Automated L1 identification in English learner essays and its implications for language transfer

Egon Stemle (EURAC Bolzano) and Alexander Onysko (University of Klagenfurt)[1]

This article focuses on automatic text classification which aims at identifying the first language (L1) background of learners of English. A particular question arising in the context of automated L1 identification is whether any features that are informative for a machine learning algorithm relate to L1-specific transfer phenomena. In order to explore this issue further, we discuss the results of a study carried out in the wake of a Native Language Identification Task. The task is based on the TOEFL11 corpus (cf. Blanchard et al. 2013), which involves a sample of 12,100 essays written by participants in the TOEFL® test from 11 different language backgrounds (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish). The article will show our results in automatic L1 detection in the TOEFL11 corpus. These results are discussed in light of relevant transfer features which turned out to be particularly informative for automatic detection of L1 German and L1 Italian.

---

[1] We would like to thank an anonymous reviewer for helpful comments on an earlier version of the paper.

1. Introduction

In the field of research on language transfer, computational means of authorship identification are a fairly recent development that can provide empirical insight in the relevance of transfer phenomena among language learners. As described in the first volume devoted to this topic (Jarvis and Crossley 2012), automatic recognition of transfer is based on the presupposition that a learner's L1 will influence the use of the learner language (Jarvis 2012: 1). Furthermore, if groups of learners sharing the same L1 background are considered, it is likely that they will show similar patterns in using the learner language. In other words, language learners of the same L1 will exhibit intragroup homogeneity while learners from different L1 backgrounds will be heterogeneous to each other (Jarvis 2012: 5). Based on this premise, computational calculations of different textual features are supposed to bring to light such intergroup differences. Thus, if other factors are sufficiently controlled for, a classification of learner texts according to the L1 of their authors becomes possible.

Apart from the obvious benefit of being able to process large amounts of language data by computational means, automated classification of learner texts also allows taking a detection-based approach to possible transfer effects. By feeding computer classifiers with general parameters for calculating textual features relating to, for example, text size, word choice,

punctuation, parts of speech, and syntactic information, the results of such classifications can provide evidence for L1-based patterns of transfer that emerge as characteristics of a specific learner group. It is intrinsic to this type of automated approach to transfer as measurable crosslinguistic influence (cf. Jarvis and Pavlenko 2008) that primarily instances of negative transfer are detected. These appear as structures and patterns of language which either stick out from native language use or from the comparison with observable patterns of other learner groups.

Embracing the potential of automated L1 identification for the investigation of language transfer, our paper reports on a study conducted as part of a Native Language Identification Task (cf. Tetreault, Blanchard, and Cahill 2013) open to participants in spring 2013. The task was carried out on the basis of 12,100 written TOEFL® texts (short argumentative essays) from learners of English involving 11 different first languages. In this paper, we will show the results of our automated classification and then focus on some of the patterns that are particularly informative for successfully detecting the L1 backgrounds of German and Italian learners of English. These patterns will be discussed in light of their origin as potential transfer effects. Before presenting the results and discussing the role of transfer, the next section will provide an overview of previous research on transfer in computational L1 classification, and Section 3 will lay out the methodology of the task and our computational approach.

2. Previous research on transfer in automated L1 identification

Research on L1 identification[2] has grown out of the field of stylometry, which is concerned with authorship attribution based on statistical calculations of textual features (cf. Barr 2003). Classifying texts by the L1 of their speakers is also generally related to research on automated text classification, which frequently applies machine learning algorithms to sort texts by their type or author attributes. For example, Baroni and Bernardini (2006) employ support vector machines to successfully differentiate between original and translated Italian texts in 86.7% of their corpus of Italian articles.

As pointed out in Jarvis (2012: 14), the first study applying means of automated text classification according to the L1 of an author is described in a paper by Mayfield Tomokiyo and Jones (2001), which aimed at distinguishing between Chinese and Japanese learners of English. Further studies focusing on other L1 backgrounds of learners of English followed

[2] The term Native Language Identification is also used synonymously by some authors in this field. When investigating transfer, however, the notion of L1 (or first language) is more adequate as it refers to language dominance rather than sequential exposure to language. In addition, the notion of L1 supports a flexible conception of a person's language competence. Thus, particularly in multilingual contexts, a person might grow up with more than one native language that can be perceived differently in terms of their proficiency or dominance. Throughout the course of one's life, proficiency and dominance of a language can shift or be flexible depending on changing usage contexts or certain situations of use. The notion of (shifting) L1 describes such dynamic situations whereas the term native language mainly refers to the language a speaker is first exposed to in life.

over the next few years (Koppel et al. 2005, Estival et al. 2007, Tsur and Rappoport 2007, Wong and Dras 2009). These studies were mostly targeted at experimenting with different types of textual features to optimize the results of computational L1 classification. Apart from an occasional mention of L1 patterns that occurred in the different classification experiments (cf. e.g. Wong and Dras 2011), these studies did not explicitly consider the role of L1 transfer in automated classifications. Jarvis and Crossley (2012) is the first volume that brings L1 transfer into the picture of text classification. The book includes five studies exploring different features of learner texts for automated classification. Four of these studies are based on texts taken from the International Corpus of English (ICLE), and one investigation focuses on written English narratives prompted by an excerpt from a Charlie Chaplin film.[3] The studies in the volume are similarly structured in that the individual research designs and the results of the classification procedures are presented before a final discussion highlights how some of the informative classification patterns are related to the L1 backgrounds of their authors.

The study by Jarvis, Castañeda-Jiménez, and Nielsen (2012) is based on a corpus of English texts written by 446 foreign-language learners of English from five L1 backgrounds (Danish, Finnish, Portuguese, Spanish,

[3] So far, almost all research on L1-based text classification has focused on learners of English. A few exceptions are Aharodnik et al. (2013), who focus on learners of Czech, as well as Golcher and Reznicek (2011), who use the Falko corpus of German as a learner language.

and Swedish). The texts are prompted by a sequence taken from a Charlie Chaplin silent movie. Relying on Linear Discriminant Analysis (for details see McLachlan 2004), the authors take a lexical approach to text classification. In detail, they select the 30 most frequent words in the texts of each L1 group. Overlap between the words reduces the final feature set for their classifier to a total of 53 function and content words. The particular distributions of these 53 words in the learner texts as determined by a stepwise feature selection integrated in standard 10-fold cross validation results in a classification accuracy of 76.9%. Among the words that have the most discriminatory value, the usage of the determiners *a* and *the* indicates L1 influence. Thus, the fact that Finnish lacks articles leads to a significantly lower use of articles in Finnish learner English while the wider distribution of the definite article in Spanish and Portuguese induces a relative overuse of *the* in comparison to Danish and Swedish speakers of English (Jarvis, Castañeda-Jiménez, and Nielsen 2012: 61).

In a second study, Jarvis and Paquot (2012) further explore the role of lexical patterns for L1 identification. This time the authors rely on learner texts in ICLE and expand the range of L1 backgrounds to twelve languages. Lexical n-grams are selected as features for text classification. Jarvis and Paquot consider the 200 most frequent lexical 1-grams (i.e. single words) and the multiword combinations of the 200 most frequent lexical 2-grams, lexical 3-grams, as well as the 122 most frequent 4-grams found in the data. Their results show that, first of all, classification by lexical 1-grams is much

more accurate than classification by multiword combinations. Secondly, while a combination of 1-grams with multiword n-grams shows consistently better results, only the combined use of unigrams and bigrams leads to a significant improvement of classification results, with the highest accuracy of 53.6% (2012: 91). When considering lexical n-grams that are powerful indicators of L1 differences, the authors provide some evidence for their possible relation to L1 transfer. The bigram, *we can*, for example, is typically overused by Spanish and Italian learners of English because of the wider usage scope of the Spanish and Italian modals *poder* and *potere*. Furthermore, the bigram *going to* is highly indicative of Spanish learners in the selected ICLE texts. In this case, Spanish has a similar construction to mark futurity, *ir a+ infinitve* (2012: 96-97).

