Linguistic Profiling based on General–purpose Features and Native Language Identification

Andrea Cimino, Felice Dell'Orletta, Giulia Venturi and Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
via G. Moruzzi, 1 – Pisa (Italy)
{name.surname}@ilc.cnr.it

Abstract

In this paper, we describe our approach to native language identification and discuss the results we submitted as participants to the First NLI Shared Task. By resorting to a wide set of general–purpose features qualifying the lexical and grammatical structure of a text, rather than to ad hoc features specifically selected for the NLI task, we achieved encouraging results, which show that the proposed approach is general–purpose and portable across different tasks, domains and languages.

1 Introduction

Since the seminal work by Koppel et al. (2005), within the Computational Linguistics community there has been a growing interest in the NLP–based Native Language Identification (henceforth, NLI) task. However, so far, due to the unavailability of balanced and wide–coverage benchmark corpora and the lack of evaluation standards it has been difficult to compare the results achieved for this task with different methods and techniques (Tetreault et al., 2013). The First Shared Task on Native Language Identification (Tetreault et al., 2013) can be seen as an answer to the above mentioned problems.

In this paper, we describe our approach to native language identification and discuss the results we submitted as participants to the First NLI Shared Task. Following the guidelines by the Shared Task Organizers based on the previous literature on this topic, Native Language Identification is tackled as a text classification task combining NLP–enabled feature extraction and machine learning: see e.g. Tetreault et al. (2013) and Brooke and Hirst (2012). Interestingly, the same methodological paradigm is shared by other tasks like e.g. author recognition and verification (see e.g. van Halteren (2004), authorship attribution (see Juola (2008) for a survey), genre identification (Mehler et al., 2011) as well as readability assessment (see Dell'Orletta et al. (2011a) for an updated survey), all relying on feature extraction from automatically parsed texts and state–of–the–art machine learning algorithms. Besides obvious differences at the level of the typology of selected linguistic features and of learning techniques, these different tasks share a common approach to the problems they tackle: i.e. they succeed in determining the language variety, the author, the text genre or the level of readability of a text by exploiting the distribution of different types of linguistic features automatically extracted from texts.

Our approach to NLI relies on multi–level linguistic analysis, covering morpho–syntactic tagging and dependency parsing. In the NLI literature, the range of features used is wide and includes characteristics of the linguistic structure underlying the L2 text, encoded in terms of sequences of characters, words, grammatical categories or of syntactic constructions, as well as of the document structure: note however that, in most part of the cases, the exploited features are task–specific. In our approach, we decided to resort to a wide set of features ranging across different levels of linguistic description (i.e. lexical, morpho–syntactic and syntactic) without any a priori selection: the same set of features was successfully exploited in NLI–related tasks, i.e. focusing on the linguistic form rather than

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 207–215,
Atlanta, Georgia, June 13 2013.

the content of texts, such as readability assessment (Dell'Orletta et al., 2011a) or the classification of textual genres (Dell'Orletta et al., 2012).

The exploitation of general features qualifying the lexical and grammatical structure of a text, rather than ad hoc features specifically selected for the task at hand, is not the only peculiarity of our approach to NLI. Following Biber (1993), we start from the assumption that "linguistic features from all levels function together as underlying dimensions of variation". This choice stems from studies on linguistic variation, in particular from Biber and Conrad (2009) who claim that linguistic varieties – called "registers" from a functional perspective – differ "in their characteristic distributions of pervasive linguistic features, not the single occurrence of an individual feature". This is to say that by carrying out the linguistic analysis of collections of essays each written by different L1 native speakers, we need to quantify the extent to which a given feature occurs in each collection, in order to reconstruct the linguistic profile underlying each L1 collection: differences lie at the level of the distribution of linguistic features, which can be common and pervasive in some L1 collections but comparatively rare in others. This approach is the basis of so–called "linguistic profiling" of texts, within which "the occurrences of a large number of linguistic features in a text, either individual items or combinations of items, are counted" (van Halteren, 2004) with the final aim of reconstructing the profile of a text.

