the basis of the rules provided.[2] The French TreeTagger is used for POS tagging. It is freely available for research purposes and is also trained for a number of other languages, including German and English (Stein and Schmidt, 1995). The TreeTagger functions on the basis of binary decision trees trained on a French corpus of 35,448 words and yields a tagging accuracy of over 94% on an evaluation dataset comprising of 10,000 word forms.

5.2. Lexicon Module

The lexicon module performs an initial language classification run based on a case-insensitive double lookup procedure (Alex, 2005) using two lexicons: one for the base-language and one for the language of the inclusions. For French, the system queries Lexique, a lexical database which contains 128,919 word forms representing 54,158 lemmas (New et al., 2004). It is derived from 487 texts (31 million words) published between 1950 and 2000. In order to detect common English inclusions, the system searches the English database of CELEX[3] holding 51,728 lemmas and their 365,530 corresponding word forms. The lexicon module was adapted to French by exploiting distinctive characteristics of French orthography. For example, words containing diacritic characters typical for French are automatically excluded from being considered as English inclusions.

5.3. Search Engine Module

Tokens which are not clearly identified by the lexicon module as either French or English are further processed by a back-off search engine module. This module relies on the number of hits returned by the search engine as an indication of the actual frequency of the query in the documents accessible by the search engine. This assumption is justified given that Zhu and Rosenfeld (2001) show that n-gram page counts and phrase counts obtained from a search engine are largely log-linear and therefore highly correlated. Moreover, Keller and Lapata (2003) demonstrate that bigram search engine counts are highly correlated to corpus counts from the British National Corpus.

The search engine module performs language classification based on $rf_{C_{web}(L)}(t)$, the maximum normalised score of the number of hits returned for two searches per token, one for each language L (Alex, 2005). As shown in the following equation, this score is determined by weighting the number of hits, i.e. the "absolute frequency" $f(t, C_{web}(L))$, by the size of the accessible Web corpus for that language, $N_{Cweb}(L)$. The notation t designates token and C refers to corpus.

$$rf_{C_{web}(L)}(t) = \frac{f(t, C_{web}(L))}{N_{C_{web}}(L)} \quad (1)$$

The size of the Web corpus for each language is estimated following a method motivated by Grefenstette and Nioche (2000). The relative frequencies of a series of common words within a standard corpus in a language, $rf_{C_{std}}(w_{1\ldots n})$, are used to make a series of n predictions on the overall size of the corpus of that language indexed by the search engine. This is done by dividing the number of hits of each word returned by the search engine by the relative frequency of the same word in the standard corpus. The total number of words in the particular language accessible through the search engine is then determined by taking the average of each individual word prediction:

$$C_{web}(L) = \frac{1}{n} \sum_{k=1}^{n} \frac{f(w_k, C_{web})}{rf_{C_{std}(L)}(w_k)} \quad (2)$$

Extending the search engine module to French merely involved adjusting the language preferences in the search engine API and incorporating the relative frequencies of representative French tokens in a standard French corpus for estimating the size of the French Web corpus. The search engine Yahoo was used instead of Google as it allows a larger number of automatic queries per day.

5.4. Post-processing Module

The final system component is a post-processing module that resolves several language classification ambiguities and classifies some single-character tokens. I invested some time in analysing the core system output of the French development data in order to generate these post-processing rules. The individual contribution of each of the following rules on the system performance on the French development data is discussed in Section 6.

The most general rules are designed to disambiguate one-character tokens and interlingual homographs. They are flagged as English if they are followed by a hyphen and an English token (*e-mail* or *joint-venture*). Furthermore, typical English function words are flagged as English, including prepositions, pronouns, conjunctions and articles, as these belong to a closed class and are easily recognisable. This also avoids having to extend the core system to these

[2] They will soon be available under GPL as LT-XML2 and LT-TTT2 at: `http://www.ltg.ed.ac.uk`

[3] `http://www.ru.nl/celex`

	Development Set				Test Set			
	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
French: Baseline (I) versus Core System (II) and Full System (III)								
I	93.91%	-	-	-	93.25%	-	-	-
II	96.74%	82.91%	62.98%	71.59	96.59%	82.07%	69.33%	75.16
III	98.44%	91.50%	84.08%	87.63	97.55%	87.55%	83.93%	85.70
German: Baseline (I) versus Core System (II) and Full System (III)								
I	93.95%	-	-	-	93.62%	-	-	-
II	97.47%	90.60%	66.32%	76.58	97.15%	87.28%	67.70%	76.25
III	98.03%	93.31%	72.68%	81.71	97.74%	92.12%	73.30%	81.64

