Advancing Linguistic Features and Insights by Label-informed Feature Grouping: An Exploration in the Context of Native Language Identification

Serhiy Bykh
Seminar für Sprachwissenschaft
Universität Tübingen
sbykh@sfs.uni-tuebingen.de

Detmar Meurers
Seminar für Sprachwissenschaft
Universität Tübingen
dm@sfs.uni-tuebingen.de

Abstract

We propose a hierarchical clustering approach designed to group linguistic features for supervised machine learning that is inspired by variationist linguistics. The method makes it possible to abstract away from the individual feature occurrences by grouping features together that behave alike with respect to the target class, thus providing a new, more general perspective on the data. On the one hand, it reduces data sparsity, leading to quantitative performance gains. On the other, it supports the formation and evaluation of hypotheses about individual choices of linguistic structures. We explore the method using features based on verb subcategorization information and evaluate the approach in the context of the Native Language Identification (NLI) task.

1 Introduction and related work

Native Language Identification (NLI) is the task of inferring the native language (L1) of writers from texts they wrote in another language. NLI started to attract attention in computational linguistics with the work of Koppel et al. (2005). Since then interest has steadily risen, leading to the First NLI Shared Task in 2013, with 29 participating teams (Tetreault et al., 2013).

NLI is usually considered as a text classification problem with the different L1s as labels. A range of features reaching from character and word n-grams to dependency- and constituency-based features have successfully been used in standard supervised machine learning setups, yielding accuracies of up to around 83% for the 11 classes in the First NLI Shared Task. Some more recent papers further advance the best result from that competition, namely 83.6% (Jarvis et al., 2013), reaching around 85% (Bykh and Meurers, 2014; Ionescu et al., 2014).

While pushing the quantitative side is one option of advancing the NLI work further, another avenue of research tries to improve our understanding of how the different feature types work and what conclusions one can draw from these observations for Second Language Acquisition (SLA) research. Swanson and Charniak (2013) utilized Tree Substitution Grammars as well as different measures of relevancy and redundancy to extract indicative linguistic patterns. Swanson and Charniak (2014) adopted the approach to dependencies. Malmasi and Dras (2014) proposed a technique to detect over- and underuse of certain patterns by writers with a particular L1-background using linear SVM weights derived from Adaptor grammar collocations or Stanford Dependencies. Meurers et al. (2014) employed verb subcategorization patterns as features and showed that there are differences in the usage patterns of verbs between native English writers and, e.g., writers with Chinese L1-background. Bykh and Meurers (2014) systematically explored constituency-based features and discussed the distinctive power of the different variants realizing lexical and phrasal categories. Malmasi and Cahill (2015) investigated the correlation between various features in a feature set commonly used in NLI.

Many of the current NLI approaches rely on large feature sets, which makes it difficult to qualitatively interpret the findings. We therefore want to explore grouping features together which behave alike to advance linguistic insight and improve classification. In place of zooming in on single features, with the potential danger of overfitting the training data, feature grouping can help make explicit underlying

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 739–749, Osaka, Japan, December 11-17 2016.

linguistic properties within a set of features. Thus, on the one hand, it can support the identification of linguistic generalizations, which is relevant for qualitative analysis and theoretical interpretation. On the other hand, we also expect quantitative benefits due to the potential reduction of data sparsity, especially in cases where the particular single feature realizations might be rare but the underlying structure captured by a group is more common.

One of the well-established techniques for building feature groups is hierarchical clustering (Park, 2013; Krier et al., 2007; Butterworth et al., 2005). It can be an effective method for capturing linguistic generalizations. For example, Pate and Meurers (2007) show in the context of PCFG parsing that contextually enriching categories followed by clustering the categories with similar distributions results in a performance improvement.

In this paper, we propose to employ hierarchical clustering for feature grouping in a way that is informed by the classification label – here the L1 of the writer. Adopting a variationist linguistic perspective that attempts to identify variants of an underlying variable (Tagliamonte, 2011), we illustrate and evaluate the feature grouping technique in detail for features encoding the different subcategorization options realized by a given verb – a feature type that is well-motivated in related SLA research (Tono, 2004; Callies and Szczesniak, 2008; Stringer, 2008). Using the technique, we first test the hypothesis, whether writers with different L1-backgrounds prefer certain subcategorization patterns when realizing particular verbs. Then we show how the technique can be used to investigate specific hypotheses about L1-transfer suggested in the SLA research; we focus on the subject in the subcategorization pattern and explore some of its realization options in L1 Chinese following Wang (2009). The results presented below confirm that feature grouping can indeed provide theoretical and practical benefits.