We carried out native language identification in two steps. The first step consisted of the identification of the set of linguistic features characterizing the essays written by different L1 native speakers, i.e. the linguistic profiling of the different sections of TOEFL11 corpus (Blanchard et al., 2013) distributed as training and development data. In the second step, the features which turned out to have highly discriminative power were used for the classification of essays written by different L1 native speakers. Essay classification has been carried out by experimenting with different approaches: i.e. a single–classifier method and two different multi–model ensemble approaches.

The paper is organised as follows: after introducing the set of used linguistic features (Section 2), Section 3 illustrates a selection of the linguistic profiling results obtained with respect to the training section of the TOEFL11 corpus; Section 4 describes the different classification approaches we followed and the feature selection process; in Section 5 achieved results are reported and discussed.

2 Features

In this study, we focused on a wide set of features ranging across different levels of linguistic description. Differing from previous work on NLI, no a priori selection of features was carried out. Instead of focusing on particular classes of errors or on different types of stylistic idiosyncrasies, we took into account a wide range of features which are typically used in studies focusing on the "form" of a text, e.g. on issues of genre, style, authorship or readability. As previously pointed out, this represents a peculiarity of our approach. This choice makes the selected features language–independent, domain–independent and reusable across different types of tasks, as empirically demonstrated in Dell'Orletta et al. (2011a) where the same set of features has been successfully exploited for readability assessment, and in Dell'Orletta et al. (2012) where the features have been used for the classification of different types of textual genre. Note that in both cases the language dealt with was Italian: for the NLI Shared Task we had to specialize the feature extraction process with respect to the English language as well as to the annotation scheme used to represent the underlying linguistic structure.

The whole set of features we started with is described below, organised into four main categories: namely, raw text and lexical features as well as morpho-syntactic and syntactic features. This proposed four–fold partition closely follows the different levels of linguistic analysis automatically carried out on the text being evaluated, i.e. tokenization, lemmatization, morpho-syntactic tagging and dependency parsing.

2.1 Raw and Lexical Text Features

Sentence Length, calculated as the average number of words per sentence.

Word Length, calculated as the average number of characters per word.

Document Length, calculated as the total number

of words per document.

Character bigrams.

Word n-grams, including both unigrams and bigrams.

Type/Token Ratio: the Type/Token Ratio (TTR) is a measure of vocabulary variation which has shown to be a helpful measure of lexical variety within a text as well as style marker in an authorship attribution scenario: a text characterized by a low type/token ratio will contain a great deal of repetition whereas a high type/token ratio reflects vocabulary richness and variation. Due to its sensitivity to sample size, TTR has been computed for text samples of equivalent length (the first 50 tokens).

2.2 Morpho–syntactic Features

Coarse grained Part-Of-Speech n-grams: distribution of unigrams and bigrams of coarse–grained PoS, corresponding to the main grammatical categories (e.g. noun, verb, adjective, etc.).

Fine grained Part-Of-Speech n-grams: distribution of unigrams and bigrams of fine–grained PoS, which represent subdivisions of the coarse–grained tags (e.g. the class of nouns is subdivided into proper vs common nouns, verbs into main verbs, gerund forms, past particles, etc.).

Verbal chunks: distribution of sequences of verbal PoS (also including adverbs). This feature can be seen as a proxy to capture different aspects of verbal predication, with particular attention to idiosyncratic usages of verbal mood, tense, person and adverbial modification.

Lexical density: ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text.

2.3 Syntactic Features

Dependency types n-grams: distribution of unigrams and bigrams of dependency types calculated with respect to *i)* the hierarchical parse tree structure and *ii)* the surface linear ordering of words.

