A Dependency Treebank of Spoken Second Language English

Kristopher Kyle, Masaki Eguchi, Aaron Miller and Theodore Sither
Learner Corpus Research and Applied Data Science Lab
https://lcr-ads-lab.github.io/LCR-ADS-Home/
Linguistics Department, University of Oregon
https://linguistics.uoregon.edu/

Abstract

In this paper, we introduce a dependency treebank of spoken second language (L2) English that is annotated with part of speech (Penn POS) tags and syntactic dependencies (Universal Dependencies). We then evaluate the degree to which the use of this treebank as training data affects POS and UD annotation accuracy for L1 web texts, L2 written texts, and L2 spoken texts as compared to models trained on L1 texts only.

1 Introduction

In the field of applied linguistics, natural language processing tools such as part of speech (POS) taggers and syntactic parsers have been and continue to be used to investigate characteristics of second language (L2) use at scale (e.g., Biber et al., 2014; Kyle & Crossley, 2017; Lu, 2010; Paquot, 2018). Although taggers and parsers are increasingly accurate (achieving F1 scores of around .98 for POS taggers and .95 for dependency annotation) when evaluated on in-domain texts (i.e., texts with similar linguistic characteristics as the training data), accuracy can drop precipitously for out of domain texts (e.g., McClosky et al., 2006). A pressing issue, then, is the availability of appropriate annotated corpora to test and train tagging and parsing models on the types of data applied linguists often use (Kyle, 2021; Meurers & Dickinson, 2017). Although a treebank of written second language (L2) English is available (Berzak et al., 2016), to our knowledge no treebanks of spoken L2 speech are publicly and freely available. In this paper, we report on the development of an annotated corpus of spoken L2 English and evaluate the accuracy of a POS tagger and dependency parser when trained on L1 texts and a combination of L1 and L2 texts.

2 Applied linguistics research and NLP

The use of NLP tools such as taggers and parsers to examine characteristics of language use has a long history in the field of applied linguistics. Early studies (e.g., Biber, 1988) focused on the analysis of lexical and lexicogrammatical variation across registers (e.g., different spoken and written language use domains). As the subfield of learner corpus research has grown, taggers and parsers have also been used to investigate how second language learners' linguistic patterns change over time (e.g., Crossley & McNamara, 2014; Kyle et al., 2021) and/or differ across proficiency levels (e.g., Biber et al., 2014; Grant & Ginther, 2000; Kyle et al., 2018; Paquot, 2018).

2.1 Application of taggers and parsers in L2 research

POS taggers and syntactic parsers have been used in L2 research for a variety of purposes, ranging from relatively simple homograph disambiguation (e.g., Jarvis & Hashimoto, 2021) to the analysis of complex linguistic phenomena such as verb argument constructions (e.g., Kyle and Crossley, 2017). An abbreviated overview of this research is outlined below.

Grammatical error correction: A number of studies have used (and developed) tagging and parsing systems for identifying and correcting grammatical errors in L2 texts (e.g., Choshen & Abend, 2010; Nagata & Sakaguchi, 2016; Sakaguchi et al., 2017.)

Homograph disambiguation: One use of POS taggers in L2 research is homograph disambiguation. Homograph disambiguation can be particularly important in the measurement of lexical diversity, where the variety of words used by L2 learners can be an indicator of proficiency (e.g., Jarvis & Hashimoto, 2021; McCarthy & Jarvis, 2010).

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 39 - 45
July 15, 2022

Lexical bigrams: The characteristics of lexical combinations that are used in L2 productions can be an important predictor of development and/or proficiency level. Research has shown that more proficient L2 writers and speakers tend to use more frequent and more strongly associated lexical bigrams than less proficient L2 users (e.g., Granger & Bestgen, 2014; Garner et al., 2019; Kyle et al. 2018). For more precise insights into linguistic development, some studies have constrained the lexical combinations that are used (e.g., adjective + noun or noun + noun combinations).

