Corpus-based language comparison: From morphology to dependencies and beyond

DOI: http://dx.doi.org/10.21165/el.v54i1.4032

Daniel Zeman[1]

Abstract

We provide an overview of the Universal Dependencies multilingual corpus collection, its current status and numerous extensions, such as the UNER annotation of named entities or the CorefUD annotation of coreference and anaphora. We discuss the utility of the data in several areas of Digital Humanities, with a particular focus on comparative linguistics and typology.

Keywords: annotated corpus; treebank; morphology; syntax; typology.

1 ÚFAL MFF, Charles University, Prague, Czechia; zeman@ufal.mff.cuni.cz; https://orcid.org/0000-0002-5791-6568

Comparação de línguas baseada em corpus: da morfologia às dependências e além

Resumo

Apresentamos uma visão geral da coleção de *corpus* multilíngue Universal Dependencies, seu estado atual e suas diversas extensões, como a anotação UNER de entidades nomeadas ou a anotação CorefUD de correferência e anáfora. Discutimos a utilidade dos dados em várias áreas das Humanidades Digitais, com foco particular em linguística comparativa e tipologia.

Palavras-chave: corpus anotado; treebank; morfologia; sintaxe; tipologia.

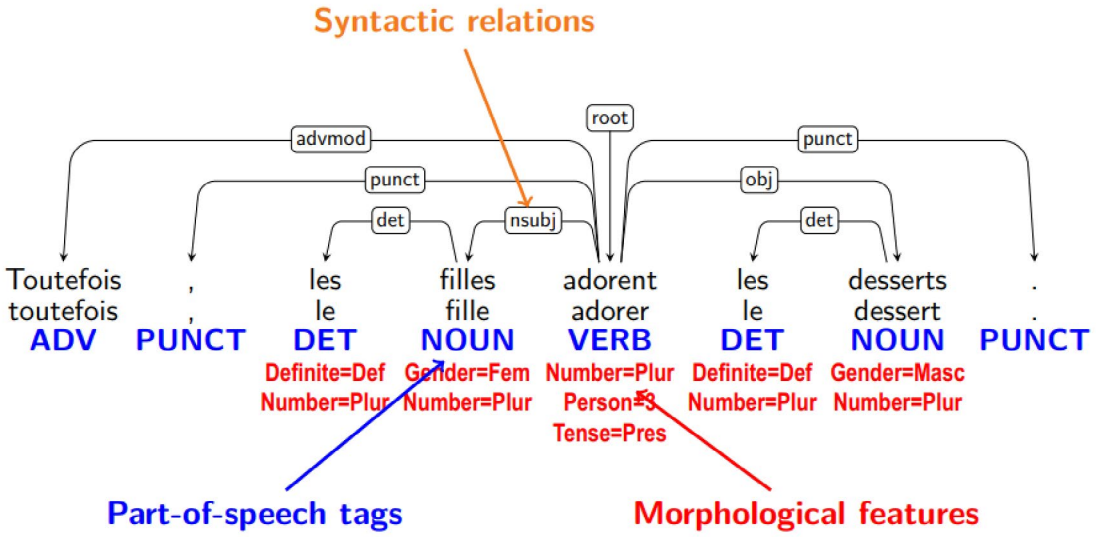
Introduction

In this contribution, we discuss the latest developments in Universal Dependencies (UD) (de Marneffe *et al*., 2021), with a particular focus on the applicability of UD in comparative linguistics and typology. Citing its website,[2] UD is "a project developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective." Nevertheless, UD is not just the name of the project. It also denotes its main outcomes, which are the annotation scheme and the data repository. Last but not least, UD is also a thriving community of researchers, providers and users of the annotated language data.

As a minimum, each UD dataset (treebank) contains manually verified annotation of part-of-speech categories and binary syntactic relations between words, organized as a rooted directed tree structure. Most treebanks also contain additional morphological annotation of lemmas and morphological features (see Figure 1; in a few cases, this part of the annotation has been generated using language-processing software and has not been fully verified by humans). Some treebanks also provide other annotation layers (more on that in Section "Extensions"). Morphosyntactic words are the basic annotation unit; UD normally does not annotate relations between morphs. Nevertheless, the extra annotation available in some treebanks includes a mechanism to indicate segmentation of words into morphs, Leipzig-style morphemic glosses, as well as full-word glosses and, where applicable, transliteration.

