Exploring the syntactic parsing of learner texts in Modern Greek: Variety trumps size, genre trumps grammaticality

Christina Klironomou[1] and **Arianna Masciolini**[2] and **Aleksandrs Berdicevskis**[2]
[1]University of Gothenburg
[2]Språkbanken; Department of Swedish, Multilingualism,
Language Technology; University of Gothenburg
[1] gusklirch@student.gu.se, [2]name.surname@gu.se

Abstract

We investigate which properties of train data affect the performance of syntactic parsers on learner texts. We train and evaluate six models for Modern Greek, using the same annotation scheme (Universal Dependencies) and the same parser (*MaChAmp*), but varying train and test sets. The data come from standard texts, learner texts and learner with errors corrected. We confirm that training and testing on the same language variety yields better results than, for instance, training on standard texts and testing on learner texts. Moreover, our experiment setup makes it possible to compare the role of variety to that of other factors and dissect the very notion *variety* into more specific factors. Most importantly, we show that training on the same variety has stronger effect than train set size, and that adding non-target variety to the train set does not necessarily decrease performance. In practical terms, it means that adding learner data (that are often scarce) to the standard train set is likely to boost the performance on learner data without damaging the performance on standard data. Further, we show that the same value of grammaticality across train and test does not have an effect, while having the same genre does.

1 Introduction

Second Language Acquisition (SLA) often relies on learner data annotated automatically for morphosyntax (see among others Kyle and Crossley 2017, Römer 2019, Hwang and Kim 2023, Römer-Barron 2024, Bozdağ et al. 2025 and Geremia et al. 2025). However, the nonstandard and sparse language phenomena that characterize this kind of material, such as transfer errors and other idiosyncrasies, are often challenging for tagging-parsing pipelines, which tend to handle them inconsistently. Annotation errors, in turn, can negatively affect the results of SLA studies. High-quality automatic annotation can further help identify linguistic patterns arising in second language use, as well as to the difficulties and strategies of learners, thus resulting in new, usage-based teaching methods. We therefore highlight the need of developing models that account for this data variability and better describe nonstandard language use.

When it comes to morphosyntax, Universal Dependencies (UD, de Marneffe et al. 2021) is one of the dominant annotation schemes across languages and varieties. Its wide selection of language-agnostic and language-specific categories allows a multi-level analysis that encourages a computer-assisted investigation of intralinguistic phenomena and cross-linguistic comparisons. Furthermore, UD data have been used to train models to perform high-accuracy automatic morphosyntactic annotation for a large number of major languages, and there has been a considerable number of attempts to extend the their capabilities to suitably characterize nonstandard data, e.g. dialectal data (Bompolas et al., 2025; Vakirtzian et al., 2025; Blaschke et al., 2024; Puddu et al., 2026) and user-generated content (Liu et al., 2018; Sanguinetti et al., 2018; Cignarella et al., 2019; Cassidy et al., 2022).

A growing number of studies on learner treebanks annotated in this framework (Berzak et al., 2016; Kyle et al., 2022; Sung and Shin, 2025; Pulido et al., 2025) report significant performance improvements upon integrating learner data into the train set. That the addition of training instances representing the target language variety produces a positive effect is expected. However, it remains to be seen which properties of the language variety are actually important, to what extent the improvement is to be attributed to *grammaticality* (presence or absence of grammatical errors) and

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 130–139, November 19–20, 2026, University of Tartu Library

genre. On the one hand, off-the-shelf parsers tend to underperform in the presence of learner errors, which makes it reasonable to expect that introducing gold-annotated ungrammatical sentences into the training mix would help the models handle them as expected by the annotators. On the other, learner treebanks typically consist of essays, a text genre with little to no presence in the reference datasets normally used to train parsing models. Finally, not all of the above mentioned studies control for train set *size*.

In this paper, we try to disentangle these three aspects. To do that, we conduct a case study on modern Greek using one of the reference large-scale datasets, the *UD_Greek-GDT* treebank (Prokopidis and Papageorgiou, 2017), and *UD_Greek-GLCII* (Klironomou et al., 2026), a parallel treebank that includes both original and corrected learner essays. The latter gives us an opportunity to quantify the effects of text genre independently from grammaticality. In addition, we investigate whether the use of learner data in training produces a decline in performance when the model is used to parse standard modern Greek.

