Bootstrapping the Annotation of UD Learner Treebanks

Arianna Masciolini
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
arianna.masciolini@gu.se

Abstract

Learner data comes in a variety of formats, making corpora difficult to compare with each other. Universal Dependencies (UD) has therefore been proposed as a replacement for the various *ad-hoc* annotation schemes. Nowadays, the time-consuming task of building a UD treebank often starts with a round of automatic annotation. The performance of the currently available tools trained on standard language, however, tends to decline substantially upon application to learner text. Grammatical errors play a major role, but a significant performance gap has been observed even between standard test sets and normalized learner essays. In this paper, we investigate how to best bootstrap the annotation of UD learner corpora. In particular, we want to establish whether Target Hypotheses (THs), i.e. grammar-corrected learner sentences, are suitable training data for fine-tuning a parser aimed for original (ungrammatical) L2 material. We perform experiments using English and Italian data from two of the already available UD learner corpora. Our results show manually annotated THs to be highly beneficial and suggest that even automatically parsed sentences of this kind might be helpful, if available in sufficiently large amounts.

Keywords: second language acquisition, learner corpora, dependency parsing, universal dependencies

1. Introduction

In recent years, Second Language Acquisition (SLA) has become more and more reliant on corpus studies, to the point of Learner Corpus Research becoming a well-established field, as attested by the founding of the Learner Corpus Association[1] and the institution of a dedicated journal[2]. Learner data, however, comes in a variety of formats depending on each corpus' original purpose. This makes such datasets difficult to reuse and hardly comparable with each other. In this sense, linguistic annotation in Universal Dependencies (UD) (de Marneffe et al., 2021) is an appealing alternative to the various existing *ad-hoc* annotation schemes. UD would in fact provide a uniform annotation layer not only across datasets, but also across languages.

In particular, Lee et al. (2017) proposed *L1-L2 parallel dependency treebanks*, consisting of UD-annotated learner sentences paired with *correction* or *target hypotheses* (henceforth THs) as a replacement for explicitly error-tagged corpora.[3] The key idea is that systematic cross-linguistically consistent morphosyntactical annotation is sufficient for retrieving grammatical errors via tree queries, as demonstrated in Masciolini (2023). In addition, UD-annotated data lends itself to comparative cross-language studies and other types of analyses, both quantitative and qualitative. L1-L2 treebanks of different sizes have been released for English (Berzak et al., 2022), Chinese (Lee et al., 2023) and Italian (Di Nuovo et al., 2023), and we have the medium-term goal of releasing a fourth one based on the Swedish Learner Language (SweLL) corpus (Volodina et al., 2019).

Nonetheless, building a high-quality UD corpus requires in-depth knowledge of the annotation guidelines and remains a time consuming task even for expert annotators. For this reason, most treebanks are, rather than annotated from scratch, the result of a process where the output of an automatic parser is used as a basis for manual validation and editing. The performance of off-the-shelf UD parsers, however, is often unsatisfactory on learner text, independent of the L2 and parser in question (Huang et al., 2018; Di Nuovo et al., 2022; Volodina et al., 2022; Sung and Shin, 2023).

In this paper, we address the problem of how to best bootstrap the annotation of UD learner corpora. More specifically, we hypothesize that part of the decline in performance observed upon evaluating standard tools on L2 material is due to differences between the training and test domain that go beyond grammaticality. Learner sentences, for instance, may be unidiomatic without necessarily containing an error (cf. Table 1 for examples in English and Italian). Our research question therefore becomes whether utilizing THs in the training of a dependency parser is helpful for parsing original learner sentences and, if so, whether automatically annotated THs suffice for this use case.

To find out, we fine-tune an array of parsers on

[1] learnercorpusassociation.org

[2] benjamins.com/catalog/ijlcr

[3] In the expression "L1-L2 parallel dependency treebank", "L2" indicates original learner material, while "L1" refers to THs, assumed to be native-liked.

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 111–117

	LEARNER SENTENCE	TARGET HYPOTHESIS
EN	For electrical goods, there will be no any kind of electrical products except computer.	Regarding electrical goods, there will not be any kind of electrical product except computers.
IT	in quello momento era lei, che diventa furiosa!	In quel momento era lei che diventava furiosa!

