Interpreting the UAS and the LAS of the parsing of Old English with Universal Dependencies

Javier Martín Arista
University of La Rioja
javier.martin@unirioja.es

Ana Elvira Ojanguren López
University of La Rioja
ana-elvira.ojanguren@unirioja.es

Sara Domínguez Barragán
University of La Rioja
sara.dominguez@aurea.unirioja.es

Abstract

This paper interprets, from a linguistic point of view, the Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS) metrics obtained in the Universal Dependencies parsing of Old English. The study assesses the performance of three distinct training methods based on the Natural Language Processing library spaCy: a baseline pipeline, a pretrained model, and a transformer-based model (MobileBERT). Using datasets ranging from 1,000 to 20,000 words, the best-performing model (pretrained model with 20,000 words) achieved 83.2% UAS and 74.2% LAS. The model performs better at identifying structural relations than at labeling specific dependency relations. There is a consistent 9 point gap between UAS and LAS accross the different structural levels, including the word, the phrase, the clause and the complex sentence. While the model shows high accuracy in morphologically marked local relations and morphological feature recognition (often over 90%), its accuracy is lower with long-distance dependencies and complex syntactic structures. Particularly problematic areas include non-projective dependencies, fixed expressions, copulative constructions, and double object constructions. The conclusion is reached that improving parsing accuracy will require larger training datasets and a fine-grained analysis of complex syntactic relations that is compatible with the strong performance reached in morphological feature recognition.

1 Old English within the framework of Universal Dependencies

Old English (650-1150 CE) is a West Germanic language characterised by a consistently Germanic lexicon with borrowings from Latin and Old Norse, remarkable semantic transparency in word-formation (Kastovsky, 1992), generalised inflection of nominal, pronominal and verbal categories (Campbell, 1987), and relatively free word order compared to the contemporary language (Fischer et al., 2000). The written records of Old English comprise a total of 3 million words, attested in around 3,000 texts. The main corpora of Old English are *The Dictionary of Old English Web Corpus* (3 million words; Healey et al., 2004) and *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (1.5 million words; Taylor et al., 2003), which provides POS tagging and constituent parsing of approximately one half of the existing texts.

Universal Dependencies (UD) is an annotation framework designed for Natural Language Processing tasks, as well as for language comparison, translation and language learning (Nivre et al., 2016; Nivre et al., 2020; Zeman, 2024). The UD model consists of a universal inventory of lexical categories, morphological features and dependency relations that are adequate for cross-linguistic analysis and that can account for language-specific phenomena (de Marneffe et al., 2014; de Marneffe et al., 2021). In UD, syntactic representation is dependency-based, in such a way that binary asymmetric relations hold between heads and their dependents (de Marneffe and Manning, 2016). The annotation scheme of UD can be broken down into three layers: universal part-of-speech tags (UPOS), morphological features (FEATS), and syntactic dependencies (DEPREL). The UPOS layer includes seventeen coarse-grained lexical categories, while the FEATS layer comprises morphological properties like gender, number, case, and tense. The DEPREL layer contains a set of universal dependency relations that can be enhanced in order to deal with language-specific constructions. Overall, priority is given to universal phenomena over language-specific ones and to content words as heads over function words.

Several recent works engage in the annotation of Old English within the framework of UD. Martín Arista (2022a, 2022b) lays the foundations of a UD parsing of this historical stage of the English language. Martín Arista (2024) incorporates the description of word-formation to the annotation of Old English, given the regularities and the points of contact with the syntax of this stage of the English language, found not only in derivations but also in nominalisations with inheritance of verbal properties (Ojanguren López 2024). Regarding the automatic annotation of dependencies, Villa and Giarda (2023) test the performance of a multilingual parser aimed to Old English data. Villa and Giarda find that the highest accuracy rates are achieved by a combination of data of Old English with German and Icelandic. These authors reach a maximal accuracy of 75% with Icelandic, German and Old English, including UAS of 68% and LAS of 59% with Icelandic and Old English. Villa and Giarda (2023) put these accuracy metrics down to linguistic characteristics of Old English such as word order and case syncretism. To be more precise, they identify areas of error that include postpositions and discontinuity in relative clauses. These authors also attribute a number of annotation errors to inaccurate POS tagging, which lead to mismarking of the dependency relations coordinating conjunction, adverbial modifier: negation, auxiliary, locative adverbial modifier and temporal adverbial modifier.