2 Data

In our experiments, we use one of the two released Standard Modern Greek treebanks, *UD_Greek-GDT* (GDT) (Prokopidis and Papageorgiou, 2017). *UD_Greek-GDT* is the UD-converted version of the *Greek Dependency Treebank*,[1] and contains multi-genre texts from the public domain. It was preferred over the other treebanks available for modern Greek due both to its larger size and to the fact that the its internal guidelines served as a starting point for the annotation of the learner treebank.

The *UD_Greek-GLCII* treebank (Klironomou et al., 2026) includes sentences stemming from the development set of the Greek Learner Corpus II (GLCII) (Tantos et al., 2023), an error-annotated corpus of texts from adult learners of Greek as a second language. The corpus covers a wide range of sociolinguistic and cultural backgrounds, with learners of different L1s and CEFR levels. A correction hypothesis for every learner text is also available in the corpus. The annotation of the treebank was manual for the dependency relations and automatic with a manual inspection for all other linguistic levels of analysis (tokenization, lemmatization, POS tagging and feature assignment). Additionally, the treebank contains the annotated correction hypotheses. The original (learner) sentences of the GLCII treebank constitute the *learner-original* data that we used in our experiments, whereas their correction hypotheses constitute the *learner-corrected* data.

3 Experiments

The data we have enable us to partly tease apart factors that can potentially affect the quality of parsing, namely language variety, grammaticality, genre and size.

By *language variety* we mean the difference between standard language and learner language. We say that in our data, language variety is a combination of *grammaticality* and *genre*. By *grammaticality* we mean the presence or absence of any grammatical errors,[2] whereas *genre* is everything else that is different between standard and learner data apart from errors. Thus, genre may include certain stylistic features of the learner language that are not errors, and it subsumes also the difference in topics and formality covered in our learner and standard treebanks.

Given this definition, we say that learner-corrected data and the original learner data belong to the same genre, but have different grammaticality value. Learner-corrected and standard data are both grammatical, but differ in genre.

Given the available data, we ask the following research questions:

1. How does grammaticality affect parsing?
2. Is adding a non-target language variety to the train data harmful? In other words, if we test on language variety A, and have A in the train set, but also add B to the train set, will the performance go down?
3. What is more important for improving performance: training on the same language variety or having a larger train set?
4. What affects the performance more: differences between train and test sets in genre or in grammaticality?

[1] The Greek Dependency Treebank is a manually annotated for morphology, syntax and semantics resource of Modern Greek.

[2] Grammatical errors are only expected in learner texts, given that GDT's sentences stem from texts occurring in established informative websites or European parliament sessions.

In section 3.1, we describe the models that we train, and in Table 1 we show which pairwise comparisons of the model performance enable us to address which questions. In section 3.2, we describe our test settings and explain further the model comparisons that attempt to answer our research questions.

3.1 Models

In order to train our models, we use the syntactic parser of the *MaChAmp* toolkit (van der Goot et al., 2021) with its default hyperparameter values. The train data for the models are selected in specific proportions, to mitigate any data size or variety imbalance and achieve model comparability. We present the trained models below:

- ***learner-small*** (learner-only model): 90 percent (80:training, 10:dev) of the GLCII treebank (8,155 + 1057 tokens).
- ***corrections-small*** (standard-only model, equal train set size —in terms of sentences— to the *learner-small* model): 90 percent (80:training, 10:dev) of the corrected version of the GLCII treebank (8,479 + 1109 tokens).
- ***standard-small*** (standard-only model, equal train set size to the *learner-small* model): Train and development sets of GDT, normalized to the size of the GLCII treebank (7,934 + 1048 tokens).
- ***standard-large*** (standard-only model): Train and development sets of the GDT treebank (42,874 + 10,542 tokens).
- ***mixed-large*** (standard + learner, equal train set size to the *standard-large* model): Combination of the train and development sets of the *standard-large* model, minus the size of the *standard-small* [71.63%], and the *learner-small* model [28.37%] (43,095 + 10,584 tokens). When interpreting the experiments where this model is compared to *standard-large*, we ignore the fact that *standard-large* has seen a little bit more standard train data. In other words, we attribute all potential differences between the models to the presence of learner data in the *mixed-large* train set.
- ***mixed-large-corrections*** (standard + corrections, equal training size to the *standard-large* and *mixed-large* models): Combination of the training and development sets of the *standard-large* model minus the size of the *standard-small* (71.63%), and the *learner-corrections* model (28.37%) (43,419 + 10,636 tokens).

