Are UD Treebanks Getting More Consistent?
A Report Card for English UD

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

Recent efforts to consolidate guidelines and treebanks in the Universal Dependencies project raise the expectation that joint training and dataset comparison is increasingly possible for high-resource languages such as English, which have multiple corpora. Focusing on the two largest UD English treebanks, we examine progress in data consolidation and answer several questions: Are UD English treebanks becoming more internally consistent? Are they becoming more like each other and to what extent? Is joint training a good idea, and if so, since which UD version? Our results indicate that while consolidation has made progress, joint models may still suffer from inconsistencies, which hamper their ability to leverage a larger pool of training data.

1 Introduction

The Universal Dependencies project[1] (de Marneffe et al., 2021) has grown over the past few years to encompass not only over 100 languages, but also over 200 treebanks, meaning several languages now have multiple treebanks with rich morphosyntactic and other annotations. Multiple treebanks are especially common for high resource languages such as English, which currently has data in 9 different repositories, totaling over 762,000 tokens (as of UD v2.11). While this abundance of resources is of course positive, it opens questions about consistency across multiple UD treebanks of the same language, with both theoretical questions about annotation guidelines, and practical ones about the value of joint training on multiple datasets for parsing and other NLP applications.

In this paper we focus on the two largest UD treebanks of English: the English Web Treebank (EWT, Silveira et al. 2014) and the Georgetown University Multilayer corpus (GUM, Zeldes 2017).[2] Although both datasets are meant to follow UD guidelines, their origins are very different: EWT was converted to UD from an older constituent treebank (Bies et al., 2012) into Stanford Dependencies (de Marneffe et al., 2006) and then into UD, while GUM was natively annotated in Stanford Dependencies until 2018, then converted to UD (Peng and Zeldes, 2018), with more material added subsequently via native UD annotation. Coupled with gradual changes and clarifications to the guidelines, there are reasons to expect systematic dataset differences, which UD maintainers (including the authors) have sought to consolidate from UD version to version.

Despite potential pitfalls, NLP tools are increasingly merging UD datasets for joint training: for example, Stanford's popular Stanza toolkit (Qi et al., 2020) defaults to using a model called `combined` for English tagging and parsing, which is trained on EWT and GUM (including the Reddit subset of GUM).[3] We therefore consider it timely to ask whether even the largest, most actively developed UD treebanks for English are actually compatible; if not, to what extent, and are they inching closer together or drifting apart from version to version? Regardless of the answer to these questions, is it a good idea to train jointly on EWT and GUM, and if so, given constant revisions to the data, since what UD version?

2 Related work

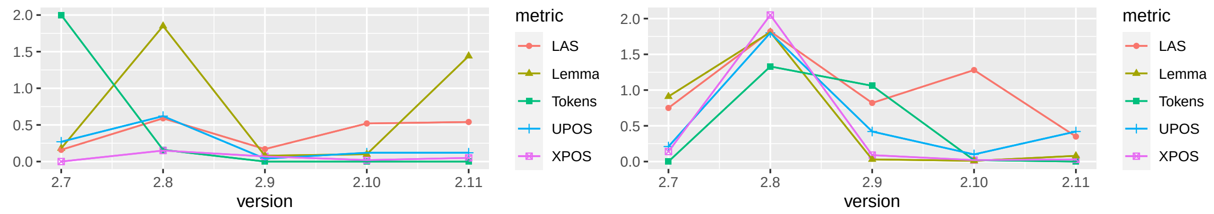
Much previous work on consistency in UD has focused on cross-linguistic comparison, and especially on finding likely errors. Some papers have taken a 'breadth-first' automatic approach to identifying any inconsistencies (de Marneffe et al., 2017), with the caveat that many types of differences are hard to detect. Others have taken a more focused approach to particular phenomena,

[1] https://universaldependencies.org

[2] Due to licensing, GUM Reddit data (Behzad and Zeldes, 2020) has a separate repo, but we merge both repos below.

[3] Though we focus on English here, the same is true for other UD languages with multiple datasets.

Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 58 - 64
March 9-12, 2023



(a) EWT

(b) GUM

Figure 1: Version-to-version changes across annotation layers in EWT and GUM. Y-values are percentages.

for example Bouma et al. (2018) showed that the `expl` relation was used differently across languages for comparable cases, using UD v2.1. Sanguinetti et al. (2022) show a broad range of practices in annotating user-generated content from the Web across UD languages in v2.6. Dönicke et al. (2020) also showed inconsistencies within UD languages using UD v2.5, including the finding that two of the top 20 most inconsistently headed relations in UD came from English, where across 7 datasets, `compound` and `csubj` behaved differently (of these, only the latter differed substantially in EWT and GUM, though the authors write it is possible that GUM 'simply contains more sentences with expletives'). Aggarwal and Zeman (2020) examined part of speech (POS) tag consistency in UD v2.5 and found that POS was relatively internally consistent within most languages.

Fewer studies have examined cross-corpus parsing accuracy (Alonso and Zeman 2016 for Spanish on UD v1.3, Droganova et al. 2018 for Russian using UD v2.2), and fewer still have looked at parsing consistency and stability (Kalpakchi and Boye, 2021). However to the best of our knowledge, no previous study has examined changes in consistency across UD versions, i.e. whether cross-treebank compatibility is increasing over time, how much so, and for which annotations?

3 How has the data changed?

To see how data in both corpora has changed across versions, we use the official CoNLL 2018 UD Shared Task (Zeman et al., 2018) scorer and compare each of the past six versions of each corpus to its next version, taking the updated version as an improved 'gold' standard.[4] This results in a score for each UD metric, such as the labeled attachment score (LAS), universal POS (UPOS) and English-specific POS (XPOS), as well as lemmatization and tokenization. Figure 1 shows the difference in score between each pair of versions for each dataset, which we discuss for each corpus below. For example, taking v2.7 of EWT as the correction for v2.6, we see a 2% rate of tokenization errors (in green), indicating a substantial change in tokenization, but less than 0.2% change to v2.8, and zero changes to tokenization moving to v2.9.

One caveat to note when working with data across versions is that unlike EWT, GUM's contents are not frozen: the corpus grows with new material every year. In the overview below, we therefore keep the evaluation fixed and limited only to documents that have existed since v2.6 (136 documents, 120K tokens). In §4 we will consider scenarios using both this fixed subset and the entire corpus (193 documents, 180K tokens in v2.11).[5]

3.1 EWT

Below we explain the main causes of the larger differentials between consecutive versions.

Tokenization Multiword tokens (MWTs) were added for most clitics (e.g. *'ll*) and contractions (*don't*) in v2.7, with some stragglers in v2.8. Essentially no changes to tokenization were made in subsequent versions.

Tagging Moderate UPOS changes occurred in 2.7 (many WH-words changed to `SCONJ`) and 2.8 (`ADJ` and `VERB` for adjectives and verbs in proper names, formerly `PROPN`, paralleling the XPOS `NNP`); this change was followed by GUM as well, see below. XPOS changes were small, peaking in 2.8 for select expressions like *of course*, *at least*, and *United States*.

Lemmatization Lemma errors were corrected throughout, but principal sources of lemma changes in v2.8 included capitalization of the content word lemmas in proper names, the lemma for the pronoun *I*, and removal of comparative or superlative

[4] Earlier comparisons are impossible since they predate GUM's conversion to UD.

[5] The subsequent release of the larger GUM v9, with 203K tokens and 213 documents, was around the same time as the camera-ready deadline for this paper, and could not be evaluated in time.

degree in the lemmas of *better* and *best*. In v2.11, a new policy for possessive pronoun lemmas was enacted to remove a key discrepancy with GUM.

Dependencies As shown in Figure 1a, the largest changes to LAS occurred in versions 2.8, when newly tagged `ADJ` tokens in names triggered `amod`; 2.10, where the analysis of the *X, so Y* construction was changed to `parataxis` (among others); and 2.11, which featured changes to nesting subjects (`nsubj:outer`), relative constructions, and clefts.