The study by Crossley and McNamara (2012) uses the more abstract textual features of cohesion, lexical sophistication, syntactic complexity, and conceptual knowledge to construct a computational model that is tested on 900 argumentative essays from ICLE written by learners of English from four L1 backgrounds (Czech, German, Finnish, and Spanish). While their model shows a success rate of 66% in correctly predicting L1, the bundle of measures used to discriminate between the learner groups is more difficult to interpret in terms of L1 transfer effects.

Similarly, the approach taken by Bestgen, Granger, and Thewissen (2012) of using error patterns for automated L1 identification emphasizes the fact that transfer is not the only reason for differences between groups of

learners. The authors apply seven error domains as features for classification, which yields an accuracy of 65% in the error tagged subset of ICLE (consisting of 223 learner essays). At the same time, the authors stress the importance of controlling for learner proficiency in line with the observation that language learners rely less on (negative) transfer the higher their level of proficiency (cf. e.g. Taylor 1975).

Apart from these studies on the relation between automated L1 classification and transfer, research on L1 identification has seen a recent boost due to an open call to participate in the first shared task in Native Language Identification (cf. Tetreault, Blanchard, and Cahill 2013). 29 teams participated in the shared task and many of the individual solutions for L1 identification are gathered in the *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. As described in more detail in Section 3, the shared task was based on the TOEFL11 corpus (cf. Blanchard et al. 2013), which was specifically designed to meet the demands of automated L1 identification. This makes TOEFL11 better suited to the task than the comparably smaller ICLE corpus (Tetreault, Blanchard, and Cahill 2013: 48).

Considering the overall results in the task, the best classification accuracies ranged between 80% and 84%, which was achieved by the submissions of 13 teams (Tetreault, Blanchard, and Cahill 2013: 53). Among these submissions, lexical features played a key role for successful classifications. Thus, the authors obtaining the highest accuracy score in the

main task conclude from their results that "the most reliable L1 specificity in the TOEFL11 is to be found simply in the words, word forms, sequential word combinations, and sequential POS [part of speech] combinations that the nonnative writers produced" (Jarvis, Bestgen, and Pepper 2013: 117). This observation emphasizes the fact that even if a whole range of parameters are used to train computer classifiers including lexical, syntactic, and stylistic features, as well as dependency parsers and grammatical errors, it is striking that high baselines of classification can be achieved by a simple combination of lexical n-grams and character n-grams using support vector machines (cf. Tetreault, Blanchard, and Cahill 2013: 54-56). Several studies report that, among a mix of features, lexical unigrams and bigrams contribute most to their classification accuracies accounting for baselines close to 80% (cf. Gebre et al. 2013, Wu et al. 2013, Brook and Hirst 2013). The importance of lexical n-grams and character n-grams for successful L1 identification has also been shown in other research on different corpora (e.g. Ahn 2011, Van Halteren 2008, Tsur and Rappoport 2007). This leads to the conclusion that a speaker's L1 background influences her/his lexical choice in English, which, in turn, could be based on certain transfer effects from the L1.

In our paper, we would like to explore this relation further. For that, we first of all build a general computational model based on simple grammatical, orthographical, and lexical features for classifying English texts in TOEFL11 according to the L1 of their authors. The results of

applying this model to the corpus are then analyzed to find the most distinctive features in the texts of L1 German and Italian learners of English. These features are finally discussed for their potential to indicate L1 transfer in the English texts.

3. Methods

A few methodological issues lie at the core of every effort in automated L1 identification. This is, first of all, the design of the database or corpus used for the task. Moreover, performance in automated text classification is also dependent on the kind of computational classifier as well as on the type and amount of information which is applied to guide the classifier in making decisions on the L1 of language learners. Below, these aspects are addressed in the context of our study.

3.1 The design of TOEFL11

As mentioned in section 2, TOEFL11 has been compiled to meet the needs of automated L1 classification better than previously used corpora and collections of texts. The compilers of the corpus describe TOEFL11 as consisting of short essays written during TOEFL® examinations in 2006 and 2007 by learners of English of eleven L1 backgrounds: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and

Turkish (Blanchard et al. 2013). The corpus comprises 1,100 English texts for each of the L1s and care has been taken to sample texts as evenly as possible among 8 topic prompts and three ratings given by human examiners for the learners' written proficiency levels (low / medium / high). According to Blanchard et al. (2013), the overall size of 12,100 texts (corresponding to the same number of learners of English) renders TOEFL11 a far bigger publicly available balanced corpus of written learner English than other learner corpora such as the *International Corpus of Learner English* (ICLE; Granger et al. 2009) or the *Cambridge First Certificate of English* dataset (cf. Yannakoudakis, Briscoe, and Medlock 2011). However, as pointed out by Jarvis, Bestgen, and Pepper (2013: 113-114), TOEFL11 is not perfectly balanced for the number of essays per prompt and language. For example, the number of essays for prompt 6[4] is particularly low for German, Hindi, Italian, Telugu, and Turkish compared to the other languages (cf. Blanchard et al. 2013: 12). In addition, there is even more variability in the distribution of texts for proficiency level and language. In an ideal case, about 33% of all texts per language would cohere with one of the three proficiency levels (low / medium / high). However, Jarvis, Bestgen, and Pepper (2013: 113) point out that the distribution is highly skewed for a few languages. In general, most of the essays in all

[4] Prompt 6: Do you agree or disagree with the following statement? The best way to travel is in a group led by a tour guide. Use reasons and examples to support your answer.

languages fall into the medium range of proficiency. By comparison, the number of texts rated as high is considerably smaller, and texts having low proficiency ratings are few and far between (cf. Blanchard et al. 2013: 13). In the examples of L1 German and Hindi merely 1.4% and 2.5% of texts are rated as low compared with 61.5% (German) and 57.6% (Hindi) rated as high (cf. Jarvis, Bestgen, and Pepper 2013: 113). Despite these imbalances among prompts and proficiency levels, TOEFL11 remains the most extensive and balanced resource for L1 identification so far and could thus provide some interesting insight into characteristic patterns of L1 transfer in learners of English.

3.2 Automated classification and feature selection

The TOEFL11 data set has been prepared for applying machine learning (ML) algorithms as it is divided into a training set (9,900 texts), a development set (1,100 texts), and a test set (1,100 texts). All texts have been tokenized. For automatic classification and data analysis, we used the Scikit-learn Python package version 0.13 (Pedregosa et al. 2011). The ML algorithm (LinearSVC with standard parameter settings) was modeled on the development set of 1,100 texts, and its performance was tested with 10-fold cross-validation on the larger set of 9,900 texts. The results of the classification for the eleven languages in the corpus are shown in Section 4. Apart from the classification, we also wanted to take a closer look at the most discriminative features for the machine learning algorithm as these

features represent L1-specific patterns in the English texts. It is interesting to investigate whether the most informative features for the ML algorithm can be related to L1-specific transfer effects that distinguish one group of learners of English from another one. For the scope of this paper, we focus on the features which proved to be particularly indicative for L1 German and L1 Italian learners of English. In detail, the amount of a feature in the 900 texts each for German and Italian is compared to the remaining body of the TOEFL11 training set consisting of 9000 learner texts from the other ten L1 backgrounds. Thus, the informative features indicating either German or Italian L1 emerge from a comparison with their distribution in other learner texts and not with a comparison of texts by L1 English speakers. The informative features were ranked according to ANOVA F-score calculations provided by Scikit-learn.

