Explicit Edge Length Coding to Improve Long Sentence Parsing Performance

Khensa Daoudi
CRISCO, University of Caen, Inria Nancy
khensa.daoudi@unicaen.fr

Mathieu Dehouk
Lattice, CNRS, ENS-PSL, USN
mathieu.dehouk@cnrs.fr

Natasha Romanova
CRISCO, University of Caen
natalia.romanova@unicaen.fr

Rayan Ziane
LLL, University of Orléans
rayan.ziane@univ-orleans.fr

Abstract

Performance of syntactic parsers is reduced for longer sentences. While some of this reduction can be explained by the tendency of longer sentences to be more syntactically complex as well as the increase of candidate governor number, some of it is due to longer sentences being more challenging to encode. This is especially relevant for low-resource scenarios such as parsing of written sources in historical languages (e.g. medieval and early-modern European languages), in particular legal texts, where sentences can be very long whereas the amount of training material remains limited. In this paper, we present a new method for explicitly using the arc length information in order to bias the scores produced by a graph-based parser. With a series of experiments on Norman and Gascon data, in which we divide the test data according to sentence length, we show that indeed explicit length coding is beneficial to retain parsing performance for longer sentences.

Introduction

As a rule, when syntactic parsing models are evaluated, the general Labeled Attachment Score (LAS) is calculated without taking into account performance for different sentence lengths. The LAS assesses the performance of a parser by considering the number of words that have been assigned both the correct syntactic head and the correct label (Nivre and Fang, 2017).

For *treebanks* of low-resourced languages or language varieties (e.g. medieval languages) where small amounts of annotated data exist, precision of the annotation is paramount for syntactic research and constitution of reliable training corpora; manual revision of automatic parsing is therefore required. When correcting automatic annotation of historical French texts (e.g. Old, Middle and sixteenth-century French), it was empirically observed by the authors that the performance of parsers is significantly reduced on longer sentences; we elaborate on this in the next paragraph. Some errors appear counter-intuitive, e.g. distance between the token and its head, the direction of the arc, especially in the case of nominal dependents such as *det* and *case*. Thus, the longer the sentence, the higher the likelihood that, for example, an article would be attached to a noun several tokens to the left when its actual head is the next token to the right.

To give an example, we tested a model trained on one type data on a similar target corpus. First, we trained a dependency parser, BertForDeprel (Guiller, 2020), an open source model, based on Dozat and Manning (Dozat et al., 2017) architecture. For the embedding layer, we used XLM-RoBERTa multilingual model. This parser was trained on Old French (UD_Old_French–PROFITEROLE@2.16 corpus (Prévost et al., 2024) and achieved a global LAS of 89% and UAS of 92%. To evaluate its performance and assess its sensitivity to sentence length, we used a small sample from the 13th-century chronicle *Histoire ancienne jusqu'à César* (HaC-Sample). The sentences were selected from the digital edition of the chapter 'Rome II' from the manuscript BnF fr. 20125 (Morcos et al., 2021) and manually annotated and validated. The language and the genre of the target corpus as well as the principles of sentence segmentation were the same as in the training corpus.

The HaC-Sample dataset was divided into ten groups based on sentence length to examine the influence of length on parsing performance. The result of the parsing presented in graph 1 shows that the parser has better performance on medium-length sentences. Performance decreases for shorter and longer sentences, however. This drop may be explained by the lack of syntactic structure for the shorter sentences and the rise of syntactic complexity in the longer ones.

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 102–110, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-011