3.2 Test settings and model comparisons

In order to measure whether the occurrence of the learner data in the train set would impact model performance due to their specific and sometimes erroneous nature, we defined two test settings. One was a **learner test setting**, composed of the remaining 10 percent of the GLCII treebank. The other one was a **standard test setting**, composed of a portion of the GDT test set, normalized to the size of the GLCII test set.

In order to test whether different values of grammaticality between the train and test set negatively affect model performance (see research question 1), we compare the results of equal-size models whose train data have dissimilar values of grammaticality; In this way, one model is trained on data that have —at least partly— the same value of grammaticality as the test data, and the other on data with —at least partly— different grammaticality than the test. Thus, we have the below pairwise comparisons:

- *corrections-small* (same grammaticality as test) VS *learner-small* (different grammaticality) on standard test data;
- *corrections-small* (different grammaticality than test) VS *learner-small* (same grammaticality) on learner test data;
- *mixed-large-corrections* (same grammaticality as test) VS *mixed-large* (partly different grammaticality) on standard test data;
- *mixed-large-corrections* (different grammaticality than test) VS *mixed-large* (partly same grammaticality) on standard test data.

In order to test whether the existence of different language varieties between the train and test set have a negative effect on model performance (research question 2), we compare the results of models trained on —at least partly— different language varieties. Thus, we have the below pairwise comparisons:

ID	Train A	Train B	Test	Question	Result
a	corrections-small	learner-small	standard	1	$A \simeq B$
b	corrections-small	learner-small	learner	1	$A \simeq B$
c	mixed-large-corrections	mixed-large	standard	1	$A \simeq B$
d	mixed-large-corrections	mixed-large	learner	1	$A \simeq B$
e	mixed-large	standard-large	standard	2	$A \simeq B$
f	mixed-large	standard-large	learner	2	$A > B$
g	mixed-large	learner-small	learner	2	$A > B$
h	standard-large	learner-small	learner	3	$A < B$
i	corrections-small	standard-small	learner	-	$A > B$
	Difference A	**Difference B**			
j	a (grammaticality)	i (genre)		4	$A < B$

Table 1: Pairwise comparisons of the models. Overview of results with respect to the research questions.

- *mixed-large* (presence of a non-target language variety) VS *standard-large* (target variety only) on standard test data;
- *mixed-large* (presence of a non-target variety) VS *standard-large* (non-target variety only) on learner test data;
- *mixed-large* (presence of a non-target variety) VS *learner-small* (target variety only) on learner test data.

In order to test which of language variety and train set size impacts model performance the most (research question 3), we compare the results of models of unequal training size that have train data of different language varieties; In this way, the small model is trained on data that are of the same variety as the test data (target variety), whereas the other is trained on data of different variety than the test (non-target), but has an important size advantage. Thus, we have the below comparison:

- *standard-large* (different language variety than test) VS *learner-small* (same variety) on learner test data.

Finally, in order to test whether the genre or the grammaticality of the train data is more impactful when it comes to model performance (research question 4), we compare the difference between the scores of two pairs of models:

- *corrections-small* (same grammaticality as test) and *learner-small* (different grammaticality) on standard test data VS *corrections-small* (same genre as test) and *standard-small* (different genre) on learner test data.

Note that this comparison is slightly more complicated than the rest: this is not a comparison of two models, but the comparison of *differences* between one pair of models and another pair of models.

4 Evaluation

4.1 Quantitative evaluation

We perform a quantitative evaluation by calculating **accuracy scores** for **UPOS tagging**, **Unlabeled Attachment Score** (assigned head of the relation) and **Labeled Attachment Score**. Results for the six trained models are presented in Table 2.

To test to what extent random initializations and local maxima can affect the differences between models, we performed a pilot experiment. We trained 10 models on the same train set with different seeds (mixed-large) and then tested them on the standard and the learner test set. A visual summary of the results is reported in Figure 1. We conclude that the variance introduced by random seeds is relatively minor, and that we do not need to perform the multiple runs for every model comparison we do.