2 Feature grouping

We propose a label-informed feature grouping technique that can be used to group structured linguistic features in line with a variationist perspective. We first introduce the variationist perspective and the nature of variationist features, before we turn to our implementation of the technique.

Variationist sociolinguistics In variationist sociolinguistics, the focus is on the possible linguistic choices made by a speaker. This makes it possible to connect the choices in the language with extralinguistic variables, such as the gender or the age of a speaker. For example, in Labov's seminal study "The Social Stratification of (r) in New York City Department Stores" (Labov, 1972), he found that the presence or absence of the consonant [r] in postvocalic position (e.g., *fourth*) correlates with the ranking of people in status (social stratification). Hence, under a variationist perspective, one observes which of the possible *variants* of a *variable* is chosen by a particular speaker (Tagliamonte, 2011). Recent research in the language learning context argues that a preference for particular variants can also be indicative of individual characteristics such as proficiency or L1-background (Lüdeling, 2011; Callies and Zaytseva, 2011; Meurers et al., 2014; Bykh and Meurers, 2014).

Variationist features To obtain *variationist features* one has to implement the logic described in the previous paragraph. It requires choosing some language *variables* that can be realized by a particular set of *variants*. For our first explorations of the proposed technique, we chose verb lemmas as variables and the different subcategorization (subcat) patterns of that lemma as variants.

Grouping technique As motivated above, we want to explore whether writers with a given L1 prefer certain subcat variants when realizing a particular verb lemma variable.

In the training corpus, we can record the relative frequencies for the different subcat variants used to realize a lemma. Individual lemmas occur quite rarely, though, so we want to group together all those lemmas that behave alike with respect to their subcat variants.

Can we also take the classification label into account when clustering the variables by the frequency proportions of their variants? In other words, how can we group those lemmas together that for a given L1 have similar proportions of subcat variant realizations? In order to incorporate the classification label into the grouping procedure, we do not generate a single vector of variants for a variable, but k vectors, where k is the number of L1 labels in the training data. Each of the k vectors contains the proportions of

the variants for a given variable, calculated using the subset of the training data for a particular L1. Then the k vectors for each variable are concatenated in order to get an instance for clustering. Hierarchical clustering then groups variables together that for writers of a specific L1 realize a similar set of variants in similar proportions, i.e., it groups lemmas together that for a specific L1 label pattern alike with respect to the realized subcat variants. Since the feature set for clustering is informed by the L1 labels, we refer to this technique as *label-informed feature grouping*. Viewed from the variationist perspective, the method is designed to group those variables together that in terms of their variants behave alike with respect to the classification label.

Let us spell this out in an example. Assume that writers with Spanish L1 prefer the subcat variant $p \in \{p, q\}$, whereas writers with Chinese L1 prefer the variant q in connection with a particular set of verbs A. That information is captured by the difference in relative frequencies for the variants p and q in connection with A in the training data subsets for the two different L1s. Using separate vectors for the different L1s, explicitly provides that relevant information to the clustering algorithm. Clustering thus can identify the group A of verbs that is indicative for the classification purposes in terms of the choice of variants made by different L1s, and also of interest from the perspective of interpreting these effects in terms of SLA research.

We cluster the variables via *agglomerative hierarchical clustering*, employing some standard parameters, namely, Euclidean distance[1] and complete-linkage. We set the number of clusters $c = 1$, which means that after clustering we obtain a *dendrogram* corresponding to a *single-rooted binary tree*.

Now, the question is, how to decide, which grouping is the most appropriate one? I.e., where do we want to cut the dendrogram? We approach that issue experimentally, by systematically applying different branch length cut-offs with step $s = 0.1$ to the dendrogram, and then evaluating every grouping via text classification using the different groupings as features.