As discussed in Section 2, the type of features that are selected for training an ML algorithm are crucial for achieving a high accuracy in text classification. The most successful approaches to determining the L1 background in the shared task based on the TOEFL11 corpus (cf. Blanchard et al. 2013) used a combination of lexical n-grams (mostly from single words up to 5 word combinations) and character n-grams. While these features cover the characteristics of learner texts to a high degree, their results are often difficult to interpret in terms of possible transfer effects from a learner's L1. Since the detection of possible transfer effects is the aim of our study, we have opted for a mixed methodology of feature

selection combining introspection with automated extraction while keeping the overall amount of features low, not exceeding 400. In detail, introspective feature selection relied on observations in a sample of 30 texts per L1 background, which led to the formation of hypotheses on group specific patterns. Automatic feature extraction, on the other hand, drew on the results of n-grams which were generated from the development set of 1,100 texts. More specifically, all combinations of tokenized items in the texts (e.g. words and punctuation marks) were automatically computed for combinations of two tokens up to five tokens. In order to build the complete set of 400 features from these two different ways of selection, we first of all included all observation-based hypothetical features and added the most discriminating n-gram combinations, leading to an overall distribution of 216 observation-based and 184 automatically extracted features. For reasons of space, the complete lists of all the features used for automatic L1 identification are available on the web.[5]

In linguistic terms, our features relate to four different characteristics of the learner texts: 1) text surface features, 2) grammatical and discourse features, 3) orthographical features, and 4) derivational and lexical features. Text surface features comprise the number of characters, digits, tokens (i.e. words and punctuation marks), sentences, and paragraphs. In addition, we

[5] The set of features, the complete classification results, and the computational implementation will be freely available at https://bitbucket.org/commul/2013_nli-st after publication.

also measured the number of words with initial capital letters, with all capital letters as well as the number of sentences starting without a capital letter and the number of hyphenated words. The last two textual features were added as the observation of the sample texts indicated that writers of different L1 backgrounds differed in their use of capitalization. Checking the amount of hyphenated words relates in particular to learners of English of an L1 German background as it is sometimes observed that compounding in German more frequently involves hyphenization of its constituents compared to English and other language standards.

Among grammatical and discourse features, we took a lexical approach to identifying any patterns that could be indicative of L1 transfer. This means that we considered specific lexical items and character combinations which are indicative of certain grammatical aspects, morphological forms, and discourse related features. Table 1 provides an overview of the grammatical and discourse features and their corresponding lexical and character combinations used in the automated text classification.

Table 1: Grammatical and discourse features implemented in automated L1 identification

Grammatical / Discourse features	**Search strings**
Adverbs	<ly> word finally

Article usage	<a>; <an>; <the>
Conjunctions / Connectives	<and>; <but>; <or>; <who>; <which>; <that>; <so>; <as>; <however>; <therefore>; <thus>; <because>; <while>; <when>; <if>; <nevertheless>; <despite>; <although>; <on the other hand>; <in fact>; <actually>; <for example>; <for instance>; <indeed>; <yet>; <also>; <furthermore>; <in addition>; <besides>; <apart from>; <whereas>; <during>; <instead>; <still>; <then>; <now>; <there>; <similarly>; <otherwise>; <in this case>; <in that case>; <consequently>; <as a result>; <before>; <after>; <until>; <in order to>; <in order that>
Clitics	<'ll>; <n't>; <'m>; <'d>; <'ve>; <'re>;
Demonstrative pronouns	<this>; <these>; <that>; <those>
Distributives / Quantifiers	<each>; <every>; <all>; <either>; <neither>; <some>; <any>; <many>; <much>; <a lot>; <few>; <several>; <both>; <such>
Intensifiers	<very>; <quite>; <extremely>; <rather>; <even>; <just>; <only>; <really>; <more>; <most>; <already>; <absolutely>
Main modals	<can>; <would>; <will>; <could>; <may>; <might>; <should>; <must>; <able to>; <have to>
Personal and possessive pronouns	<I>; <you>; <he>; <she>; <it>; <we>; <they>; <me>; <my>; <mine>; <your>; <yours>; <his>; <her>; <hers>; <him>; <our>; <ours>; <them>; <their>; <theirs>
Prepositions	<of>; <off>; <for>; <in>; <at>; <on>; <out>; <from>; <about>; <with>; <into>; <onto>; <under>; <within>; <without>; <by>; <underneath>; <beneath>; <above>; <below>; <through>;

	<across>; <along>; <away>; <up>; <down>; <over>; <in front of>; <behind>; <between>; <among>
Progressive aspect	<ing> word finally
Reflexive pronouns	<self>; <myself>; <yourself>; <himself>; <herself>; <itself>; <ourselves>; <yourselves>; <themselves>
Saxon genitive	<'s>
Verbal infinitive	<to>
Verb *be*	<be>; <is>; <are>; <am>; <was>; <were>; <been>; <being>

As Table 1 shows, an approach purely based on lexical and certain character string combinations allows capturing some of the peculiar grammatical and discourse related aspects of the English language. However, this approach also has some limitations. First of all, it is far from being exhaustive of grammatical and discourse characteristics of English. Secondly, the lexical mapping of grammatical and discourse features can also bear the danger that not all of the chosen indicators are actually representative of a specific grammatical or discourse pattern. In some cases there is also an overlap between the categories and their indicators. This, for example, occurs in the use of clitics and the Saxon genitive. In our study, we take the string <'s> as indicative of the possessive construction. However, the same string also designates the cliticized version of *is* or *has*, which has to be considered when interpreting results on <'s>. Similarly, counting the amount of <to> might not only be indicative of whether an L1 prefers verbal constructions over a nominal style, but it can also relate to the prepositional usage of *to*. In

order to minimize the skewing of results when probing for the use of progressive aspect, we excluded some common words ending in <ing> from the results (e.g. *thing*, *something*, *everything*, *according*, *evening*, *morning*, *sing*, *king*, and so on). Furthermore, the lists of conjunctions/connectives, distributives/quantifiers, and intensifiers merely represent some common members of these categories. Neither are they to be regarded as closed and comprehensive lists, nor can all of the items be considered as only carrying this particular discourse function. An interpretation involving any of these categories will have to be made on an individual basis.

The limitations for the grammatical and discourse features also apply to some extent to the orthographical features and particularly to the derivational features selected for this study. The orthographical features consist primarily of the regular set of punctuation marks (<.>; <,>; <;>; <:>; <->; </>; <?>; <!>) and the amount of misspelled words per text as indicated by the widely used and freely available spell checker *Hunspell* (cf. http://hunspell.sourceforge.net/). The derivational features are based on the hypothesis that learners from a Romance L1 background might use more English words of historically Romance roots. This is why we added up the numbers of words ending in <ment>, <ion>, and <ize> as an indicator of Romance vocabulary in English.

Whenever a single feature represents a category of features, we counted the single instances for each item and the summary score of all items in a category. For example, the category of article use consists of the

items <a>, <an>, and <the>, each of which are counted individually. Furthermore, a summary score was also calculated for all three articles and compared among the different learner populations.

With the exception of a few lexical items (e.g. *particular/particularly* and *special/especially*), the lexical features in our classification scheme relate to the automatically generated set of n-grams (2-grams to 5-grams) from the development set of the corpus. The complete list of 184 features included in the classification task can be found on the web (cf. note 5). These features are most informative for distinguishing the 11 languages in the development set. To highlight just a few characteristics of the automatically generated discriminators, it is notable that there is quite a substantial number of bigrams involving a punctuation mark and a lexical item (e.g. <. indeed>; <however ,>; <, i>; <, the>). In addition, the list consists only of bigram combinations with the exception of 8 trigrams (<in a group>; <a successful people>; <to conclude ,>; <now a days>; <knowledge of all>; <as i am>; <person who is>; <one specific subject>) and one 4-gram (<. for example ,>). This outcome ties in with the findings by Jarvis and Paquot (2012), who report that lexical unigrams and bigrams contribute most to the accuracy of L1 identification of English learner texts.

