Syntactic parsing: where are we going?

Lucelene Lopes[1], Thiago Alexandre Salgueiro Pardo[1], Magali S. Duran[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos-SP, Brazil

Abstract. *In this review & opinion paper, we discuss the options and challenges for syntactic parsing. Despite significant advances in recent years, driven primarily by neural network architectures, parsing accuracy appears to be approaching a plateau. This paper proposes a reflection on the factors that may possibly be influencing such results and suggests some future paths.*

Motivation

The importance of good part of speech tagging and parsing annotation tools for downstream Natural Language Processing (NLP) tasks is acknowledged by several publications in the history of the area, including both more classic (symbolic and statistic) approaches and new (usually neural-based) ones. In particular, the rise of "Universal Dependencies" (UD) framework[1] [Nivre et al. 2016, de Marneffe et al. 2021] has sparked renewed interest in dependency parsing, driving new efforts in syntax studies and parsing in NLP.

This review & opinion paper attempts to draw a landscape of more recent parsing efforts that align to UD standards, trying to figure out the potential limits of the task with current methods and what other strategies might be adopted for keeping improving the achieved results in the area. Such initiative is bold and naturally subject to failure, as natural languages have diverse characteristics and there are always new NLP methods emerging. Knowing this, this article makes a selection of works from the literature, choosing relatively recent and widely cited approaches in the area in order to draw some tentative (and certainly temporally anchored) conclusions.

Besides the possibly interesting work selection and overview that supported this paper, our contribution includes an exercise of "keeping the head above water", showing how far we have come and the imperfections of the landscape.

On current parsing techniques

The use of neural networks for detection of patterns, and consequently, the prediction of part of speech tags and dependency relations became the preferred method in the area [Goldberg 2016]. Within neural networks, several techniques as Long Short-Term Memory (LSTM) in its various versions [Van Houdt et al. 2020], together with other deep learning techniques [Dozat and Manning 2016], have been employed in the last decade with consistent advances for well resourced languages. The latest evolution brought by the self-attention methods [Vaswani et al. 2017], based on the famous Transformers, goes back a few years now, but it is still one of the main reasons for recent improvements.

[1]`https://universaldependencies.org/`

Overall, although different criteria could be used, in this paper we distinguish the parsing efforts according to the generic parsing tools or specific language parsing initiatives; and basic technology employed (e.g., BiLSTM, Deep Biaffine, and Self-Attention).

The more popular parsing tools, within UD standard, are the UDPipe in its versions 1.3 and 2.0 [Straka et al. 2016, Straka 2018], Stanza pipeline [Qi et al. 2020], UDify [Kondratyuk and Straka 2019], and AllenNLP pipeline [Dozat and Manning 2016]. Other less popular tools were developed, but apparently had fewer number of users, as the Diaparser [Attardi et al. 2021], UDapter [Üstün et al. 2020], UUParser [de Lhoneux et al. 2017], LAL-Parser [Mrini et al. 2019], and Hierarchical Pointer Network algorithm [Fernández-González and Gómez-Rodríguez 2023].

These parsers usually focus their efforts to cover several languages, being clearly multilingual. Some of these tools were specifically designed to cover the large set of languages available at the UD repository (which currenlty includes over 150 languages). However, from a technological point of view, the tools have considerable differences, although all of them make use of neural network models.

The technology of Bidirectional LSTM (BiLSTM) [Van Houdt et al. 2020] is frequently employed by many systems, including UDPipe 2.0, Stanza, and Hierarchical Pointer Networks algorithm. The Deep Biaffine technology [Dozat and Manning 2016] is found in AllenNLP pipeline, but also in tools as Diaparser and UDapter. Self-attention [Vaswani et al. 2017] is found in LAL-Parser and UDify tools. Additionally, the mentioned tools show differences on offering a static model or the possibility to perform model construction through a training set and/or to adopt pre-trained word embeddings.

Parsing results

The best values reported for each of the previously cited parsing methods are shown in Table 1. We chose to report only the Label Attachment Score (LAS), as this is usually the most adopted evaluation metric and also one of the most punitive metrics, as it measures the accuracy of the dependency relation identification and the tokens related as head and dependent. The table also indicates the language for which the highest LAS was reported.

Table 1. Highest LAS reported for the generic parsing tools.

parsing system	highest LAS	language	cited technology	publication
UUParser	87.34%	Portuguese	BiLSTM	2017
Stanza	90.01%	Spanish	Deep Biaffine	2020
UDPipe 1.3	91.20%	Hindi	NN Classifier	2016
UDapter	92.60%	Italian	Deep Biaffine	2020
Diaparser	93.65%	Italian	Deep Biaffine	2021
UDify	93.70%	Russian	Self-Attention	2019
UDPipe 2.0	94.53%	Russian	BiLSTM	2018
AllenNLP pipeline	94.60%	English	Deep Biaffine	2016
Hier. Pointer Networks	96.15%	English	BiLSTM	2023
LAL-parser	96.29%	English	Self-Attention	2019

The performance of the parsing methods vary considerably according to the language to which they are applied, as the scientific literature has shown. For example, for UDPipe 2, the reported LAS for Spanish and Italian can be as low as 80.68% and 77.34%, respectively. For AllenNLP pipeline, LAS for Chinese and Spanish was 85.38% and 91.65%, respectively. The values shown in the table may also reflect the number of tested languages. While UDify and UDPipe test over more than 70 languages, AllenNLP pipeline, UUParser, and LAL-parser test for only 6, 5, and 2 languages, respectively.