2 https://universaldependencies.org/; all webs cited were accessed in February 2025.

Figure 1. Example UD annotation of a French sentence meaning "However, girls love desserts." The annotation comprises lemmas, POS tags, morphological features and syntactic relations



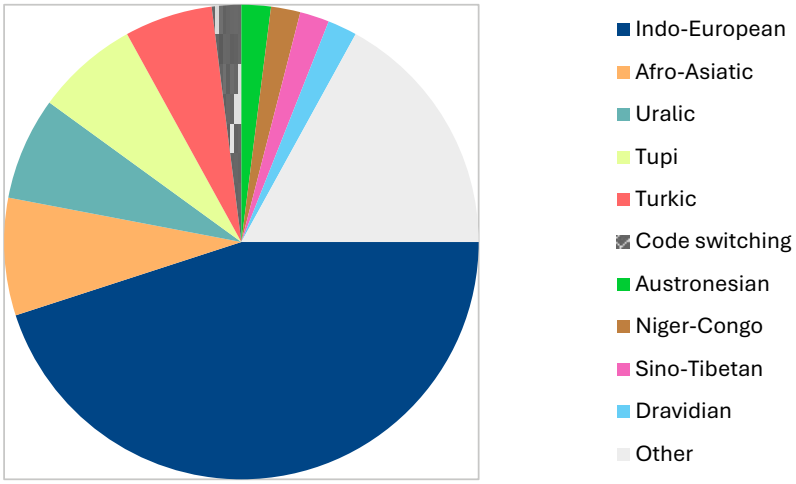
Source: Own elaboration

The treebanks contain various types of data (called 'genres' in the metadata, although it is a simplistic classification). They range from news through Wikipedia, religious texts (Bible) and fiction, to user-generated content in social networks, and spoken data (including fieldwork data). There are many languages that have more than one treebank in UD, and often the treebanks differ in genres covered. On the other hand, there are many treebanks that contain data of more than one genre. Some of them allow for filtering of the genres by sentence ids, but unfortunately this option is not available in all cases (Müller-Eberstein et al. 2021; Danilova; Stymne, 2023).

Some UD treebanks contain parallel texts in different languages, making them particularly suitable for comparative studies (e.g., Alves et al. 2023). The largest parallel set are the PUD (standing for "Parallel Universal Dependencies") treebanks (Zeman et al. 2017) with 1000 sentences of online news and Wikipedia, currently translated to 21 languages. Over 10 treebanks contain excerpts from the Bible, not always the same sections, but for example the Gospels from the New Testament are represented in several languages and can be used as a parallel corpus, well aligned at the level of verses. A few other treebanks have been based on parallel texts, e.g., LinES for English and Swedish, or SETS for Croatian and Serbian.

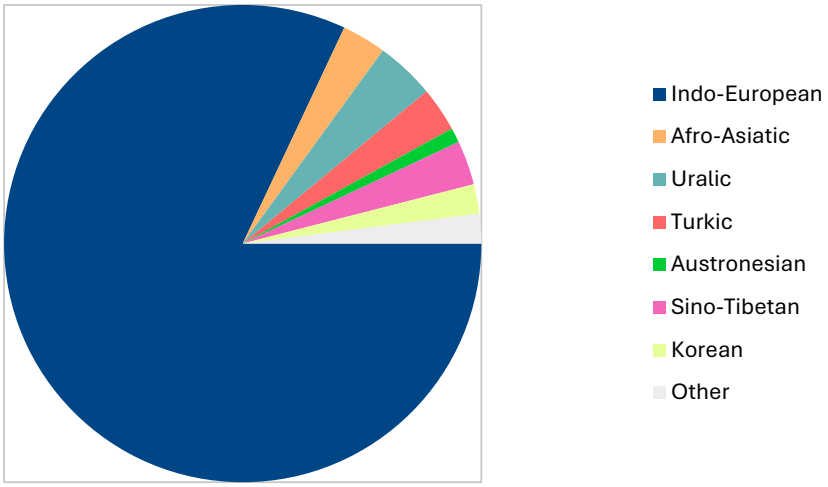
The UD data collection has a release cycle of two releases per year. The most recent[3] version, 2.15, covers 168 languages from 33 language families.[4] In total the annotated text amounts to almost 33 million words (1.9 million sentences).