In connection with using subcat variants, realized by groups of verbs, as features there are two more points to clarify. First, how to merge clustered variables with different sets of variants? Here, we simply take the union of the variant sets as the resulting variant set of the variables group. Second, how to compute the feature values for the groups? For that we use the *micro average* measure adapted to the variationist perspective (Krivanek, 2012; Meurers et al., 2014). In sum, we apply the following steps:

1. For each of the n variables V_i and each of the m variants v_j occurring in the whole training data, calculate the matrices M_{ij}^k using the corresponding label-distinct data subset l_k:

$$M_{ij}^k = \frac{f(v_j, V_i, l_k)}{\sum_{q=1}^{m} f(v_q, V_i, l_k)}$$

   where $f(v, V, D)$ yields the frequency of the variant v realizing the variable V in the data D. Here $D = l_k$. If a variant v does not occur in the context of V using data D, $f(v, V, D) = 0$.

2. Perform a horizontal matrix concatenation of the k matrices M_{ij}^k resulting in a single matrix M_{ij} containing the variable based instances for clustering (the rows of M_{ij}).

3. Perform hierarchical clustering with the number of clusters $c = 1$, Euclidean distance, and complete-linkage as parameters.

4. Systematically apply different branch length cut-offs r using a suitable step s (here we employed $s = 0.1$). Evaluate the resulting clusters, i.e., groups of variables by using them as features in a classification setup.

[1] We also explored using other distance measures, such as the Manhattan or the Hellinger distance. Especially, the latter is supposed to be more suitable for probability-based features. However, the Euclidean distance performed best.

(a) Merging groups: Let a group (cluster) C contain x variables, each realized by a particular variant set X_i. Then the resulting variant set X^c for the variables group C is defined as the union of all variant sets X_i:

$$X^c = \bigcup_{i=1}^{x} X_i$$

(b) Encoding groups as features for classification: Calculate *micro average* for each variant $v \in X^c$ associated with the group $C = \{V_1, ..., V_n\}$, using data t:

$$mic(v, C) = \frac{\sum_{i=1}^{n} f(v, V_i, t)}{\sum_{i=1}^{n} \sum_{j=1}^{m} f(v_j, V_i, t)}$$

where $f(v, V, D)$ is defined as above (1), and $D = t$ is a given text.

5. Terminate evaluation (at the latest) after a particular cut-off r yielded one single group containing the whole variables set.

Note that the same technique can also be used to group *variants* based on their frequency proportions for all variables using the k label-based data subsets – an option we here do not go into further to keep things clear and within the space constraints.

3 Data

For the experiments described in this paper, we use the TOEFL11 corpus (Blanchard et al., 2013) introduced for the NLI Shared Task 2013 (Tetreault et al., 2013), which has become a common frame of reference for NLI research. It consists of essays written by English learners with 11 L1-backgrounds (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish) at three proficiency levels (low, medium, high). Each of the 11 L1s is represented by 1,100 essays (900 training, 100 development, 100 test). We use the union of the training and development sets for training and the standard test set for testing. In total for all L1s, we thus train on 11,000 and test on 1,100 essays.

4 Tools

We utilized the MATE tools[2] (Björkelund et al., 2010) for data preprocessing (tokenization, lemmatizing, POS-tagging) and the MATE dependency parser (Bohnet, 2010) to identify the arguments of a verb realized in a sentence, i.e., the subcat frame that was realized. For hierarchical clustering we employed WEKA (Hall et al., 2009). To process the resulting dendrograms we used the Libnewicktree[3] tree parser. Finally, classification was carried out using L2-regularized Logistic Regression from the LIBLINEAR package (Fan et al., 2008) accessed through WEKA.

5 Features

The hypothesis we are testing is whether writers with different L1s prefer different subcat variants. To systematically explore the potential benefits of feature grouping, we start with *simple features*, where every variable, i.e., verb lemma, is considered separately. We then infer sets of *complex features*, i.e., sets of various groups of variables using the proposed technique, abstracting from individual verb lemmas to classes of verbs. All feature values are calculated using *micro average* as introduced in section 2 (4b), with simple features being a special case of the complex ones, where the group C consists of a single variable ($C = \{V_1\}$).

