Word Etymology as Native Language Interference

Vivi Nastase
University of Heidelberg
Heidelberg, Germany
nastase@cl.uni-heidelberg.de

Carlo Strapparava
Fondazione Bruno Kessler
Trento, Italy
strappa@fbk.eu

Abstract

We present experiments that show the influence of native language on lexical choice when producing text in another language – in this particular case English. We start from the premise that non-native English speakers will choose lexical items that are close to words in their native language. This leads us to an etymology-based representation of documents written by people whose mother tongue is an Indo-European language. Based on this representation we grow a language family tree, that matches closely the Indo-European language tree.

1 Introduction

In second-language writing, vestiges of the native language such as pronunciation, vocabulary and grammar are well-attested, and the phenomenon is called native language interference (Odlin, 1989). At the lexical level, the choice as well as the spelling can be indicative of the native language, through the choice of cognates, true or false friends – e.g. a writer with native language German may choose *bloom* cognate with *blume*, while a French one may choose *flower*, cognate with *fleur*. Misspellings – *cuestion* instead of *question* are also indicative, as the writer will tend to spell words close to the form from her original language (Nicolai et al., 2013).

In this paper we also look at native language interference starting from the lexical level, but abstract away from the actual word forms, and focus instead on the language of the etymological ancestors. The hypothesis we investigate is that the collective evidence of etymological ancestor languages are indicative of the language of the native speaker, and that this effect is sufficiently strong to allow us to rebuild an Indo-European language family tree. We use a corpus of essays written by English language learners, whose native language cover the languages from the Indo-European family. Etymological information is provided by an etymological dictionary extracted from Wiktionary.

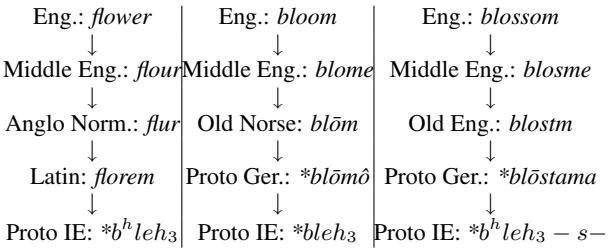

Figure 1: Examples of etymological ancestry from Wiktionary

The fact that word etymologies are included in the native-language interference phenomena could be used in various ways, i.a.: (i) to influence the selection of material for language learning, by focusing on vocabulary closer etymologically to the native language of the student, thus facilitating lexical retention; (ii) to reveal lexical choice errors caused by "etymological interference"; (iii) together with other interference phenomena, for the automatic corrections of language errors.

2 Related Work

English is a widespread common language for communication in a variety of fields – science, news, entertainment, politics, etc. A consequence is that numerous people learn English as a second (or indeed n^{th} language). The study of native language interference with the learning of English can be used in multiple ways, including devising methods to make the learning easier and correcting language errors (Leacock et al., 2014; Gamon, 2010; Dahlmeier and Ng, 2011).

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2692–2697
Copenhagen, Denmark, September 7–11, 2017.

Massung and Zhai (2016) present an overview of approaches to the task of natural language identification (NLI). Various surface indicators hold clues about a speaker's native language, that make their way into language production in a non-native language. Nagata and Whittaker (2013),Nagata (2014) find that grammatical patterns from the native language seep into the production of English texts. Tsur and Rappoport (2007) verify the hypothesis that lexical choice of non-native speakers is influenced by the phonology of their native language, and Wong and Dras (2009) propose the idea that (grammatical) errors are also influenced by the native language. One could draw the inference that character n-grams then could be indicative of the native language, and this was shown to be the case by Ionescu et al. (2014).

The natural language identification task (Tetreault et al., 2013) attracted 29 participating teams, which used a variety of features to accomplish the NLI task as a classification exercise: n-grams of lexical tokens (words and POS tags), skip-grams, grammatical information (dependency parses, parse tree rules, preference for particular grammatical forms, e.g. active or passive voice), spelling errors.

Apart from morphological, lexical, grammatical features, words also have an etymological dimension. The language family tree itself is drawn based on the analysis of the evolution of languages. Language evolution includes, or starts with, word etymologies. Word etymologies have been under-used for tasks related to NLI. They have been used implicitly in work that investigates cognate interference (Nicolai et al., 2013), and explicitly by (Malmasi and Cahill, 2015) who use words with Old English and Latin etymologies as unigram features in building classifiers for the TOEFL11 dataset. Etymological information is obtained from the Etymological WordNet (de Melo and Weikum, 2010).

