# Learning with Learner Corpora: using the TLE for Native Language Identification

**Allison Adams and Sara Stymne**
Linguistics and Philology
Uppsala University
aadams297@gmail.com,sara.stymne@lingfil.uu.se

## Abstract

This study investigates the usefulness of the Treebank of Learner English (TLE) when applied to the task of Native Language Identification (NLI). The TLE is effectively a parallel corpus of Standard/Learner English, as there are two versions; one based on original learner essays, and the other an error-corrected version. We use the corpus to explore how useful a parser trained on ungrammatical relations is compared to a parser trained on grammatical relations, when used as features for a native language classification task. While parsing results are much better when trained on grammatical relations, native language classification is slightly better using a parser trained on the original treebank containing ungrammatical relations.

## 1 Introduction

Native Language Identification (NLI), in which an author's first language is derived by analyzing texts written in his or her second language, is often treated as a text classification problem. NLI has proven useful in various applications, including in language-learning settings. As it is well-established that a speaker's first language informs mistakes made in a second language, a system that can identify a learner's first language is better equipped to provide learner-specific feedback and identify likely problem areas.

The Treebank of Learner English (TLE) is the first publicly available syntactic treebank for English as a Second Language (Berzak et al., 2016). One particularly interesting feature of the TLE is its incorporation of an annotation scheme for a consistent syntactic representation of grammatical errors. This annotation system has the potential to be useful to native language identification, as the ability to parse ungrammatical and atypical dependency relations could improve the informativeness of dependency-based features in such a classification task.

Assessing this potential has been accomplished by training a parser on the original treebank and using it to extract dependency relations in a learner English corpus. Those dependency relations were then used as features in a machine learning classification task. The success of this classification was then assessed by comparing the results to a classification on features extracted by a parser trained on the error-corrected version of the treebank, based on the assumption that the original version of the treebank will more accurately handle grammatical errors in learner texts. This is a novel approach in that other similar experiments have used dependency parsers trained on grammatical treebanks to extract dependency relations.

We found that using the original version of the corpus gave slightly better results on native language classification than using the error-corrected version. However, when we investigated parsing results, the original version gave much lower results on parsing both for original and error-corrected texts. This seems to suggest that there is useful information in the types of errors made by this parser.

## 2 Related Work

### 2.1 L1 Identification in L2 Texts

As mentioned in the previous section, the task of native language identification (NLI) involves determining a writer's first language (L1) by analyzing texts produced in their second language (L2). Language learner data is used to train clas-



Allison Adams and Sara Stymne 2017. Learning with learner corpora: Using the TLE for native language identification. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017.* Linköping Electronic Conference Proceedings 134: 1–7.

sifiers, such as support vector machines (SVM), for predicting the L1 of unseen texts. One of the first studies carried out in automatic L1 detection (Koppel et al., 2005) classified L2 texts using features such as function words, part-of-speech bigrams, and spelling and grammatical errors. The features were evaluated on a corpus of learner English, and the researchers ultimately found that by combining all of the features using a SVM, they could achieve an accuracy of 80.2% on the International Corpus of Learner English (ICLE) (Granger et al., 2002). Wong and Dras (2011) extended this study to include the use of syntactic features for this task by extracting features from parse trees produced by a statistical parser. In doing this, they incorporated production rules from two parsers: the Charniak parser as well a CFG parser. Other studies such as Swanson and Charniak (2012) make use of tree substitution grammars as a source of features for NLI. Several studies, such as Tetreault et al. (2012), Brooke and Hirst (2012), and Swanson (2013) have tested a range of features including dependency features, as well as combinations of features to ascertain which feature or ensemble of features is most useful. In doing so, they demonstrated the value of dependency features in classifying the L1 of texts.

In the case of Brooke and Hirst's study (2012), when running their system on both the FCE and the ICLE, after testing the usefulness of a range of different types of features, they found dependency features to provide a muted benefit to their system, with cross-validation resulting in accuracy scores of 61.4% for the ICLE and 45.1% on the FCE. They noted, however, that other features were more useful. Tetreault et al (2012) also tested a wide range of different types of features, testing their system also on the ICLE as well as the TOEFL11 corpus. By increasing the dependency relation feature set by including several different types of back-off dependency representations (described in section 3.2), they were able to raise accuracy of classification on the ICLE corpus to 77.1%, and reported an accuracy of 70.9% on the TOEFL11 corpus. Furthermore, the authors of the study found that classification accuracy was lowest for languages in the corpus with a high concentration of high-proficiency test responses, and best for higher concentrations of medium proficiency responses.