3.2 GUM

Similarly to EWT, GUM has become more stable across layers, with little change to XPOS or lemmas since v2.9. However earlier versions show several substantial revisions. Many changes are again simply due to error corrections, but some systematic changes include the following.

Tokenization saw major changes in v2.8, with the introduction of MWTs to match EWT changes. Additional major changes in v2.9 resulted from changing word tokenization to match EWT and other recent LDC corpora, which tokenize hyphenated compounds (e.g. v2.7 has *data-driven* as one token, but v2.8 has three tokens, like EWT).

Tagging shows a similar shift in v2.8 due to introduction of the `HYPH` tag for hyphens in compounds like 'data-driven', but also the removal of special XPOS tags for square brackets (`-LSB-` and `-RSB-` for left/right square brackets were collapsed with the round bracket tags `-LRB-`/`-RRB-`, again matching EWT). Changes to UPOS, by contrast, are more substantial, primarily due to verbs/adjectives in proper names, as in EWT above. Later changes to UPOS in v2.9 and 2.11 result from retagging some pronominal determiners (XPOS `DT`) as `DET` and not `PRON` (*some*, *all*, *both*), and changing WH subordinators from `SCONJ` to `ADV` respectively, again in harmony with changes to EWT.

Lemmatization largely reflects the hyphenation change (since e.g. 'data-driven' is no longer a lemma in v2.8) and the change from `PROPN` to `VERB` or `ADJ` in names, since the lemma for '*Glowing*' in '*the Glowing Sea*' was changed from '*Glowing*' (based on being `PROPN`) to '*Glow*' (as a `VERB`).

Dependencies Here too, transition to new tokenization and tagging names created changes in v2.8, but we also see a peak in v2.10, primarily due to consolidation of proper name dependencies (changing `flat` to syntactically transparent analyses), more aggressive identification of ellipsis (with promoted arguments) and `orphan` relations, and removal of some uses of the `dep` relation.

4 Cross-corpus parsing

Cross-corpus results To test whether EWT and GUM are becoming more internally and mutually consistent, we train parsers on each version of each corpus, and test them against both corpora. If each corpus is becoming more consistent, we expect higher scores in each version; and if cross-corpus model scores are increasing, we infer that the data is becoming more consistent across corpora. To ensure a fair comparison, we keep training and test data from GUM fixed to those documents that have been available since v2.6.

The results in Table 1 show that within-corpus scores are indeed improving slightly with each version (all scores are 3-run averages using Diaparser, Attardi et al. 2021, a recent transformer-based biaffine dependency parser). Cross-corpus scores are substantially lower, but also improving: in v2.6, EWT in-domain LAS was 90.24, which has improved slightly to 90.9 in v2.11, but scores training on GUM and testing on EWT have gone up from 81.78 to 84.27. In the opposite direction, GUM in-domain scores improved from LAS=87.9 to 89.48, or for a parser trained on EWT, from 83.89 to 84.74. The macro-average of both corpora also shows a steady increase, more so on GUM. In all cases, v2.11 is the best version yet for all metrics.

However, since the experiments are limited to the smaller subset of UD GUM v2.6 documents, they do not reflect current NLP tools (which train on all documents in the current UD repos), nor do they tell us whether joint training is a good idea.

Joint training results In this series of experiments we train on both corpora jointly, comparing two scenarios: the SUBSET scenario limits GUM training data to the v2.6 subset, while ALL uses all available GUM documents for training at each version; for fairness, scores are always limited to documents in the v2.6 test set, which are a subset of all subsequent release test sets.[6]

Table 2 shows that here too, there is only improvement over time. Using all GUM documents is superior to just the subset on GUM, but actually leads to a slight degradation on EWT, presumably due to the inclusion of more out-of-domain data

[6]Note that no new documents were added to GUM in v2.7, hence scores are identical for SUBSET and ALL until v2.8.