We also investigate here the impact of etymological information, but unlike previous work, we do not extract unigram/n-gram features for classification, but we look at the collective evidence captured by the etymological "fingerprint" for each document and set of essays.

3 Etymological fingerprints

To investigate the influence of etymological ancestor languages, we represent each essay through etymological features, based on which we also built language vectors for each Indo-European language represented in the corpus. Essay vectors are then used to test native language identification potential, and the language vectors are used to grow a language family tree.

3.1 Word etymologies

Dictionaries customarily include etymological information for their entries. Wikipedia's Wiktionary has amassed a considerable number of entries that joined this trend. The etymological information can, and indeed has been extracted and prepared for machine consumption (de Melo and Weikum, 2010): Etymological WordNet[1] contains 6,031,431 entries for 2,877,036 words (actually, morphemes) in 397 languages. Because we process essays written in English, we use only the entries that give etymological origins for English words – 240,656. Figure 1 shows as an example of the kind of information formalized in the Etymological WordNet the etymological ancestry for the words *flower, blossom, bloom*.

3.2 Document/collection modeling

Essay representation After tokenization, POS tagging and lemmatization, each essay in the dataset is represented as a vector of etymological ancestry proportions obtained through the following processing steps:

1. for each token that has an entry in the etymological dictionary, we replace it with the language of its etymological ancestor – e.g. *sight* will be replaced by *ang* (Old English), *vision* by *lat* (Latin) (Table 1 shows the number of words with etymological ancestors in the subsets corresponding to each language);

2. compute the proportion of each etymological language in the essay, and represent the essay as a vector of language proportions[2]. We experimented with using etymological information going back through several ancestry levels, but using the first level led to the best results.

$$e_i \sim \langle p_{il_1}, ..., p_{il_n} \rangle \quad \textit{where} \quad p_{il_k} = \frac{n_{il_k}}{N_{i,etym}}$$

[1] http://www1.icsi.berkeley.edu/~demelo/etymwn/

[2] Using automatically corrected typos (first option of ispell) did not change the results significantly.

n_{il_k} is the number of words with etymological ancestor l_k in the i-th essay, and $N_{i,etym}$ is the number of words with etymological information in essay i.

Language vectors For each subcollection corresponding to one (student native) language L_j, we build the language vectors by averaging over the essay vectors in the subcollection:

$$V_{L_j} = \langle p_{L_j l_1}, ..., p_{L_j l_m} \rangle$$

$$\textit{where} \quad p_{L_j l_k} = \frac{\sum_{lang(e_i)=L_j} p_{il_k}}{|\{e_i | lang(e_i)=L_j\}|}$$

$p_{L_j l_k}$ is the proportion of etymological ancestor language l_k in all essays whose author has as native language L_j.

The essay and language vectors are filtered by removing etymological languages whose corresponding values in the language vectors are less than 10^{-4}.

4 Experiments

We investigate the strength of the etymological "fingerprint" of individual and collective essays written by non-native speakers of English, through two tasks – native language identification and language family tree construction. Towards this end, we work with a collection of essays written by contributors whose native language is an Indo-European language. The dataset is described in Section 4.1. For etymological information we rely on an etymological dictionary, described briefly in Section 3.1. Data modeling and the experiments conducted are described in Section 3.2.

4.1 Data

We used the ICLE dataset (Granger et al., 2009), consisting of English essays written by non-native English speakers. We filter out those that were written by people whose mother tongue is not from the Indo-European family (i.e. Chinese, Japanese, Turkish and Tswana). Table 1 shows a summary of the data statistics, including the number of words for which we have found ancestors in the etymological dictionary used. The corpus consists entirely of essays written by students. Two types of essay writing are present: argumentative essay writings and literature examination papers. Table 2 displays a list of topics in the corpus. The essays should be at least 500 words long and up to 1,000, and contain all the spelling mistakes made by their authors.

Following Nagata and Whittaker (2013), who also built the Indo-European family tree based on n-grams composed of function words and open-class parts of speech, essays that do not respect one of the following rules are filtered out: (i) the writer has only one native language, (ii) the writer has only one language at home; (iii) the two languages in (i) and (ii) are the same as the native language of the subcorpus to which the essay belongs. Table 1 shows a summary of the data statistics after filtering, including the number of words for which we have found ancestors in the etymological dictionary used.