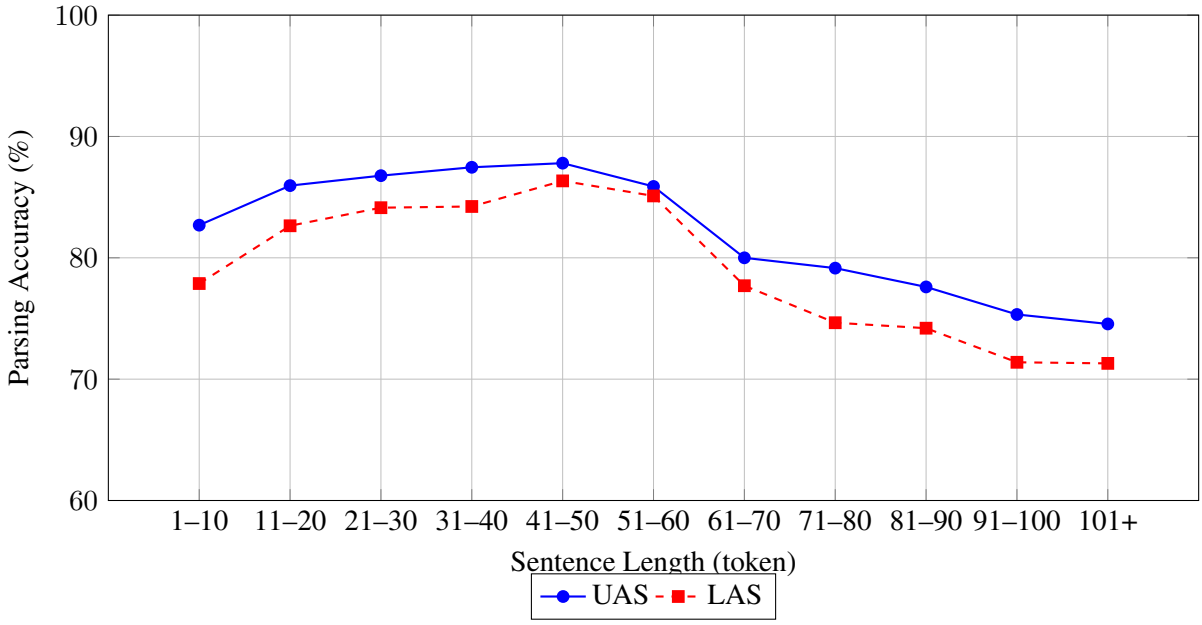

Figure 1: LAS and UAS parsing performance by sentence length group on HaC-Sample.

Related Work

The impact of sentence length on the accuracy of dependency parsers has been highlighted in different studies. (Gulordava and Merlo, 2016) conducted multilingual evaluation using artificially-generated *treebanks*, demonstrating that word variability and longer dependencies significantly degrade parser performance independently of the language or the *treebank* size. (Anderson and Gómez-Rodríguez, 2020) introduced the concept of Inherent Dependency Displacement Bias, which shows the bias of the parsing algorithm in handling the distance and direction of syntactic arcs. The authors found a strong correlation between sentence length and parsing accuracy. (Ajusha and Ajees, 2024) investigated the challenges in Malayalam, southern Dravidian language, where they found that the parsers struggled on long distance dependencies. These studies emphasize sentence length as a linguistic factor affecting parser performance.

At the same time, to address the problem of improving parser accuracy, previous researchers focused on the incorporation of morphosyntactic features into parsing models. (Nguyen and Verspoor, 2018) showed that high-quality PoS (Part-of-speech) tagging can improve parsing accuracy in biomedical texts. In the context of low-resource languages, (Anderson et al., 2021) demonstrated that predicted Universal PoS tags can significantly enhance the parsing, even in the absence of gold tags. (Ziane and Romanova, 2024) explored pre-finetuning of a parser with PoS tagging, thus biasing the parser's behaviour to improve its learning algorithm. On the other hand, (Altıntaş and Tantuğ, 2023) 's approach focused on architectural enhancement of the parser. By injecting global sentence embedding and CNN-based local context features into the arc scoring layer, this method empowered the graph based parser.

In this work, we aim to address specifically the problem of sentence length in dependency parsing in the context of low-resource historical texts. The method is based on the idea of biasing the scores produced by the parser to reflect the arc length information in 16th-century Norman (Guernsey) and medieval Gascon *treebanks*.

Corpus

For the experiments described below we used two of the corpora of the latest release of the Universal Dependencies collection (@2.16 released 15 May 2025) with the longest average sentence length. We selected corpora of medieval Romance languages, both belonging to the legal genre.

The "Norman" corpus, UD_French-ALTS@2.16 (42,832 tokens; 1,269 sentences) (Romanova et al., 2025) is a corpus of court proceedings from the island of Guernsey (1563-1569) transcribed from the manuscript of the register *Crime I* preserved at Guernsey Greffe (court archives of the island).[1]

[1] https://github.com/UniversalDependencies/UD_French-ALTS.