train	version	EWT test UAS	(sd)	EWT test LAS	(sd)	GUM test UAS	(sd)	GUM test LAS	(sd)	Macro-Avg UAS	(sd)	Macro-Avg LAS	(sd)
EWT	v2.6	92.82	0.132	90.24	0.066	87.81	0.073	83.89	0.023	90.31	0.059	87.07	0.025
	v2.7	92.84	0.037	90.25	0.173	87.87	0.088	84.19	0.074	90.35	0.062	87.22	0.114
	v2.8	92.93	0.060	90.42	0.090	87.97	0.078	84.90	0.028	90.45	0.065	87.66	0.042
	v2.9	92.88	0.107	90.41	0.131	87.57	0.148	84.36	0.105	90.23	0.098	87.38	0.117
	v2.10	93.06	0.082	90.70	0.158	87.81	0.084	84.72	0.138	90.44	0.082	87.71	0.088
	v2.11	**93.18**	0.142	**90.90**	0.139	**88.05**	0.260	**84.74**	0.289	**90.62**	0.196	**87.82**	0.207
GUM	v2.6	86.53	0.357	81.78	0.397	91.37	0.201	87.90	0.141	88.95	0.187	84.84	0.209
	v2.7	86.69	0.336	82.28	0.322	91.66	0.156	88.24	0.284	89.18	0.242	85.26	0.299
	v2.8	87.02	0.133	82.90	0.214	91.88	0.132	88.86	0.159	89.45	0.002	85.88	0.041
	v2.9	87.42	0.143	83.43	0.025	91.88	0.300	88.78	0.281	89.65	0.219	86.11	0.140
	v2.10	87.53	0.190	83.79	0.191	92.16	0.216	89.24	0.191	89.85	0.203	86.51	0.191
	v2.11	**88.23**	0.198	**84.27**	0.095	**92.28**	0.137	**89.48**	0.224	**90.26**	0.121	**86.88**	0.132

Table 1: Cross-corpus parsing scores (three run averages with standard deviations)

train	version	EWT test UAS	(sd)	EWT test LAS	(sd)	GUM test UAS	(sd)	GUM test LAS	(sd)	Macro-Avg UAS	(sd)	Macro-Avg LAS	(sd)
JOINT$_{subset}$	v2.6	92.38	0.044	89.59	0.108	90.08	0.366	86.80	0.326	91.23	0.177	88.20	0.146
	v2.7	92.31	0.078	89.61	0.072	90.15	0.311	86.96	0.360	91.23	0.122	88.29	0.148
	v2.8	92.49	0.159	89.99	0.128	90.51	0.351	87.86	0.449	91.50	0.154	88.92	0.195
	v2.9	92.39	0.324	89.80	0.278	90.63	0.392	87.91	0.415	91.51	0.086	88.85	0.114
	v2.10	92.62	0.034	90.24	0.058	90.51	0.418	87.86	0.381	91.56	0.192	89.05	0.163
	v2.11	**92.92**	0.072	**90.58**	0.052	**90.75**	0.073	**87.94**	0.059	**91.83**	0.064	**89.26**	0.045
JOINT$_{all}$	v2.6	92.38	0.044	89.59	0.108	90.08	0.366	86.80	0.326	91.23	0.177	88.20	0.146
	v2.7	92.31	0.078	89.61	0.072	90.15	0.311	86.96	0.360	91.23	0.122	88.29	0.148
	v2.8	92.07	0.277	89.55	0.312	91.26	0.267	88.72	0.247	91.66	0.077	89.14	0.066
	v2.9	92.27	0.154	89.77	0.287	90.81	0.084	88.12	0.123	91.54	0.110	88.95	0.176
	v2.10	92.18	0.018	89.86	0.010	91.54	0.170	88.99	0.211	91.86	0.092	89.43	0.110
	v2.11	**92.54**	0.259	**90.11**	0.240	**91.71**	0.426	**89.11**	0.534	**92.13**	0.147	**89.61**	0.181

Table 2: Joint training parsing scores (three run averages with standard deviations)

from the EWT perspective. Nevertheless, JOINT$_{all}$ performance on EWT also improves over time.