To avoid overinterpretation of small differences, we set a threshold of **three percentage points**. This threshold is an educated guess based on the random-seed pilot. There are other potential sources of randomness we do not account for, such as the choice of a parser or the choice of treebanks. In most cases, all three measures either pass or do not pass the threshold. In three cases (g-i), only LAS and/or UAS do, but the difference is in the same direction. We therefore treat these cases as above-threshold.

It can be easily seen from Table 2 that training and testing on the same language variety (target

	learner			standard		
	UPOS	**UAS**	**LAS**	**UPOS**	**UAS**	**LAS**
standard-small	90.06	82.43	75.40	95.23	87.55	83.55
corrections-small	92.17	86.75	81.33	92.39	80.25	74.79
learner-small	92.87	86.85	80.42	90.47	79.40	74.40
standard-large	90.36	85.24	77.21	98.69	92.47	89.93
mixed-large	94.38	89.26	83.73	98.39	93.31	90.47
mixed-large-corrections	93.47	88.55	83.43	98.23	92.47	89.78

Table 2: UPOS, UAS, and LAS scores for the six trained models, evaluated on learner and standard data.

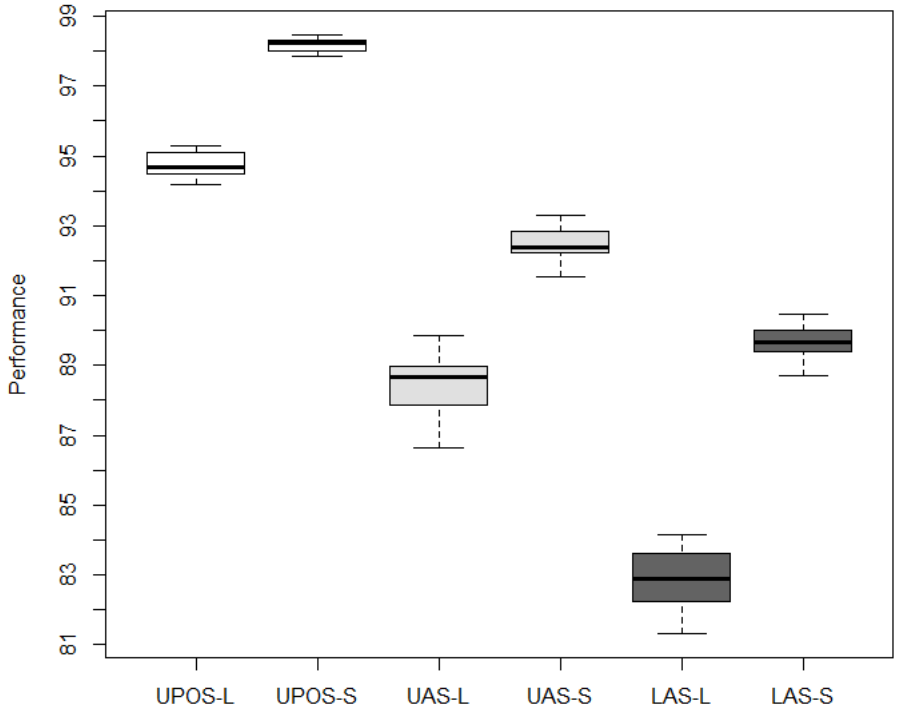


Figure 1: Variance in performance introduced by random seeds for the *mixed-large* model across the three measures. L = learner test set, S = standard test set.

variety) yields best results. Standard models perform better on standard data, while learner models perform better on learner data. Obviously, larger size is also beneficial: large models perform better than small models, all things being equal.

The *corrections-small* model performed similarly to *learner-small* when tested on either standard or learner data (with the *corrections-small* model performing slightly better —but not above the threshold— for all annotation layers on standard test data (score differences below 1 percentage point for UAS and LAS and 1.92 for UPOS) and the *learner-small* model slightly better than *corrections-small* on learner test data) (score differences below 1 percentage point). In addition, the *mixed-large-corrections* model performed almost equally well as *mixed-large* when tested either on learner or standard data (score differences below 1 percentage point). These results indicate that different values of grammaticality do not reduce parsing performance importantly.