Even more recently, researchers have begun to use dependency parses to analyze lexical items in particular grammatical relationships (e.g., verb + object; Kyle & Eguchi, 2021; Paquot, 2018, 2019; Rubin, 2021).

Lexicogrammatical features: A number of studies have investigated the relationship between L2 proficiency and the use of lexicogrammatical features that are common in academic writing such as various types of noun phrase elaboration (e.g., Biber et al., 2014; Grant & Ginther, 2000; Picoral et al., 2021). A related line of research has explored the relationship between characteristics of verb argument construction use and L2 writing proficiency (e.g., Kyle & Crossley, 2017; Kyle et al., 2021).

Syntactic complexity: A particularly common use of NLP tools in second language research is the calculation of classic syntactic complexity indices such as mean length of clause and dependent clauses per clause (e.g., Lu, 2010, 2011) or more fine-grained indices such as the number of dependents per nominal (e.g., Kyle & Crossley, 2018; Díez-Bedmar & Pérez-Paredes, 2020). NLP tools have allowed research to examine relationships between syntactic complexity and language proficiency and/or development at a scale that would be infeasible for most researchers if manual analyses were used.

2.2 Evaluations of system performance on L2 data

The literature indicates that L2 researchers are fully aware of potential issues with tagger and parser performance (e.g., Meurers & Dickinson, 2017). However, most accuracy analyses have been small in scale and have not resulted in publicly available treebanks that can be used to improve future models (c.f., Berzak et al., 2016). Lu (2010), for example, which introduced the second language syntactic complexity analyzer (L2SCA), evaluated the accuracy of the tool using a 30-essay subset of texts used in a validation study. Polio and Yoon (2018) independently evaluated the accuracy of L2SCA using a different sample of texts. Kyle et al. (2021) evaluated the accuracy of verb argument construction identification using a sample of 100 sentences from a corpus of L2 essays. Similar procedures have been used in a number of other studies (e.g., Díez-Bedmar & Pérez-Paredes, 2020; Paquot, 2019; Rubin, 2021). While small-scale accuracy analyses are important for establishing the effectiveness of particular linguistic analysis tools for a particular data set, these datasets are rarely made publicly available and do not necessarily follow the annotation guidelines or formatting conventions of well-known treebanks. One exception to this pattern is the Treebank of Learner English (Berzak et al., 2016) which includes written L2 English sentences annotated for Penn POS tags and Universal Dependencies. While this is an important resource, no treebanks of spoken L2 English are currently available.

2.3 Contributions of this study

In this study, we introduce a freely and publicly available treebank of Spoken L2 English that includes gold standard annotations for Penn POS tags and Universal Dependencies. We then evaluate the performance of tagger and parser models on both L1 and L2 data when the training set includes only L1 data and when the training set includes both L1 and L2 data.

3 Method

3.1 Dependency Treebank of Spoken L2 English (SL2E)

The Dependency Treebank of Spoken L2 English (SL2E) consists of a random sample of sentences from the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) corpus (Izumi et al., 2004). NICT JLE includes transcripts of oral proficiency interviews (OPI). Prior to sampling, all interviewer language was removed, leaving only utterances produced by second language speakers. The corpus includes a range of L2 English proficiency levels (mid-beginner to lower advanced). In total, the annotated portion of the corpus includes 7,412 sentences (70,016 tokens) annotated for Penn POS tags (Santorini et al., 1990), of which 2,320 sentences (21,312 tokens) are also annotated for Universal Dependencies (Nivre et al, 2020).