Dependency triples: distribution of triplets representing a dependency relation consisting of a syntactic head (h), the dependency relation type (t) and the dependent (d). Two different variants of this feature are distinguished, based on the fact that either the coarse–grained PoS or the word–form of h and d is considered: we will refer to the former as *Coarse grained Part-Of-Speech dependency triples* and to the latter as *Lexical dependency triples*. In both cases, the relative ordering of h and d, i.e. whether h precedes or follows d at the level of the linear ordering of words within the sentence, is also considered.

Dependency Subtrees: distribution of dependency subtrees consisting of a dependency relation (represented as the dependency triple $\{h, t, d\}$), the head father and the dependency relation linking the two. As in the previous case, two different variants of this feature are distinguished, based on the fact that either the coarse grained PoS or the word–forms of the nodes in the dependency subtree are considered.

Parse tree depth features: this set of features is meant to capture different aspects of the parse tree depth and includes: *a)* the depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf; *b)* the average depth of embedded complement 'chains' governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers; *c)* the probability distribution of embedded complement 'chains' by depth. These features represent reliable indicators of sentence complexity, as stated by, among others, Yngve (1960), Frazier (1985) and Gibson (1998), and they can thus allow capturing specific difficulties of L2 learners.

Coarse grained Part-Of-Speech of sentence root: this feature refers to coarse grained POS of the syntactic root of a sentence.

Arity of verbal predicates: this feature refers to the number of dependencies (corresponding to either subcategorized arguments or modifiers) governed by the same verbal head. In the NLI context, it can allow capturing improper verbal usage by L2 learners due to language transfer (e.g. with pro–drop languages as L1).

Subordination features: this set of features is meant to capture different aspects of the use of subordination and includes: *a)* the distribution of subordinate vs main clauses; *b)* the average depth of 'chains' of embedded subordinate clauses and *c)* the probability distribution of embedded subordinate clauses 'chains' by depth. Similarly to parse tree depth, this set of features can be taken to reflect the structural complexity of sentences and can thus be indicative of specific difficulties of L2 learners.

Length of dependency links: measured in terms

of the words occurring between the syntactic head and the dependent. This is another feature which reflects the syntactic complexity of sentences (Lin, 1996; Gibson, 1998) and which can be successfully exploited to capture syntactic idiosyncracies of L2 learners due to L1 interferences.

2.4 Other features

Two further features have been considered for NLI purposes, which were included in the distributed datasets. For each document, we have also considered i) the English language proficiency level (high, medium, or low) based on human assessment by language specialists, and ii) the topic of the essays.

3 Linguistic Profiling of TOEFL11 Corpus

In this section, we illustrate the results of linguistic profiling carried out on the training and development sets extracted from the TOEFL11 corpus. This corpus, described in Blanchard et al. (2013), contains 1,100 essays per 11 languages (for a total of 12,100 essays) sampled as evenly as possible from 8 prompts (i.e., topics) along with score levels (low/medium/high) for each essay. The considered L1s are: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. For the specific purposes of the NLI Shared Task, a total of 9,900 essays has been distributed as training data (900 essays per L1), 1,100 as development data (100 per L1) and the remaining 1,100 essays have been used as test data.

We started from the automatic linguistic annotation of training and development data whose output has been searched for with respect to the features illustrated in Section 2.

3.1 Linguistic Pre–processing

Both training and development data were automatically morpho-syntactically tagged by the POS tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser using Multi–Layer Perceptron as learning algorithm (Attardi et al., 2009), a state–of–the–art linear–time Shift–Reduce dependency parser. Feature extraction is carried out against the output of the multi–level automatic linguistic analysis carried out during the pre–processing stage: lexical and grammatical patterns corresponding to the wide typology of selected features are looked for within each annotation layer and quantified.