The *mixed-large* model performed slightly better than *standard-large* when tested on standard data (less than 1 percentage point for all annotation layers), and robustly better on learner test data (more than 4 percentage points for all annota-

tion layers). Given that *mixed-large* and *standard-large* are comparable in size, we conclude that the presence of learner data (non-target variety) in the train set of *mixed-large* does not harm the performance on standard test data. Conversely, the presence of only standard data (non-target variety) in the train set of *standard-large* reduced performance on learner test data. Moreover, *mixed-large* performed better than *learner-small* when tested on learner data (importantly, more than 3 percentage points for UAS and LAS). Thus, the addition of data of a different language variety (standard) did not harm performance. Given the above results, we can suggest that adding data of a non-target variety does not harm performance (since more data can mean better performance), as long as there are at least some data of the target variety in the train set.

The *learner-small* model performed better than *standard-large* on learner test data (importantly, more than 3 percentage points for LAS). This result points towards that same language variety is more important than size for model performance.

When comparing the results of the *corrections-small* model VS the *learner-small* model to standard test data with the results of the *corrections-small* model VS the *standard-small* model to learner test data, we observe that the score differences between *corrections-small* and *learner-small* are minor (and below-threshold) juxtaposed with the much better performing *corrections-small* against *standard-small* (importantly, more than 4 percentage points for UAS and LAS). Thus, we suggest that differences in genre between the train and test set seem to be more critical than grammaticality for parsing.

Our results regarding research question 1 seem to align with Berzak et al. (2016)'s observation of a "negative, albeit limited effect of grammatical errors on parsing". In addition, model performance, calculated for UPOS, UAS and LAS on two different test settings, one on original learner data and one the corrected versions of the learner data, seem to display similar tendencies to what is encountered in our experiments (grammaticality seems to affect model performance less than domain,[3] as well as size less than domain, when there is no presence of target domain data in the train set).

[3]What multiple authors (e.g. Berzak et al., Di Nuovo et al.) call as *domain* corresponds to our definition of genre.

In Di Nuovo et al.'s experiments, pre-trained models on standard Italian are tested on learner Italian data (original, corrected and a combination of both); best performance is reported for the learner-corrected test data, and worst for the learner-original test data, pointing to a negative influence of ungrammaticality to parsing. However, the authors highlight the influence of domain as more determinate to parsing than grammaticality, which agrees to our results regarding research question 4 (genre trumps grammaticality).

The results of Pulido et al. (2025) initially seem to indicate something different, since the standard-only model performs slightly worse than the standard + learner model, but better than the learner-only model (except for the case of UPOS) when tested on learner data, implying an advantage of size over language variety. However, by performing an experiment on learner test data using a standard + learner model that favors the presence of learner data in the train set (by using subcorpora of the standard-only train dataset), the same authors report that when the learner data are proportionally more prominent, better results are achieved, as the increase of standard train data improves the performance of the model up to a certain point (approx. 40k tokens). This seems to point towards the direction that train data size is less important than variety for model performance, which is compatible to our findings with regards to research question 3.

When it comes to comparing the different annotation layers, robust differences between the paired models are most often observed for LAS followed by UAS, whereas score variation for UPOS is usually minor.

Overall, scores ranged from 74 to 90% accuracy, suggesting a noteworthy performance for all the models across test settings. This can be an indicator of no radical differences among the various train data, independently of their variety, genre or value of grammaticality; even though these variables can noticeably affect the parsing, it seems that the overall similarity among the train sets is enough to achieve adequately high scores.

Furthermore, Klironomou et al. (2026) point out that the language in the annotated learner texts does not deviate dramatically from standard language, possibly due to the high level of proficiency of the majority of the learners (see also their discussion in section 5). We consider this, along with

the consistency of the *GLCII* and *GDT* annotation choices, to be partly responsible for the relatively high scores obtained, even when no target variety data was involved in training.

4.2 Error analysis

To further analyze and interpret the above performance scores, we also conduct a qualitative evaluation, by manually inspecting the model predictions. To facilitate this process, we inspect the predictions for UPOS tagging and dependency relation of each model in comparison to the gold annotation of the test sets (learner or standard).

Overall, predictions of the different models which diverge from the gold annotation often refer to the same structures and sentences. Below, we present some patterns of annotation of the various models, which is different from the gold standard.