Against this background, this paper deals with the evaluation of UD parsing. More specifically, its aim is to provide a linguistic interpretation of the LAS (Labelled Attachment Score) and UAS (Unlabelled Attachment Score) metrics of a task involving the UD parsing of a 25,000 word Old English dataset. At this point, a terminological note is needed. LAS (Labelled Attachment Score) and UAS (Unlabelled Attachment Score) are evaluation metrics used in dependency parsing (Nivre et al., 2007; Kübler et al., 2009; Manning, 2011). The UAS measures the percentage of tokens that are assigned the correct syntactic head, that is to say, it focuses on the structural dependency (attachment), not on the type of relation. The LAS, for its part, measures the percentage of tokens that have both the correct syntactic head and the correct dependency label. The LAS, therefore, is more selective than the UAS and is always lower or equal to the UAS.

2 The automatic parsing of Old English UD

In Section 2, we describe the task that we carried out for automatically parsing Old English texts within the UD framework. We compare three distinct training methods, including (i) a baseline tok2vec model (henceforth *Baseline*), (ii) a pretrained tok2vec model (hereafter *Pretrained*), and (iii) a MobileBERT-based model (henceforth *MobileBERT*), each of which replaces or extends different components in our SpaCy pipelines. In addition, we discuss the motivation for testing multiple dataset sizes and provide details on how long each pipeline was trained.

We relied on the standard pipeline stages of spaCy, including Tokeniser, tok2vec/Transformer, Tagger, Morpho (for UPOS and FEATS), Lemmatiser, and Parser. The first training method (Baseline) was a basic pipeline with default configuration, using spaCy default tok2vec component initialised with random weights. The Tagger, Morpho, Lemmatiser, and Parser draw on these default embeddings. The second training method (Pretrained) also used the tok2vec component but initialised its weights through a pretraining phase on an unannotated Old English corpus of about three million words (*The Dictionary of Old English Corpus*). This step allowed the system to learn semantic distances and token co-occurrences before the main training phase. Then, the Tagger, Morpho, Lemmatiser, and Parser began training with these pre-initialised weights. For the third method (MobileBERT), we trained a new tokeniser and a language model from scratch using approximately three million words (about 17 MB of text). We replaced the tok2vec component with a custom-trained MobileBERT transformer. Then, the Tagger, Morpho, Lemmatiser, and Parser relied on the MobileBERT embeddings. The MobileBERT architecture (25.3 million parameters) was selected to match the limited size of available Old English training data. Additionally, Old English includes certain graphemes (e.g., <æ/Æ>, <ʒ/Ʒ>, <ð/Đ>, <þ/Þ>, and <ƿ/Ƿ>) that contemporary English models cannot handle, while ransfer learning from contemporary English BERT models was not viable due to lexical differences.

Each of the three pipelines was trained on four dataset sizes: approximately 1,000, 5,000, 10,000, and 20,000 words. To ensure a fair comparison, we used the same split for all three models. Given that Old English texts are limited in number, we wanted to assess how well the models could learn from small training sets. In practical terms, researchers are often constrained by the availability of only very small amounts of manually annotated historical Old English data. Our goal in this respect was to compare performance under different training sizes and to discuss the feasibility of small datasets.

The test set, comprising 4,887 words across 288 sentences, was randomly selected and remained constant throughout all tests to ensure comparable results. For the training data, four sets were selected: a small set of about 1,000 words (995 words/59 sentences), a medium set of approximately 5,000 words (4,992 words/283 sentences), a larger set of around 10,000 words (9,982 words/562 sentences), and the largest set of about 20,000 words (19,991 words/1,134 sentences). The evaluation set was put aside first and then the training sets were created, in order to ensure no overlap between training and test data. The source texts of the datasets, extracted from *ParCorOEv3. An open access annotated parallel corpus Old English-English* (Martín Arista et al., 2023), included *Ælfric's Catholic Homilies I*, *The Anglo-Saxon Chronicle A*, *Anglo-Saxon Laws*, *St. Mark's Gospel* and *King Alfred's Orosius*.