[2] https://code.google.com/p/mate-tools

[3] https://github.com/cjb/libnewicktree

5.1 Simple features

We dependency parsed the data and extracted the corresponding argument realization patterns, i.e., the realized subcat variants, for all verbs occurring in the data. We consider the following labels as arguments:

- *sbj*: subject
- *lgs*: logical subject
- *obj*: (in)direct object or clause complement
- *bnf*: benefactor in dative shift
- *dtv*: dative in dative shift
- *prd*: predicative complement
- *oprd*: object complement
- *put*: locative complements of the verb put
- *vc*: verb chain

Utilizing the verb lemmas with their extracted subcat variants, we generated features, such as:

- *believe*_sbj
- *believe*_sbj_obj
- *may*_sbj_vc
- *put*_sbj_put
- *help*_sbj_obj
- *make*_sbj_obj_oprd

Feature reduction We performed the following three feature reduction steps due to some theoretical and practical considerations:

1. Verbs as features are rather rare. In order to reduce data sparsity issues, at this point we opted for ignoring the *different permutations* of arguments within a subcat variant. This step reduced the number of distinct variants from 355 to 218.

2. Some of the subcat variants are still rather specific and unlikely to occur frequently enough in the data. Some of them also suffer from tagging or parsing errors. So, in a second reduction step we grouped all *argument labels into three coarse-grained classes* in order to get more general patterns and to cope with data sparsity:

   - $\{sbj, lgs\} \rightarrow s$ (subject)
   - $obj \rightarrow o$ (object)
   - $\{bnf, dtv, prd, oprd, put, vc\} \rightarrow x$ (rest group)

   The number of distinct subcat variants reduced from 218 to 48. Applied to the examples listed above, we obtain features of the following form:

   - *believe*_sbj $\rightarrow$ *believe*_s
   - *believe*_sbj_obj $\rightarrow$ *believe*_s_o
   - *may*_sbj_vc $\rightarrow$ *may*_s_x
   - *put*_sbj_put $\rightarrow$ *put*_s_x
   - *help*_sbj_obj $\rightarrow$ *help*_s_o
   - *make*_sbj_obj_oprd $\rightarrow$ *make*_s_o_x, etc.

3. The last reduction step is conceptually different from the first two. It is based on theoretical considerations in connection with the variationist perspective: We are interested in the linguistic choices made by a speaker. If there is only a single variant for using a verb, we cannot observe a *choice* being made. We therefore dropped all features for verb lemmas that only occur with a single subcat variant in the training data. That reduced the number of distinct verb lemmas from originally 11,401 to 3,785.

Feature reduction clearly also means a loss of potentially indicative subcat information. A more fine-grained reduction therefore constitutes an important topic for future work.

As a result of feature reduction, we obtain 3,785 distinct verb lemmas (variables) with 14,389 subcat variants as features. That means that on average there are roughly four subcat variants per variable.

5.2 Complex features

We refer to the features defined by grouping the verb lemmas as proposed in section 2 as *complex features*. The purpose of these complex features is to abstract away from individual verb lemmas to more general classes. In contrast to Meurers et al. (2014), where verb lemmas realizing *exactly the same* subcat variants were grouped together, the technique proposed here makes it possible to systematically explore a range of different groupings of lemmas based on the *similarity* of the realized subcat variants, and to take into account the classification label. Each group of verb lemmas C and the corresponding set of subcat variants X^c constitutes a complex feature in the sense of (4a) of section 2.

6 Results

An overview of the results is presented in Figure 1. On the x-axis, the leftmost point (marked "s") corresponds to using only simple features (every verb lemma with the corresponding subcat variants is considered separately), i.e., no clustering. With increasing x-values, we go up in the dendrogram to obtain groups of verbs. The x-values are the branch length cut-offs applied to the dendrogram using a step of 0.1. The y-axis represents the accuracy of the classification on the test set, using the training-test split described in section 3 and the classifier spelled out in 4. The random baseline is 9.1%.

Model with simple and complex features ([s/c]): This is the basic setting using simple and complex features. The figure shows that 44.5% at point "s" is the highest accuracy, so feature grouping does not provide a quantitative edge. For settings incorporating complex features, the best result is 44.2%, obtained for the cut-off 0.3. The clustering technique groups verb lemmas in terms of the proportion of

