Word Formation Variation as Features for Native Language Identification

Julia Krivanek, Detmar Meurers
Universität Tübingen
{jkrvnk,dm}@sfs.uni-tuebingen.de

The task of native language identification can be useful for theoretical studies of language transfer (Jarvis 2012) and it can inform applications, e.g., by informing learner models for intelligent language tutoring systems to support different feedback depending on the L1 (Amaral & Meurers 2008). Current L1-classification approaches (e.g., Brooke & Hirst 2012; Bykh & Meurers 2012; Jarvis et al. 2012) achieve high accuracy with surface-based features, such as word and part-of-speech n-grams. However, surface-based approaches make use of large feature sets, which are hard to interpret qualitatively in terms of linguistic insight. In addition, surface features are directly dependent on the genre and topic of the texts being classified, so that results degrade significantly for out-of-domain classification (Brooke & Hirst 2011). Other approaches (Wong & Dras 2009; Bestgen, Granger & Thewissen 2012) make use of error patterns, which capture one conceptually interpretable characteristic of learner language, but typically require manual error annotation.

In this paper, we propose to shift the focus to a new class of features for L1-classification: linguistic variation. In many situations, language offers a range of options for formulating a given message. Indeed, in variationist sociolinguistics, the choices speakers make have successfully been used to identify relevant speaker properties (cf. Tagliamonte 2011). Adapting this perspective, we propose to make use of variation features for native language identification. We make use of the variationist method observing where speakers make choices in the language system, but different from variationist sociolinguistic research we then investigate the impact of the L1 (rather than the social properties focused on in sociolinguistics). Making this general idea concrete, we describe an experiment we carried out on German learner texts using word formation variation as features for L1-identification.

We use the term word formation to refer to the range of processes through which new words are formed. Typically a given language offers several options. Which options get used when and how the options are realized differs across languages. New words can be formed with the help of derivational morphemes or without them, the process can change a word's category or not, and so on. Accordingly, we can define variables such as the ones in Figure 1 and use their variants as features for L1-classification.

Variables	**Variants**	**Examples**
Morpheme alternation	no affix	*Frau<NN> + Welt<NN> → Frauenwelt<NN>*
	suffix	*Feminist<NN>* + ***in<SUFF>*** *→ Femimistin<NN>*
	prefix	***un<PREF>*** *+ gerecht<ADJ> → ungerecht<ADJ>*
	verb particle	***auf<VPART>*** *+ geben<V> → aufgeben<V>*
Derived category alternation	noun	*anerkennen<V> + ung<SUFF> →* ***Anerkennung<NN>***
	verb	*auf<VPART> + geben<V> →* ***aufgeben<V>***
	adjective	*entsprechen<V> →* ***entsprechend<ADJ>***
	adverb	*möglich<ADJ> + weise<SUFF> →* ***möglicherweise<ADV>***
Source category alternation	noun	***Feminist<NN>*** *+ in<SUFF> → Femimistin<NN>*
	verb	***anerkennen<V>*** *+ ung<SUFF> → Anerkennung<NN>*
	adjective	***möglich<ADJ>*** *+ weise<SUFF> → möglicherweise<ADV>*

Figure 1: Some word formation variables

For example, the *morpheme alternation* allows us to distinguish word formation without affix from that using suffixes or using prefixes The *derived category alternation* supports distinguishing derived from basic variants. The *source category alternation* supports identifying which source categories undergo a word formation process.

As learner corpus data we used 185 essays from the Falko learner corpus of German (Reznicek et al. 2012), written by learners with five native languages (English, Polish, Russian and Danish, and a native German control group), with an average length of 470 words. The data was annotated using the RFTagger (Schmid & Laws 2008) providing part-of-speech and morphological information.

As features, we took four categories (noun, verb, adjective and adverb) and compiled out all possible variations of the word formation variables described above. These variations were then counted for each text and normalized by the derived category. After removing features which did not occur in the data set, we obtained 29 features.

For classification, we used the WEKA SMO classifier (Witten & Frank 2005) and report the results of leave-one-out evaluation. Using only the 29 word formation features, we obtained a classification accuracy of 55.1%, which is encouraging given the random baseline of 20% for this balanced five class problem. Just as in the current NLI approaches for English, the accuracy can be increased by introducing a combination of different feature types, as we demonstrate in Bykh et al. (2013); we here instead provide an analysis of the word formation variation features as the focus of this paper.

An analysis of the confusion matrix shows that the German control group data is most clearly singled out, whereas many confusions arise within the Slavic group (8 Polish texts are identified as Russian, 12 Russian ones as Polish). We therefore are exploring the use of cascading classification to first distinguish language families (e.g., Slavic vs. others) followed by a second classification trained only on the subdistinctions within a language family (e.g., Polish vs. Russian). We expect that the features which are most effective at these different stages will differ clearly and meaningfully, in line with the findings of Vajjala & Loo (2013), who used a cascading classifier in a proficiency classification task.