For the analysis of the results in the next section, only the most significant observation-based and automated features are taken into account for classifying English texts written by L1 German and L1 Italian speakers. Finally, it has to be emphasized that this approach of testing relevant

features for classifying the L1 background of German and Italian learners is one that generates hypotheses on possible L1 transfer patterns. Ideally, each of the hypotheses would have to be empirically tested in follow-up studies.

4. Results

In this section, we would like to present two sets of results. Accuracies in classification and a confusion matrix will show the performance of our specific set of features in automatically detecting the L1 background of learner texts in TOEFL11. This is followed by an overview of the most informative features for classifying texts written by L1 Italian and L1 German learners of English.

4.1. Results of the classification task

The overall performance of the ML algorithm fed with our selection of 400 features on the training set of TOEFL11 is displayed in Table 2. Separate scores for each of the 11 L1 learner groups indicate the amount of successfully classified texts according to the standard measures of precision[6]

[6] Precision is calculated as the number of correctly identified texts divided by the sum of correctly and incorrectly retrieved texts.

and recall.[7] The harmonic mean of precision and recall (i.e. the f1-score) is taken as the indicator of classification accuracy. The average f1-score across all languages represents the overall accuracy in L1 identification. On the whole, our classification system performed at an accuracy of 0.59, which means that 59% of all texts have been accurately identified according to their authors' L1 backgrounds from the pool of 11 languages.

Table 2: Accuracy of automated L1 identification in TOEFL11

L1	**Precision**	**Recall**	**F1-score**	**Number of texts**
Arabic	0.60	0.62	0.61	900
Chinese	0.61	0.63	0.62	900
French	0.61	0.57	0.59	900
German	0.66	0.72	0.69	900
Hindu	0.53	0.53	0.53	900
Italian	0.62	0.66	0.64	900
Japanese	0.60	0.60	0.60	900
Korean	0.57	0.52	0.54	900
Spanish	0.53	0.49	0.51	900
Telugu	0.63	0.63	0.63	900
Turkish	0.55	0.55	0.55	900
Average/total	**0.59**	**0.59**	**0.59**	**9900**

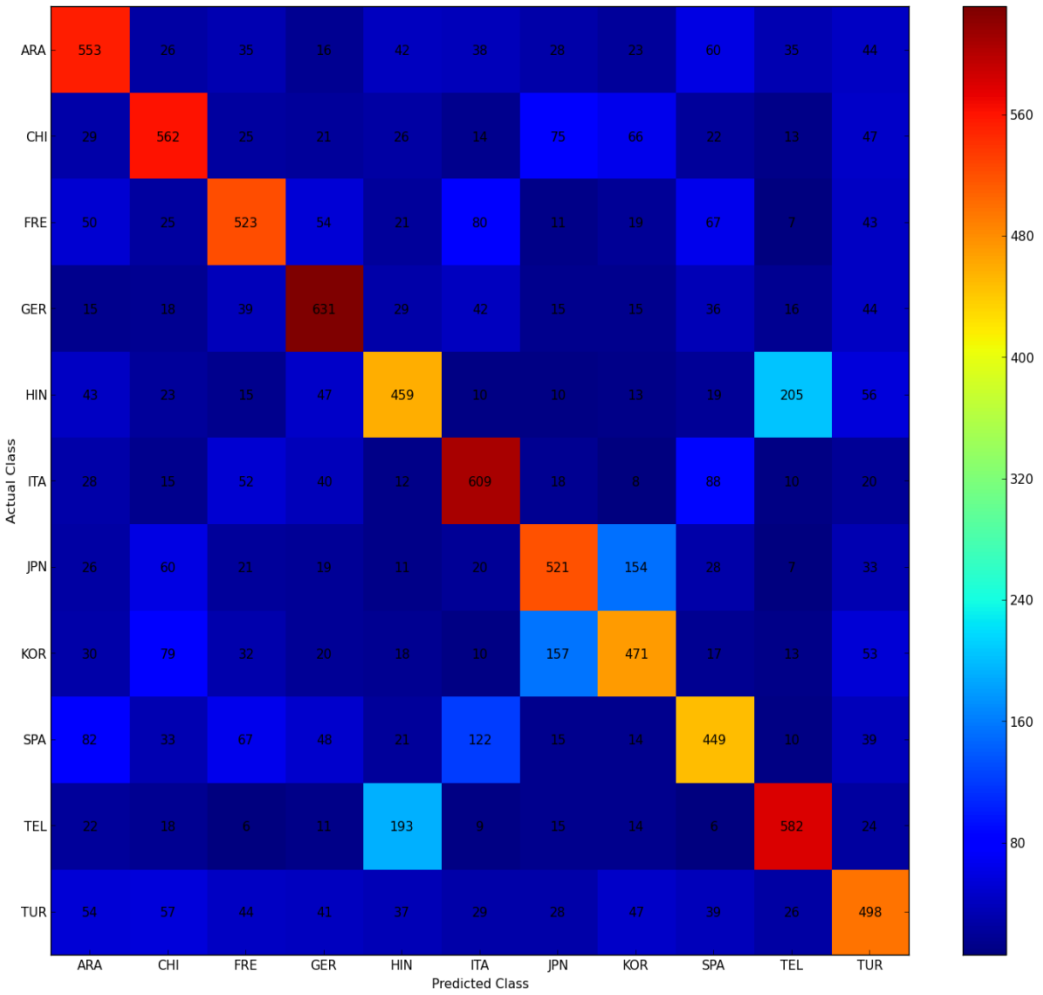
The results show that there is some variation in the accuracy of L1 classification as measured by f1-scores. Identifying L1 German and L1 Italian authors is most accurate while TOEFL essays written by Spanish

[7] Recall represents the ratio of the number of correctly retrieved texts from the set of relevant texts.

learners of English exhibit the lowest rate of identification. This can be interpreted as an indication that the selected set of features is particularly relevant for capturing L1 Italian and L1 German profiles in English learner texts; or, to put it in more cautious terms, we can say that English texts written by L1 German and L1 Italian learners in TOEFL11 are most likely to be discriminated from all other learner languages in the corpus according to the selected features for text classification.

To get a more precise picture of the classification results, it is interesting to consider the distribution of correctly and incorrectly assigned texts for each of the languages. This information is given in the confusion matrix shown in Figure 1.

Figure 1: Confusion matrix of L1 classification in TOEFL11



Read horizontally, the cells in the confusion matrix show how many of the texts in a language were classified correctly and how many were incorrectly assigned to the other languages in the corpus. The values forming the diagonal of the matrix represent the number of correctly assigned texts for each language. These values are consistently and by far the highest of the other cells in the matrix, highlighting the overall accuracy in L1 identification. As indicated by their f1-scores in Table 2, German and Italian

achieve the highest numbers of correctly identified texts. The confusion matrix also shows some interesting language family and areal effects among the learners' first languages represented in the corpus. Thus, automatic classification confuses English texts by Japanese and Korean learners to a great extent. Similarly, texts written by learners pertaining to the Romance language family show a higher number of misattributions among each other. This is particularly noticeable in English texts composed by Italian and Spanish learners. Although they do not belong to the same language family, the areally related languages of Hindu and Telugu show the largest amount of mix-up in automatic classification. This is a token of their long history of intense language contact, which has led to some convergence between Indic and Dravidian languages (cf. Thomason 2001: 116).[8] Altogether, the emerging effects of language family and language contact in the confusion matrix emphasize the fact that the L1 backgrounds of learners of English leave an identifiable footprint in their use of the English language as represented in TOEFL® test essays. Since this footprint transcends the boundaries of individual languages and holds for learners of areally and typologically related languages, the hypothesis is strengthened that L1 transfer underlies some of the patterns that are helpful in automatic classification. In order to explore the role of transfer further, the next section

[8] Widespread multilingualism among learners of English in India might be a further factor heightening convergence between the learner populations.

will focus on important features for identifying L1 German and L1 Italian as they achieved the best results in the classification.