The best in-domain numbers from Table 1 are always better than the best joint training numbers, indicating that the added data cannot quite compensate for the distraction of different genres in each corpus, and possible remaining annotation inconsistencies. This is not surprising given the importance of genre for NLP performance (Zeldes and Simonson, 2016; Müller-Eberstein et al., 2021). In fact, similar tradeoffs of a helpful increase in data size vs. a harmful increase in heterogeneity have been observed for UD parsing in other languages (see Zeldes et al. 2022 for Hebrew, León 2020 for Spanish) and similarly for other tasks (e.g. for discourse parsing, Peng et al. 2022; Liu and Zeldes 2023).

However, the gap is narrowing: the joint model has gained about a point on EWT, placing it only 0.32 points behind the best in-domain model, and it has gained 2.31 points on GUM for the best ALL scenario in v2.11. Perhaps more importantly, the macroaverage, which may better reflect 'real-world' applicability of the parser model to any unseen genre data (since the macro-test set contains the most target genres), is now at LAS=89.61, within one point of the best models for each corpus.

Since the best joint result is also for v2.11, it seems fair to answer the questions posed at the beginning of this paper as follows: it has never been a better idea to train jointly than now; joint training always lags closely behind in-domain training, but the gap has been narrowing and is now very small; and for totally unseen new data, the joint model now looks like a very good idea. The joint SUBSET model is a close second on EWT, and the joint ALL model is the runner-up on GUM. That said, the fact that more data in the form of a second corpus does not outperform in-domain training alone suggests that there are still inconsistencies between the corpora, on which the next UD versions can hopefully improve.

In terms of concerns about what current jointly trained parsers are actually getting wrong, we direct readers to the confusion matrices in Figure 2 in the Appendix, which indicates that despite training on distinct datasets, the most common errors on both test sets are invariably confusing `nmod` and `obl`, which usually corresponds to a PP attachment error. Other systematic errors are rare, and largely concern notable subcategories of names and other types of terms. One recurring subtype is GUM's `dep` label being confused with EWT `nummod` for numeric modifiers which are not count-modifiers (e.g. 'Page **3**' has '3' as `dep` in GUM but `nummod` in EWT; GUM only uses `nummod` for counting cases like '3 pages'). Several of these discrepancies are discussed in Schneider and Zeldes (2021) and form a target for further consolidation.

Also of possible concern are `compound` relations, which a GUM-trained model predicts for various gold-standard relations in EWT, and an EWT model predicts for various gold-standard relations in GUM. It seems likely that these are remaining artifacts from the automatic conversion of the EWT gold constituent annotations to dependencies, in which various complex nominals were analyzed as compounds, for example for names such as *Sri Lanka* or *Hong Kong* (right-headed `compound` in EWT, but left-headed `flat` in GUM) and borrowed foreign words or phrases such as *cordon-blu* (*sic*) (again right-headed in EWT, would be flat in GUM), or also in complex nested phrases which are analyzed as left branching in EWT, e.g. *Marvel Consultants, Inc.* is headed by *Inc.* with two `compound` dependents in EWT. In GUM it would be headed by *Consultants* with *Inc.* as `acl`, or `flat` for lexicalized cases (attested in GUM for the film *Monsters Inc.*). Similarly, capitalized adjectival modifiers with XPOS `NNP` are sometimes labeled as `compound` in EWT, leading to `amod` predictions in the GUM-trained model and vice versa (e.g. *Islamist officers* or *Baathist saboteurs*).

5 Discussion

In this paper we surveyed progress in consolidating the largest UD English corpora, EWT and GUM. Results show data is moving closer together: single-corpus training still beats joint training by a hair, but joint models are nearly as good, and likely much more robust. As consolidation continues, we hope to see joint models overtake in-domain training, and more consistency expanding to other English datasets and other UD languages.

References

Akshay Aggarwal and Daniel Zeman. 2020. Estimating POS annotation consistency of different treebanks in a language. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 93–110, Düsseldorf, Germany. Association for Computational Linguistics.