POS Annotation: The annotation was conducted in multiple stages. Undergraduate Linguistics majors who had taken upper-level courses related to linguistic structure were recruited to work on the project. POS annotation training sessions were conducted with annotators, followed by the annotation of sample sentences. Feedback was provided based on performance on the sample sentences. After training, sentences were annotated independently by at least two annotators using the browser-based application WebAnno (Eckart de Castilho et al., 2016). Any disagreements between annotators were checked by a third annotator. In rare cases where the third annotator disagreed with both of the first annotators, adjudication between annotators was conducted. During the annotation period, annotators had access to the original Penn POS tagging guidelines, the Berzak et al. (2016) tagging guidelines, and gold standard corpora (which were accessed using AntConc; Anthony, 2019). We also had weekly meetings to discuss difficult cases, a Discord server to report and discuss difficult cases asynchronously, and an extended tagging guidelines manual that was created based on these discussions. Initial annotation agreement for POS tags (prior to third ratings and adjudication) was 95.1%.

Dependency Annotation: After sentences were annotated for POS tags, annotators were trained for dependency annotation using procedure outlined above. Annotators had access to the Universal dependencies guidelines (Version 2; Nivre, 2020), gold standard corpora (accessed via Tündra; Martens, 2013), weekly meetings, a Discord server, and updated guidelines. After POS and dependency annotation was complete, POS tags and dependency annotations were checked again for consistency, resulting in some minor corrections. Initial annotation agreement (prior to third ratings and adjudication) was 86.5% (labeled attachment agreement).

3.2 Other corpora used

In this study, we decided to use data that was publicly and freely available. Accordingly, we used selected annotated corpora from the UD project, each of which are outlined briefly below.

Treebank of Learner English (TLE): TLE (Berzak et al., 2016) consists of data from the CLC FCE Dataset (Yannakoudakis et al., 2011), which includes writing samples from the Cambridge ESOL First Certificate in English (FCE) exam. The FCE includes written responses across 5 registers (letter, report, article, composition, and short story) that prototypically range from 200-400 words. The TLE sample includes sentences from upper-intermediate learners of English across 10 first language (L1) backgrounds. TLE includes 97,681 tokens annotated for POS tags and universal dependencies.

Data Splits Used			
Data	**Train**	**Dev**	**Test**
EWT	204,579	25,149	25,097
GUM	103,400	16,270	16,216
PUD	21,176	n/a	n/a
UDEP	1,705	n/a	n/a
TLE	78,541	9,549	9,591
SL2E POS	55,873	6,815	7,328
SL2E UD	16,879	2,167	2,266
Training Data Summary			
Data	**L1**	**L1+L2**	**L1+L2e**
POS	432,826	567,240	435,624
UD	432,826	528,246	434,951

Table 1: Number of tokens in each split

English Web Treebank (EWT): The Universal Dependency (UD) version (Silveira, et al., 2014) of the EWT (Bies, et al., 2012) consists of annotated data divided roughly evenly across five web genres (weblogs, newsgroups, emails, reviews, and Yahoo! answers). The UD version includes 254,825 tokens annotated for Penn POS tags and Universal dependencies.

Georgetown University Multilayer Treebank (GUM): GUM (Zeldes, 2017) consists of annotated data from various online sources, including interviews, news stories, and forum discussions (among many others). In this study, we use the versions of GUM included in the UD 2.9, which also includes sentences from Reddit (Behzad & Zeldes, 2020). In total, the version of GUM used in this study includes 135,886 tokens annotated for POS tags and universal dependencies.

Parallel Universal Dependencies Treebank (PUD): PUD (Zeman et al., 2017) includes sentences from the news section of Wikipedia and comprises 21,312 tokens annotated for POS tags and universal dependencies. In this project, PUD was used as training data only.

UD-English Pronouns (UDEP): UDEP (Monarch, 2021) includes sentences designed to mitigate biases (e.g., gender biases) that exist in extant treebanks by including sentences with pronouns that are rare in other treebanks (e.g., *hers*). UDEP includes 1,705 tokens. In this project, UDEP is used as training data only.