We ran training in epochs such that each complete pass over a given training set counted as one epoch. Table 1 provides an overview of our training and test splits. On average, we conducted up to 20 epochs, though training often converged earlier for smaller datasets. For the largest (20,000 word) training sets, we allowed the training to continue for up to 30 epochs, although minimal additional gains in accuracy were observed after around 15-20 epochs.

	Train	Test	Total
1,000 words			
Tokens	995	4,987	5,982
Sentences	59	288	347
5,000 words			
Tokens	4,992	4,887	9,879
Sentences	283	288	571
10,000 words			
Tokens	9,982	4,887	14,969
Sentences	562	288	850
20,000 words			
Tokens	19,991	4,887	24,978
Sentences	1,134	288	1,422

Table 1. The training and test datasets.

The performance of the models was assessed by means of metrics for all the stages of the pipeline, including TAG_ACC (XPOS, tagger stage), POS_ACC (UPOS, morphologiser stage), MORPH_ACC (FEATS, morphologiser stage), LEMMA_ACC (LEMMA, trainable lemmatiser stage), DEP_UAS (Unlabelled Attachment Score, based on HEAD assigment, parser stage), DEP_LAS (Labelled Attachment Score, based on HEAD and DEPREL assignment, parser stage), and SENTS_F (sentence tokenisation, parser stage). Model performance is measured in terms of accuracy, as standard UAS/LAS reporting typically uses this metric.

The results show outstanding differences across different annotation tasks and training methods. The part-of-speech tagging proves the most reliable of all metrics. The Pretrained model achieves accuracy rates of 93.2% for POS tagging. Morphological feature annotation returns similar results, although with slightly lower accuracy rates. The Pretrained model achieves 84.2% accuracy with the 20,000 word dataset. However, the results present considerable variation in the accuracy of the various morphological features. While some features such as polarity and reflexivity are above 95%, others, like degree marking, present accuracy rates below 75%. The results of lemmatisation, which reach approximately 80% accuracy with the 20,000 word dataset, suggest that the benefits of pretraining may be less significant for lemmatisation tasks than for other types of annotation. Figure 1 shows the training curves (accuracy vs. epochs and loss vs. epochs) of the Pretrained model with the 20,000 word dataset: DEP_UAS (Unlabelled Attachment Score), DEP_LAS (Labelled Attachment Score), and SENTS_F (sentence tokenisation).

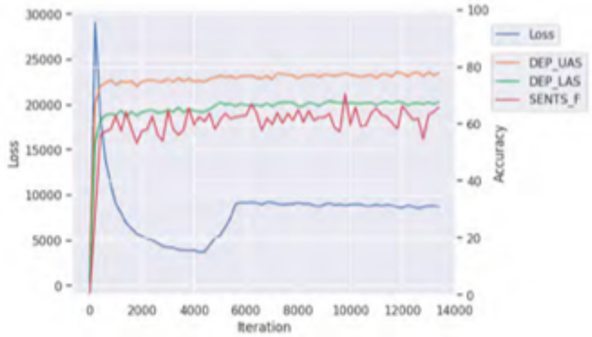

Figure 1. Loss function and accuracy metrics (UAS; LAS and SENT_F) of the Pretrained model with the 20,000 word dataset.

The general trend across all models is that the learning curve flattens after only a few epochs when the dataset is small (e.g., 1,000 words), whereas larger datasets (e.g., 20,000 words, shown in Figure 1) continue to provide improvements longer. The loss function is crucial because it guides the model toward more accurate parsing decisions. In the Pretrained model presented in Figure 1, thanks to the pretraining phase, the initial weights of the tok2vec component are closer to expected results than random weights. Consequently, the loss decreases more rapidly, in such a way that the model requires fewer training epochs to return higher UAS and LAS scores.

Dependency parsing is the most demanding taks in the annotation, although differences arise between UAS and LAS. The Pretrained model gets its best UAS of 83.2% with the 20,000 word dataset. The UAS is notably lower at 74.2%. This gap indicates the particular difficulty of correctly labeling dependency relations compared to simply identifying their existence by selecting their heads and dependents. Table 2 focuses on UAS and LAS metrics. The best results are highlighted in bold face.