Native language	*# essays*	*# tokens (with etym)*
Bulgarian	302	226,407 (149,151)
Czech	243	226,895 (148,391)
Dutch	263	264,981 (169,040)
French	347	256,749 (161,136)
German	437	259,967 (170,056)
Italian	392	253,798 (165,500)
Norwegian	317	238,403 (156,764)
Polish	365	263,223 (172,319)
Russian	276	259,510 (167,938)
Spanish	251	225,341 (139,565)
Swedish	355	224,948 (146,143)

Table 1: Statistics on the subset of ICLE dataset used.

1	Crime does not pay.
2	The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.
3	Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.
4	A man/woman's financial reward should be commensurate with their contribution to the society they live in.
5	The role of censorship in Western society.
6	Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.
7	All armies should consist entirely of professional soldiers : there is no value in a system of military service.
8	The Gulf War has shown us that it is still a great thing to fight for one's country.
9	Feminists have done more harm to the cause of women than good.
10	In his novel Animal Farm, George Orwell wrote "All men are equal: but some are more equal than others". How true is this today?
11	In the words of the old song "Money is the root of all evil".
12	Europe.
13	In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed. "Do you think it is still true nowadays ?
14	Some people say that in our modern world, dominated by science technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion ?

Table 2: Topics in the ICLE dataset.

The suitability of the dataset above for NLI was questioned by Brooke and Hirst (2012). They have shown that the fact that the corpus consists of sets of essays on a number of topics causes an overes-

timation of the results of NLI when random splitting, particularly for groups of contributors that were presented with very different topics – e.g. students from Asia vs. students from Europe. We have analyzed the distribution of essays into topics using the essay titles, and observed that contributions from Europe (which are our focus) have similar distributions across the featured topics.

Growing the language tree To grow the language tree from the language vectors built from the English essays, we use a variation of the UPMGA – Unweighted Pair Group Method with Arithmetic Mean – algorithm. Starting with the language vectors V_{L_j}, we compute the distance between each pair of vectors using a distance metric algorithm. At each step we choose the closest pair (L_a, L_b) and combine them in a subtree, then combine their corresponding sub-collection of essays, and build the language vector for the "composite" language $L_{a,b}$, and compute its distance to the other language vectors.

4.2 Results

We test whether etymological information surfaces as native language interference that is detectible through the tasks of native language identification and reconstruction of the language family tree. Table 3 shows results on the multi-class classification of essays according to the native language of the author, in the form of F-score average results using SVM classification in 5-fold cross-validation (using Weka's SMO implementation[3] with polynomial kernel and default parameters). The baseline corresponds to the language distribution in the dataset. We use as additional comparison point another set of features used to reconstruct the language family tree – the (closed-class) word and POS 3grams Nagata and Whittaker (2013), such as *the NN of; a JJ NN; the JJ NN*. We build all such patterns for the data, and keep the top 1000 by overall frequency.

Adding etymological features that capture the distribution of etymological ancestors for each essay led to improved results for all languages, varying from a non-significant improvement of 0.2% point for Russian, to a significant and high 5.3% improvement for German. Using only words, the accuracy is 73.2%, which increases marginally to 73.7 when etymology information is added. Using a full complement of standard features – word, lemma and character ngrams (n=1..3) (built following (Lahiri and Mihalcea, 2013)) – gives an average accuracy (over 5 fold cross-validation) of 85.7%. Adding etymology does not lead to improvements when added to this set.

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

Language	Baseline	Etym.	Patt.	both
Bulgarian	8.52%	32.4	51.7	54.3
Czech	6.85%	21.9	53.4	54.4
Dutch	7.41%	11.7	50.4	51.1
French	9.78%	30.0	58.8	62.9
German	12.31%	45.4	47.4	52.7
Italian	11.04%	34.3	66.3	67.3
Norwegian	8.93%	35.5	57.0	59.3
Polish	10.28%	42.5	59.9	62.1
Rusian	7.78%	12.7	46.9	47.1
Spanish	7.07%	24.6	57.9	59.6
Swedish	10.00%	23.1	44.8	45.7
Accuracy		31.7	54.2	56.3

Table 3: 5-fold cross-validation F-scores and accuracy for language classification

Despite the rather low results when etymology is used on its own for language identification, the cumulative evidence leads to a language family tree that closely matches the gold standard (Figure 2). The tree on top is the gold standard cf. (Nagata and Whittaker, 2013; Crystal, 1997). The tree is grown by computing the euclidean distance between pairwise vectors, and then iteratively grouping together the closest vectors at each step as described in Section 4.1.