Table 2: Evaluation of the best French and German system on the development and unseen test data versus the baseline.

categories which not only prevents some output errors but also improves the performance of the POS tagger as that is often unable to process foreign function words correctly. In the post-processing step, their POS tags are therefore corrected. Any words in the closed class of English function words that are ambiguous with respect to their language such as *an* (in French *year*) or *but* (in French *goal*) are only flagged as English inclusions if their surrounding context is already classified as English by the system. Similarly, the possessive marker *'s* if preceded by an English token is flagged as English. Moreover, I devised several rules in order to automatically deal with names of currencies (e.g. *Euro*) and units of measurement (e.g. *Km*). Such instances are prevented from being identified as English inclusions.

As the system classifies each token individually, a further post-processing step was implemented to relate language information between abbreviations or acronyms and their definitions. Firstly, they are identified by means of an abbreviation extraction algorithm (Schwartz and Hearst, 2003). Subsequently, post-processing is applied in order to guarantee that each pair and earlier plus later mentions of either the definition or the abbreviation/acronym are assigned the same language tag within a document.

When analysing the errors which the system made in the development data, it was also observed that foreign person names (e.g. *Graham Cluley*) are frequently identified as English inclusions. At this point, the system is merely evaluated on identifying actual inclusions. These are defined as any English words in the text except for person and location names. Therefore, the evaluation data does not contain annotations of foreign, or specifically English person names in the gold standard. In order to improve the performance of recognising real English inclusions, further post-processing rules were implemented to distinguish between the latter and English person names that are incorrectly classified as English inclusions. The aim is to increase precision against the development set without reducing recall. Based on a careful error analysis on the development data, I generated patterns that signal person names in French text, e.g. "Mme X" or "X, directeur", and excluded these instances from the English inclusions. It should also be noted that for a potential task-based evaluation of the system output, for example via TTS synthesis, the language information provided by the system for person names could prove beneficial for generating correct pronunciations.

After implementing the post-processing rules described above, the balanced F-score amounts to 87.63 points (91.50% precision and 84.08% recall). This represents an overall performance improvement of 16.04 points in F-score, 8.59% in precision and 21.10% in recall over the core system (see Table 2). The results show that post-processing is mainly aimed at identifying false negatives, i.e. English inclusions which are missed by the core system. The precision of the core system is already relatively high.

6. French and German Systems Evaluation

This section provides information on the performance of the French system, compared to the German one when evaluating on data from a similar domain.[4] Furthermore, it presents some additional results illustrating the improvement gained from the various post-processing rules.

Table 2 shows the results of the core and full French and German systems on the development and test data versus the baseline. The baseline accuracies are determined assuming that the system found none of the English inclusions in the data and believes that it is all written in either French or German, respectively. As precision, recall and F-score are determined in relation to the English tokens in the gold standard, they are essentially zero for the baseline. For this reason, we only report the accuracy baseline scores.

The German core system (without post-processing) performs similarly on both the development set and the test set at approximately 76 points in F-score. The French core system actually performs almost 4 points better in F-score on the test set (75.16) than on the development set (71.59). This means that the core systems do not overfit on new data in the same domain and language. Comparing the results of the core systems across languages shows that they perform relatively similarly in F-score but vary slightly in terms of precision and recall. These differences can be attributed to some system internal differences resulting from language-specific characteristics or pre-processing tools. For example, 13.7% of all tokens in the French development set contain diacritics compared to only 7.8% of all tokens in the German development set. As information about diacritics

[4]The results for the German system reported in Alex (2005) have been recently improved as a result of a series of parameter tuning experiments. The current German system incorporates output of the TnT POS tagger trained on the TIGER Treebank (Brants et al., 2002) and uses Yahoo in the search engine module. Moreover, the post-processing of the system was enhanced.