Model	dataset	UAS	LAS
Baseline	1k words	57.44%	27.32%
Baseline	5k words	69.05%	53.17%
Baseline	10k words	73.76%	60.95%
Baseline	20k words	78.26%	68.10%
Pretrained	1k words	68.48%	34.19%
Pretrained	5k words	76.47%	62.68%
Pretrained	10k words	80.07%	69.10%
Pretrained	20k words	**83.24%**	**74.23%**
MobileBERT	1k words	42.78%	14.47%
MobileBERT	5k words	50.42%	29.84%
MobileBERT	10k words	55.58%	38.83%
MobileBERT	20k words	60.17%	45.51%

Table 2. UAS and LAS metrics by model and dataset.

As can be seen in Table 2, increased corpus size consistently improves performance across all metrics and methods. This means that the limit of potential accuracy has not been reached yet and that a larger dataset may turn out better metrics. The more accurate results of the pretrained model across nearly all metrics indicate that this method is particularly adequate for Old English annotation tasks. The continuous underperformance of the transformer model evidences that this architecture may require larger training corpora to be effective for the processing of historical languages like Old English.

3 UAS and LAS

As shown in Table 2, the performance of the model on Old English texts poses some issues in both structural attachment (UAS) and relation labeling (LAS). While problems of structural attachment are significant, the additional complexity of relation labeling results in lower LAS scores for most structural levels and syntactic configurations. Table 3 displays the accuracy metrics of DEPREL with the best performing method (the Pretrained model with the 20,000 word dataset).

DEPREL	Accuracy
Root (root)	55.1%
Nominal Subject (nsubj)	67.9%
Passive Nominal Subject (nsubj:pass)	53.8%
Object (obj)	41.4%
Indirect Object (iobj)	60%
Clausal Complement (ccomp)	25.5%
Adverbial Clause Modifier (advcl)	32.3%
Adjectival Modifier (amod)	68.8%
Adverbial Modifier (advmod)	71.4%
Nominal Modifier (nmod)	57.1%
Appositional Modifier (flat)	62.8%
Auxiliary (aux:pass)	76.1%

Copula (cop)	46.3%
Case Marking (case)	87.8%
Determiner (det)	69.4%
Possessive Determiner (det:poss)	90.4%
Coordinating Conjunction (cc)	73.8%
Conjunction (conj)	51.6%
Marker (mark)	76.3%
Oblique Nominal (obl)	54.1%
Relative Clause Modifier (acl:relcl)	77.7%
Orphan (orphan)	12.5%

Table 3. DEPREL metrics with the 20,000 word dataset (pretrained model).

As can be seen in Table 3, the dependency relation accuracy results show notable variations across different dependency relations. Function words and grammatical markers generally achieve higher accuracy: possessive determiners (90.4%), case markers (87.8%), relative clause markers (77.7%), and passive auxiliaries (76.1%) perform very well. Among core syntactic relations, subject identification (67.9% for active subjects) performs moderately well, while objects show lower accuracy (41.4% for direct objects, though indirect objects reach 60%). Modifier relations show mixed results: adverbial (71.4%) and adjectival modifiers (68.8%) also perform reasonably well, while clausal modifiers lag behind (adverbial clauses at 32.3% and clausal complements at 25.5%). Root identification achieves moderate accuracy (55.1%), which points to issues in the identification of the main predicate. Coordination structures show a wide gap between coordinating conjunction identification (73.8%) and the actual conjunct relation (51.6%). The notably low performance on orphaned elements (12.5%) could be attributed to their structural complexity in elliptical constructions. These results clearly indicate that the model performs better on morphosyntactically marked local relations than on semantic long-distance dependencies. This is confirmed by the metrics of accuracy of morphological features, which are tabulated in Table 4.

FEAT	**Accuracy**
Case	80.1%
Degree	63.6%
Gender	74.9%
Invariable	98.2%
Mood	91.4%
Number	90.5%
Person	94.9%
Polarity	100%
Possessive	96.9%
Pronoun Type	96.9%
Reflexive	100%
Tense	83.8%
Uninflected	95.9%
Verb form	92.0%

Table 4. FEAT metrics with the 20,000 word dataset (pretrained model).