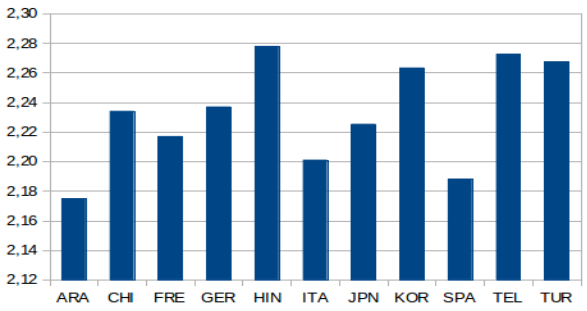
3.2 Linguistic Profiling

Generally speaking, linguistic profiling makes it possible to identify (groups of) texts which are similar, at least with respect to the "profiled" features (van Halteren, 2004). In what follows we report the results of linguistic profiling obtained with respect to the 11 L1 sub–corpora considered in this study. Figure 1 shows the results obtained with respect to a selection of the features described in Section 2. These results refer to the combined training and development data sets: note, however, that we also calculated the values of these features in the two datasets separately and it turned out that they do not vary significantly between the two sets. This fact can be taken as a proof both of the reliability of our approach to linguistic profiling and of the relevance of these features for NLI purposes.

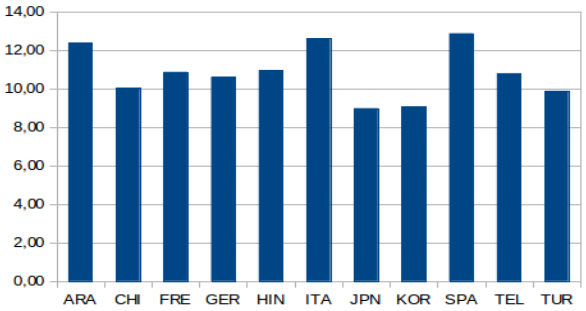
Starting from raw textual features (Figures 1(a) and 1(b)), both average sentence length and average word length vary significantly across L1s. In particular, if on the one hand the essays written by Arabic and Spanish L1 speakers contain the shortest words and the longest sentences, on the other hand the Hindi and Telugu L1 essays are characterized by the longest words; the L1 Japanese and Korean corpora contain the shortest sentences.

Let us focus now on the distribution of unigrams of coarse grained Parts–Of–Speech. If we consider the distributions of determiners and nouns, two features typically used for NLI purposes (Wong and Dras, 2009) which also represent stylistic markers associated with different linguistic varieties (Biber and Conrad, 2009), it can be noticed (see Figures 1(c) and 1(d)) that for Japanese and Korean the essays show the lowest percentage of determiners, while for Hindi and Telugu they are characterized by the highest percentage of nouns.

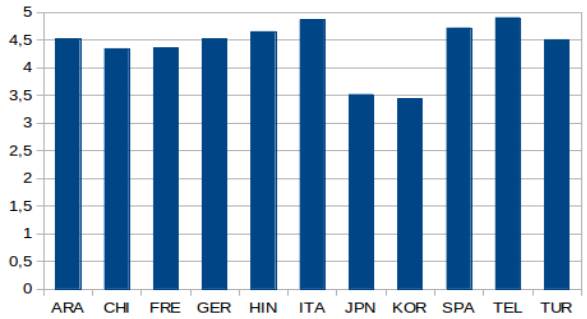
For what concerns syntactic features, we observe that essays by Japanese and Korean speakers are characterized by quite a different distribution with respect to the other L1 corpora. In particular, they show the shallowest parse trees, the shortest dependency links as well as the shortest 'chains' of embedded complements governed by a nominal head. On the other hand, the essays by Spanish and Ara-