4.2 Features for classifying English texts in TOEFL11 for L1 German and L1 Italian

When investigating possible transfer effects emerging from automated classification, it is important to focus on the features that are most helpful for the classifier in discriminating L1 German and L1 Italian from the other languages in the corpus. Therefore, we carried out ANOVA F-score calculations for the 400 features of the classifier, comparing the relative amount and distribution of each feature in L1 German and in L1 Italian texts with the other 10 languages. All the features whose distributions show a significant difference for L1 German and L1 Italian are closely considered for whether they possibly reflect a particular pattern of language use in the L1 of the learners.

For L1 German, 84 automatically generated lexical n-grams and 97 observation-based features turn out to be significant indicators ($p<0.05$). Similarly, 86 automatically generated lexical n-grams and 98 observation-based features achieve significance ($p<0.05$) for L1 Italian. The complete tables of significant features are available on the web (cf.note 5). For the sake of illustration, only the 50 most significant features for each L1 are shown in Table 3 and 4. The features are ranked from 1 to 50 according to

their F-score, and their p-values are close to zero ($p<0.01$) indicating highly language specific effects for L1 German and L1 Italian.

In addition to their score and ranking, the tables show the average rate of each feature in an L1 text (Num/900). This figure is compared to the average occurrence of the feature in a text pertaining to the other L1s in the corpus (Num/9000). To show the dispersion of a feature across the learner texts for L1 German and for L1 Italian, the tables give further information on the number and ratio of texts in which the feature occurs. The overall number of texts containing the feature is furthermore split into texts of a medium and of a high proficiency level. The rightmost column in each table indicates the features, which appear as actual search strings or class names as shown in appendices 1 and 2. To differentiate between automatically generated n-gram features and the observation-based criteria, a difference has been made in the capitalization of the letter *n* (number) preceding the feature name: a small *n* flags the automatically generated token n-grams while a capital *N* signifies observation-based features.

Table 3: The 50 most significant features (of 181) for classifying L1 German

Rank	**F-score**	**Num/ 900**	**Num/ 9000**	**Num texts**	**Num texts (med)**	**Num texts (high)**	**Feature name**
1	538.88	0.56	0.13	283 / 0.31	112 / 0.33	167 / 0.30	n_, that
2	253.05	0.40	0.08	165 / 0.18	50 / 0.15	113 / 0.21	N_-

3	212.74	0.86	0.44	486 / 0.54	200 / 0.59	278 / 0.51	n_. but
4	195.92	0.32	0.11	221 / 0.25	79 / 0.23	140 / 0.26	N_ESPECIALLY_VOC_all
5	185.61	0.25	0.08	187 / 0.21	47 / 0.14	139 / 0.25	n_of course
6	183.17	0.64	0.28	339 / 0.38	118 / 0.35	220 / 0.40	n_able to
7	168.98	0.20	0.05	129 / 0.14	29 / 0.09	100 / 0.18	n_a certain
8	166.31	0.94	0.49	429 / 0.48	162 / 0.48	264 / 0.48	n_have to
9	164.34	0.10	0.02	93 / 0.10	41 / 0.12	51 / 0.09	n_one hand
10	141.79	2.47	1.70	780 / 0.87	283 / 0.84	489 / 0.89	N_or
11	125.92	8.42	6.58	893 / 0.99	333 / 0.99	549 / 1.00	N_a
12	122.25	0.73	1.84	262 / 0.29	94 / 0.28	163 / 0.30	N_we
13	120.69	7.55	6.16	896 / 1.00	334 / 0.99	550 / 1.00	N_INTENSIFIERS_VOC_all
14	119.44	15.70	13.53	899 / 1.00	336 / 1.00	550 / 1.00	N_BES_VOC_all
15	117.41	0.45	0.19	224 / 0.25	90 / 0.27	129 / 0.24	n_you have
16	116.19	0.17	0.05	115 / 0.13	53 / 0.16	61 / 0.11	N_special
17	116.15	1577.45	1441.17	900 / 1.00	337 / 1.00	550 / 1.00	N_CHARS
18	111.45	0.33	0.15	221 / 0.25	65 / 0.19	156 / 0.28	N_still
19	109.84	0.53	0.25	273 / 0.30	80 / 0.24	193 / 0.35	N_might
20	108.87	377.81	347.06	900 / 1.00	337 / 1.00	550 / 1.00	N_TOKS
21	107.80	0.89	0.55	480 / 0.53	175 / 0.52	303 / 0.55	n_on the
22	107.33	1.00	0.62	494 / 0.55	171 / 0.51	320 / 0.58	n_to be
23	104.39	0.10	0.30	79 / 0.09	30 / 0.09	48 / 0.09	n_. for example ,
24	103.95	0.33	0.79	184 / 0.20	51 / 0.15	133 / 0.24	n_, and
25	100.38	7.16	9.58	883 / 0.98	331 / 0.98	540 / 0.98	N_MISSPELLED
26	100.31	3.50	2.64	824 / 0.92	297 / 0.88	519 / 0.94	N_be
27	98.53	0.43	0.23	286 / 0.32	95 / 0.28	190 / 0.35	n_of a
28	97.00	0.40	0.20	236 / 0.26	101 / 0.30	131 / 0.24	n_, because
29	95.37	0.61	0.31	278 / 0.31	105 / 0.31	165 / 0.30	n_if you
30	94.60	0.05	0.33	43 / 0.05	16 / 0.05	24 / 0.04	n_we can
31	93.87	0.44	0.25	321 / 0.36	123 / 0.37	195 / 0.36	n_the statement
32	90.57	1.24	0.83	578 / 0.64	197 / 0.59	375 / 0.68	N_an
33	88.54	2.23	1.64	731 / 0.81	271 / 0.80	453 / 0.82	N_this
34	85.07	28.72	25.50	899 / 1.00	336 / 1.00	550 / 1.00	N_PREPOSITIONS_VOC_all
35	84.57	3.42	2.70	835 / 0.93	308 / 0.91	515 / 0.94	N_for
36	81.09	0.10	0.37	75 / 0.08	23 / 0.07	52 / 0.10	n_, we
37	79.64	2.01	1.47	723 / 0.80	267 / 0.79	452 / 0.82	N_on
38	75.54	3.62	2.25	544 / 0.60	220 / 0.65	315 / 0.57	N_you
39	73.36	12.69	11.14	897 / 1.00	336 / 1.00	550 / 1.00	N_to
40	73.36	12.69	11.14	897 / 1.00	336 / 1.00	550 / 1.00	N_TO_all

41	70.55	0.53	0.33	353 / 0.39	123 / 0.37	224 / 0.41	n_in my
42	68.69	17.02	15.43	900 / 1.00	337 / 1.00	550 / 1.00	N_.
43	68.18	29.38	26.62	899 / 1.00	336 / 1.00	550 / 1.00	N_CONJUNCTIONS_VOC_all
44	67.80	3.80	3.01	822 / 0.91	308 / 0.91	506 / 0.92	N_it
45	67.20	17.77	16.19	900 / 1.00	337 / 1.00	550 / 1.00	N_SENTS
46	64.07	4.83	4.41	900 / 1.00	337 / 1.00	550 / 1.00	N_PARAS
47	62.61	0.06	0.01	52 / 0.06	24 / 0.07	28 / 0.05	n_anymore .
48	61.83	0.26	0.12	166 / 0.18	64 / 0.19	97 / 0.18	n_. you
49	61.58	0.21	0.40	159 / 0.18	50 / 0.15	108 / 0.20	n_example ,
50	60.82	0.62	0.41	360 / 0.40	109 / 0.32	249 / 0.45	N_even

As can be gleaned from Table 3, both automatically generated lexical n-grams and observation-based features are quite evenly dispersed among the most highly significant indicators of L1 German. While the most significant lexical n-grams consist to a large extent of bigrams involving the combination of a punctuation mark and a lexical item, many of the observation-based features relate to grammatical and discourse phenomena such as articles, prepositions, pronouns, and certain discourse related lexical choices. There are also a few punctuation marks that emerge as peculiar features of L1 German. Moreover, merely formal features emerge as characteristics such as the number of characters, words (i.e. tokens), sentences, and paragraphs. The comparatively higher amount of each of these indicators merely emphasizes the fact that German learners of English have composed longer test essays than their peers from other L1 backgrounds. In turn, the reason why these purely formal features of the learner texts are relevant for automatic classification could be related to the

composition of the TOEFL11 corpus as the selection of texts was not evenly balanced across proficiency levels.