Table 1: Example sentence-correction pairs from the two datasets used in our experiments, the ESL and VALICO-UD treebanks. The Italian sentence can be translated as "In that moment, she was the one who was getting furious!". Note how both THs are grammatically correct but might be perceived as unidiomatic: a more proficient English speaker would probably use the word *electronics* rather than the expression *electrical goods/products*, while native-like Italian speakers tend to use the inchoative verb *infuriarsi* more than the construction *diventare furiosi* (literally "becoming furious").

both manually and automatically annotated THs from two largest available L1-L2 treebanks, the English as Second Language (ESL) treebank (Berzak et al., 2022), and the VALICO-UD treebank of learner Italian (Di Nuovo et al., 2023). We then evaluate their performance on normative data, unseen THs and, crucially, original learner sentences, comparing it with that of baselines trained on large-scale reference treebanks.

2. Related work

Nonstandard language in general and learner language in particular still pose significant challenges for automatic annotation tools. Early experiments using the Turbo parser (Martins et al., 2013) on L1-L2 English data showed that grammatical errors negatively affect parser performance (Berzak et al., 2016). This was confirmed by a systematic study on dependency parsing for learner English, which concluded that, despite often misleadingly high overall scores, all tools considered were vulnerable to grammatical errors (Huang et al., 2018).

More recently, Di Nuovo et al. (2022) evaluated a UDPipe 2 model trained on standard Italian on an L1-L2 treebank. They reported a substantial decline in performance on L2 originals, but also a more modest one on THs. Similarly, Volodina et al. (2022) assessed the accuracy of the Sparv annotation pipeline (Borin et al., 2016) on both original and normalized L2 Swedish sentences from the Swedish Learner Language corpus (SweLL) (Volodina et al., 2022) as well as on a corpus of Swedish course books, COCTAILL (Corpus of CEFR[4]-based Textbooks as Input for Learner Levels' modelling) (Volodina et al., 2017). They observed both an 11-percentage-points performance gap between the original L2 Swedish sentences and the coursebook material, and a significant - although smaller - discrepancy between the latter and normalized learner data. In addition, they reported a strong correlation between the parsers' performance on L2 texts - both normalized and not - and their authors' CEFR-based proficiency level.

Work on parsers specifically meant for L2 material is limited, although notably Sakaguchi et al. (2017) combined dependency parsing with Grammatical Error Correction (GEC), building an error-repairing parser for learner English. To the best of our knowledge, however, all previous studies have focused specifically on dealing with ungrammatical input, while no attempts have been made to adapt parsers to the broader domain of learner essays.

3. Parsing experiments

As mentioned in the introduction, our goal is to find out whether corrections are suitable data for fine-tuning a parser aimed for original learner texts. To do that, we use MaChAmp (Massive Choice, Ample tasks) (van der Goot et al., 2021) to train and compare an array of models on both manually annotated (gold) and automatically parsed (silver) THs from two of the available L1-L2 treebanks.

MaChAmp is a toolkit that allows easy fine-tuning of deep contextualized embeddings for a variety of linguistic annotation tasks. It has been shown to be especially beneficial in cases where multiple datasets are available for the same task. This is exactly our case, as we want to combine large-scale UD treebanks of standard language with smaller, domain-specific training sets derived from the aforementioned learner treebanks.

In a nutshell, our approach consists of selecting a suitable BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2018) and fine-tuning it for dependency parsing on the largest UD treebank available for the language at hand until its performance is comparable to that of off-the-shelf tools. This leaves us with strong baselines to compare our specialized models with. We then continue training on silver- and, when available, gold-annotated THs. In this further fine-tuning step, the baseline pre-trained dependency parser is specialized on the specific domain of normalized learner essays. An alternative to this kind of sequential training would be building a single, larger

[4]Common European Framework of Reference for languages.

treebank	language	# sentences		
		train	dev	test
EWT	standard en	12544	2001	2077
ESL	learner en	2×5124	2×100	2×5024
ISDT	standard it	13121	564	482
VALICO-UD	learner it	2×1613	2×233	2×398

Table 2: Summary of the datasets used in our experiments. Note that ESL and VALICO-UD consist of L1-L2 sentence pairs and that VALICO-UD's development set was sampled from its training set.

training set by mixing the reference treebanks with the THs. Creating and experimenting with different mixes, however, requires training multiple models largely on the same reference data, with the energy and time costs this implies. Our approach, on the other hand, only adds a few epochs of domain-specific tuning to the more resource-intensive training of the baselines, which is only carried out once.