One can also anylze the results of our approach in terms of *overuse/underuse* (Lüdeling et al. 2011). In order to detect distinctive features, one compares the frequencies of a variant of a given variable across the L1 groups. Comparing the L1-German control group with the other L1 groups, we, for example, found that the phrasal verb feature "verb particle + verb" (e.g., *auf<VPART>geben<V>*) was underused by all learners, with native Danish learners being the closest to native German usage. Native speakers of Slavic languages, lacking phrasal verbs, and English, where particles follow different distributional patterns than in German, showed the strongest underuse.

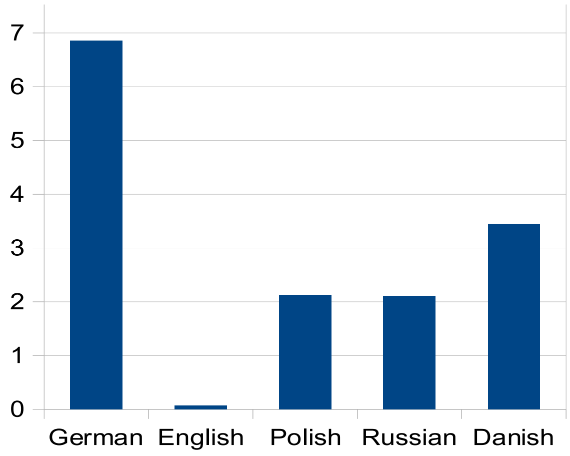

Figure 2: Relative frequency of phrasal verbs in German texts across different L1s

In conclusion, an analysis of variation in word formation provides an effective and insightful perspective for L1-classification. As such it further populates the landscape of data-driven and theory-driven approaches (Meurers et al. 2013) in a way yielding qualitatively interpretable features. At the same time, it can also be integrated into ensemble classifiers combining different sources of information for L1-classification (Bykh et al. 2013) to further improve the quantitative state-of-the-art in terms of classification accuracy.

References

Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning* 21(4), 323–338. URL http://purl.org/dm/papers/amaral-meurers-call08.html.

Bestgen, Y., S. Granger & J. Thewissen (2012). Error Patterns and Automatic L1 Identification. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 127–153.

Brooke, J. & G. Hirst (2011). Native Language Detection with 'Cheap' Learner Corpora. In *Learner Corpus Research 2011 (LCR 2011)*. Louvain-la-Neuve.

Brooke, J. & G. Hirst (2012). Robust, Lexicalized Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 391–408.

Bykh, S. & D. Meurers (2012). Native Language Identification Using Recurring N-grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 425–440. URL http://purl.org/dm/papers/bykh-meurers-12.html.

Bykh, S., S. Vajjala, J. Krivanek & D. Meurers (2013). Combining Shallow and Linguistically Motivated Features in Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA. URL http://purl.org/dm/papers/Bykh.Vajjala.ea-13.html.

Jarvis, S. (2012). The Detection-Based Approach: An Overview. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 1–33.

Jarvis, S., G. Castañeda-Jiménez & R. Nielsen (2012). Detecting L2 Writers' L1s on the Basis of Their Lexical Styles. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 34–70.

Lüdeling, A., H. Hirschmann & A. Zeldes (2011). Variationism and Underuse Statistics in the Analysis of the Development of Relative Clauses in German. In Y. Kawaguchi, M. Minegishi & W. Viereck (eds.), *Corpus Analysis and Diachronic Linguistics*, Amsterdam: John Benjamins.

Meurers, D., J. Krivanek & S. Bykh (2013). On the Automatic Analysis of Learner Corpora: Native Language Identification as Experimental Testbed of Language Modeling between Surface Features and Linguistic Abstraction. In *Proceedings of 4th International Conference on Corpus Linguistics (CILC 2012)*. To appear.

Reznicek, M., A. Lüdeling, C. Krummes & F. Schwantuschke (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Ver. 2.0*. URL http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko.

Schmid, H. & F. Laws (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*. Stroudsburg, PA, vol. 1, pp. 777–784.

Tagliamonte, S. A. (2011). *Variationist Sociolinguistics: Change, Observation, Interpretation*. John Wiley & Sons.

Vajjala, S. & K. Loo (2013). Role of Morpho-syntactic features in Estonian Proficiency Classification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8), Association for Computational Linguistics*. URL http://aclweb.org/anthology/W13-1708.pdf.

Witten, I. H. & E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam; Boston, MA: Morgan Kaufmann, 2nd ed.

Wong, S.-M. J. & M. Dras (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*. pp. 53–61.