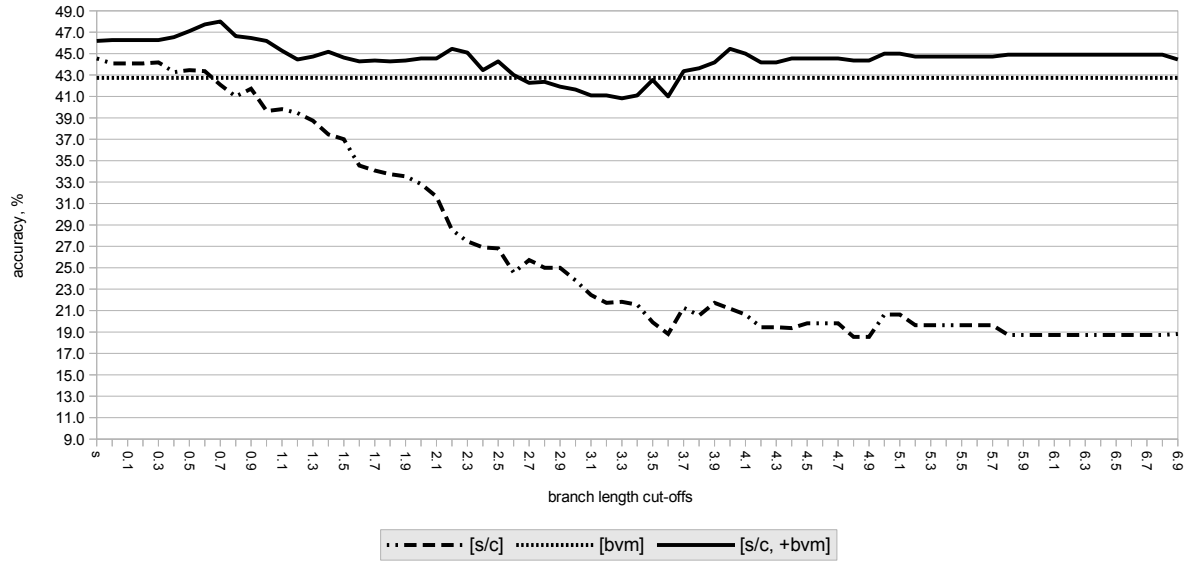

Figure 1: Accuracy of classification for different feature sets

their subcat variants. The particular verb lemmas, i.e., surface forms, which are part of a group, are not by themselves encoded in the feature space any more. For complex features the classifier therefore does not have an access to the potentially highly indicative surface based properties. Even different misspellings of the verbs can be indicative of the native language – distinctions that the basic method counting variants glosses over.[4] The disadvantage of loosing indicative surface information seems to outweigh the potential advantage of the generalization in terms of reducing data sparsity. We can validate that assumption by creating a binary verb lemma model and then combining it in a model with the simple and complex variant features.

Binary verb lemma model ([bvm]): To be able to identify the contribution of the subcat variants of individual verbs and groups of verbs, we need a way to separately quantify the information provided by the verb lemma itself (i.e., the presence of the variable, as separate from the choice of variants). In the bvm model, we thus only encode the presence/absence of the 3,785 verb lemmas for each text. The accuracy for that model is 42.7%. We included that result as a line in Figure 1 to visualize the performance relative to the other two models, which make use of subcat pattern information. Comparing the other models to that one shows the benefits of incorporating the subcat variants as features. Indeed, the other curves discussed below show better results, confirming that such features are useful.

Binary verb lemma combined with simple and complex features ([s/c, +bvm]): To validate our assumption regarding the role of surface properties, we tried a combined setup, where the [*s/c*] setting was used in combination with the binary verb lemma model [*bvm*] described above. We assume the [*bvm*] model to restore the surface information lost due to the generalization, which should improve the classification performance. Indeed, the classification performance increased compared to the basic [*s/c*] setting. For simple features, the accuracy is 46.2%, and thus 1.7% higher. Including [*bvm*] therefore is beneficial even for settings not involving clustering. For settings including complex features, the difference depends on the actual cut-off. The best performance is 48.0% obtained using the cut-off 0.7 (273 complex and 3016 simple features). The difference between the best complex feature results is 3.8%. In most of the cases, the difference is even much higher, as seen when comparing the [*s/c*] and the [*s/c, +bvm*] curves at corresponding cut-offs in Figure 1. That result supports our assumption regarding the role of the surface properties. It is also supported by the shape of the [*s/c, +bvm*] curve only. The best model using complex features (cut-off 0.7) outperforms the model solely based on simple features ("s") by 1.8%[5]. Thus, when built on top of a surface-based model such as [*bvm*], the proposed grouping and generalization technique shows practical advantages in terms of accuracy. The findings confirm the hypothesis that in general learners with different L1s seem to prefer different subcat patterns. Finally, using more data or some more frequent variables resulting in more reliable frequency distributions of the variants, is expected to increase the quantitative gains. We plan to explore this issue in our future work.