Figure 2. Language families in UD by number of languages



Source: Own elaboration

Figure 3. Language families in UD by number of words in the data



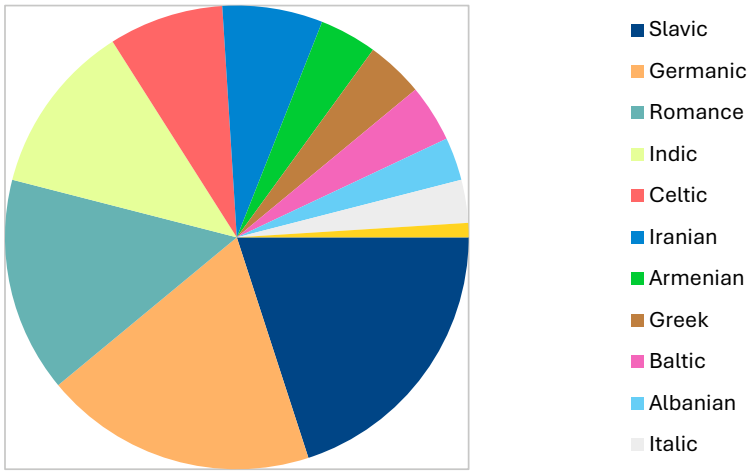
Source: Own elaboration

3 The present paper is based on the author's talk at a GEL 70 round table in July 2024; by that time, UD 2.14 was the most recent release. In the paper, the numbers have been updated to UD 2.15 (released in November 2024).

4 Including three special families for creoles, sign languages, and code switching.

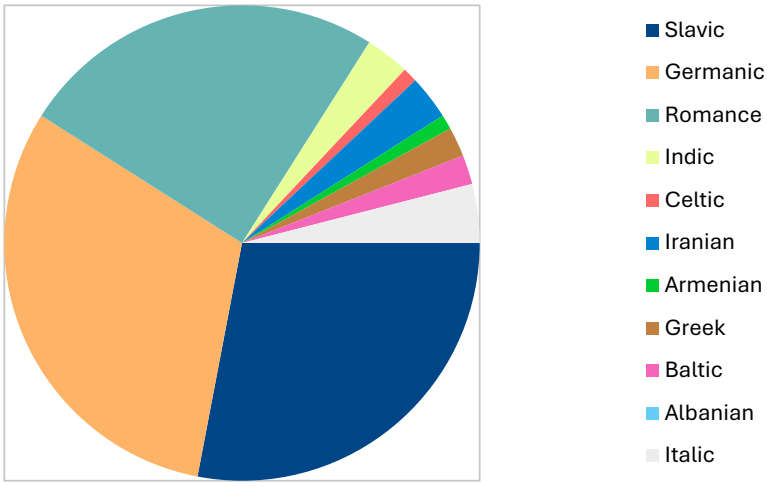
The collection is still quite unbalanced: The largest family, Indo-European, is represented by 76 languages, while 18 families have only 1 language each (Figure 2). UD is even less balanced when it comes to the size of the data (Figure 3): Indo-European languages together amount to 24 million words, Uralic have 1 million, 8 other families have over 100 thousand words each (Afro-Asiatic, Turkic, Sino-Tibetan, Korean, Japanese, Austronesian, Creole, Basque), while at the other end of the scale there are 11 families that have not reached 10 thousand words yet. Similar imbalance can be observed also within the Indo-European family, with Slavic, Germanic, and Romance being represented much better than the rest (Figures 4 and 5). These are limitations of UD that have to be taken into account when working with the data. UD cannot serve as a proportional reflection of the trends in the world's languages. Nevertheless, it is a useful resource for comparative studies of the languages that are already included.