A legal text, it contains many long formulaic sentences, complex sentences and lists. The register is written in French, the language of the court of justice on Guernsey in the sixteenth century. However, since the island was under the British rule, the scribes were not obliged to follow the ordinances of Villers-Cotterêts (1539), which imposed the use of standard French in the official documents of the Kingdom of France. The language of *Crime I* therefore exhibits numerous dialectal (Norman) features such spellings and morphological characteristics of Northern French dialects. Like Old and Middle French, it is characterised by high degree of variation of forms and word orders. The average sentence length for the dev part of the corpus is 40,22 tokens, for the test part 36,16 tokens.

The "Gascon" corpus, UD_Occitan-CorAG@2.16 (1,094 sentences; 37,585 tokens) (Francioni et al., 2025) contains two medieval (one thirteenth-century and one fifteenth-century) legal manuals.[2] Gascon is a dialect of Old Occitan. This is the first available UD-annotated corpus in any medieval variety of Occitan. The average sentence length for the dev part of the corpus is 29,13 tokens, for the test part 35,24 tokens.

Both corpora were annotated in Parts-of-Speech (PoS), syntactic functions and heads in the Universal Dependencies (UD) framework (de Marneffe et al., 2021) by progressively adapting a model for Old and Middle French based on Profiterole corpus (Prévost et al., 2024) using ArboratorGrew software (Guibon et al., 2020) and built-in BertForDeprel parser (Guiller, 2020). Automatic annotation was manually checked.

The data for the experiments described below was split into three groups 70% train, 20% test and 10% dev, then the test group was divided into ten groups by length of the sentence.

Methodology and results

Length-Biased Graph Parser

As mentioned above, graph based parsers suffer a drop in quality as sentence length increases. There are several compounding factors leading to this. Longer sentences tend to be more syntactically complex with several levels of subordinated clauses for example. Moreover, longer sequences tend to be harder to handle for recurrent neural networks. The number of potential governors simply increases with the length of the sentence and, whereas the number of valid dependencies increases linearly with the length of the sentence, the number of invalid dependencies increases quadratically with it.

However, we noticed that even very simple errors appear in very long sentences, such as determiners attaching to nouns tens of tokens away. The most likely explanation for this is the difficulty for the biaffine layer to use the relative distance between tokens in order to reduce the score of unlikely long distance dependencies. We therefore propose to add a biasing mechanism beside the biaffine layer to help the parser avoid invalid long dependencies.

The basic idea is to add a multiplicative bias to the biaffine layer in order to boost or diminish the scores of arcs based on their signed distances. However, since different syntactic relations can have very different lengths and directions, we need to add extra information about each arc.

Therefore, we hypothesized that learning a bias for each triplet of governor PoS-tag, dependent PoS-tag and signed length of the arc should help the parser select better heads for words that have very local relations such as determiners or adjectives.

The biases for the selection of the relation label are based on the pair of governor-dependent PoS-tags and the dependency relation.

We also experimented with biasing over the signed length of the relation, however the results did not seem to improve. This may be due to the small size of our training data, and maybe with a bigger training set results would become more interesting.

In order to easily experiment with different biasing methods, we worked with our own re-implementation of (Dozat et al., 2017)'s graph-parser.[3]

We now describe the arc's length biasing mechanism. Given a sentence x of length n, the base parser produces the arc score matrix $S \in \mathbb{R}^{n \times n}$ and the relation label score tensor $R \in \mathbb{R}^{n \times n \times r}$, where r is the number of dependency relation labels.

Let $P\{0,1\}^{n \times p}$ be the matrix of one-hot encoded PoS-tags corresponding to x, where p is the number of PoS-tags types. Let l be the maximum arc length we want to consider, every longer edges will be cast to $\pm l$. Then, let $D\{0,1\}^{n \times n \times (2l+1)}$

[2] https://github.com/UniversalDependencies/UD_Occitan-CorAG.

[3] Code can be downloaded at https://github.com/MathieuDehouck/LowRes-Parser.

be the tensor encoding signed edge lengths in a one-hot manner:

$$D_{ijk} = \begin{cases} 1 \text{ if } k = \max(\min(i-j, l), -l) + l \ , \\ 0 \text{ otherwise.} \end{cases}$$

The arc biases B^{arc} and relation biases B^{rel} are then computed as follows:

$$B^{arc} = (P \otimes P^T \otimes D)\Theta^{arc},$$

$$B^{rel} = (P \otimes P^T \otimes \mathbf{1}^r)\Theta^{rel},$$

where $\Theta^{arc} \in \mathbb{R}^{p \times p \times (2l+1)}$ and $\Theta^{rel} \in \mathbb{R}^{p \times p \times r}$ are learnable parameters, $\mathbf{1}^r$ is the vector of length r where each entry is a 1, and where $\otimes$ notes the Kronecker product.