Inaccurate predictions concern:

1. Some proper nouns, which are not assigned the UPOS tag `PROPN`. We assume that capitalization errors may have influenced the model's choice of UPOS tag.
2. Some abbreviations that contain punctuation (full stop), which are assigned the UPOS tag `PUNCT`.
3. Some passive past participles (ending in -μένος-η-ο), which are assigned the UPOS tag `ADV` or `ADJ` instead of `VERB`.
4. Some numerical tokens, which are assigned the UPOS tag `NOUN` or `PROPN` instead of `NUM`.
5. Some punctuation, which is assigned the `NOUN` UPOS tag instead of `PUNCT`, when there are consecutive instances of punctuation in the sentence.
6. Some adjectives denoting quantities (e.g. πολύς 'many/much') and some indefinite pronouns (e.g. κάθε 'every'), which are assigned the `ADJ` UPOS tag instead of `DET`.
7. The auxiliary είμαι 'to be', which is attributed a few times the UPOS tag `VERB` instead of `AUX`.
8. The `parataxis` relation, which is either over- or under-used, and generally used interchangeably with the `conj` relation, especially in cases where where rare conjunctions occur in (learner) sentences.
9. The `appos` relation, which is often under- or overused.
10. The `reparandum` and `compound` relations, which are not assigned at all, with the exception of the use of `compound` by the *corrections-small* model to the learner test set (`reparandum` occurs in gold learner test and `compound` in gold standard test).
11. The `expl` relation, which is is under-used (`expl` only occurs in gold learner test).
12. The `vocative`, `discourse` and `flat` relations, which are over- or under-used or not used at all (though `flat` is mostly used correctly by all models in the standard test set).
13. The `orphan` relation, which is overused by *standard-small*.

We identify the following possible causes of errors:

- a learner sentence contains non-standard usage (for instance, wrong stress or spelling, wrong use of case or missing content word that hardens the syntactic analysis)
- a syncretic form is difficult to disambiguate;
- a sentence has an infrequent construction or contains a non-compositional meaning;
- the correct UPOS tag or dependency relation is infrequent in the training data;
- it is difficult to choose between two or more alternatives because the distinction is very subtle or underspecified in the guidelines (cf. cases 5-6 in Table 4 in (Manning, 2011);
- there are several possible annotation choices;
- there are errors in the gold standard.

For subtle or under-specified distinctions, see points 8-9 in the list above as examples.

For infrequent UPOS tags or dependency relations in the training data, see `reparandum` relation and `discourse` in points 9 and 11 above. More specifically, the `reparandum` relation does not occur at all in the standard train set, and `discourse` only a couple of times.

For several possible annotation choices, consider the following examples, where we consider

model predictions which differ from the gold annotation but are still compliant with the UD guidelines: The orphan relation (see point 13) is used differently in standard train (almost for any case of head ellipsis) than learner train (only for elements that cannot be safely connected to another element in terms of syntax or meaning).[4]

Moreover, in the standard treebank, foreign names (often in other alphabets) are left unanalyzed through the attribution of the X UPOS tag (partly following the Universal guidelines for foreign expression and code-switching[5] where this approach is termed "foreign analysis"). In the learner treebank, on the other hand, they are annotated fully, following the guidelines for code-switched analysis (described in the same page of the Universal guidelines); Thus, the models trained on learner or learner-corrected data only do not assign the X tag at all, even though there are occurrences of the tag in the gold annotation of standard test.

For errors in the gold standard, consider the following example: The word θεμελιώδης 'fundamental' is correctly assigned the UPOS tag ADJ by all the models in standard test, whereas in the gold annotation it is assigned the tag NOUN.

5 Conclusions

There are multiple factors that are involved in automatic parsing and can affect model performance. In this study, we tackled factors concerning the nature of the train data in relation to the test data, namely language variety (and more specifically, genre and grammaticality) and size; in order to test to what extent these different variables influence performance, we trained six models for morphosyntactic parsing with standard, learner-original and learner-corrected data.

Models trained on data with different values of grammaticality (*corrections-small* VS *learner-small* and *mixed-large-corrections* VS *mixed-large*) performed similarly across test settings. We therefore suggest that ungrammatical train data do not necessarily affect parsing quality.

Moreover, a comparison of models with train data of (partly) different language varieties (*mixed-large* VS *standard-large* and *mixed-large* VS *learner-small*) tested on data of (partly) the same or different variety indicated an increasing performance trend when more data were added to the training independently of their variety, but *only* with the presence of at least some same-variety data in the train set.