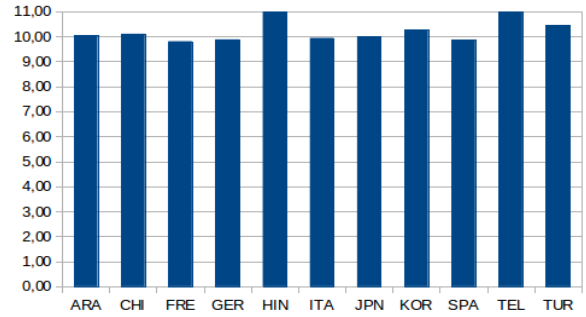
(a) Average word length



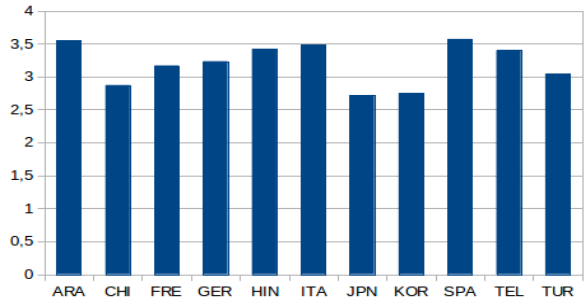
(b) Average sentence length



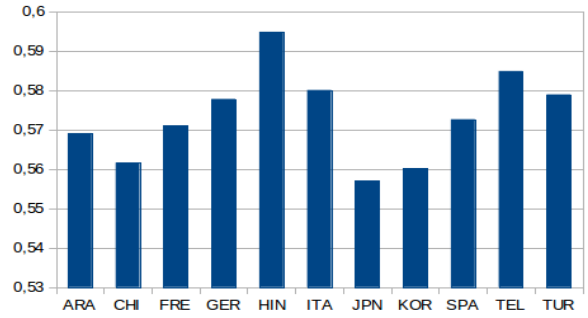
(c) Distribution of Determiners



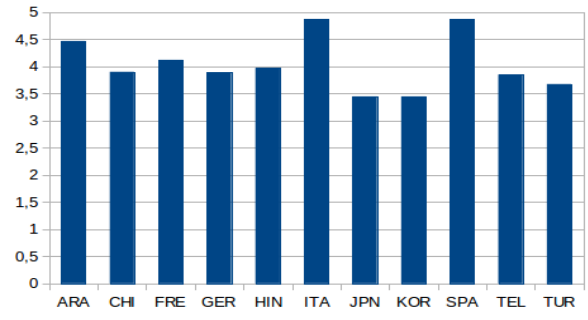
(d) Distribution of Nouns



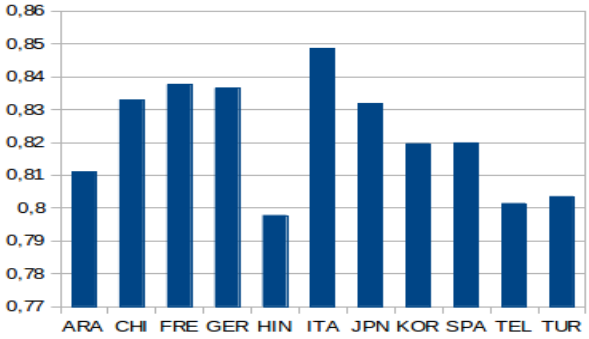
(e) Average parse tree depth



(f) Average depth of embedded complement 'chains'



(g) Average length of the longest dependency link



(h) Arity of verbal predicates

Figure 1: Results of linguistic profiling carried out on the combined training and development sections of the TOEFL11 corpus.

bic speakers contain the deepest parse trees, for Italian and Spanish we observe the longest dependency links and for Hindi and Telugu the longest sequences of embedded complements. Moreover, while the essays by Italians are characterised by the highest value of arity of verbal predicates, for Hindi, Telugu and Korean essays much lower values are recorded.

Interestingly, these linguistic profiling results show similar trends across the 11 languages at different levels of linguistic analysis. For instance, it can be noted that Japanese and Korean or Italian and Spanish, which belong to two different language families, show similar distributions of features. Similarities have also been recorded in the sub–corpora by Hindi and Telugu speakers, even if these languages do not belong to the same family; we can hypothesize that this might originate from language contact phenomena.