The final scores are then $S \odot B^{arc}$ and $R \odot B^{rel}$, where $\odot$ notes the Hadamard product.

Since we chose to work with multiplicative biases, values bigger than 1 are positive biases and values below are negative biases.

Experiments

In order to test the capacity of arc biasing to increase parsers' ability to handle longer sentences, we experimented with four parsing scenarios.

We trained parsers using only word embeddings taken from an encoder large language model as a simple baseline (Embedding). We then trained parsers using concatenated word and PoS-tag embeddings (+ PoS). This is a stronger baseline. Then, we trained parsers that only bias the arcs' scores based on their lengths, but do not bias the relations' scores (+ Arc bias). This is equivalent to setting Θ^{rel} to 1 and not updating it. Finally, we trained parsers that bias both arcs' and relations' scores as described in previous section (+ Rel bias).

For the embedding layer, we use the BERTrade language model (Grobol et al., 2022) trained specifically for Medieval French for both Norman and Gascon text since the only natively Occitan encoder we found had a too short context length to represent our sentences. While Occitan and Medieval French are closely related languages, this is obviously a sub-optimal situation and will explain the relative quality of the Gascon parser. When a word is split into multiple tokens by the encoder's tokenizer, we only keep the representation of the first token.

The PoS-tags embeddings are learned alongside the rest of the parser's parameters. In order to see the influence of the biases on the parsing quality of longer sentences we split the Norman and Gascon test sets into subsets of similarly sized sentences. The detail of the splits are reported in table 1 for the Norman data and in table 2 for the Gascon data.

Sentences length	Number of sentences	Number of tokens
5 - 10	24	195
11 - 20	121	1973
21 - 30	82	2008
31 - 40	49	1731
41 - 50	25	1140
51 - 60	22	1196
61 - 80	19	1291
81 - 137	5	505
All	347	6673

Table 1: Sizes of the Norman test subsets based on sentence length.

Results are thus reported for the whole test set and for each length based subset. They are averaged over 5 runs initialized with different random seed.

Results

Results for the Norman parsing experiment are reported in table 3 and those for the Gascon experiment are reported in table 4.

As we can see from table 3, adding PoS-tags embeddings already improves a lot the parsing capacity of the models.

However, while the models with and without arc and relation biasing are on par for sentences of length up to 60 tokens when they can use PoS-tags, for longer sentences, the biased models have a clear advantage. For sentences of length between 61 and 80 tokens, biased parsers show a 1.25 unlabeled attachment score point (UAS) increase and a 1.20 labeled attachment score point (LAS) increase. For sentences beyond 81 tokens, it reaches 3.09 UAS and 2.89 LAS points increase.

Thus arc biasing indeed seems to help maintaining a better parsing accuracy for longer sentences.

Table 4 gives a very similar picture.

However, since the parsers are of an overall lower quality due to the mismatch between the pre-training language of the encoder and the language it is applied to, the effects are even more marked. Here, even for reasonably sized sentences (less than 60 tokens) the biased models already show an advantage over the non biased ones.

Sentences length		Number of sentences	Number of tokens
5	10	27	227
11	20	88	1347
21	30	52	1316
31	40	32	1124
41	50	26	1173
51	60	16	886
61	70	6	388
71	80	14	1063
81	90	6	520
91	100	4	371
101	125	5	535
126	150	3	391
151	175	2	304
176	200	2	355
	All	285	10007

Table 2: Sizes of the Gascon test subsets based on sentence length.

We go from +0.38 UAS point for sentences of lengths between 21 and 30 tokens, to +2.49 UAS for sentences between 71 and 80 tokens, to up to +9.67 UAS for sentences of lengths between 151 and 175 tokens.

Figure 2 and figure 3 represent the evolution of the percentage of UAS error reduction for different models with respect to the baseline, embedding only, parser for the ALTS Norman and the CorAG Gascon test sets respectively.