### 2.2 Universal Dependencies and the Treebank of Learner English

Dependency parsing has been rapidly gaining popularity over the past decade and differs from the older traditional constituency parsing in that in a dependency tree, the words are connected to each other by directed links (Kübler et al., 2009). The main verb in a clause assumes the position of the head, and all other syntactic units are connected to the verb by their links or dependencies to the head (Kübler et al., 2009). Annotated treebanks are typically used to generate dependency parsing models. The Universal Dependencies (UD) Project is a recent effort aimed at facilitating cross-lingual parsing development through the standardization of dependency annotation schemes across languages (Nivre et al., 2016). A central aspect to the UD project is the creation of open-source treebanks in a variety of languages that can be used to facilitate cross-lingual parsing research. All of the treebanks have been annotated according to the UD annotation scheme, in order to ensure consistency in annotation across treebanks. These guidelines have been developed with the goal of maximizing parallelism between languages (Nivre et al., 2016).

The Treebank of Learner English (TLE) is a part of the UD project and is a manually annotated syntactic treebank for English as a Second Language (Berzak et al., 2016). It includes PoS tags and UD trees for 5,124 sentences from the Cambridge First Certificate in English (FCE) corpus (Yannakoudakis et al., 2011). The treebank is split randomly in to a training set of 4,124 sentences, a development set of 500 sentences and a test set of 500 sentences. Ten different language backgrounds are represented in this corpus: Chinese, French, German, Italian, Japanese, Korean, Portuguese,Spanish, Russian and Turkish. For each language background, the TLE contains 500 randomly sampled sentences from the FCE data set, in order to ensure even representation. All sentences included in the TLE were selected so that they contain grammatical errors of some kind. The creators of the treebank exploit a pre-existing error annotation scheme in the FCE, adapting it to fit UD guidelines. In this scheme, full syntactic analyses are provided for the error corrected and original versions of each sentence. This in conjunction with additional ESL annotation guidelines provide for a consistent syntactic treatment of ungrammat-

| Language | Low | Medium | High |
|---|---|---|---|
| Arabic | 296 | 605 | 199 |
| Chinese | 98 | 727 | 275 |
| French | 63 | 577 | 460 |
| German | 15 | 412 | 673 |
| Hindi | 29 | 429 | 642 |
| Italian | 164 | 623 | 313 |
| Japanese | 233 | 679 | 188 |
| Korean | 169 | 678 | 253 |
| Spanish | 79 | 563 | 458 |
| Telugu | 94 | 659 | 347 |
| Turkish | 90 | 616 | 394 |
| **Total** | 1330 | 6568 | 4202 |

Table 1: Score level distributions in TOEFl11

ical English.

### 2.3 TOEFL11 Corpus

The TOEFL11 corpus was designed specifically with the task of NLI in mind, and comprises 12,100 learner essays written as a part of the standardized English language test, TOEFL (*Test of English as a Foreign Language*) (Blanchard et al., 2013). As the name of the corpus implies, 11 language backgrounds are included in the corpus: Arabic, German, French, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish, and Chinese. These language backgrounds are distributed evenly across the corpus, with 1,100 essays per language, and an even sampling across responses to eight different prompts. All essays have been graded according to proficiency as *high*, *medium*, or *low*, and have not been sampled evenly across L1s. The thought behind this is that proficiency score distributions in the corpus ought to correspond to the real-life score distributions in test results, as this information may be relevant and useful to L1 classification. The distribution of score levels per language can be found in Table 1.