Relative performance In comparison with previous research, the proposed automatic feature grouping method outperforms the approach presented in Meurers et al. (2014), where only verb lemmas with equal subcat variant sets constituted a group. A replication of that approach employing the verb subcat features and the data setup used in this paper, showed an accuracy of 38.7%, and after adding the *[bvm]* model, we obtained 44.4%. This result is 3.6%[6] lower than our best accuracy obtained in [*s/c, +bvm*].

In order to further investigate the potential quantitative advantages of the proposed features and the clustering method, we combined a range of features using the meta-classifier approach described in Bykh and Meurers (2014). The results are summarized in Table 1. First, we combined the core 16 feature types employed in Bykh et al. (2013), namely features based on n-grams, dependencies, local trees, suffix information, linguistic complexity and lemma realization, with the best performing model in Bykh and Meurers (2014), which is based on constituency variation features plus 40 different types of n-grams, i.e., word- and lemma-based n-grams as well as two types of n-grams incorporating POS,

[4]For example, we discovered that some of the clusters contained misspelled versions of the same verbs, such as *commmunicate/communicat*, *tounderstand/ubderstand*, *exaggrate/exxagarate*, etc.

[5]$p < 0.05$ using McNemar's test.

[6]$p < 0.001$ using McNemar's test.

with $1 \leq n \leq 10$ each (see also Bykh and Meurers, 2012). This model yielded an accuracy of 85.2%.[7] Second, we added the best performing system in this paper, namely [*s/c, +bvm*] with cut-off 0.7[8] to that ensemble, and alternatively the basic setting of [*s/c, +bvm*] without clustering, i.e., using simple features only ("s"). Both options showed an increase in accuracy by the same value of 0.2%, resulting in 85.4%. To the best of our knowledge, the highest outcome obtained so far on the same data was 85.3%, reported by Ionescu et al. (2014). Thus, by using the new features explored in this paper, it was possible to slightly outperform the already very high best previous accuracy on the standard TOEFL11 data setup. However, in the comprehensive ensemble model used here, there was no quantitative difference between adding the best performing [*s/c, +bvm*] setting (cut-off 0.7), which incorporates complex features, and the lower performing version containing simple features only ("s"). Yet, there is a difference in terms of the feature counts for the two models, namely, 16,841 vs. 18,174 respectively. Thus, the version incorporating complex features is more efficient, providing the same quantitative advantage with a more compact model. This supports the assumption that the generalizations made by the technique are reasonable. The findings suggest that the approach can further advance the already high performance of the state-of-the-art NLI systems.

Rank	Id	System	Accuracy
1	A	D + E + F/G	85.4%
2	B	Ionescu et al. (2014)	85.3%
3	C	D + E	85.2%
4	D	Bykh and Meurers (2014)	84.8%
5	E	Bykh et al. (2013)	82.5%
6	F	[*s/c, +bvm*], cut-off 0.7	48.0%
7	G	[*s/c, +bvm*], s	46.2%

Table 1: Relative performance. System B is the best previously reported system based on the same data.

7 Qualitative explorations

In the previous sections we explored in detail the quantitative gains of the proposed technique using verb subcat features. In this section we sketch, how the method can be used to advance the qualitative analysis in the context of the SLA research. In particular, we investigate the hypothesis by Wang (2009), suggesting that learners with L1 Chinese overuse *pronoun-subjects* over *noun-subjects* in Chinese-English translations.[9] The findings of Wang (2009) based on translations by 81 students support the hypothesis.

In order to investigate this hypothesis, we slightly modify our features, i.e., we use only the subject part of the verb subcat, and in addition we consider only those subjects, which are tagged as personal pronouns (*prp*) or nouns (*nn*). So, based on our training data, we extract features such as *believe*_sbj+prp and *believe*_sbj+nn, or *study*_sbj+prp and *study*_sbj+nn, etc. Then we run *one vs. rest* classifiers with L1 Chinese vs. the western L1s in our set, namely French, German, Italian and Spanish. The classifiers follow the logic of the [*s/c*] and [*s/c, +bvm*] settings discussed in section 6. To determine distinctive patterns, we used the weights assigned to the features by the classifier (Malmasi and Dras, 2014).