Focusing only on the highest LAS accuracy as presented in Table 1, it is noticeable that the majority of the highest scores are over 90% of accuracy. These numbers suggest that the State Of The Art (SOTA) for LAS is attainable despite of the technology employed, date of publication, and even specificity of each parsing development. Observing the three best reported results, we see different techniques and that English shows the best scores (probably because English is the best resourced language).

This fact suggests that, after the spread of neural network-based models, the quality of the training model plays a more important role than the specific technology employed. As such, the variations for different languages seem to reflect the quality of the training data for each language. For example, LAS for UDify for a low resourced language as Breton is as low as 40.19%, which is much lower than the 93.70% maximum attained for Russian.

Fortunately, the literature is abundant in terms of efforts for specific languages. These works usually are presented either with the construction of a specific corpus for the target language, or transferring learning from a better resourced language towards the low resourced one. Observing the works dedicated to specific languages, we found a reasonable number of publications, some of which are summarized in Table 2.

Table 2. Highest LAS reported by specific language efforts.

work	LAS	language	overall approach
[Dione 2021]	31.43%	Yoruba	Transfer learning
[Brigada Villa and Giarda 2023]	58.70%	Old English	Transfer learning
[Cassidy et al. 2022]	59.34%	Indonesian	Transfer learning
[Lusito and Maillard 2021]	60.74%	Ligurian	Corpus building
[Baig et al. 2021]	62.90%	Urdu	Corpus building
[Dione 2021]	67.83%	Wolof	Transfer learning
[Türk et al. 2022]	76.04%	Turkish	Corpus building
[Ghiffari et al. 2023]	79.22%	Irish	Corpus building
[Pedrazzini and Eckhoff 2021]	79.66%	Old Slavic	Transfer learning
[Sánchez-Rodríguez et al. 2024]	84.31%	Galician	Corpus building
[Alves et al. 2021]	89.09%	Croatian	Transfer learning
[Branco et al. 2022]	92.54%	Portuguese	Corpus building
[Kabiri et al. 2022]	92.68%	Persian	Corpus building
[Gamba and Zeman 2023]	94.61%	Latin	Corpus building
[Lopes and Pardo 2024]	94.70%	Portuguese	Corpus building

The examples summarized in Table 2 show efforts that can be grouped into attempts to serve very low resourced languages (as Old English, Old Slavic, Ligurian, Urdu, Bambara, Wolof, and Indonesian) and low resourced languages (as Turkish, Croatian, Galician, Irish, Persian, Latin, and Portuguese). While the very low resourced languages attempts are mostly based on transfer learning, the languages better resourced mostly center the efforts in building better corpora to be used to train specific models.

The observation of LAS in Table 2 shows that the best reported results are also above the 90% score of the generic parsing methods (Table 1). Obviously, the hard cases, as Yoruba and Old English, show low accuracy despite the efforts, probably because they are low-resourced languages. However, it is noticeable the accuracy achieved by transfer learning for Old Slavic and Croatian, as well as the high values for Persian, Latin, and Portuguese with the production of high quality training corpora.

Where can we head to?

The advent of popular neural network methods in the last decade has brought impressive progress in several areas of NLP, bringing Artificial Intelligence to the center of topics in all areas of the human knowledge. For parsing tasks, specifically, using UD standards, we notice the increase of quality since 2016. However, improvements seem to reach a limit up to 96% accuracy, and it is noticeable that no specificity show a clear predominance.

It is also well known that languages with few resources may not be able to benefit from the advantages of SOTA methods. It would be better for theses languages to invest in more classic methods or in the improvement of resources through corpora building including careful annotation. Specific techniques like data augmentation and joint task resolution may also be interesting ways (see, e.g., the work of [Yshaayahu Levi and Tsarfaty 2024] for Hebrew parsing). Such paths may also be relevant for languages already reaching accuracy around 95%, i.e., already delivering SOTA results.

Another relevant question is if the search for a better accuracy (over 96%) is a realistic goal. Should we make our peace with these missing 4% due to a natural inaccuracy of dependency annotation? Looking at the best method for a specific language (Portuguese), the authors [Lopes and Pardo 2024] [Duran et al. 2023a] [Duran et al. 2023b] discuss some reasons for the remaining errors that are also cited in the literature: under-represented phenomena in the training corpus (that might be solved by data augmentation and/or more corpus annotation) and difficult annotation issues (as to decide which is the head of a prepositional phrase) that sometimes may challenge even the humans. Personally, we believe that the above 99% accuracy already achieved for part of speech tagging may be achieved for parsing too. However, it may require to simplify some syntactic distinctions or to look for new approaches to the parsing problem.