## 3 System

### 3.1 Parsing the corpus

For the purposes of this paper, five dependency parsers were trained using MaltParser (Nivre et al., 2007). Three parsers were trained using the TLE as a training corpus. We also trained two contrastive parsers on the English Web Treebank (EWT), a UD treebank of English containing documents from five genres: weblogs, newsgroups, emails, reviews, and Yahoo! Answers (Silveira et al., 2014). The sizes of the treebanks are shown in Table 2. The EWT is substantially larger than the TLE. In order to investigate the effect of corpus size to some extent, we also used half of the EWT to train a parser.

| | Sentences | Words |
|---|---|---|
| TLE | 4124 | 78541 |
| EWT | 12544 | 204586 |
| EWT 50% | 6272 | 101101 |

Table 2: Size of the training corpora for the parsers.

Of the three parsers trained on TLE, the first parser was trained on the original version of the TLE (containing grammatical errors), while the second parser was trained on the corrected version of the TLE. A third parser was trained on a hybrid version of the original and corrected treebanks, the driving idea behind this being that while the corrected version of the treebank would be ill-equipped to model grammatical errors in dependency parse trees, the original version of the treebank, in which every sentence contained at least one error, would be hard-pressed to accurately model entirely grammatical sentences. To keep the size of all three treebanks consistent, the merged treebank was created by taking every other sentence from the original and corrected treebanks. In this scheme, the same sentences (save for the minor differences in the corrected sentences) are represented in all three treebanks. Because MaltParser requires texts to be part-of-speech-tagged in order to be parsed, the HunPos part-of-speech tagger (Halácsy et al., 2007), trained on the EWT was used to acquire PoS tags for each document in the TOEFL11 corpus. All three parsers were then run on each part-of-speech-tagged document using default parameter settings, resulting in three individual parsed data sets.

In order to estimate the accuracy of the parsing models, we evaluated them on three test sets from the TLE, original, corrected, and merged, created by applying the same process described earlier. The accuracy of the parsers is assessed by means of labeled and unlabeled attachment scores (LAS and UAS), the results of which can be found in Table 3. As established by Berzak et al. (2016), parsers trained on both the corrected and original versions of the TLE outperform the parser trained on a standard English treebank, with the merged

| Train Set | Test Set | LAS | UAS |
|---|---|---|---|
| $TLE_{corr}$ | *corrected* | **94.5** | **95.6** |
| | *original* | **90.1** | **92.2** |
| | *merged* | **92.5** | **94.1** |
| $TLE_{orig}$ | *corrected* | 85.7 | 88.5 |
| | *original* | 85.1 | 88.0 |
| | *merged* | 85.2 | 88.0 |
| $TLE_{merged}$ | *corrected* | 85.0 | 88.1 |
| | *original* | 85.0 | 88.0 |
| | *merged* | 85.4 | 88.0 |
| EWT | *corrected* | 80.7 | 86.0 |
| | *original* | 80.6 | 86.1 |
| | *merged* | 80.8 | 86.0 |
| EWT 50% | *corrected* | 79.8 | 85.4 |
| | *original* | 79.3 | 85.0 |
| | *merged* | 80.0 | 85.5 |

Table 3: Parser accuracies for all three test sets.

version of the TLE following this trend as well. Interestingly, however, and contrary to assumptions made in the beginning of this paper, the parser trained on the corrected version of the treebank considerably outperformed both the original and merged versions of the treebank on all three test sets. The two parsers trained on the EWT had considerably lower scores than any parser trained on TLE. The difference in training data size between the two EWT parsers was small, in comparison.

### 3.2 Using dependency arcs as features

Similar to most other NLI systems, in this paper, the task of native language identification is approached as a text classification problem. In order to solve this classification problem, dependency relations were extracted from each document to be used as frequency-based features. To do this, a system similar to the one presented in (Tetreault et al., 2012) was used, with the main difference being that MaltParser, rather than the Stanford Dependency parser was used to obtain them. This system, represented below in Table 4, can be described as follows: each basic dependency relation, consisting of the dependency label, the parent node, and the child node is extracted from the sentence. To mitigate sparsity, each dependency in the document was represented in several different ways. In the first representation, the lemmas for the root and child node were used to form the dependency relation. Secondly, part-of-speech tags were considered instead of lemmas, with dependency relations consisting of the dependency label, one lemma, and one PoS tag, or a dependency label and two PoS tags. Lastly, the corresponding dependency relations without labels were also incorporated into the feature set. In this work we only used parsing-based features and do not combine them with other feature sets. From the parsing output for each parsing model on the 12,100 essays in TOEFL11 corpus, we extracted on average just over 1.5 million features. Once the feature set was established, a support vector machine (SVM) was used to classify the data set. Scikit Learn's LinearSVC (Pedregosa et al., 2011), which is powered by liblinear (Fan et al., 2008), set with default parameter settings was used to carry out the classification.