Figure 4. Indo-European genera in UD by number of languages



Source: Own elaboration

Figure 5. Indo-European genera in UD by number of words in the data

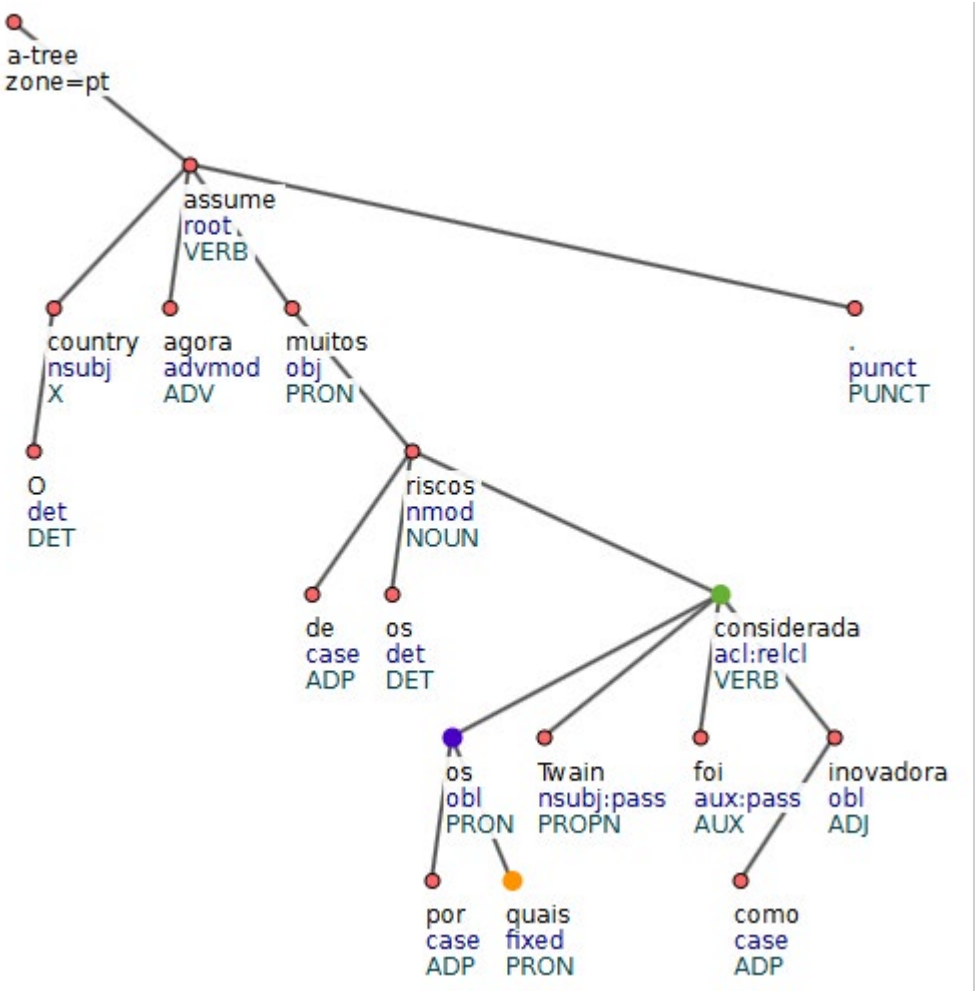


Source: Own elaboration

UD in Digital Humanities

The annotation harmonization efforts, which eventually led to UD, started in the parsing community as an attempt to increase interoperability of datasets that were used to train parsing models. Nevertheless, as the treebank collection grows, it has become an invaluable resource for comparative linguistics, as well as various other fields of digital humanities.

Figure 6. Search result in the Porttinari treebank of Portuguese. Meaning: "Country now takes many of the risks for which Twain was considered an innovator."



Source: Own elaboration

Linguists can access the data through specialized online search engines that do not require any programming skills. The queries can be much more complex than just asking for a particular word form or part-of-speech category. For example, one can search for examples of relative clauses where the relative pronoun is not a direct dependent of the predicate and on the path between the predicate and the pronoun there is no other relation labeled acl:relcl (nested related clause) or conj (coordination). Figure 6 shows an example found by this query[5] in the Porttinary treebank of Portuguese in UD 2.15. It is also possible to aggregate the results of a query and obtain statistics about their features.

5 Permanent link to the query: http://hdl.handle.net/11346/PMLTQ-DNLA

For example, the query on relative clauses could be altered to provide an overview of the relative pronouns and their dependency relation types, ordered by number of occurrences:

Relative word	Dependency relation	Frequency
qual	fixed	24
que	fixed	18
quais	fixed	13
cuja	det	10
cujo	det	4
cujas	det	4
cujos	det	1

Besides querying the manually annotated treebanks, users can also take one of the parsing models trained on the UD treebanks and use it to analyze their own data (which can be subsequently searched for interesting patterns, too). The accuracy of automatic parsing depends on the size and properties of the training treebank, but for many languages it exceeds 85% of correctly attached words, making it a useful tool especially in scenarios where the pre-selected parse can be subsequently verified by a linguist. Morphosyntactic parsers such as UDPipe[6] (Straka et al. 2016) or Stanza (Qi et al. 2020) come with models pre-trained on UD data[7] and can be used out-of-the-box to parse running text.