4 System Description

4.1 Machine Learning Classifier

Our approach to Native Language Identification has been implemented in a software prototype, i.e. a classifier operating on mopho–syntactically tagged and dependency parsed texts which assigns to each document a score expressing its probability of belonging to a given L1 class. The highest score represents to the most probable class. Given a set of features and a training corpus, the classifier creates a statistical model using the feature statistics extracted from the training corpus. This model is used in the classification of unseen documents. The set of features and the machine learning algorithm can be parameterized through a configuration file.

For each feature, we have implemented three different variants, depending on whether the feature value is encoded in terms of: *i*) presence/absence of the feature (*binary* variant), *ii*) the normalized frequency (*normalized frequency* variant), and *iii*) the normalized *tf*idf* value (*normalized tf*idf* variant). Since the binary feature variant outperformed the other two, in all the experiments carried out on the development set reported in Section 5 we illustrate the results obtained using this variant only. This is in line with the results obtained by Brooke and Hirst (2012) and Tetreault et al. (2013). According to (Brooke and Hirst, 2012), a possible explanation is that "in these relatively short texts, there is high variability in normalized frequencies, and a simpler metric, by having less variability, is easier for the classifier to leverage". Support Vector Machines (SVM) using LIBSVM (Chang and Lin, 2001) and Maximum Entropy (ME) using MaxEnt[1] have been used as machine learning algorithms.

We experimented two classification approaches: a single classifier method and two ensemble systems, combining the output of several classifiers.

The single classifier uses the set of features resulting from the feature selection process described in Section 4.2 and the SVM using linear kernel as machine learning algorithm. This choice is due to the fact that in all the experiments the linear SVM outperformed the SVM using polynomial kernel. There are two possible explanations for this fact, namely: a) the number of features is much higher than the number of training instances, accordingly it might not be necessary to map data to a higher dimensional space, therefore the nonlinear mapping does not improve the performance; b) Weston et al. (2000) showed that SVMs can indeed suffer in high dimensional spaces where many features are irrelevant. Note that in Section 5, we report the results of this classifier using different sets of features corresponding to the lexical, morpho–syntactic and syntactic levels of linguistic analysis.

The two ensemble systems combine the outputs of the component classifiers following two different strategies. The first one is based on the majority voting method (henceforth, *VoteComb*): the combination strategy is seen as a classical voting problem where for each essay is assigned the L1 class that has been selected from the majority of classifiers. In case of ties, the L1 class predicted from the best individual model (as resulting from the experiments carried out on the development set) is selected. The second strategy combines the outputs of the component classifiers via another classifier (henceforth referred to as *meta–classifier*): we will refer to this second strategy as *ClassComb*. The meta–classifier uses as a feature the probability score predicted from each component classifier for each L1 class. Differently from the component classifiers, the meta–classifier is based on polynomial kernel SVM. In both en-

[1] https://github.com/lzhang10/maxent#readme

semble systems, the component classifiers use linear SVM and ME as machine learning algorithms and exploit different sets of features among the ones resulting from the feature selection process described below.

4.2 Features Selection Process

Since our approach to NLI relies on a wide number of general–purpose features, a feature selection process was necessary in order to prune irrelevant and redundant features which could negatively affect the classification results. The selection process starts taking into account all the n features described in Section 2. In each iteration, for each feature f_i we generate a configuration c_i such that f_i is disabled and all the other features are enabled. When an iteration finishes, we obtain for each c_i a corresponding accuracy score $score(c_i)$ which is computed as the average of the accuracy obtained by the classifier on the development set (a_d) and on an internal development set (a_i), corresponding to the 10% of the training set, used in order to reduce the overfitting risk. Being c_b the best configuration among all the c_i configurations, if $score(c_b) \leq$ of the accuracy scores resulting from the previous iterations the process stops. Otherwise:

1. store in F the pair $\langle f_b, disabled \rangle$;
2. for each configuration c_i, if $score(c_i) \leq$ of the accuracy scores resulting from the previous iterations, we store in F the pair $\langle f_i, enabled \rangle$;
3. set $C = \langle c_b, score(c_b) \rangle$

where F is a map containing elements $feature \rightarrow \{disabled, enabled\}$ and C is a pair that contains the current best configuration c_b and the corresponding score $score(c_b)$. In each iteration, we consider only the features which do not occur in F. At the initialization step F is empty and C contains the configuration where all the considered features are enabled.