In fact, when looking at the most significant features for classifying L1 Italian learner texts in Table 4, the same formal features indicating number of paragraphs, sentences, tokens, and characters pop up again. Only that this time Italian learners of English produce significantly shorter texts than the average across all other learner populations. A comparison of L1 German and L1 Italian shows the relation between text length and proficiency level. The set of 1,100 German texts in TOEFL11 consists of 15 low, 412 medium, and 673 high while L1 Italian is represented by 164 low, 623 medium, and 313 high (cf. Blanchard et al. 2013: 9).

Table 4: The 50 most significant features (of 185) for classifying L1 Italian

Rank	**F-score**	**Num/ 900**	**Num/ 9000**	**Num texts**	**Num texts (med)**	**Num texts (high)**	**Feature name**
1	640.29	1.07	0.35	522 / 0.58	328 / 0.64	125 / 0.48	n_think that
2	551.19	0.72	0.18	352 / 0.39	205 / 0.40	118 / 0.46	N_:
3	483.16	1.32	0.55	578 / 0.64	365 / 0.71	138 / 0.54	n_i think
4	451.86	11.92	15.94	900 / 1.00	516 / 1.00	258 / 1.00	N_.
5	425.51	0.30	0.06	188 / 0.21	106 / 0.21	61 / 0.24	n_in fact
6	388.55	12.92	16.67	900 / 1.00	516 / 1.00	258 / 1.00	N_SENTS
7	273.53	0.24	0.04	139 / 0.15	89 / 0.17	32 / 0.12	N_'m
8	195.69	0.91	0.48	475 / 0.53	277 / 0.54	143 / 0.55	n_, but
9	186.30	0.48	0.16	224 / 0.25	141 / 0.27	52 / 0.20	n_it 's
10	173.01	9.54	7.58	895 / 0.99	514 / 1.00	258 / 1.00	N_DEMONS_VOC_all
11	168.43	1.65	0.88	513 / 0.57	315 / 0.61	123 / 0.48	N_CLITICS_VOC_all
12	141.63	2.37	1.63	761 / 0.85	442 / 0.86	221 / 0.86	N_this

13	140.32	0.15	0.04	106 / 0.12	70 / 0.14	18 / 0.07	n_people that
14	137.75	6.66	5.26	880 / 0.98	508 / 0.98	256 / 0.99	N_that
15	137.75	6.66	5.26	880 / 0.98	508 / 0.98	256 / 0.99	N_that
16	131.07	0.38	0.17	244 / 0.27	147 / 0.29	52 / 0.20	n_and i
17	114.53	0.30	0.12	170 / 0.19	104 / 0.20	50 / 0.19	N_(
18	104.11	0.27	0.09	143 / 0.16	92 / 0.18	38 / 0.15	N_!
19	100.53	1.58	1.12	675 / 0.75	390 / 0.76	191 / 0.74	N_but
20	97.51	1339.94	1464.92	900 / 1.00	516 / 1.00	258 / 1.00	N_CHARS
21	91.52	0.10	0.31	72 / 0.08	32 / 0.06	36 / 0.14	n_however ,
22	86.35	1.29	0.88	595 / 0.66	359 / 0.70	162 / 0.63	n_. i
23	84.77	0.12	0.33	90 / 0.10	42 / 0.08	37 / 0.14	n_. however
24	83.50	26.58	29.98	900 / 1.00	516 / 1.00	258 / 1.00	N_IPUNCTS_VOC_all
25	81.06	1.23	0.76	451/ 0.50	283 / 0.55	104 / 0.40	N_n't
26	80.14	0.14	0.47	90 / 0.10	43 / 0.08	38 / 0.15	N_may
27	76.62	0.55	0.96	281 / 0.31	143 / 0.28	123 / 0.48	N_which
28	74.36	2.10	3.15	610 / 0.68	355 / 0.69	185 / 0.72	N_they
29	72.36	5.98	7.27	878 / 0.98	510 / 0.99	258 / 1.00	N_MODALS_VOC_all
30	71.50	327.16	352.13	900 / 1.00	516 / 1.00	258 / 1.00	N_TOKS
31	66.39	0.21	0.51	133 / 0.15	75 / 0.15	46 / 0.18	n_. they
32	65.25	0.65	0.40	315 / 0.35	208 / 0.40	66 / 0.26	n_lot of
33	64.45	0.36	0.21	243 / 0.27	145 / 0.28	71 / 0.28	n_at the
34	63.20	7.94	6.63	888 / 0.99	515 / 1.00	258 / 1.00	N_a
35	61.07	0.03	0.16	28 / 0.03	19 / 0.04	4 / 0.02	n_. because
36	61.06	0.55	0.32	289 / 0.32	181 / 0.35	62 / 0.24	n_do n't
37	60.23	0.14	0.30	112 / 0.12	62 / 0.12	41 / 0.16	n_. for example ,
38	60.17	0.74	0.47	352 / 0.39	230 / 0.45	83 / 0.32	n_a lot
39	56.93	1.42	2.20	467 / 0.52	264 / 0.51	153 / 0.59	N_their
40	53.77	1.02	0.72	462 / 0.51	267 / 0.52	134 / 0.52	N_very
41	52.26	0.86	1.27	433 / 0.48	245 / 0.48	160 / 0.62	N_by
42	51.91	0.13	0.29	102 / 0.11	57 / 0.11	41 / 0.16	n_. as
43	51.45	0.96	0.65	401 / 0.45	235 / 0.46	121 / 0.47	N_SAXONGEN_VOC_all
44	51.45	0.96	0.65	401 / 0.45	235 / 0.46	121 / 0.47	N_'s
45	50.46	1.14	1.60	460 / 0.51	246 / 0.48	189 / 0.73	N_as
46	49.94	0.12	0.24	112 / 0.12	52 / 0.10	53 / 0.21	n_. first
47	46.24	0.04	0.13	36 / 0.04	16 / 0.03	19 / 0.07	n_first ,
48	44.83	1.76	2.31	636 / 0.71	352 / 0.68	207 / 0.80	N_more
49	44.50	7.89	9.51	886 / 0.98	512 / 0.99	252 / 0.98	N_MISSPELLED
50	44.40	0.17	0.32	127 / 0.14	68 / 0.13	44 / 0.17	N_up

Apart from the important, albeit opposite, effect of formal textual features for identifying L1 German and L1 Italian texts in the corpus, Table 4 shows that the classification of L1 Italian relies on rather different significant features than those relevant for L1 German. Thus, the informative features for automatic classification highlight different profiles of L1 German and L1 Italian learners of English. The next section will take a closer look at some of the relevant features for each learner group and discuss possible relations to L1 specific habits of language use.

5. Possible transfer effects from L1 German and L1 Italian

Before the features in Table 3 and 4 are discussed in terms of their possible grounding in L1 transfer, it is important to stress some limitations of this approach. First and foremost, an interpretation of the significant features for automatically classifying L1 German and L1 Italian texts can at best generate hypothesis of L1 transfer into English if a particular phenomenon can be related to an L1-specific pattern of language use. Secondly, the features arise from comparing learners of English from different L1 backgrounds and not from a comparison with L1 English speakers. This might skew the relevance of certain features which might not be informative for the automatic classifier if compared with L1 English texts. Finally, the general discussion of the results in the previous section has made it clear

that the relevance of the features can be dependent on the design of the corpus.