The treebanks (as well as automatically parsed data) are useful not just for linguistic research but also for the broader audience of language learners and teachers; educative applications can use both the morphological and the syntactic annotation encoded in the data. In addition, some UD treebanks are learner corpora and can be used to study typical errors made by non-native users of the language (Berzak *et al*., 2016; Lee *et al*., 2017; Hana; Hladká 2018; Di Nuovo *et al*., 2022; Sung and Shin, 2024).

UD is also a resource for historical linguistics. It has a growing number of corpora of classical languages and historical language varieties, ranging from Old Egyptian through Sanskrit and Classical Chinese to Middle French or Ottoman Turkish. In the synchronic perspective, UD is often used in documentation of endangered minority languages. It has

6UDPipe is also available as a web service that can be accessed through the browser at https://lindat.mff.cuni.cz/services/udpipe/

7 Not all UD treebanks are large enough to provide training material for parsers. Following UD release 2.15, UDPipe has at least one model for 77 languages.

samples of varying sizes from indigenous languages of all inhabited continents, including 20 languages native to the Americas. Syntactically annotated data from moribund languages contribute to documenting linguistic diversity and in some cases they may be used in revitalization efforts. Thanks to the ready-to-use treebanking infrastructure and ready-to-adapt annotation guidelines, it is now quite easy to start a treebank of a new language, even for field linguists who have not necessarily worked on treebanks before.

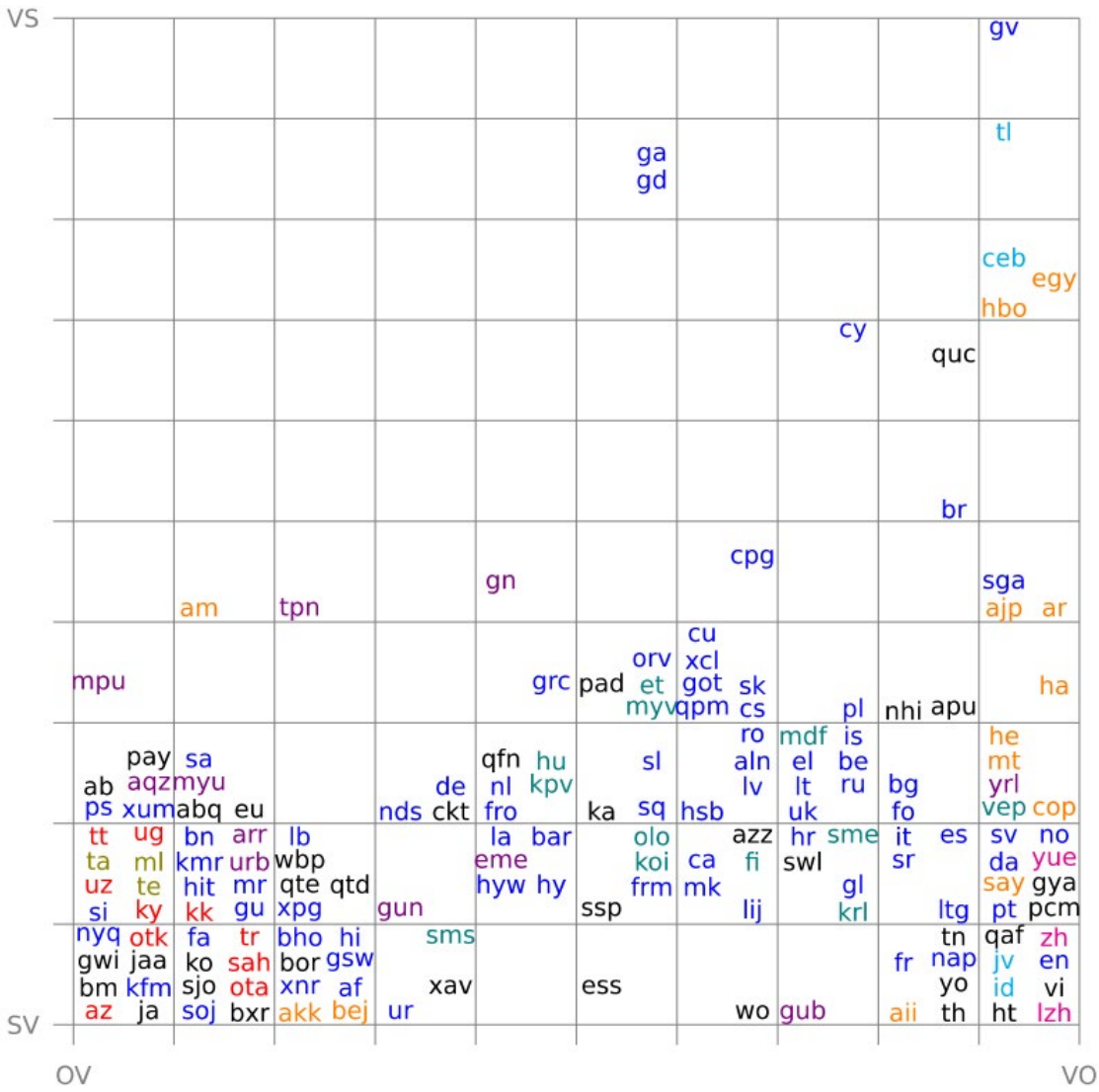
Finally, and obviously, a collection like UD is very useful for comparison of languages. A few examples of such studies are presented in the next section.