First, we explored the general usage pattern for *sbj+prp* and *sbj+nn* variants, detached from particular verb lemmas. That was done by cutting off the dendrograms for the [*s/c*] settings at the root (cut-off 3.0), which results in having all of the considered verbs in a single cluster and hence, just the two variants, i.e., *sbj+prp* and *sbj+nn*, encoded by the relative frequency for each text. It turned out that both weights are negative, showing that there do not seem to be any pattern indicative for L1 Chinese compared to the

[7] Best ensemble optimization parameters for all ensembles in this paper: *+all, -opt* (Bykh and Meurers, 2014).

[8] This cut-off turned out to yield best results in our evaluations on both, the TOEFL11 *test* and *development* sets, thus it seems to be relatively reliable for TOEFL11 data.

[9] "Chinese people always hold the idea that human being and nature are mingled together, so Chinese people intend to make themselves as the start to narrate object things and are used to taking the pronoun as the subject. However, in the western philosophy, object is emphasized and it is believed that human being and nature are separated. So western people intend to express things from an object view and are used to taking non-pronoun such as things or abstract concept as the subject. The choice of pronoun-subject or non-pronoun-subject between Chinese and English will lead to negative transfer of mother tongue, which will make the translation of subject an improper one." (Wang 2009, p. 139)

western L1s. Second, we bring the actual verbs back into the equation, and explore the best performing [*s/c, +bvm*] setting (cut-off 2.6). Here, all verb lemmas are grouped into five clusters. The findings are summarized in Table 2.

Cluster id	# Verb lemmas	Pattern indicative for L1 Chinese	Weight
1	166	sbj+nn	< 0.01
2	100	-	-
3	312	-	-
4	65	sbj+prp	0.73
5	68	sbj+prp	0.19

Table 2: Usage pattern for L1 Chinese at the best performing dendrogram cut-off yielding five clusters.

For the cluster 1 there is some very weak (a positive weight ≈ 0) indication for the variant *sbj+nn*, which essentially can be ignored. For the two clusters 2 und 3 there is no indicative pattern for L1 Chinese, whereas for the two clusters 4 and 5, there is a clear indicative preference for the variant *sbj+prp*. In sum, for most of the verbs in our data set, there is no indicative usage pattern for L1 Chinese compared to the western L1s. However, in connection with some particular verb groups, there is an indicative preference indeed, namely, for the variant *sbj+prp*, supporting the given hypothesis. Interestingly, the method does not simply support a known hypothesis, but it makes it possible to observe subsets of verbs for which the characteristics emerge. Studying what the 65 verbs grouped in the most indicative cluster have in common thus provides the opportunity for a more fine-grained qualitative analysis in SLA research.

8 Conclusions

In this paper, we proposed and explored a grouping technique for linguistic features. The method is inspired by a variationist linguistic perspective and uses hierarchical clustering on the basis of label-informed feature representations. The approach emphasizes how the underlying linguistic structure informs the classification label, reducing potential problems arising from idiosyncrasies and sparsity of individual features. We evaluated the approach in the context of NLI using a linguistic feature type well-suited to a variationist perspective, verb subcategorization patterns, treating the verb lemmas as variables and the different patterns as variants.

We motivated why we consider the technique to be of interest from a theoretical and a practical perspective. Grouping verb lemmas based on the subcategorization information, and thus abstracting from individual occurrences to the underlying linguistic structure, resulted in a significant improvement in terms of accuracy, confirming the hypothesis that in general learners with different L1s seem to prefer different subcategorization patterns. We then turned to investigating a particular hypothesis from SLA regarding the usage of the subject as part of the verb subcategorization information. We showed that the method can discover differences in the variant realization patterns in connection with different automatically induced classes of verbs, supporting a fine-grained qualitative analysis. Linking the analysis to a theoretical perspective informed by traditional SLA research, the method seems well capable of advancing the qualitative insights in NLI – a primary concern in that field of research today. Combining features obtained by the approach proposed in this paper with a set of previously used features resulted in an accuracy of 85.4%, which is the best result reported so far on the standard TOEFL11 data setup.