It must be kept in mind that our current aim is not to build a general-purpose parser robust to learner language, but to develop a simple method to maximize parsing performance on a highly specific domain, even at the cost of a significant performance drop on standard language. This is because the resulting parser is meant to be used to speed up a single annotation effort. At the same time, however, we are interested in assessing whether and to what degree the introduction of THs negatively affects model performance on the standard test sets. We also want to compare the results obtained on original L2 sentences, which remain at least partially out-of-domain, with the performance on unseen THs. For these reasons, we test all of our models on all three evaluation sets at our disposal: that of the reference treebank and, when it comes to the learner corpora, both the L1 (TH) and L2 portions of their respective test splits.

Even though our models are trained in a multi-task setting,[5] we focus on dependency annotation in its strictest sense. This is both for the sake of compactness and due to the fact that, when it comes to learner language, dependency parsing has been shown to be more problematic than most other linguistic annotation tasks (Volodina et al., 2022). We therefore evaluate our models only in terms of Labelled and Unlabelled Attachment Scores (henceforth LAS and UAS) (Kübler et al., 2009), computed with the official CoNLL-18 evaluation script (Zeman et al., 2018).

[5]The Italian model produces complete CoNNL-U files, while the English one is only trained for dependency parsing and POS tagging, as the ESL treebank does not provide any information regarding lemmatization or morphological features.

3.1. English

In our first experiment, we fine-tune the original monolingual English BERT model (Devlin et al., 2018). We train our baseline on the UD English Web Treebank (EWT), the gold standard dependency corpus for English (Silveira et al., 2023), using MaChAmp's default hyperparameters. As can be seen in Table 3, the resulting performance even slightly surpasses the LAS and UAS scores reported for the UDPipe 2.12 model trained on EWT we use for comparison (Straka, 2023).[6]

The THs used in the additional fine-tuning passes come from the English as a Second Language (ESL) treebank (Berzak et al., 2022),[7], which is in turn based on the First Certificate in English (FCE) corpus (Yannakoudakis et al., 2011). The latter is a collection of short essays produced by learners with widely different language backgrounds, all taking the FCE exam, which assesses English at an upper-intermediate level (B1 in terms of the CEFR) . As indicated in Table 2, the ESL treebank consists of 10000+ manually annotated L1-L2 sentence pairs, pre-split in a roughly same-sized training and test set and a smaller development set. Unlike most medium- to large-scale treebanks, ESL is manually annotated completely from scratch, with the goal of avoiding any potential annotation biases. Annotation is however limited to dependency labels and Part-of-Speech tags.

The default 20 training epochs were enough for the baseline to learn from the standard-language treebank. Consequently, on account of the training set sizes, we do not expect this further fine-tuning step to require more than 8 epochs. As MaChAmp allows for epoch-wise monitoring of development set performance scores, as well as because overfitting is not the main concern for our use case, however, we set the limit to 10. We then train our first specialized model, FT-GOLD, using the THs from the gold-annotated train and development splits of the ESL treebank. Indeed, most of the learning happens during the first 7 training epochs and scores start oscillating slightly after epoch 8, but peak performance on the development set is reached after training for all 10 epochs. As a consequence, we use the same settings for the FT-SILVER model. The only difference between the two is the training data: the latter uses automatically parsed versions of the same sentences, obtained by re-annotating them

[6]Note, however, that the comparison between MaChAmp UDPipe 2 scores is not exact, as the UDPipe 2 model was trained and evaluated on the latest versions of the treebank, which is in a format not yet fully supported by MaChAmp. For this reason, all data was preprocessed with the cleanup script provided as part of the MaChAmp toolkit before using it with our models.

[7]This treebank is also known as the Treebank of Learner English (TLE).