Corpus-based Typology

As a massively multilingual collection following a common annotation schema, UD has been recognized and used as a resource in various comparative studies (e.g., Futrell *et al*., 2015; Alzetta *et al*., 2018; Levshina, 2019; Gerdes *et al*., 2021; de Marneffe *et al*., 2024).

We can demonstrate the utility of UD data on one of the central questions in linguistic typology, the order of subject, verb, and object. The data confirm the widely accepted observation that in some languages word order plays a more significant role in coding core participants than in others. However, we can now say more than just assign a language to a category like SOV or SVO (or to say that the language has no dominant word order). We can quantify the trends and say how strongly a language leans towards a particular word order type. Figure 7 plots all UD languages to a two-dimensional space according to their word order frequencies. One can clearly see that some families, such as Turkic and Dravidian, strongly prefer the SOV/OSV type. Sino-Tibetan languages all end up as strongly SVO, although here we must be more careful with generalizations, as the family is currently represented only by three closely related languages (Mandarin Chinese, Classical Chinese, and Cantonese). All four Austronesian languages are positioned at the VO end, but the Philippine branch (represented by Tagalog and Cebuano) is also mostly VS, while Indonesian and Javanese are strongly SV. Other families, such as Indo-European, Afro-Asiatic and Tupian are more spread across the space.

Figure 7. The proportion of SV and VS orders (the y axis) vs. OV and VO orders (the x axis) for nominal arguments (i.e., nsubj is counted but csubj is not). Strictly SOV or OSV languages are in the bottom left corner, strictly SVO are bottom right, strictly VSO or VOS are in the top right corner. Here V refers to the position of the main verb (not AUX). Subjectless clauses are counted as the middle between SV and VS; similarly, objectless clauses pull the language towards the center on the OV–VO axis. Languages are represented by their ISO 639 code. Selected language families are colored



One could modify the query and the plot in various ways. For example, some languages (e.g., German) apply different word order rules in subordinate clauses as opposed to main clauses. This could be the reason why German, Low Saxon and Dutch appear near the center in Figure 7, yet slightly closer to the SOV side. In UD we can easily identify subordinate clauses and compute the language positions separately for each clause

type. Similarly, we can identify the effect of other factors, such as nominal vs. clausal arguments, subjectless and objectless clauses, or the mutual position of the main verb and its auxiliaries.

Extensions

While UD proper focuses on morphology and syntax, some of the UD treebanks contain additional layers of annotation. Moreover, there are separate projects that attempt to provide corpora with other annotation types in a similarly multilingual and "universal" manner as UD, some of them inspired by UD or even working on top of UD data.

First and foremost, UD has a symbiotic relationship with a sister project called Surface Syntactic UD or SUD[8] (Gerdes et al. 2018) that defines a different perspective on certain syntactic relations, which is preferred by some users, while maintaining full convertibility between SUD and UD. All UD-released corpora are also available in SUD, and some UD corpora were originally annotated in SUD and then converted.