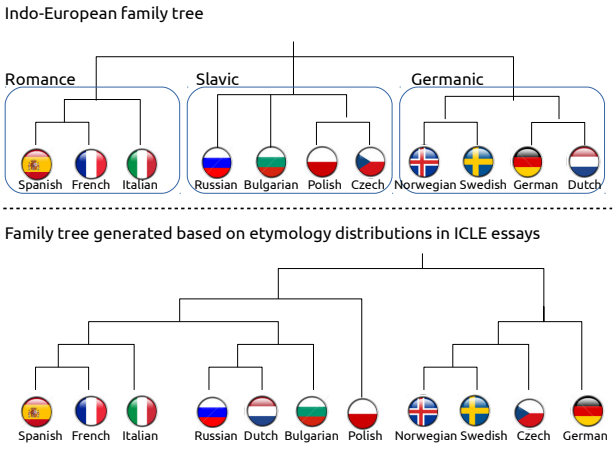

Figure 2: Language family trees – the gold standard and the automatically generated one

The two wrongly placed languages in our language family tree are Czech and Dutch. Czech

is grouped with the Germanic languages. Historically, the country which is now the Czech Republic has been under German occupation for long periods of time. We propose the hypothesis that this has influenced the Czech language at the lexical level, and our etymological fingerprinting characterizes mostly the lexical aspects of language. We plan to verify this theory as etymological information for Czech and German becomes more readily available in sufficient quantities. We have not yet found an explanation for the grouping of Dutch with the Slavic languages. Like mentioned before, the language vectors we built rely exclusively on lexical information, and it is possible that Dutch's grammatical structure is what defines it best as being a part of the Germanic language family, as opposed to the lexical dimension.

5 Conclusion

In this paper we have shown an exploration of a novel indicator of native language interference in second language learning, particularly etymology. While cross-linguistically related words (cognates, false and true friends) have been part of the repertoire of features for native language identification and cross-language studies, we have focused here on the language of etymological ancestors of words, and in particular their distribution in documents. Experiments in recreating the Indo-European family tree have shown that the composition of a document in terms of etymological languages is indicative of the native language of the writer to the extent that fundamental characteristics of languages – typological relatedness between languages – emerge.

References

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *Proceedings of COLING 2012*, pages 391–408, Mumbai, India. The COLING 2012 Organizing Committee.

David Crystal. 1997. *The Cambridge Encyclopedia of Language*. Cambridge University Press.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Correcting semantic collocation errors with l1-induced paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 107–117, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Gamon. 2010. Using mostly native data to correct errors in learners' writing: A meta-classifier approach. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 163–171, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International Corpus of Learner English v2 (ICLE).

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. Can characters reveal your native language? a language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1363–1373, Doha, Qatar. Association for Computational Linguistics.

Shibamouli Lahiri and Rada Mihalcea. 2013. Using n-gram and word network features for native language identification. In *Proc. of BEA@NAACL-HLT 2013*, pages 251–259.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. Automated grammatical error detection for language learners, 2nd edition. In Graeme Hirst, editor, *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool.

Shervin Malmasi and Aoife Cahill. 2015. Measuring feature diversity in native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–55, Denver, Colorado. Association for Computational Linguistics.

Sean Massung and Chenkxiang Zhai. 2016. Non-native text analysis: A survey. *Natural Language Engineering*, 22(2):163186.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*, pages 149–156, New Delhi, India.

Ryo Nagata. 2014. Language family relationship preserved in non-native english. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1940–1949, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ryo Nagata and Edward Whittaker. 2013. Reconstructing an indo-european family tree from non-native english texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1147, Sofia, Bulgaria. Association for Computational Linguistics.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Lei Yao, and Grzegorz Kondrak. 2013. Cognate and misspelling features for natural language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 140–145, Atlanta, Georgia. Association for Computational Linguistics.

Terence Odlin. 1989. *Language Transfer*. Cambridge University Press.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, CACLA '07, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sze-meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop*, pages 53–61, Sydney, Australia.