In spite of the fact that the described selection process does not guarantee to obtain the global optimum, it however permitted us to obtain an improvement of about 8% with respect to the starting model indiscriminately using all features.

Table 1 lists the features resulting from the feature selection process. It can be noted that some of them coincide with those typically used for NLI purposes: this is the case of n–grams of words, Parts-Of-Speech and syntactic dependencies. Interestingly, to our knowledge, other features such as arity of verbal predicates, length of dependency links as well as subordination and parse tree depth features have not been used for NLI so far, in spite of their being widely exploited in the syntactic complexity literature (as discussed in Section 2).

Lexical features:
Word n-grams
Morpho–syntactic features:
Coarse grained Part-Of-Speech unigrams
Fine grained Part-Of-Speech bigrams
Syntactic features:
Dependency types unigrams
Lexical dependency triples
Parse tree depth features
Coarse grained Part-Of-Speech of sentence root
Arity of verbal predicates
Subordination features
Length of dependency links

Table 1: Features resulting from the feature selection process.

5 Results

Table 2 reports the overall Accuracy achieved with the different classifier models in the NLI classification task on the official test set as well as the F-measure score recorded for each L1 class. The first two lines show the accuracies of the two combination models, while the last three report the results obtained by the single classifier using i) the set of features resulting by the features selection process (*Best_Single*), ii) the selected lexical features only (see Table 1) (*Lexical*) and iii) the lexical and morpho–syntactic features (*Lex+Morph*).

The two combination models outperform all the single model classifiers: note that *ClassComb* achieved much better results with respect to *VoteComb*. By comparing these results with the F-measure scores obtained on the distributed development data (see Table 3), it can be seen that the ranking of the scores achieved by the different classifiers remains the same even if on the test data we obtained a performance of -2,2% with respect to the develop-

	Accuracy	ARA	CHI	FRE	GER	HIN	ITA	JAP	KOR	SPA	TEL	TUR
ClassComb	**77,9**	73,8	77,5	**83,2**	87,3	71,1	**86,0**	78,8	**74,2**	**70,8**	76,2	**78,0**
VoteComb	77,2	74,3	77,0	80,0	87,0	72,8	81,6	79,6	73,8	67,7	77,6	77,6
Best_Single	76,6	71,9	**77,6**	75,8	85,7	73,2	82,0	**80,0**	74,0	69,0	76,9	76,5
Lex+Morph	76,4	**77,2**	76,2	78,6	85,9	72,1	80,4	76,8	71,9	68,0	76,4	76,4
Lexical	76,2	71,1	76,5	79,0	**87,6**	**74,5**	80,8	77,7	70,8	66,7	**79,2**	73,4

Table 2: Classification results of different classifiers on official test data.

ment test set.

Let us consider now the results obtained by the single model classifiers. In all cases the *Best_Single* outperforms the other two models demonstrating the reliability of the features selection process and that a combination of lexical, morpho–syntactic and syntactic features leads to better results.

Although the best performing model is the *ClassComb*, this is not true for all the 11 languages. In Table 2, the best results for each L1 are bolded. Interestingly, even though *Lexical* is the worst model, it is the best performing one for three L1s while the best model, i.e. *ClassComb*, for five only.

It can be noted that with respect to the development data set the syntactic features used by the *Best_Single* model allow an increment of +1% as opposed to the *Lexical* model: this represents a much higher increase if compared with the result obtained on the test data, which is +0,4%. This is an unexpected result since the feature selection described in Section 4.2 was carried out on an internal development set in order to prevent the risk of overfitting on the distributed development data.