We see that on both figures, the curves representing the UAS error reduction for the two arcs' length-biased models (with and without relation label biasing) stay close together around the 40 % line, while the curve corresponding to the unbiased model starts departing from the other two for longer sentences (more than 60 tokens) getting below the 30 % line.

It is also interesting to note that despite the Norman and Gascon models having very different performances, the error reduction of the PoS-tag informed and the arcs' length-biased models are surprisingly similar.

However, we do not know if it is a meaningful phenomenon or if it is just a coincidence and thus it needs further investigation.

These results indeed seem to support the ability of arc and relation biasing to improve accuracy of longer sentences parsing.

This is true even with respect to models that use

Group test set	Parser	UAS	LAS
5 - 10	Embedding	90.15	83.49
	+ PoS	**97.85**	**94.46**
	+ Arc bias	97.64	**94.46**
	+ Rel bias	97.23	94.05
11 - 20	Embedding	92.60	89.72
	+ PoS	95.43	93.75
	+ Arc bias	95.45	93.75
	+ Rel bias	**95.57**	**94.01**
21 - 30	Embedding	89.28	85.64
	+ PoS	92.30	90.13
	+ Arc bias	92.59	90.54
	+ Rel bias	**92.87**	**90.98**
31 - 40	Embedding	88.39	84.84
	+ PoS	**93.19**	91.18
	+ Arc bias	92.92	91.00
	+ Rel bias	92.96	**91.22**
41 - 50	Embedding	86.79	83.60
	+ PoS	**91.68**	**89.72**
	+ Arc bias	91.54	89.70
	+ Rel bias	91.61	89.61
51 - 60	Embedding	86.76	83.70
	+ PoS	91.69	90.27
	+ Arc bias	**91.99**	**90.72**
	+ Rel bias	91.62	90.27
61 - 80	Embedding	87.02	84.32
	+ PoS	89.70	88.23
	+ Arc bias	90.84	89.14
	+ Rel bias	**90.95**	**89.43**
81 - 137	Embedding	83.92	80.55
	+ PoS	87.60	86.26
	+ Arc bias	90.38	88.36
	+ Rel bias	**90.69**	**89.15**
All	Embedding	88.65	85.37
	+ PoS	92.46	90.64
	+ Arc bias	92.78	90.96
	+ Rel bias	**92.85**	**91.14**

Table 3: Results of the experiments on Norman data. Performance metrics (UAS and LAS) for different test subsets, grouped by sentence length, across four parser variants: word **Embedding** alone, **+ PoS** tags embedding, **+ Arc bias**, and **+Rel bias**.

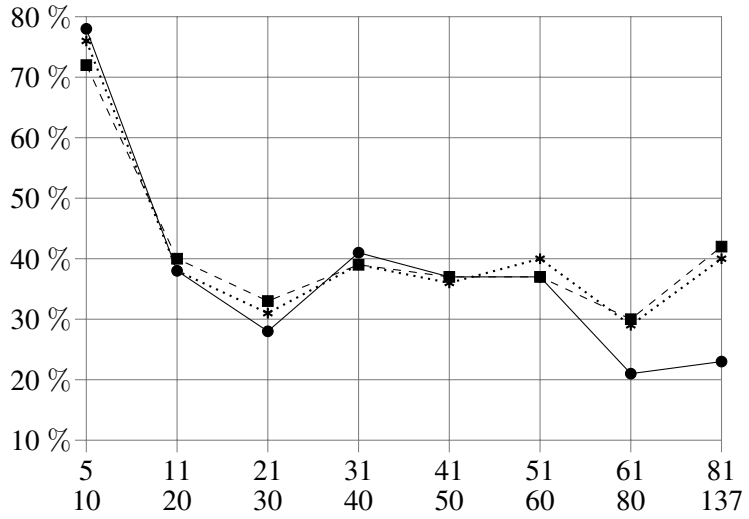

Figure 2: Graphical representation of the error reduction with respect to the baseline, embedding only, model UAS score for each sentence length-based ALTS Norman test subset. Bullets (•) represent the **+ PoS** model. Asterisks (∗) represent the **+ Arc bias** model. Squares (■) represent the **+ Rel bias** model.