Despite these limitations, a few of the patterns identified as peculiar to L1 German and L1 Italian call for an explanation which makes a case for L1 influence in the use of written English. The ensuing discussion will highlight some of these candidate constructions and patterns without attempting to be exhaustive for all the features given in Table 3 and 4. These might contain more features revealing L1 transfer.

5.1 Hypotheses of transfer from L1 German

To start with the peculiar features helping the automatic identification of L1 German in TOEFL11, the highest ranked indicator, the bigram <, that>, invites a straightforward explanation of transfer. Thus, German comma rules foresee the obligatory use of a comma in front of the equivalent final clause conjunction *dass* whereas commas are generally ruled out in front of the conjunction *that* in English. The same type of L1 influence is also evident in the bigram <, because>, which mirrors the German conventions of placing a comma in front of subordinating conjunctions. This, however, does not usually occur in English. The comparatively lower rate of <, and> is a further indication of L1 influence on comma use as it follows the German convention of not putting a comma in front of the coordinating conjunction in contrast to English.

Transfer of another orthographic convention is most likely at the root of the second most distinctive feature of L1 German texts in the corpus. The relative overuse of hyphens can be motivated by their common use as phrase connectors (i.e. as a dash, which is realized in the corpus data as a hyphen). On the other hand, there is no measurable effect concerning the occurrence of hyphens for connecting compound constituents.

Apart from orthographical transfer, a few instances of lexical usage reflect German lexical choices compared to learners of English from other L1 backgrounds. Ranked fourth in Table 3, the class feature <ESPECIALLY_VOC_all> combines the uses of the terms *special* and *especially*, both of which show a significantly higher rate in texts written by L1 German learners of English. This is most likely due to the fact that its German cognate form *speziell* is very frequently used in German, also in the function of a discourse marker. Similarly, the high relative frequency of the bigram <of course>, can be related to the very common German discourse marker *natürlich*, which literally translates as *of course* but is more versatile than its English equivalent. Another peculiar lexical construction which is quite highly ranked in Table 3 is the bigram <a certain>, which translates literally from German *ein gewisser/eine gewisse*. Furthermore, the more frequent use of adverbials subsumed in the study under the label of intensifiers (<INTENSIFIERS_VOC_all>) is another lexical pattern that characterizes L1 German learners of English in the corpus. This is particularly evident in the items <still>, <even>, <just>, and <only>.

However, it is difficult to argue for a transfer hypothesis in this case as the use of intensifiers and adverbials generally increases with higher proficiency in English as a learner language. Since there is a bias towards high level texts in the German component of TOEFL11, the higher rate of these adverbials compared to other L1 backgrounds might simply be a sampling effect of the corpus.

A case for a possible transfer can be made for the more frequent use of the conjunction *or* in texts of L1 German learners of English. While *or* is typically used in an exclusive sense in English, its German equivalent conjunction *oder* can also be used in a loosely coordinating sense similar to the function of *und* ('and') when connecting the final element in listings.

Another potential instantiation of transfer can be found in the elevated occurrence of indefinite articles and prepositions by German learners of English. However, the significance of these features might also be a consequence of the comparison between learner groups and might not show in comparison to L1 English speakers. This is due to the fact that the set of learner languages contains a few languages which lack definite and/or indefinite articles (e.g. Korean and Hindi; cf. Dryer 2013a). For the use of prepositions, comparatively lower figures in texts from learners of Turkish, Telugu, Hindi, Korean, and Japanese L1 backgrounds falls in line with the observation that these languages use postpositions instead of prepositions (cf. Dryer 2013b).

Finally, there is evidence of another possible transfer effect concerning the encoding of impersonal reference. In German, particularly in the genre of argumentative essays, general statements are frequently built around the use of the impersonal referent *man* which can be represented by using the pronoun *you* with impersonal reference in English. In this respect, it is quite striking to observe that texts written by L1 German learners of English show a significant overuse of the pronoun *you* and of its combinations in particular bigrams. In detail, automatic classification has established the following order of relevance among the significant features containing *you*: <you have>; <if you>; <you>; <. you>; <you are>; <you can>; <that you>; <you do>; <you will>; <when you>. Some of these bigrams reverberate common combinations of modals and conjunctions with impersonal *man*, which are particularly used in argumentative prose such as *wenn man* ('if you'/ 'when you'), *man kann* ('you can'), *dass man* ('that you'), and *man wird* ('you will').

5.2 Hypotheses of transfer from L1 Italian

Among the indicators that make up the automatic feature profile of L1 Italian texts in TOEFL11, a few can be related to Italian patterns, which shine through in the English essays. The bigrams <think that> and <i think> as well as the consistently more frequent occurrence of other features containing the first person pronoun (e.g. <and i>; <. i>; <my>; <in my>) are a token of a personal style in the argumentative texts written by L1 Italian

learners of English. As with German learners of English, there is also evidence for some transfer of comma conventions by L1 Italian speakers, as indicated in the frequent use of the bigrams <, that> and <, because>. In a similar vein, <, but> appears as one of the most significant discriminators of L1 Italian texts. This most likely relates to the fact that the Italian equivalent conjunction *ma* is very versatile and frequently used in Italian, stimulating its relative overuse also in English argumentative essays.

The significantly higher rate of colons, brackets, and exclamation marks is difficult to motivate by an explanation involving transfer. It might be tempting to connect the higher rate of exclamation marks with a more emphatic style of argumentation; however, the lack of support from related indicators of the group of intensifiers does not allow any further speculation. A clear indication of transfer on the orthographic level, on the other hand, can be gleaned from the prominence of clitics. This is not only indicated by the more frequent use of all clitics considered as such (<CLITICS_VOC_all>) but also by the relatively high number of n-grams such as <’ m>, <it ’ s>, <n ’ t>, and <’ s>. As Italian regularly uses clitics in standard orthography, it seems as if Italian learners of English more readily embrace cliticized constructions even if they are marked as informal variants of their full forms in written English.

On the lexical level, a few significant indicators call for an explanation in terms of transfer. The overuse of <in fact> finds a model in the semantically close Italian expression *infatti*, used as a conjunction and

discourse marker. Interference of the Italian term is emphasized by numerous examples of the spelling *infact* found in texts by Italian learners of English. Boosts in the use of *probably* and *particular/ly* due to their Italian cognates of *probabilmente* and *in particulare*/*particularmente* are further candidates of lexical transfer.

Finally, two interesting hypotheses arise for transfer involving grammatical aspects. On the more speculative side, Italian learners of English show a significantly lower use of prepositions in TOEFL11. Tentatively, this difference coheres with the characterization of Italian as a verb-framed language, where path and manner of motion is more likely to be encoded in the lexical verb. English and Germanic languages, by contrast, are described as satellite-framed languages, which rely more on prepositions and adverbials to encode path and manner of motion (cf. Talmy 2000, Slobin 2004). The hypothesis for potential L1 influence in this area would have to be explored in separate studies devoted to this aspect.

A stronger case for transfer can be made for the use of modals among Italian learners in TOEFL11. In general, texts of L1 Italian learners show a markedly low rate of modal verbs as captured in the class feature <MODALS_VOC_all>. A more detailed look at the individual modals provides a diversified picture. The modals <may>, <might>, <would>, <should>, and <will> exhibit a significantly low rate whereas <must> and <could> are significantly overrepresented. This maps nicely onto the inventory of modal verbs in Italian, which essentially consists of the two

modal verbs of *potere* ('can / could') and *dovere* ('must / have to'). Thus, the results of the automatically generated feature profile of Italian learners gives numerical evidence that the expression of modality in English can be subject to transfer from Italian.