	EWT		ESL L1		ESL L2	
	LAS	UAS	LAS	UAS	LAS	UAS
BASELINE	**91.79**	**93.64**	86.43	90.18	85.21	89.38
FT-GOLD	84.32	88.67	**98.92**	**99.65**	**95.28**	**97.05**
FT-SILVER	86.61	90.55	90.70	93.44	89.32	92.46
UDPIPE 2	90.56	92.62	90.70	93.44	89.42	92.51

Table 3: LAS and UAS scores for all three evaluation sets for the full-scale English experiment.

with the same UDPIPE 2 model used as a reference for the baseline.

Results, summarized in Table 3, clearly show gold-annotated THs to produce an important performance improvement on ESL data over both the MACHAMP baseline and the UDPIPE pretrained model. Fine-tuning on automatically annotated THs results in a more modest improvement over the MACHAMP baseline, but is substantially equivalent to using the UDPIPE 2 EWT model. The latter, in fact, seems to have much better cross-domain generalization capabilities than our baseline, to the point that it performs slightly better on the THs than on its own test set. Finally, we note that the scores on the EWT evaluation set are higher for the FT-SILVER model than for FT-GOLD. This is unexpected, but possibly due to the fact that the silver-annotated THs follow the exact same annotation conventions as the EWT, as the UDPIPE 2 model has been trained on the EWT itself.

3.2. Italian

We repeat the same experiment with Italian data. This time, the starting pretrained model is an Italian BERT (MDZ Digital Library team at the Bavarian State Library, 2021) and the baseline trained on the Italian Standard Dependency Treebank (ISDT) (Simi et al., 2023).

Learner data comes from the VALICO-UD corpus (Di Nuovo et al., 2022), a UD-annotated subset of the VALICO (*Varietà Apprendimento Lingua Italiana Corpus Online*, "online corpus of learner varieties of the Italian language"), an L2 Italian learner corpus elicited by comic strips (Corino et al., 2017). VALICO-UD comprises 237 texts written by L2 Italian learners, all native speakers of one of four Western European languages (English, French, German and Spanish). While there is no mention of CEFR levels, proficiency can be to some extent inferred from reported years of study, ranging from 1 to 4. VALICO-UD is therefore more homogeneous than the ESL treebank in terms of L1 backgrounds, but much more heterogeneous when it comes to proficiency. As displayed in Table 2, VALICO-UD is over four times smaller than its English counterpart in terms of total size. Furthermore, only its test set is manually validated, while the rest of the data is automatically parsed with the UDPIPE 2.12 ISDT model, meaning that it is not impossible to fine-tune on gold-annotated THs.

	ISDT		VALICO-UD L1		VALICO-UD L2	
	LAS	UAS	LAS	UAS	LAS	UAS
BASELINE	**93.64**	**95.21**	89.25	91.86	85.99	89.94
FT-SILVER	89.96	93.15	88.49	91.46	85.59	89.77
UDPIPE 2	93.34	94.96	**90.22**	**92.86**	**87.69**	**91.61**

Table 4: LAS and UAS scores for all three evaluation sets for the Italian experiment.

Nonetheless, we are interested in seeing whether the improvement over the MACHAMP baseline observed upon fine-tuning on silver THs in the English experiment can be replicated on a different dataset and, most importantly, with less training instances at our disposal. Since VALICO-UD does not come with a development set, we build one by randomly sampling sentences from the training data. The resulting development set is 10% of the total size of the corpus. In terms of hyperparameters, we stick to the same values used for the English experiment.

Unsurprisingly, results for this smaller dataset are less conclusive. Table 4 shows a pattern that is only partially similar to that of Table 3. On the one end, the performance of the fine-tuned model does decrease on the standard-language treebank while staying relatively high on the L1 and L2 evaluation sets. At the same time, however, none of the MACHAMP-based models outperforms UDPIPE on learner data, even if the MACHAMP baseline is marginally better on standard Italian.