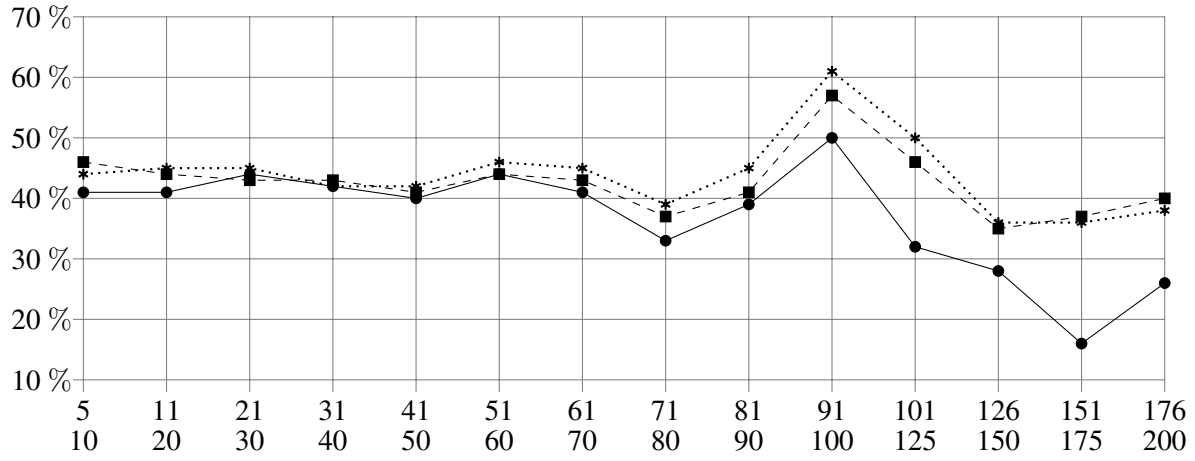

Figure 3: Graphical representation of the error reduction with respect to the baseline, embedding only, model UAS score for each sentence length-based CorAG Gascon test subset. Bullets (•) represent the **+ PoS** model. Asterisks (∗) represent the **+ Arc bias** model. Squares (■) represent the **+ Rel bias** model.

Group test set	Parser	UAS	LAS
5 - 10	Embedding	66.61	57.53
	+ PoS	80.26	76.30
	+ Arc bias	81.23	76.12
	+ Rel bias	**82.11**	**77.36**
11 - 20	Embedding	65.66	55.99
	+ PoS	79.82	75.77
	+ Arc bias	**81.14**	**76.60**
	+ Rel bias	80.88	76.38
21 - 30	Embedding	67.57	57.95
	+ PoS	81.73	**77.63**
	+ Arc bias	**82.11**	77.31
	+ Rel bias	81.63	77.25
31 - 40	Embedding	65.36	56.94
	+ PoS	79.80	75.21
	+ Arc bias	79.95	75.27
	+ Rel bias	**80.32**	**75.53**
41 - 50	Embedding	60.90	52.23
	+ PoS	76.47	73.08
	+ Arc bias	**77.24**	**73.59**
	+ Rel bias	76.83	73.23
51 - 60	Embedding	57.47	47.88
	+ PoS	76.00	71.74
	+ Arc bias	**77.20**	**72.37**
	+ Rel bias	76.32	71.38
61 - 70	Embedding	60.31	53.40
	+ PoS	76.49	73.20
	+ Arc bias	**78.30**	**74.85**
	+ Rel bias	77.32	74.18
71 - 80	Embedding	59.12	48.62
	+ PoS	72.47	67.70
	+ Arc bias	**74.96**	**69.80**
	+ Rel bias	74.43	69.13
81 - 90	Embedding	59.00	47.77
	+ PoS	74.96	71.31
	+ Arc bias	**77.42**	**72.58**
	+ Rel bias	75.88	71.58
91 - 100	Embedding	62.26	52.56
	+ PoS	81.08	76.33
	+ Arc bias	**85.39**	**80.97**
	+ Rel bias	83.83	80.11
101 - 125	Embedding	53.83	43.63
	+ PoS	68.75	64.45
	+ Arc bias	**77.05**	**71.78**
	+ Rel bias	75.18	70.88
126 - 150	Embedding	56.21	48.49
	+ PoS	68.59	65.17
	+ Arc bias	**71.76**	68.34
	+ Rel bias	71.61	**68.59**
151 - 175	Embedding	53.03	49.21
	+ PoS	60.33	58.68
	+ Arc bias	69.87	67.43
	+ Rel bias	**70.20**	**67.89**
176 - 200	Embedding	49.75	36.85
	+ PoS	62.87	58.31
	+ Arc bias	68.79	**63.38**
	+ Rel bias	**69.69**	62.59
All	Embedding	61.30	51.97
	+ PoS	76.01	71.97
	+ Arc bias	**78.17**	**73.65**
	+ Rel bias	77.71	73.31