Classifier	Accuracy
ClassComb	**80,1**
VoteComb	79,3
Best_Single	78,8
Lex+Morph	78,2
Lexical	77,8

Table 3: Classification results of different classifiers on distributed development data.

6 Conclusion

In this paper, we reported our participation results to the First Native Language Identification Shared Task. By resorting to a wide set of general–purpose features qualifying the lexical and grammatical structure of a text, rather than to ad hoc features specifically selected for the task at hand, we achieved encouraging results. After a feature selection process, new features which to our knowledge have never been exploited so far for NLI purposes turned out to contribute significantly to the task. Interestingly, the same set of features we started from has been previously successfully exploited in other related tasks, such as readability assessment and genre classification, operating on the Italian language. The obtained results suggest that our approach is general–purpose and portable across different domains and languages. Further directions of research currently include: i) comparison of results obtained with general purpose features and with NLI–specific features (e.g. typical errors or different types of stylistic idiosyncrasies specific to L2 learners), with a view to combining them to achieve better results; ii) design and development of new ensemble classification methods as well as new feature selection methods considering not only classes of features but also individual features; iii) testing our approach to NLI on different L2s (e.g. Italian) .

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, Italy.

Douglas Biber. 1993. Using Register–diversified Corpora for General Language Studies. *Computational Linguistics Journal*, 19(2): 219–241.

Douglas Biber and Susan Conrad. 2009. Genre, Register, Style. Cambridge: CUP.

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

Julian Brooke and Graeme Hirst. 2012. Robust, Lexicalized Native Language Identification. In *Proceedings of COLING 2012*, Mumbai, India, 391–408.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines. *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*

Walter Daelemans. 2012. Explanation in Computational Stylometry. In A. Gelbukh (ed.) *CICLing 2012, Part II*, LNCS 7817, Springer–Verlag, 451–462.

Felice Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. In *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi. 2011a. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Workshop on "Speech and Language Processing for Assistive Technologies" (SLPAT 2011)*, Edinburgh, July 30, 73–83.

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi. 2012. Genre–oriented Readability Assessment: a Case Study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education (SLP-TED)*, 91–98.

Lyn Frazier. 1985. Syntactic complexity. In D.R. Dowty, L. Karttunen and A.M. Zwicky (eds.), *Natural Language Parsing*, Cambridge University Press, Cambridge, UK.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. In *Cognition*, 68(1), pp. 1–76.

Patrick Juola. 2008. *Authorship Attribution*. Now Publishers Inc.

Moshe Koppel, Jonathan Schler and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, vol. 3495, LNCS, Springer–Verlag, 209–217.

Dekan Lin. 1996. On the structural complexity of natural language sentences. In *Proceedings of COLING 1996*, pp. 729–733.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the Errors of Data-Driven Dependency Parsing Models. In *Proceedings of EMNLP-CoNLL, 2007*, 122–131.

Alexander Mehler, Serge Sharoff and Marina Santini (Eds.). 2011. *Genres on the Web. Computational Models and Empirical Studies*. Springer Series: Text, Speech and Language Technology.

Sze–Meng Jojo Wong and Mark Dras. 2009. Contrastive Analysis and Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop*.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics (ACL04)*, 200–207.

Joel Tetreault, Daniel Blanchard, Aoife Cahill and Martin Chodorow. 2012. Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. In *Proceedings of COLING 2012*, Mumbai, India, 2585–2602.

Joel Tetreault, Daniel Blanchard and Aoife Cahill. 2013. Summary Report on the First Shared Task on Native Language Identification. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NL*, Atlanta, GA, USA.

Victor H.A. Yngve. 1960. A model and an hypothesis for language structure. In *Proceedings of the American Philosophical Society*, 444–466.

Jason Weston, Sayan Mukherjee, Oliver Chapelle, Massimiliano Pontil, Tomaso Poggio and Vladimir Naumovich Vapnik. 2000. Feature selection for SVMs. In *Advances in Neural Information Processing Systems 13*, MIT Press, 668–674.