In terms of future work, we plan to explore different syntactic and morphological features under a variationist perspective, to extend the qualitative analysis, and to establish a firm enough link between the data-induced patterns and the traditional insights into L1-transfer to be able to test specific SLA hypotheses. Regarding the label-informed feature grouping technique, we are also considering applying it to NLP tasks other than NLI in order to obtain a more comprehensive assessment of the method.

References

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Demonstration Volume of the 23rd International Conference on Computational Linguistics (COLING)*, pages 23–27, Beijing, China. `https://code.google.com/p/mate-tools/`.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. Technical report, Educational Testing Service.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23th International Conference on Computational Linguistics (COLING)*, pages 89–97, Beijing, China.

Richard Butterworth, Gregory Piatetsky-Shapiro, and Dan A. Simovici. 2005. On feature selection through clustering. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'2005)*.

Serhiy Bykh and Detmar 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)*, pages 425–440, Mumbay, India.

Serhiy Bykh and Detmar Meurers. 2014. Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1962–1973, Dublin, Ireland.

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek, and Detmar 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.

Marcus Callies and Konrad Szczesniak. 2008. Argument realization, information status and syntactic weight – a learner-corpus study of the dative alternation. In Maik Walter and Patrick Grommes, editors, *Fortgeschrittene Lernervarietäten*, pages 165–187. Niemeyer.

Marcus Callies and Ekaterina Zaytseva. 2011. The corpus of academic learner English (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties. In *Hedeland, Hanna and Thomas Schmidt and Kai Wörner*, pages 51–56. Multilingual Resources and Multilingual Applications (Hamburg Working Papers in Multilingualism B 96).

R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874. Software available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

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. ACL.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy 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.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05)*, pages 624–628, New York.

C. Krier, D. François, F. Rossi, and M. Verleysen. 2007. Feature clustering and mutual information for the selection of variables in spectral data. In *Proceedings of the ESANN'2007*, pages 157–162, Bruges, Belgium.

Julia Krivanek. 2012. Investigating syntactic alternations as characteristic features of learner language. Master's thesis, University of Tübingen, April.

William Labov. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

Anke Lüdeling. 2011. Corpora in linguistics: Sampling and annotation. In Karl Grandin, editor, *[Nobel Symposium 147] Going Digital: Evolutionary and Revolutionary Aspects of Digitization*, pages 220–243, New York. Science History Publications.

Shervin Malmasi and Aoife Cahill. 2015. Measuring feature diversity in native language identification. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-10) at NAACL-HLT 2015*, pages 49–55, Denver, Colorado.

Shervin Malmasi and Mark Dras. 2014. Language transfer hypotheses with linear SVM weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1385–1390. ACL.

Detmar Meurers, Julia Krivanek, and Serhiy Bykh. 2014. On the automatic analysis of learner corpora: Native language identification as experimental testbed of language modeling between surface features and linguistic abstraction. In *Diachrony and Synchrony in English Corpus Studies*, pages 285–314, Frankfurt am Main. Peter Lang.

Cheong Hee Park. 2013. A feature selection method using hierarchical clustering. In R. Prasath and T. Kathirvalavakumar, editors, *Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration (MIKE 2013), LNAI 8284*, pages 1–6, Virudhunagar, Madurai, India. Springer International Publishing Switzerland.

John Pate and Detmar Meurers. 2007. Refining syntactic categories using local contexts – experiments in unlexicalized pcfg parsing. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)*, Bergen, Norway.

David Stringer. 2008. What else transfers? In Roumyana Slabakova et al., editor, *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007)*, pages 233–241, Somerville, MA. Cascadilla Proceedings Project.

Ben Swanson and Eugene Charniak. 2013. Extracting the native language signal for second language acquisition. In *Proceedings of NAACL-HLT*, pages 85–94. ACL.

Ben Swanson and Eugene Charniak. 2014. Data driven language transfer hypotheses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 169–173. ACL.

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

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*, Atlanta, GA, USA, June. ACL.

Yukio Tono. 2004. Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In Guy Aston, Silvia Bernardini, and Dominic Stewart, editors, *Corpora and Language Learners*, pages 45–66. John Benjamins.

Xiaoru Wang. 2009. Exploring the negative transfer on english learning. *Asian Social Science*, 5(7):138–143.