3.3 Splits used

In this study, we used 80/10/10 splits for the Spoken L2 Treebank, and the extant splits in all treebanks available in UD release 2.9. Because more data was annotated for POS tags than for dependency relations, we created training/dev/test sets separately for POS annotation and dependency annotation. For each, we tested three versions of training data. The first (L1) included only L1 data (EWT, GUM, PUD, and UDEP). The second (L1+L2) included the L1 data plus the L2 data (TLE + SL2E). Given the relatively small datasets (and the positive relationship between the amount of training data and model accuracy), we also included a third version of the training data (L1+L2e) in which the number of tokens in the L1+L2 training data was made roughly equal to that of the L1 training data by excluding a random sample of L1 sentences. See Table 1 for the splits used in POS annotation and dependency annotation.

3.4 NLP pipeline

For this study, we used Spacy version 3.2 (Honnibal et al., 2020) to train transformer-based POS and dependency annotation models (L1, L1+L2, and L1+L2e models for each task). Spacy is freely available, easy to use, and has achieved state-of-the-art performance for both POS and dependency annotation (Honnibal et al., 2020). The models used pre-trained weights from RoBERTa-base (Liu, 2019). The POS and dependency layers listen to the transformer embedding, and they were optimized using Adam optimizer. The same hyperparameter settings were used for training all models. Training scripts, models generated during training, and evaluation scripts are available at (https://github.com/LCR-ADS-Lab/l2-nlp-training-spacy). POS annotation accuracy was measured using sentences with gold standard splits and tokenization. Dependency annotation accuracy was measured using gold standard splits and tokenization, and model-based POS tags (using the best-performing POS model).

POS Models			
Data	**L1**	**L1+L2**	**L1+L2e**
EWT	0.958	0.965	0.964
GUM	0.973	0.975	0.977
SL2E	0.936	0.970	0.966
TLE	0.953	0.969	0.966
Dependency Models			
Data	**L1**	**L1+L2**	**L1+L2e**
EWT	0.884	0.895	0.895
GUM	0.884	0.897	0.895
SL2E	0.876	0.935	0.938
TLE	0.886	0.920	0.918

Table 2: F1 scores for lexical tags

4 Results

4.1 POS annotation results

Despite the relatively small amount of training data used, all three models resulted in relatively high tagging accuracy for the L1 corpora (EWT and GUM), ranging from F1 scores of 0.958 to 0.977 on the test set (see Table 2). Somewhat surprisingly, the highest F1 scores for the L1 corpora were achieved when L2 data was added during the training (even when the number of tokens in the training data was held constant in L1+L2e), and these gains were modest (see Table 2). The lowest tagging accuracy was observed when the L1-trained model was applied to the L2 spoken test set (F1 = 0.936). However, when L2 data was included in the training set, the F1 scores for the L2 spoken test set (F1 = 0.970) were similar to those for L1 corpora.

4.2 Dependency annotation results

Labeled attachment scores (LAS) for test set data ranged from F1 scores of 0.876 (Spoken L2 data, L1 model) to F1 scores of 0.938 (Spoken L2 data, L1+L2e model). Accuracy for all models increased with the inclusion of L2 data in the training set (even when the total amount of training data was held constant). However, the most dramatic increases were for both written and spoken L2 data.

5 Discussion and conclusion

5.1 Summary of findings

The results of this study suggest that substantial improvements in POS tagging and dependency parsing performance on L2 texts can be made through the use of training sets that include L2 data, even when the total amount of training data is held constant. Following previous research (e.g., Berzak et al., 2016), these improvements were observed for written L2 data. However, the improvements were particularly marked for the spoken L2 data introduced in this study. It should also be clearly noted that the highest dependency annotation accuracy was observed with L2 spoken data (followed by L2 written data).