is exploited in the lexicon module for both languages, the French system is expected to perform better at that stage.
A further core system difference lies in the POS tag sets for the two languages. The German system makes use of TnT (Brants, 2000) trained on the TIGER Treebank (Brants et al., 2002) to assign POS tags. Earlier experiments showed that this POS tagger yields best results for a set of German datasets in different domains. TnT assigns STTS tags to German text (Schiller et al., 1995). The English inclusion classifier is set up to process any token in the German data with the tag: NN (common noun), NE (proper noun), ADJA or ADJD (attributive and adverbial or predicatively used adjectives) as well as FM (foreign material). The French data, on the other hand, is tagged with the TreeTagger whose POS tag set differs to STTS. Although it also differentiates between common nouns (NOM) and proper names (NAM), it only has one tag for adjectives (ADJ). Moreover, the French tag set contains an additional abbreviation tag (ABR). It does not, however, contain a separate tag for foreign material. Despite the fact that TnT is not very accurate in identifying foreign material in the German data, I suspect that this additional information has a positive effect on the overall performance of the German system.
The full system scores show that the post-processing improves the overall system performance considerably for both languages. The improvements are relatively similar on both the development set and the test set for each language. The full system scores for German are almost identical at approximately 81.7 points in F-score. The full French system performs only 1.93 points lower in F-score (85.7) on the test set compared to the development set (87.63).
Table 3 presents lesion studies showing the individual contribution of post-processing rules to the overall performance of the full French system on the development data. The results show that the biggest improvement is due to the post-processing of single-character tokens which are not classified by the core system. The second largest increase in F-score is achieved by the post-processing rules dealing with ambiguous words, i.e. those that are classified as either French or English by the core system. Identifying the language of such tokens based on the language of their surrounding context helps to improve the overall performance considerably. Comparing Tables 2 and 3 also shows that most post-processing rules are designed to improve recall. The only post-processing rule implemented to improve precision without deteriorating recall is that for person names. In the final run of the full French system on the test data, the post-processing module results in a large performance increase of 10.66 points in F-score. Therefore, it can be concluded that the post-processing is designed well enough to apply to new data in the same domain and language.

Post-Processing	Precision	Recall	F-score	Δ F
Single letters	90.39%	72.52%	80.47	-7.16
Ambiguous words	91.60%	74.14%	81.95	-5.68
Person names	86.44%	84.08%	85.24	-2.39
Function words	91.26%	81.54%	86.13	-1.50
Currencies etc.	90.98%	81.85%	86.17	-1.46
Abbreviations	90.77%	83.77%	87.13	-0.50

Table 3: Evaluation of the post-processing module with one rule removed at the time on the French development data

Overall, the full French system performs slightly better than the German one. Table 2 illustrates that this difference is mainly due to the larger gap between recall and precision for the full German system. Even though the full German system performs better in precision than the French one, its recall is much lower, causing the overall F-score to drop. This discrepancy is due to language-specific post-processing differences as post-processing rules are generated on the basis of error analysis on the development data. However, comparing the results of the two systems is not entirely straightforward because they are not completely identical in parts of their components. Despite these differences, the fact that the both systems yield considerably high F-scores demonstrates that the underlying concept of identifying English inclusions in text can be applied to different language scenarios, particularly those with Latin alphabets.

7. Discussion and Conclusion

The English inclusion classifier was successfully converted to a new language: French. The extended system is able to process either German or French text for identifying English inclusions. The system is a pipeline made up of several modules, including pre-processing, a lexicon, a search-engine and a post-processing module. The extension of the core system was carried out in only one person week and resulted in a system performance of 71.59 in F-score on the French development data. A further week was spent on implementing the post-processing module which boosted the F-score to 87.63. A third week was required to select external language resources plus collect and annotate French evaluation data in the domain of internet and telecoms. The performance drop between the development set and the unseen test sets is relatively small which means that the system does not seriously over-fit for this domain and will result in an equally high performance on new data. This paper demonstrates that the English inclusion classifier is easy to extend to a new language in a relative short period of time and without having to rely on expensive manually annotated training data. Therefore non-recoverable engineering costs for extending and updating the classifier are kept to a minimum. Not only can the system be easily applied to new data from the same domain and language without a serious performance decrease, it can also be extended to a new language and produce similarly high scores.
The English inclusion classifier described in this paper is designed particularly for languages composed of tokens separated by white space and punctuation and with Latin-based scripts. A system that tracks English inclusions occurring in languages with non-Latin based scripts necessitates a different setup as the inclusions tend to be transcribed in the alphabet of the base language of the text (e.g. in Russian). The English inclusion classifier is also not designed to deal with languages where words are not separated by white space. An entirely different approach would be required for such a scenario.
In future work, the aim is to extend this system to identify English inclusions with base language inflections (e.g. *Scannern*, the word *scanner* in the German dative plu-