3.3. Reducing the training set size

A simple explanation for the differences observed between the ESL and VALICO-UD-tuned silver models could be that the size of the Italian training set is too small to learn from THs. To test this hypothesis, we rerun the English experiment on a smaller sample of the ESL treebank, identical to the VALICO-UD training set in terms of number of sentences. Results, reported in Table 5, support this

	EWT		ESL L1		ESL L2	
	LAS	UAS	LAS	UAS	LAS	UAS
BASELINE	**91.79**	**93.64**	86.43	90.18	85.21	89.38
FT-GOLD-SAMPLE	84.11	88.70	**94.53**	**96.50**	**92.21**	**94.82**
FT-SILVER-SAMPLE	86.27	90.32	90.46	93.24	88.95	92.18
UDPIPE 2	90.56	92.62	90.70	93.44	89.42	92.51

Table 5: Scores for a smaller-scale English experiment, conducted by fine-tuning on a 1613-sentence ESL sample. We invite the reader to compare these results with those reported in Table 3 (same language, different training set size) and Table 4 (different language, same training set size).

only in part. If, as expected, both fine-tuned models are negatively affected by the lower amount of training instances, with FT-SILVER-SAMPLE never reaching UDPIPE 2 performance, the FT-GOLD-SAMPLE model still performs better than UDPIPE 2 on ESL data, although by a smaller margin than its fully-tuned counterpart, FT-GOLD. Furthermore, the difference between FT-SILVER-SAMPLE and FT-SILVER is almost negligible, suggesting that 1613 sentences should be sufficient to observe an improvement at least over the MACHAMP baseline.

We therefore speculate that the differences observed between the full-scale English and Italian experiments may also depend on the fact that VALICO-UD, includes even beginner-level written productions, making the gap between ISDT and VALICO-UD generally wider than that between EWT and ESL. The more significant performance gap between standard and learner data observed when testing the UDPIPE model on Italian data seems to confirm this second hypothesis.

4. Concluding remarks

In this paper, we tried to establish whether fine-tuning a dependency parser on THs results in better performance on learner language. This was based on the hypothesis that the performance drop usually observed when applying an off-the-shelf parser on L2 data might not be exclusively due to the presence of grammatical errors, but also to the fact that standard tools are generally not trained on learner essays, which are therefore out-of-domain even when grammatically correct.

The results of our experiments on ESL data strongly suggest that gold-annotated THs are indeed helpful, although the generalizability of this finding can only be confirmed by repeating them on a different, fully manually annotated dataset, which is however not available at the time of writing.

Based on the results of this first experiment, in any case, we recommend initiating the annotation of a new L1-L2 corpus by validating the THs (or, if time allows it, by manually annotating them from scratch). While still requiring skilled UD annotators, this is a relatively straightforward task compared to annotating actual learner language, as the latter requires the development of new guidelines to deal with grammatical errors consistently. The resulting gold-annotated THs can then be used to fine-tune a parser that should help bootstrap the more challenging process of analyzing L2 originals. In the best of cases, this would leave the annotators with a treebank where only the ungrammatical segments require manual editing.[8] In the near future, we plan on testings this strategy on the Swedish data at our disposal.

Whether silver THs are useful is unclear. While the English experiments seems to indicate that automatically annotated corrections can benefit a MACHAMP model and therefore help in the absence of a good pretrained parser, the results on VALICO-UD seem to contradict this finding in a way that cannot be explained solely by differences in dataset size. In this sense, further experiments with other L1-L2 treebanks are necessary, but not immediately possible. The aforementioned CFL (Chinese as a Foreign Language), the only other manually annotated treebank of this kind, consists of a mere 451 sentences, which makes it to small to generate training, development and test splits. At the same time, none of the larger learner corpora with target hypotheses comes with any extent of manual UD-annotation, which is however crucial for experiments like the ones described in this paper at least for the evaluation step. This further motivates us to proceed with the creation of a high-quality Swedish L1-L2 treebank.

An interesting byproduct of our parser evaluation is the observation that the ability to generalize to out-of-domain data appears to be much better for UDPIPE 2 models than for MACHAMP-based parsers, even if no overfitting is observed when evaluating the latter on an in-domain test set. This deserves further investigation, possibly in the context of a more systematic comparison of the cross-domain generalization capabilities of several mainstream UD parsers. When training a highly domain-specific tool, however, MACHAMP, is a powerful, easy-to-configure alternative, as exemplified by the excellent performance obtained with the FT-GOLD EWT + ESL model, whose training did not even require a hyperparameter search. Building development sets that combine standard and non-standard language should also make it possible to train more robust MACHAMP models.