| | |
|---|---|
| dep(lemma, lemma) | (lemma, lemma) |
| dep(PoS, lemma) | (PoS, lemma) |
| dep(lemma, PoS) | (lemma, PoS) |
| dep(PoS, PoS) | (PoS, PoS) |

Table 4: Types of dependency relations used in feature set

### 3.3 Results

To evaluate the three systems, we used 10-fold cross-validation. As the classification report featured in Table 5 shows, differences between the three models trained on TLE were negligible, with the model based on the original version of the TLE slightly outperforming the other two models across all metrics, but to only a very marginal degree (a couple of tenths of a percentage point most often). The model trained on the full EWT preformed as well as the model trained on the original TLE, whereas the model trained on half EWT had the lowest core of all models. This indicates that the size of the corpora is indeed important, and that considerably more out-of-domain data is needed to have a performance on par with smaller in-domain data.

The hybrid model, which contained features extracted by a parser trained on a merged version of the original and corrected treebanks performed nearly as well as the model based on the original treebank. Contrary to our hypothesis that higher parser accuracy ought to correlate to a higher classification accuracy, despite having LAS and UAS scores nearly five points above the other TLE two models, the corrected model had the lowest classification performance of the three. The full EWT model with a much lower parsing accuracy also

| | Acc | P | R |
|---|---|---|---|
| Original | 70.5 | 70.7 | 70.6 |
| Corrected | 70.2 | 70.3 | 70.3 |
| Merged | 70.5 | 70.6 | 70.5 |
| EWT | 70.5 | 70.7 | 70.6 |
| EWT 50% | 70.0 | 70.1 | 70.0 |

Table 5: Accuracy, precision, recall for native language identification with the three parser models.

| Language | Original | Corrected | Merged |
|---|---|---|---|
| Arabic | **68.0** | 66.1 | 67.0 |
| Chinese | 74.3 | **74.7** | 73.5 |
| French | 70.1 | 71.0 | **71.2** |
| German | 81.5 | **82.5** | 81.7 |
| Hindi | **64.7** | 64.2 | 64.3 |
| Italian | 75.9 | **76.2** | 75.8 |
| Japanese | **71.3** | 70.5 | **71.3** |
| Korean | 63.7 | 62.7 | **64.5** |
| Spanish | 62.2 | 62.7 | **62.9** |
| Telugu | **71.5** | 71.1 | 71.3 |
| Turkish | **72.1** | 70.7 | 71.6 |

Table 6: Accuracy scores by language for all three models

performed on par with the best TLE model. This can also be compared to the 70.9% classification accuracy obtained using dependency relations as features in the study carried out by Tetreault et al. (2012), in which a standard English treebank was used, which however used both a different parser and different dependency relations. However, it still indicates that although all three TLE models perform relatively well, under this experimental set-up, using dependency features based on those found in the TLE does not improve results compared to using larger standard treebanks. On the contrary, these results point toward a negative correlation between parser and classification accuracy. This could indicate that, to some degree, the classification may actually be aided by the differences in types of errors the parser makes when it encounters ungrammatical syntactic constructions.

A more detailed breakdown of the model accuracies by language (found in Table 6) provides a limited degree of insight into why this is the case. Most accuracies within languages across the models varied only by a few tenths of a percentage point, with the largest deviations found in Arabic (with a 1.9 percentage point difference between the original and corrected models), Turkish, (1.4 percentage point difference between original and corrected models), and German (with a 1 percentage point difference between the original and corrected models). It had been expected that the most accurate parsing model would be best equipped to classify the languages with the highest concentration of low and medium proficiency scores, and would result in a less accurate classification for the languages in the corpus with a higher number of high scoring documents. This intuition is based on the notion that the former set of languages would have a greater percentage of erroneous dependency structures that would be consistently captured by the parsing model. The results, however, show this not to be the case. For example, Arabic and Turkish, both of which had a relatively low number of high scoring responses, preferred the original model, which had a much lower parser accuracy. This is further reinforced by the German classification accuracies, which had the highest concentration of high scoring responses, and was one of the only languages for which the corrected model performed best. It is also interesting to note that with German being the most accurately classified language, this goes against the findings of Tetreault et al. (2012), that high-profiency texts are generally harder to classify, suggesting that this trend does not hold for dependency-based classification. This also supports the notion that parser errors made due to ungrammatical dependency relations may help classification.