5.2 Limitations and future directions

While this study demonstrated accuracy gains in L2 tagging and parsing through the use of L2 training data, there are still a few limitations that should be addressed in future studies. First, although this study added to the amount of annotated data available for training, the total amount of publicly available gold standard data annotated for universal dependencies remains rather small. Future research should focus on providing more gold standard data across a variety of English domains (including L2 domains). Second, in this study we did not fully account for strength and weaknesses of each model with regard to particular lexical items or annotations. While overall F1 scores are a helpful gauge, many L2 researchers are interested in particular grammatical features (e.g., main verb + direct object pairs), and more precise accuracy figures should be considered in future research.

5.3 Conclusion

This study introduced a new gold standard treebank of spoken L2 English annotated with Penn part of speech tags and universal dependencies. Furthermore, this study has demonstrated that the addition of a relatively small amount of in-domain data can substantively improve tagging and parsing accuracy in L2 texts. The SL2E Treebank is publicly available for non-commercial purposes (https://github.com/LCR-ADS-Lab/SL2E-Dependency-Treebank).

Acknowledgments

This project was supported by a *Language Learning* Early Career Research Grant.

References

Anthony, L. (2019). *AntConc* (3.5.8) [Computer software]. Waseda University.

Behzad, S., & Zeldes, A. (2020). A cross-genre ensemble approach to robust Reddit part of speech tagging. *ArXiv Preprint ArXiv:2004.14312*.

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016). Universal dependencies for learner English. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 737–746.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D., Gray, B., & Staples, S. (2014). Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Bies, A., Mott, J., Warner, C., & Kulick, S. (2012). English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Choshen, L. & Abend O. (2018). Reference-less Measure of Faithfulness for Grammatical Error Correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 124–129, New Orleans, Louisiana. Association for Computational Linguistics.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Díez-Bedmar, M. B., & Pérez-Paredes, P. (2020). Noun phrase complexity in young Spanish EFL learners' writing: Complementing syntactic complexity indices with corpus-driven analyses. *International Journal of Corpus Linguistics*, *25*(1), 4–35.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. https://www.aclweb.org/anthology/W16-4011

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. Advanced non-native writers: A bigram-based study. *International*

Review of Applied Linguistics in Language Teaching, *52*(3), 229–252.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123–145.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* [Python]. https://doi.org/10.5281/zenodo.1212303 (Original work published 2014)

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of The Computer, the Internet and Management*, *12*(2), 119–125.

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*(1), 163–194.

Kyle, K. (2021). Natural language processing for learner corpus research. *International Journal of Learner Corpus Research*, *7*(1), 1–16.

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

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Berger, C. M. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, *43*(4), 781–812.

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In *Perspectives on the L2 Phrasicon: The View from Learner Corpora* (pp. 126–151). Multilingual Matters.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62.

Martens, S. (2013). *TüNDRA: A web application for treebank search and visualization*. *133*.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McClosky, D., Charniak, E., & Johnson, M. (2006). *Reranking and self-training for parser adaptation*. 337–344.

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, *67*(S1), 66–95.

Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Nagata, R. & Sakaguchi, K. (2016). Phrase Structure Annotation and Parsing for Learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1837–1847, Berlin, Germany. Association for Computational Linguistics.

Nivre, J., Marneffe, M.-C. de, Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). *Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection*. 4034–4043. https://aclanthology.org/2020.lrec-1.497

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly*, *15*(1), 29–43.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Polio, C., & Yoon, H. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*, *28*(1), 165–188.

Rubin, R. (2021). Assessing the impact of automatic dependency annotation on the measurement of phraseological complexity in L2 Dutch. *International Journal of Learner Corpus Research*, *7*(1), 131–162.

Sakaguchi, K., Post, M., and Van Durme, B. (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.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing). *Ms., Department of Linguistics, UPenn. Philadelphia, PA*.

Silveira, N., Dozat, T., De Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., & Manning, C. (2014). A Gold Standard Dependency Corpus for English. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2897–2904.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). *A new dataset and method for automatically grading ESOL texts*. 180–189.

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

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., & Potthast, M. (2017). *CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*. 1–19.