5. Acknowledgements

This work is preparatory to the development of a treebank and parser for L2 Swedish, both of which are intended to enrich the Swedish national research infrastructure. As such, the research presented in this paper is supported by the Swedish national research infrastructure Nationella Språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. Special thanks go to several colleagues at Språkbanken Text and to the anonymous reviewers for their constructive and attentive feedback at different stages of the writing process.

[8] As long as a good GEC pipeline is in place to generate the THs, this strategy should also be applicable to L2-only treebanks.

6. Bibliographical 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.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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).

Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1):28–54.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*, chapter 6. Springer International Publishing, Cham.

John Lee, Keying Li, and Herman Leung. 2017. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy. Association for Computational Linguistics.

André Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective Turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria. Association for Computational Linguistics.

Arianna Masciolini. 2023. A query engine for L1-L2 parallel dependency treebanks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 574–587, Tórshavn, Faroe Islands. University of Tartu Library.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. Error-repair dependency parsing for ungrammatical texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–195, Vancouver, Canada. Association for Computational Linguistics.

Hakyung Sung and Gyu-Ho Shin. 2023. Towards L2-friendly pipelines for learner corpora: A case of written production by L2-Korean learners. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 72–82, Toronto, Canada. Association for Computational Linguistics.

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.

Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Susanna Lauriala, and Daniela Helena Piipponen. 2022. Reliability of automatic linguistic annotation: native vs non-native texts. In *Selected papers from the CLARIN Annual Conference 2021*. Linköping University Electronic Press (LiU E-Press).

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

7. Language Resource References

Berzak, Yevgeni and Kenney, Jessica and Spadine, Carolyn and Wang, Jing Xian and Lam, Lucia and Mori, Keiko Sophie and Garza, Sebastian and Katz, Boris. 2022. *English-ESL/TLE-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.11. PID http://hdl.handle.net/11234/1-5150.

Elisa Corino, Carla Marello, and Simona Colombo. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*, volume 6. Guerra. The corpus can be queried at `valico.org/index.html`.

Di Nuovo, Elisa and Sanguinetti, Manuela and Bosco, Cristina and Mazzei, Alessandro. 2023. *Italian-VALICO-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID http://hdl.handle.net/11234/1-5287.

Lee, John and Leung, Herman and Li, Keying. 2023. *Chinese-CFL-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID http://hdl.handle.net/11234/1-5287.

MDZ Digital Library team at the Bavarian State Library. 2021. *Italian BERT*. Bavarian State Library. Distributed via HuggingFace.

Silveira, Natalia and Dozat, Timothy and de Marneffe, Marie-Catherine and Bowman, Samuel and Connor, Miriam and Bauer, John and Manning, Chris. 2023. *English-EWT-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID http://hdl.handle.net/11234/1-5287.

Simi, Maria and Bosco, Cristina and Montemagni, Simonetta. 2023. *Italian-ISDT-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID http://hdl.handle.net/11234/1-5287.

Straka, Milan. 2023. *Universal Dependencies 2.12 models for UDPipe 2*. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.12. PID http://hdl.handle.net/11234/1-5200.

Volodina, Elena and Granstedt, Lena and Matsson, Arild and Megyesi, Beáta and Pilán , Ildikó and Prentice, Julia and Rosén, Dan and Rudebeck, Lisa and Schenström, Carl-Johan and Sundberg, Gunlög and Wirén, Mats. 2022. *SweLL-gold*. Språkbanken Text. Distributed via SBX/CLARIN. PID https://hdl.handle.net/10794/846.

Volodina, Elena and Pilán, Ildikó and Eide, Stian Rødven and Heidarsson, Hannes. 2017. *COCTAILL*. Språkbanken Text. Distributed via SBX/CLARIN. PID https://hdl.handle.net/10794/130.

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*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics. The dataset can be downloaded at `ilexir.co.uk/datasets/index.html`.