The surprising consistency across all three models may in fact show the degree of influence that part-of-speech tags have on MaltParser's output, regardless of the parsing model used to parse the data set. This might also reflect an underlying problem in the methodology. Due to factors of both convenience, and concerns about sparsity, HunPos, the part of speech tagger used to generate the PoS tags needed to be able to parse the corpus, was trained on the EWT, a Standard English corpus. As a result, the part of speech tags used were the same across all three models, which may have resulted in a larger degree of similarity across the dependency relations than had been anticipated. Furthermore, because the tagger was trained on texts generated by largely L1 speakers, the distribution and make-up of the part of speech tags projected on to TOEFL11 corpus might not

be reflective of those found in the TLE. Furthermore, in their study, Berzak et al. (2016) note that systematic differences in the EWT annotation of various parts of speech compared to the Universal Dependencies guidelines might also negatively affect performance. As Berzak et al. (2016) also found that combining the TLE with the EWT improved parsing accuracy and PoS tagging accuracy on their test set, an interesting point for future research could be applying that technique to this study, to see if results could be improved. In particular, it could be interesting to see if using this model to acquire part of speech tags has any affect on classification accuracy.

An additional possibility for future research, which lies outside of the scope of NLI, relates to the results of the parser accuracy tests described in section 3.1. The considerable improvement in parser accuracy on the uncorrected learner essays when trained on the corrected version of the treebank has intriguing implications for the automatic annotation of learner data. This should be further explored in future work, including a detailed error analysis of these results.

There are several ways in which this study could be improved with regard to NLI. In this work we did not optimize any of the models used. A further possibility is to combine our parse features with previously suggested features, such as language model features (Tetreault et al., 2012) or character n-grams (Ionescu et al., 2014). It would also be interesting to investigate if unlabeled learner data can be used to improve both the parsing results on learner texts and NLI.

## 4 Conclusion

This study investigated the potential of the use of the Treebank of Learner English to improve Native Language Identification. To do this, we proposed using the original version of the TLE, the corrected version, as well as a hybrid version consisting of sentences from both versions of the treebank to train three dependency parsing models using MaltParser. Each of those models was used to extract dependency relations from the TOEFL11 corpus, which were in turn used as features in a text classification task. While the classification model using features obtained using the original version of the TLE had better scores than the other two models, the differences in accuracy scores across all three models were small. It is interesting that even though the parser trained on the original model had slightly better classification results, it also had substantially lower parsing results than the parser trained on the corrected model. We also trained a contrastive system on the much larger English Web Treebank, which had even lower accuracy on parsing learner data, but performed on par with the TLE system on native language classification, while a parser trained on 50% of this treebank did not perform well. This provides an indication that both the size and domain of the training corpus are important.

## Acknowledgment

We would like to thank the three reviewers for their valuable comments, as well as Yevgeni Berzak for his help with providing the updated version of the corrected and original versions of the TLE. We would also like to thank the participants of the course Language Technology: Research and Development in Uppsala for their feedback on a first version of this paper.

## References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 737–746. Association for Computational Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2002. *International corpus of learner English*. Presses universitaires de Louvain.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Hunpos: an open source trigram tagger. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 209–212. Association for Computational Linguistics.

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

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous authors native language. In *International Conference on Intelligence and Security Informatics*, pages 209–217. Springer.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 193–197. Association for Computational Linguistics.

Ben Swanson. 2013. Exploring syntactic representations for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 146–151.

Joel R 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 the 24th International Conference on Computational Linguistics*, pages 2585–2602.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and 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-Volume 1*, pages 180–189. Association for Computational Linguistics.