A comparison of models of unequal size (*standard-large* VS *learner-small*), with the small model having the advantage of being trained on same-variety data as the test data, showed a clear prominence of language variety over size; *learner-small* achieved a substantially better performance, despite *standard-large* being five times larger.

Performance differences between the two small models trained on learner data (*corrections-small* and *learner-small*) on the standard test set were larger than those between the two that were trained on grammatical data (*corrections-small* and *standard-small*) and tested on learner data. We interpret this as to indicate that genre affects performance more than grammaticality.

Our qualitative analysis revealed minor annotation differences between the models, some of which relate to learner-specific phenomena. However, we note that not all divergences from the gold standard imply that parser predictions are non-compliant with the UD annotation guidelines. A few possible sources of annotation error were given as well.

All in all, the practical takeaway is that, when it comes to learner data, parsing accuracy benefits from enriching the train set with same-variety data, whose (potential) ungrammaticality does not negatively affect performance on standard data. In fact, the highest attachment scores were obtained upon evaluating the *mixed-large* model on the standard test set. Differences with other large models, however, are too small to draw any conclusions: Whether the addition of (ungrammatical) learner data also improves the overall robustness of the parser remains to be assessed in future studies.

Acknowledgments

CK has been supported throughout the realization of this project by Onassis Foundation, under the Scholarship Program for Greek students 2025-2026. AM and AB have been supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161).

[4] For the UD guidelines concerning the orphan relation, see https://universaldependencies.org/u/dep/orphan.

[5] https://universaldependencies.org/foreign

AI Disclosure

☒ Not applicable (no AI tools used)
☐ Assistive use
☐ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: No AI tools were used in preparing this work.

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 (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Stavros Bompolas, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2025. Crossing dialectal boundaries: Building a treebank for the dialect of lesbos through knowledge transfer from standard Modern Greek. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 39–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Fatih Ünal Bozdağ, Junhua Mo, and Gareth Morris. 2025. A corpus-based analysis of noun modifiers in L2 writing: The respective impact of L2 proficiency and L1 background. *Plos one*, 20(3):e0320092.

Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster. 2022. TwittIrish: A Universal Dependencies treebank of tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6869–6884, Dublin, Ireland. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France. Association for Computational Linguistics.

Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1).

Sara Geremia, Thomas Gaillat, Nicolas Ballier, and Andrew J. Simpkin. 2025. Exploring the cross-lingual influence of linguistic complexity in second language writing assessment. *Assessing Writing*, 66:100951.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Haerim Hwang and Hyunwoo Kim. 2023. Automatic analysis of constructional diversity as a predictor of efl students' writing proficiency. *Applied Linguistics*, 44(1):127–147.

Christina Klironomou, Thelka Pasparaki, Arianna Masciolini, Alexandros Tantos, Despoina Ourania Touriki, Konstantinos Tsiotskas, and Eleni Tsourilla. 2026. Towards Universal Dependencies for L2 learners of modern Greek: Annotation and challenges. In *Proceedings of the Ninth workshop on Universal Dependencies (UDW, LREC 2026)*.

Kristopher Kyle and Scott Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'11, page 171–189, Berlin, Heidelberg. Springer-Verlag.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Nicoletta Puddu, Manuela Sanguinetti, and Luigi Talamo. 2026. Introducing Universal Dependencies for Sardinian: the UD ContSar treebank. In *Proceedings of the Ninth Workshop on Universal Dependencies (UDW 2026)*, pages 242–251, Palma de Mallorca, Spain. ELRA Language Resources Association (ELRA).

Emiliana Pulido, Robert Pugh, and Zoey Liu. 2025. I speak for the árboles: Developing a dependency treebank for Spanish L2 and heritage speakers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 814–822, Vienna, Austria. Association for Computational Linguistics.

Ute Römer. 2019. A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, 24(3):268–290.

Ute Römer-Barron. 2024. How do constructions with modal verbs develop in second language learners of English? *Journal of Second Language Studies*, 7(2):198–226.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hakyung Sung and Gyu-Ho Shin. 2025. Towards robust morphosyntactic analysis of l2 korean: Evaluating and fine-tuning a korean language model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Alexandros Tantos, Nikolaos Amvrazis, and Eleni Drakonaki. 2023. Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics*, 36.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.