The comparison between dependency relation accuracy (Table 3) and morphological feature accuracy (Table 4) shows a wide contrast in the performance of the model. Morphological features generally achieve much higher accuracy rates. As a matter of fact, many features exceed 90% accuracy: Polarity and Reflexive show full accuracy (100%), while Invariable (98.2%), Pronoun Type (96.9%), Possessive (96.9%), and Uninfected (96%) are all in the region of full accuracy. Verbal features like Number (90.6%), Mood (91.4%), Person (95%), and Verb Form (92.1%) also achieve high accuracy. Even lower-performing nominal features like Gender (75%) and Case (80.2%) still outperform most dependency relations. As can be seen in Table 2, only a few specific relations like Possessive Determiner (90.4%) and Case Marking (87.8%) reach similar levels of accuracy. This difference reinforces the idea that the model is more successful at learning local patterns than long-distance relations, which call for broader contextual information. The wide gap in performance (most morphological features perform over 80% while many dependency relations perform under 60%) highlights the difficulties of accurate syntactic parsing as compared to morphological analysis.

That said, the comparatively lower accuracy of dependency relations requires further explanation, which is couched in the remainder of this section in terms of UAS and LAS metrics by structural level. At the structural level of the word, the parsing of negative contractions raises issues of both attachment and labeling. The failure to properly decompose and attach elements in negative contractions like *na* 'never' (*ne* 'not' + *a* 'always') affects both UAS and LAS metrics. None of the 27 occurrences of negative contractions received a correct structural analysis. Consequently, their dependency relations were incorrectly labelled. Similar issues arise in the analysis of contracted verbs like *nǽfð* 'hasn´t' (*ne* 'not'+ *hǽfð* 'has'), with the misanalysis of the dependency relation. At the structural level of the phrase, the model shows weaknesses in both structural attachments and relation labeling. All three instances of multiword proper names received incorrect analyses. Cases like the proper noun *Marcus Curtius* show errors in both head-dependent relations (UAS) and relation labeling (LAS). Complex numerals present similar problems, as all instances show incorrect head attachments and the subsequent error of relation labeling. The parsing of titles and honorifics present better results, in such a way that 13 out of 25 instances receive the correct analysis of both structure and labeling. At the structural level of the clause, significant problems arise regarding UAS and LAS scoreshi. In double object constructions, none of the 31 instances got the correct structural analysis, which led to failures in both attachment and relation labeling. The parsing of oblique nominals shows that even when morphological cases should guide both attachment and labeling, the model frequently fails on both. In this respect, only 10 out of 36 instances have received the correct analysis. The treatment of indirect objects is particularly problematic, as only 7 out of 45 instances are analysed correctly. Turning to the structural level of the complex sentence, the performance results underline the idea that LAS scores consistently lag behind UAS scores. In the analysis of clausal complements, only 24 out of 50 instances received correct relation labels, even in cases where the basic structural attachment was correct. Open clausal complements show even poorer performance, given that only 3 out of 36 xcomp relations are labelled correctly. The parsing of adverbial clause modifiers present similar problems because only 12 out of 61 advcl relations were assigned correct labels, even if the basic clause attachment was accurate. In cross-level syntactic constructions, the treatment of non-projective dependencies (occurring when a word depends on another word that is not adjacent to it in the sentence structure, which results in a crossing relation; Decatur, 2022) is undoubtedly the weakest aspect in terms of both UAS and LAS metrics. None of the 34 instances of non-projective structures received correct analysis. The performance of the model is particularly poor in relative clause modifiers (13 instances), clausal modifiers of nouns (5 instances), and conjunctions (4 instances). Fixed expressions and copulative constructions also reflect the gap between UAS and LAS performance. Only 4 out of 16 fixed expressions received correct analysis in both structural attachment and labeling. Copulative sentences show a relatively better performance, with 14 out of 32 correctly analysed instances. These constructions often require understanding of both structural relations and specific syntactic functions, and the performance of the model experiences difficulties in both areas. Non-projective dependencies, fixed expressions and copulative constructions require a fine-grained treatment of both attachment and labeling and failure in these cases points out the limitations of the model to handle complex syntactic relations.