Héctor Martínez Alonso and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, 57:91–98.

Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for Enhanced Universal dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188, Online. Association for Computational Linguistics.

Shabnam Behzad and Amir Zeldes. 2020. A cross-genre ensemble approach to robust Reddit part of speech tagging. In *Proceedings of the 12th Web as Corpus Workshop*, pages 50–56, Marseille, France. European Language Resources Association.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA.

Gosse Bouma, Jan Hajic, Dag Haug, Joakim Nivre, Per Erik Solberg, and Lilja Øvrelid. 2018. Expletives in Universal Dependency treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 18–26, Brussels, Belgium. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the Universal Dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115, Pisa,Italy. Linköping University Electronic Press.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*, pages 449–454, Genoa, Italy.

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

Tillmann Dönicke, Xiang Yu, and Jonas Kuhn. 2020. Identifying and handling cross-treebank inconsistencies in UD: A pilot study. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 67–75, Barcelona, Spain (Online). Association for Computational Linguistics.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian UD treebanks. In *Proceedings of the 17th International Workshopon Treebanks and Linguistic Theories (TLT 2018)*, pages 53–66,, Linköping, Sweden. Linköping University Electronic Press.

Dmytro Kalpakchi and Johan Boye. 2021. Minor changes make a difference: a case study on the consistency of UD-based dependency parsers. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 96–108, Sofia, Bulgaria. Association for Computational Linguistics.

Fernando Sánchez León. 2020. Combining different parsers and datasets for CAPITEL UD parsing. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020*, volume 2664 of *CEUR Workshop Proceedings*, pages 39–44. CEUR-WS.org.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? A thorough investigation of the impact of data diversity. In *Proceedings of EACL 2023*, Dubrovnik, Croatia.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Siyao Peng and Amir Zeldes. 2018. All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 167–177, Santa Fe, NM.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Manuela Sanguinetti, Lauren Cassidy, Cristina Bosco, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2022. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*.

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172, Sofia, Bulgaria. Association for Computational Linguistics.

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)*.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

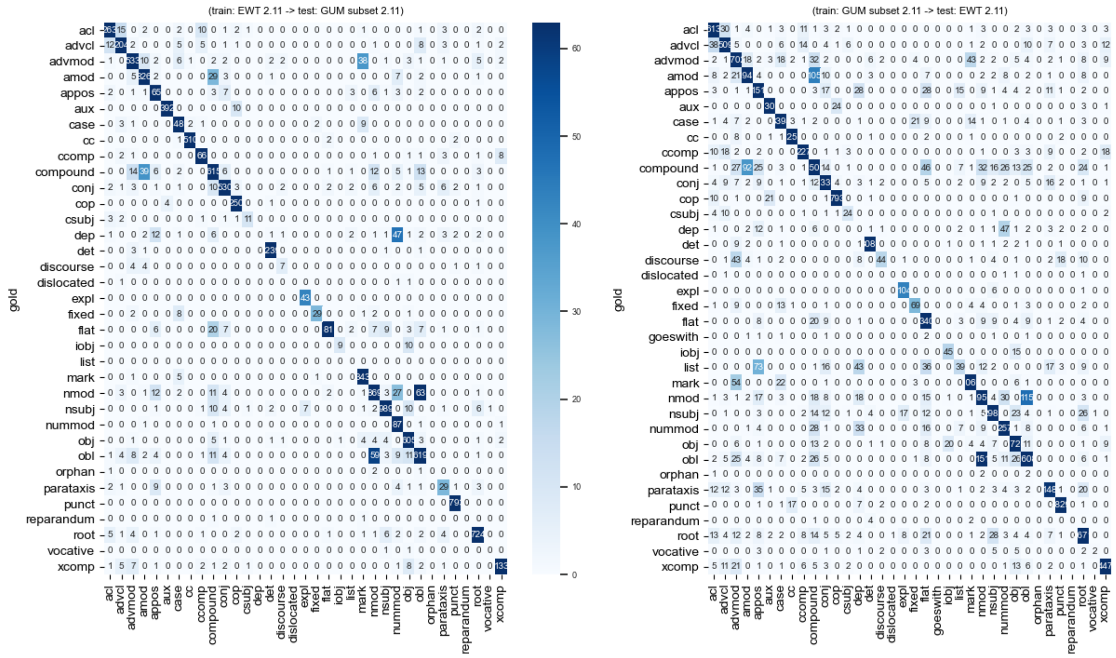
Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of EMNLP 2022*, Abu Dhabi, UAE.