Table 4: Results of the experiments on Gascon data. Performance metrics (UAS and LAS) for different test subsets, grouped by sentence length, across four parser variants: word **Embedding** alone, **+ PoS** tags embedding, **+ Arc bias**, and **+Rel bias**.

the same overall input features (word embeddings and PoS-tags) suggesting that a proper encoding of arcs' length is beneficial for longer sentences.

Discussion

In addition to increasing the parser's accuracy, PoS-tag based biases are easily interpretable by humans. Since these are multiplicative biases, a value above 1 is a positive bias and a value smaller than 1 is a negative bias. Figure 4 represents the value of length biases for a selection of pairs of PoS-tags. Biases corresponding to the NOUN-DET pairs are represented by black bullets.

The positions -1 and -2 are the only ones with a positive bias (1.23 and 1.20 respectively). This aligns perfectly with the fact that, in Medieval and early Modern French, determiners come right before their nouns, save a potential adjectival phrase. The biggest negative bias appears at position -5 with a value of 79.

There are a number of constructions where a determiner appears five tokens before a noun while not being governed by this very noun. Here we give just a few examples with English glosses below.

Le	*sabmedy*	*.xe.*	*jour*	*du*	*moes*
the	*saturday*	*10th*	*day*	*of_the*	*month*

de	*l'*	*uylle*	,	*du*	*pain*
of	*the*	*oil*	,	*of_the*	*bread*

son	*filz*	*venoient*	*en*	*sa*	*maison*
their	*son*	*came*	*in*	*their*	*house*

Overall there are 346 such instances in the training data and not a single one where a determiner would attach to a noun four tokens away.

On the same figure, we represent the biases learnt for the VERB-PRON pairs with crosses. Here we see that contrary to the NOUN-DET arcs, there are positive biases corresponding to both left and right arcs.

This too, aligns well with Medieval and Modern French grammar. In Modern French, pronouns tend to appear before their verb, but inversion is common in orders (direct and indirect object pronouns follow imperative verbs) and questions as well as a way to introduce reported speech.

Furthermore, pronouns were more mobile in Medieval French.

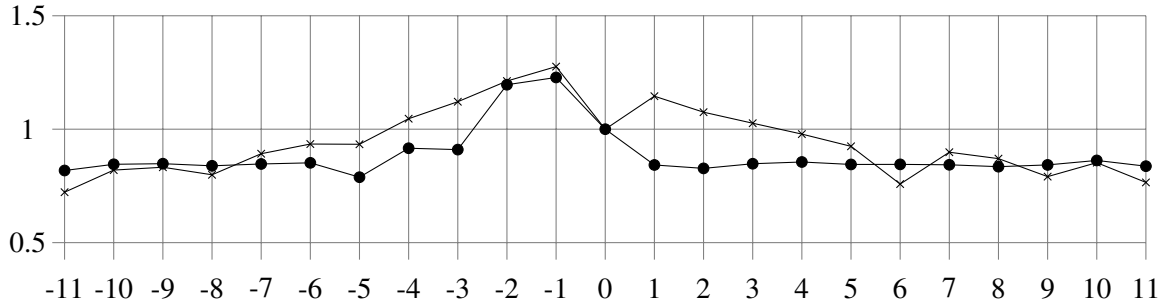

Figure 4: Samples of arc biases learnt on the ALTS Norman treebank. Bullets (•) represent the NOUN-DET arcs and crosses (×) represent the VERB-PRON arcs.

Future Work

Arc and relation label biasing can be easily applied to any parser that gives access to the score tensors on top of the actual structure prediction. So it would be interesting to see how this can be used in order to do a very light weight form of fine-tuning of already trained models.

Indeed some graph-based parsers actually take raw text as input and predict PoS-tags at the same time as the arcs' scores, so we would need to wait for this PoS-tag prediction in order to bias the arcs' scores. Furthermore, we still need to perform a more complete investigation of the learnt biases and we also intend to investigate their usability for transfer and language comparison, since they encode grammatical rules in a very simple format.