ral case) and English inclusions occurring within mixed-lingual compounds (e.g. *Scannerknopf*, in English *scanner button*). A further goal is to test the hypothesis that the additional language information provided by the English inclusion classifier can improve synthesis quality produced by a polyglot TTS system.

8. Acknowledgements

I would like to thank Claire Grover and Frank Keller for their comments. This research is supported by grants from the Scottish Enterprise Edinburgh-Stanford Link (R36759) and ESRC as well as the University of Edinburgh.

9. References

Beatrice Alex. 2005. An unsupervised system for identifying English inclusions in German text. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), Student Research Workshop*, pages 133–138, Ann Arbor, Michigan.

Kenneth R. Beesley. 1988. Language Identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, pages 47–54, Medford, New Jersey.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*, pages 24–41, Sozopol, Bulgaria.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, pages 224–231, Seattle, Washington.

Jean Carletta, Stefan Evert, Ulrich Heid, Jonathan Kilgour, Judy Robertson, and Holgar Voormann. 2003. The NITE XML toolkit: flexible annotation for multimodal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3):353–363.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)*, pages 161–175, Las Vegas, Nevada.

Estelle Dumont. 2005. Anti-spam: Microsoft choisit de passer en force pour imposer sender id. *ZDNet France*, 23rd of June. `http://zdnet.fr/actualites/internet/0,39020774,39235930,00.htm`.

Ted Dunning. 1994. Statistical identification of language. Technical report, Computing Research Laboratory, New Mexico State University.

Gregory Grefenstette and Julien Nioche. 2000. Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO (Recherche d'Informations Assistée par Ordinateur) 2000*, pages 237–246, Paris, France.

Gregory Grefenstette. 1995. Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT 1995)*, pages 263–268, Rome, Italy.

Claire Grover, Colin Matheson, Andrei Mikheev, and Moens Marc. 2000. LT TTT - a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1147–1154, Athens, Greece.

Stephen Johnson. 1993. Solving the problem of language recognition. Technical report, School of Computer Studies, University of Leeds.

Frank Keller and Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):458–484.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. *Behaviour Research Methods Instruments & Computers*, 36(3):516–524.

Patricia Newman. 1987. Foreign language identification: First step in the translation process. In *Proceedings of the 28th Annual Conference of the American Translators Accociation*, pages 509–516, Medford, New Jersey.

Diane Nicholls. 2003. False friends between French and English. *MED Magazine*, (9). `http://www.macmillandictionary.com/med-magazine/July2003/09-French-English-false-friends.htm`.

Beat Pfister and Harald Romsdorfer. 2003. Mixed-lingual analysis for polyglot TTS synthesis. In *Proceedings of Eurospeech 2003*, pages 2037–2040, Geneva, Switzerland.

Christopher Rollason. 2005. Language borrowings in a context of unequal systems: Anglicisms in French and Spanish. *Lingua Franca, Le Bulletin des Interprètes du Parlement Européen*, 8(2):9–14.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen.

Ariel Schwartz and Marti Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing (PSB 2003)*, pages 451–462, Kauai, Hawaii.

Achim Stein and Helmut Schmidt. 1995. Étiquetage morphologique de textes français avec un arbre de décisions. *Traitement Automatique des Langues*, 36(1-2):23–35.

Henry S. Thompson, Richard Tobin, David McKelvie, and Chris Brew. 1997. LT XML. Software API and toolkit for XML processing. `http://www.ltg.ed.ac.uk/software/`.

Xiaojin Zhu and Ronald Rosenfeld. 2001. Improving trigram language modeling with the World Wide Web. In *Proceedings of the International Conference on Acoustic Speech and Signal Processing (ICASSP 2001)*, pages 533–536, Salt Lake City, Utah.