Amir Zeldes and Dan Simonson. 2016. Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 68–78, Berlin, Germany. Association for Computational Linguistics.

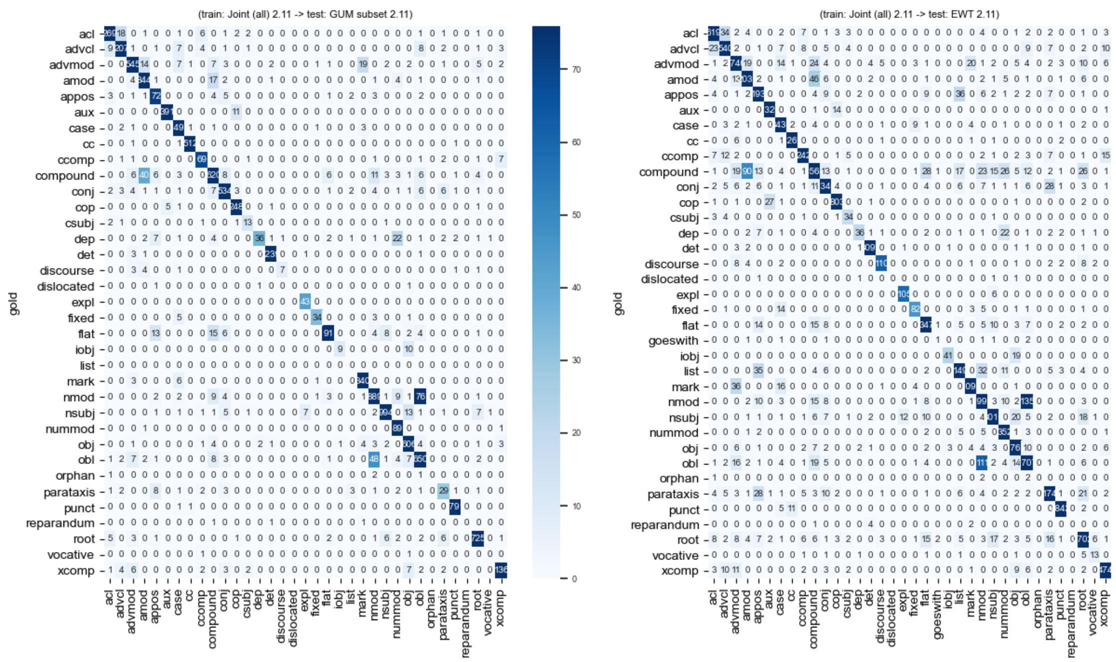
Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal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

A Confusion matrices

Figure 2 gives confusion matrices for dependency relation predictions (disregarding correct/incorrect attachment) for the joint and cross-corpus scenarios, testing on GUM (left) and EWT (right). In all cases, the most frequently confused errors are `obl` and `nmod` in both directions, largely corresponding to PP attachment ambiguity errors (i.e. high attachment to the verb for 'eat a pizza with a fork' versus low attachment to the object noun in 'eat a pizza with anchovies'). These errors are encouraging in that they are unlikely to reflect annotation practice differences between the corpora.



(a) Dependency relation errors for cross-corpus training



(b) Dependency relation errors for joint training

Figure 2: Confusion matrices for cross-corpus (a) and joint-corpus (b) dependency relation predictions on both test sets, using the GUM v2.6 document subset for GUM and the average performing parser model from each experiment.