Eventually, since taking inspiration from the models that predict PoS-tags and dependency scores at the same time, in a multi-task learning spirit, teaching parsers to predict the signed length of an arc based on its governor's and dependent's representations could help them avoiding invalid long dependencies better, maybe even without having to bias.

Conclusion

We have presented first experiments towards tackling reduced performance of syntactic parsing in longer sentences: directly biasing the scores of the arcs in order to reflect their length. This is especially relevant when working with historical written texts, particularly of the administrative and legal types. These experiments point to the necessity to learn the length and direction of arc between the syntactic function and its head and the direction of the arc from the training corpus. We have seen that the experiments presented allow beginning to improve the scores.

Limitations

A more detailed analysis of these trends is needed, including a detailed error analysis, evaluating statistic significance of the results, testing on a wider variety of corpora and using bootstrapping scenarios. We believe that in order to improve performances on longer sentences a hierarchical approach to parsing may be beneficial.

Acknowledgements

This work was partially funded by the AUTOMATED project (2023-2025, University of Caen, France, PI Professor Pierre Larrivée; funded by Normandy Region). The staff at the Guernsey Greffe archives and the Guernsey Museum & Art Gallery gave us access to the manuscript and digital images of the Crime I register that were used for the Guernsey Norman corpus. Our thanks go to student transcribers who collaborated on the transcription. We thank Barbara Francioni who annotated the Gascon CorAG corpus.

References

P.V. Ajusha and A.P. Ajees. 2024. Morphological and syntactic challenges in malayalam: A dependency parsing perspective. *SSRG International Journal of Electrical and Electronics Engineering*, 11(12):375–385.

Mücahit Altıntaş and A. Cüneyd Tantuğ. 2023. Improving the performance of graph based dependency parsing by guiding bi-affine layer with augmented global and local features. *Intelligent Systems with Applications*, 18:200190.

Mark Anderson, Mathieu Dehouck, and Carlos Gómez-Rodríguez. 2021. A falta de pan, buenas son tortas: The efficacy of predicted UPOS tags for low resource UD parsing. 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 78–83, Online. Association for Computational Linguistics.

Mark Anderson and Carlos Gómez-Rodríguez. 2020. Inherent dependency displacement bias of transition-based algorithms. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5147–5155, Marseille, France. European Language Resources Association.

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

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Barbara Francioni, Natasha Romanova, Rayan Ziane, Khensa Daoudi, and Pierre Larrivée. 2025. Corag: Corpus d'ancien gascon. https://github.com/UniversalDependencies/UD_Occitan-CorAG.

Loïc Grobol, Mathilde Regnault, Pedro Ortiz Suarez, Benoît Sagot, Laurent Romary, and Benoit Crabbé. 2022. BERTrade: Using contextual embeddings to parse Old French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1104–1113, Marseille, France. European Language Resources Association.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Kirian Guiller. 2020. Analyse syntaxique automatique du pidgin-créole du nigeria à l'aide d'un transformer (bert): Méthodes et résultats. Master's thesis, Sorbonne Nouvelle.

Kristina Gulordava and Paola Merlo. 2016. Multilingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4:343–356.

Hannah Morcos, Simon Gaunt, Simone Ventura, Maria Teresa Rachetta, Henry Ravenhall, Natasha Romanova, Geoffroy Noël, Paul Caton, Ginestra Ferraro, and Marcus Husar. 2021. The histoire ancienne jusqu'à césar: A digital edition; bnf, fr20125 (interpretive edition): Eneas (6) and assyrian kings (6bis), rome i (7), rome ii (10) and caesar (11). https://http://www.tvof.ac.uk/textviewer/.

Dat Quoc Nguyen and Karin Verspoor. 2018. From POS tagging to dependency parsing for biomedical event extraction. *CoRR*, abs/1808.03731.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev, and Serge Heiden. 2024. Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval. *Corpus*, 25:15 pp.

Natasha Romanova, Rayan Ziane, and Khensa Daoudi. 2025. Alts: Automated sixteenth-century corpus. https://github.com/UniversalDependencies/UD_French-ALTS.

Rayan Ziane and Natasha Romanova. 2024. Pistes pour l'optimisation de modèles de parsing syntaxique. In *LIFT 2 – 2024: Journées de lancement*, Orléans, France. LIFT (Linguistique Informatique, Formelle et de Terrain).