4 Conclusion

This study has highlighted the relevance of the detailed linguistic interpretation of parsing metrics over raw accuracy scores only. The interpretation of the UAS and LAS metrics of Old English parsing with UD allows us to gain insight into parsing models and to increase our understanding of how to process historical languages within up-to-date computational frameworks. The analysis demonstrates that the Pretrained model turns out promising results with the 20,000 word dataset, including 83.2% UAS and 74.2% LAS. These findings also indicate that improving LAS scores will require an increase in the size of the training dataset, given that the accuracy limit has not been reached yet. The performance gap between UAS and LAS metrics (approximately 9 percentage points) is telling us that the model is more accurate when it comes to identifying structural relations than when labeling specific dependency relations. This is true of the different levels of structural complexity (word, phrase, clause, and complex sentence) and is particularly evident in cases where morphological information should guide both structural and relational decisions, which suggests that the model cannot always integrate the morphology into both aspects of its analysis. Moreover, the model performs particularly well with morphologically marked local relations, but long-distance dependencies and complex syntactic structures do not always receive the correct analysis. This is clearly the case with syntactic configurations such as non-projective dependencies, fixed expressions and copulative constructions. Double object constructions, clausal complements, adverbial clause modifiers, contracted negative forms, and multi word expressions, including proper names and complex numerals, also presented different degrees of dificulty in accurate parsing.

To conclude, the significant contrast between morphological and syntactic parsing accuracy, on the one hand, and between structural attachment (UAS) and relation labeling (LAS), on the other, suggests that future research should focus on the improvement of the analysis of complex syntactic relations while maintaining the performance of morphological feature recognition. This may require to enrich morphological cues during parsing, to expand annotated training data, to refine the analysis of non-projective dependencies with transition-based methods, and to improve cross-level alignment so that morphological marking guides syntactic annotation by, for instance, validating morphological case against the expected syntactic functions.

Acknowledgments

Grant PID2023149762NB-100MCIN, funded by AEI /10.13039/501100011033.

References

Campbell, A. 1987. *Old English Grammar*. Oxford University Press.

Decatur, J. 2022. *A Survey of Non-Projective Dependencies and a Novel Approach to Projectivization for Parsing*. Ph.D. thesis, Uppsala Universitet.

de Marneffe, Marie-Catherine and Christopher D. Manning. 2008. *Stanford typed dependencies manual*. Technical report, Stanford University.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585-4592.

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

Fischer, Olga, Ans van Kemenade, Willem Koopman, and Wim van der Wurff. 2000. *The Syntax of Early English*. Cambridge University Press.

Healey, Antonette diPaolo, John Price Wilkin, and Xin Xiang. 2004. *The Dictionary of Old English Web Corpus*. Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Kastovsky, Dieter. 1992. Semantics and Vocabulary. In Richard M. Hogg, editor, *The Cambridge History of the English Language 1*, pages 290-408. Cambridge University Press.

Kübler, Sandra, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Manning, Christopher D. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 171-189. Springer.

Martín Arista, Javier. 2022a. Old English Universal Dependencies: Categories, Functions and Specific Fields. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022)*, volume 3, pages 945-951.

Martín Arista, Javier. 2022b. Toward the morpho-syntactic annotation of an Old English corpus with Universal Dependencies. *Revista de Lingüística y Lenguas Aplicadas*, 17:85-97.

Martín Arista, Javier. 2024. Toward a Universal Dependencies Treebank of Old English: Representing the Morphological Relatedness of Un-Derivatives. *Languages*, 9(3):76.

Martín Arista, Javier (ed.), Sara Domínguez Barragán, Luisa Fidalgo Allo, Laura García Fernández, Yosra Hamdoun Bghiyel, Miguel Lacalle Palacios, Raquel Mateo Mendaza, Carmen Novo Urraca, Ana Elvira Ojanguren López, Esaúl Ruíz Narbona, Roberto Torre Alonso and Raquel Vea Escarza. 2023. *ParCorOEv3. An open access annotated parallel corpus Old English-English*. Nerthus Project, Universidad de La Rioja, www.nerthusproject.com.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95-135.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, pages 1659-1666.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of LREC*, pages 4027-4036.

Ojanguren López, Ana Elvira. 2024. Structuring the Lexicon of Old English with Syntactic Principles: The Role of Deverbal Nominalisations with Aspectual and Control Verbs. In Javier Martín Arista and Ana Elvira Ojanguren López, editors, *Structuring Lexical Data and Digitising*

Dictionaries. Grammatical Theory, Language Processing and Databases in Historical Linguistics, pages 327-394. Brill.

Taylor, Ann, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. University of York.

Villa, Luca Brigada and Martina Giarda. 2023. Using modern languages to parse ancient ones: A test on Old English. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023)*, pages 30-41.

Zeman, Daniel; et al., 2024, Universal Dependencies 2.15, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-5787.