The interested reader may find more information at the POeTiSA project web portal: `https://sites.google.com/icmc.usp.br/poetisa`

Acknowledgements

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - `http://c4ai.inova.usp.br/`), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

Alves, D., Bekavac, B., and Tadić, M. (2021). Typological approach to improve dependency parsing for Croatian language. In Dakota, D., Evang, K., and Kübler, S., editors, *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest)*, pages 1–11, Sofia, Bulgaria. Association for Computational Linguistics.

Attardi, G., Sartiano, D., and Simi, M. (2021). Biaffine dependency and semantic graph parsing for EnhancedUniversal dependencies. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D., editors, *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 184–188, Online. Association for Computational Linguistics.

Baig, A., Rahman, M. U., Shah, A. S., and Abbasi, S. (2021). Universal dependencies for urdu noisy text. *International Journal of Advanced Trends in Computer Science and Engineering*.

Branco, A., Silva, J. R., Gomes, L., and António Rodrigues, J. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 5617–5626, Marseille, France. European Language Resources Association.

Brigada Villa, L. and Giarda, M. (2023). Using modern languages to parse ancient ones: a test on Old English. In Beinborn, L., Goswami, K., Muradoğlu, S., Sorokin, A., Kumar, R., Shcherbakov, A., Ponti, E. M., Cotterell, R., and Vylomova, E., editors, *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 30–41, Dubrovnik, Croatia. Association for Computational Linguistics.

Cassidy, L., Lynn, T., Barry, J., and Foster, J. (2022). TwittIrish: A Universal Dependencies treebank of tweets in Modern Irish. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6869–6884, Dublin, Ireland. Association for Computational Linguistics.

de Lhoneux, M., Stymne, S., and Nivre, J. (2017). Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In Miyao, Y. and Sagae, K., editors, *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104, Pisa, Italy. Association for Computational Linguistics.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Dione, C. M. B. (2021). Multilingual dependency parsing for low-resource African languages: Case studies on Bambara, Wolof, and Yoruba. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D., editors, *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on*

Parsing into Enhanced Universal Dependencies, pages 84–92, Online. Association for Computational Linguistics.

Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

Duran, M., das Graças Nunes, M., and Pardo, T. A. (2023a). Construções sintáticas do português que desafiam a tarefa de parsing: uma análise qualitativa. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 424–433, Porto Alegre, RS, Brasil. SBC.

Duran, M. S., Nunes, M. d. G. V., and Pardo, T. A. S. (2023b). Avaliação qualitativa do analisador sintático udpipe 2 treinado sobre o córpus jornalístico porttinari-base. Technical report, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Fernández-González, D. and Gómez-Rodríguez, C. (2023). Dependency parsing with bottom-up hierarchical pointer networks. *Information Fusion*, 91:494–503.

Gamba, F. and Zeman, D. (2023). Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In Grobol, L. and Tyers, F., editors, *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Ghiffari, F. A. A., Alfina, I., and Azizah, K. (2023). Cross-lingual transfer learning for Javanese dependency parsing. In Li, D., Mahendra, R., Tang, Z. P., Jang, H., Murawaki, Y., and Wong, D. F., editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9, Nusa Dua, Bali. Association for Computational Linguistics.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *J. Artif. Int. Res.*, 57(1):345–420.

Kabiri, R., Karimi, S., and Surdeanu, M. (2022). Informal Persian Universal Dependency treebank. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 7096–7105, Marseille, France. European Language Resources Association.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Lopes, L. and Pardo, T. (2024). Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing* *of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Lusito, S. and Maillard, J. (2021). A Universal Dependencies corpus for Ligurian. In de Lhoneux, M. and Tsarfaty, R., editors, *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest)*, pages 121–128, Sofia, Bulgaria. Association for Computational Linguistics.

Mrini, K., Dernoncourt, F., Bui, T., Chang, W., and Nakashole, N. (2019). Rethinking self-attention: An interpretable self-attentive encoder-decoder parser. *CoRR*, abs/1911.03875.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666, Portorož, Slovenia. ELRA.

Pedrazzini, N. and Eckhoff, H. M. (2021). Oldslavnet: A scalable early slavic dependency parser trained on modern language data. *Software Impacts*, 8:100063.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Sánchez-Rodríguez, X., Sarymsakova, A., Castro, L., and Garcia, M. (2024). Increasing manually annotated resources for Galician: the parallel Universal Dependencies treebank. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 587–592, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Türk, U., Atmaca, F., Özateş, c. B., Berk, G., Bedir, S. T., Köksal, A., Başaran, B. O., Güngör, T., and Özgür, A. (2022). Resources for turkish dependency parsing: introducing the boun treebank and the boat annotation tool. *Lang. Resour. Eval.*, 56(1):259–307.

Üstün, A., Bisazza, A., Bouma, G., and van Noord, G. (2020). UDapter: Language adaptation for truly Universal Dependency parsing. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 2302–2315, Online. Association for Computational Linguistics.

Van Houdt, G., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yshaayahu Levi, D. and Tsarfaty, R. (2024). A truly joint neural architecture for segmentation and parsing. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1408–1420, St. Julian's, Malta. Association for Computational Linguistics.