6. Conclusion

This article has explored the current topic of how texts written by learners of English can be automatically classified according to a learner's L1 background. For this we have equipped an ML algorithm with a mixed set of features combining indicators based on observation and automatically generated n-grams. The overall set of 400 features was used on the TOEFL11 corpus, yielding an average classification accuracy of 59% of correct assignments to one of 11 different L1s. As pointed out above, this is not a very high accuracy rate compared with the thirteen best results of the NLI Shared Task that range between 80% and 84% of classification accuracy on the same dataset. Rather than tuning our machine learning algorithm for higher accuracy, we pursued the aim of investigating potential transfer effects. This, first of all, guided our limited selection of lexical features, and then made us test whether the most discriminating features for classifying L1 German and L1 Italian texts might take their origin in transfer from these languages. A discussion of the most significant features has highlighted a few specific patterns and habits of L1 use. At the same

time, explanations motivating these features due to the existence of model words, structures, and conventions in the L1 remain hypothetical claims for L1 transfer effects. It would be interesting to investigate these hypotheses further and to carry out additional empirical tests of their relevance. For the time being, evidence from the TOEFL11 corpus emphasizes the conclusion that learners of English from different language backgrounds indeed show distinguishable L1 profiles in their English prose. At least part of these profiles reflects L1 specific patterns of language use, highlighting the observation that transfer plays a substantial role.

7. References

Aharodnik, Katsiaryna, Marco Chang, Anna Feldman & Jirka Hana. 2013. Automatic Identification of Learners' Language Background based on their Writing in Czech. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJNCLP 2013)*, 1428-1436. Nagoya, October 2013.

Ahn, Charles S. 2011. *Automatically Detecting Authors' Native Language*. M.A. thesis. Monterey, CA: Naval Postgraduate School.

Baroni, Marco & Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21: 259-274.

Barr, G.K. 2003. Two styles in the New Testament epistles. *Literary and Linguistics Computing* 18: 235-248.

Bestgen, Yves, Sylvaine Granger & Jennifer Thewissen. 2012. Error pattern and automatic L1 identification. In *Approaching Language Transfer through Text Classification*, Scott Jarvis & Scott A. Crossley (eds.), 127-153. Bristol / Buffalo / Toronto: Multilingual Matters.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill & Martin Chodorow. 2013. *TOEFL11: A Corpus of Non-Native English*. Princeton, NJ: Educational Testing Service.

Brooke, Julian & Graeme Hirst. 2013. Using other learner corpora in the 2013 NLI Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 188-196. Atlanta, Georgia. Association for Computational Linguistics.

Crossley, Scott A. & Danielle S. McNamara. 2012. Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In *Approaching Language Transfer through Text Classification*, Scott Jarvis & Scott A. Crossley (eds.), 106-126. Bristol / Buffalo / Toronto: Multilingual Matters.

Dryer, Matthew S. 2013a. Definite Articles. In *The World Atlas of Language Structures Online*, Matthew S. Dryer & Martin Haspelmathm (eds.), Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/37, Accessed on November 19, 2013).

Dryer, Matthew S. 2013b. Order of Adposition and Noun Phrase. In: *The World Atlas of Language Structures Online*, Matthew S. Dryer & Martin Haspelmathm (eds.). Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/37, Accessed on November 19, 2013).

Estival, Dominique, Tanja Gaustad, Son Bao Pham, Will Radford & Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, 31-39. Melbourne, Australia.

Gebre, Binyam Gebrekidan, Marcos Zampieri, Peter Wittenburg & Tom Heskes. 2013. Improving Native Language Identification with TF-IDF Weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 216-223. Atlanta, Georgia. Association for Computational Linguistics.

Golcher, Felix & Marc Reznicek. 2011. Stylometry and the interplay of topic and L1 in the different annotation layers in the Falko corpus. In

Proceedings of Quantitative Investigations in Theoretical Linguistics 4, Amir Zeldes & Anke Lüdeling (eds.), 29-34. Berlin, March 2011.

Granger, Sylvie, Estelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *The International Corpus of Learner English. Handbook and CD-ROM. Version 2*. Louvain-la Neuve: Presses Universitaires de Louvain.

Jarvis, Scott & Aneta Pavlenko. 2008. *Crosslinguistic Influence in Language and Cognition*. New York: Routledge.

Jarvis, Scott. 2012. The detection-based approach: An overview. In *Approaching Language Transfer through Text Classification* Scott Jarvis & Scott A. Crossley (eds.), 1-33. Bristol / Buffalo / Toronto: Multilingual Matters.

Jarvis, Scott & Scott A. Crossley (eds.). 2012. *Approaching Language Transfer through Text Classification*. Bristol / Buffalo / Toronto: Multilingual Matters.

Jarvis, Scott & Magali Paquot. 2012. Exploring the role of n-grams in L1 identification. In *Approaching Language Transfer through Text Classification*. Scott Jarvis & Scott A. Crossley (eds.), 71-105. Bristol / Buffalo / Toronto: Multilingual Matters.

Jarvis, Scott, Gabriela Castañeda-Jiménez & Rasmus Nielsen. 2012. Detecting L2 writers' L1 on the basis of their lexical styles. In *Approaching Language Transfer through Text Classification*, Scott

Jarvis & Scott A. Crossley (eds.), 34-70. Bristol / Buffalo / Toronto: Multilingual Matters.

Jarvis, Scott, Yves Bestgen & Steve Pepper. 2013. Maximizing classification accuracy in Native Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 111-118. Atlanta, Georgia. Association for Computational Linguistics.

Koppel, Moshe, Jonathan Schler & 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*, 624-628. Chicago: Association for Computing Machinery.

Mayfield Tomokiyo, Laura & Rosie Jones. 2001. You're not from 'round here, are you'? Naïve Bayes detection of non-native utterance text. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL '01)*. Electronic document. Cambridge, MA: The Association for Computational Linguistics.

McLachlan, Geoff .J. 2004. *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken, NJ: Wiley.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas,

Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825-2830.

Slobin, Dan. 2004. The many ways to search for a frog: linguistic typology & the expression of motion events. In *Relating Events in Narrative: Vol. 2. Typological and contextual perspectives*, Sven Strömqvist & Ludo Verhoeven (eds.), 219-257. Mahwah, NJ: Lawrence Erlbaum Associates.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics. Volume 2: Typology and Process in Concept Structuring.* Cambridge, MA: MIT Press.

Taylor, Barry P. 1975. The use of overgeneralisation and transfer learning strategies by elementary and intermediate students of ESL. *Language Learning* 25: 73-107.

Tetreault, Joel, Daniel Blanchard & 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*, 48-57. Atlanta, Georgia. Association for Computational Linguistics.

Thomason, Sarah. 2001. *Language Contact*. Edinburgh: Edinburgh University Press.

Tsur, Oren & 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, 9-16. Prague. Association for Computational Linguistics.

Van Halteren, Hans. 2008. Source language markers in EUROPARL translations. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 937-944. Manchester, August 2008.

Wong, Sze-Meng Jojo & Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association*, 53-61. Cambridge, MA: The Association for Computational Linguistics.

Wong, Sze-Meng Jojo & Mark Dras. 2011. Exploiting parse structures for Native Language Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1600-1610. Edinburgh, July 2011.

Wu, Ching-Yi, Po-Hsiang Lai, Yang Liu & Vincent Ng. 2013. Simple yet powerful Native Language Identification on TOEFL11. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 152-156. Atlanta, Georgia. Association for Computational Linguistics.

Yannakoudakis, Helen, Ted Briscoe & Ben Medlock. 2011. A new dataset and method for automatically grading Esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: *Human Language Technologies*, 180-189. Portland, Oregon, USA: Association for Computational Linguistics.