The extension that is most tightly bound to UD because it is even defined in the UD v2 guidelines (although completely optional and available only in a few dozen treebanks) is Enhanced Universal Dependencies (EUD) (Schuster & Manning 2016; Nivre et al. 2020). EUD defines a non-tree dependency representation that features reentrancies and abstract nodes representing elided words. It is a step from the syntactic structure towards meaning, but only a small step; further steps in that direction have been proposed in the Deep UD[9] initiative (Droganova; Zeman 2019, 2024).

Universal Proposition Bank[10] (Jindal *et al*., 2022) is another project with semantics in mind, inspired by UD and by the English PropBank. On a subset of UD treebanks, it adds the annotation of semantic roles for arguments of verbs, automatically projected from English.

CorefUD[11] (Nedoluzhko *et al*., 2022) combines UD-style morphosyntactic annotation with annotation of coreference relations in multiple languages. It naturally includes annotation of entities (both named and denoted by other means, such as common nouns and pronouns). A related initiative is UNER[12] (Mayhew *et al*., 2024), which enriches selected

8 https://surfacesyntacticud.github.io/

9 https://ufal.mff.cuni.cz/deep-universal-dependencies

10 https://universalpropositions.github.io/

11 https://ufal.mff.cuni.cz/corefud

12 https://www.universalner.org/

UD treebanks with the annotation of non-nested named entities (but without coreference, i.e., not marking entities referred to by other means than names).

Finally, Uniform Meaning Representation (UMR)[13] (Bonn *et al*., 2024) intends to be the ultimate semantic corpus collection, incorporating arguments and semantic roles, temporal relations, coreference and entity linking, among other phenomena. Unlike the above mentioned projects, UMR does not currently build on top of UD data, but it is similar to UD in trying to be typologically informed and suitable for data in any natural language; in release 2.0, it provides data for English, Czech, Latin, Chinese, and four indigenous languages of the Americas: Arapaho, Navajo, Kukama, and Sanapana.

In a different direction, the PARSEME corpus[14] (Savary *et al*., 2023) provides annotations of idiomatic multiword expressions, once again combined with UD morphology and dependency trees.

There are also highly multilingual morphological databases, most notably UniMorph[15] (Batsuren *et al*., 2022) and Universal Derivations[16] (Kyjánek *et al*., 2020).

Conclusion

We have discussed Universal Dependencies (UD), a large collection of morphosyntactically annotated treebanks spanning many languages and language families. Thanks to the unified annotation framework, UD is an invaluable resource for research of individual languages, as well as their comparison. The UD project is opportunistic in its selection of languages, which naturally leads to imbalance in representation: Well studied and supported languages, such as the national languages of European countries, have much more data available in UD than languages from other continents and minority languages. Nevertheless, many such languages do already have at least a small treebank in UD, and new ones are added every year.

We have shown examples how UD is useful in Digital Humanities: syntactic searches in the treebanks and in automatically parsed data, educational use and learner corpora, historical linguistics, documentation of endangered languages, contrastive studies on parallel treebanks, and linguistic typology in general.

13 https://umr4nlp.github.io/web/

14 https://gitlab.com/parseme/corpora/-/wikis/home

15 https://unimorph.github.io/

16 https://ufal.mff.cuni.cz/universal-derivations

We have also stressed that while UD proper focuses on the layers of morphology and surface syntax, additional types of annotation are available for many of the UD languages, including enhanced and deep syntax, PropBank-style semantic roles, multiword expressions, named entities, and coreference.

All in all, UD and other similarly unified multilingual annotation efforts have proven indispensable in computational linguistics and natural language processing, and their value will grow as they gradually cover more languages and more linguistic phenomena.

Acknowledgments

The author is grateful to the organizers of the GEL Seminar, and specifically to Livia Oushiro and Filomena Sandalo, for their hospitality and for the opportunity to participate in the Round Table.

References

ALVES, D.; BEKAVAC, B.; ZEMAN, D.; TADIĆ, M. *Analysis of corpus-based word-order typological methods*. *In:* Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023). Washington, DC, USA, 2023. p. 36-46.

ALZETTA, C.; DELL'ORLETTA, F.; MONTEMAGNI, S.; VENTURI, G. *Universal dependencies and quantitative typological trends. A case study on word order*. *In:* Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan, 2018.

BATSUREN, K.; GOLDMAN, O.; KHALIFA, S.; HABASH, N.; KIERAŚ, W.; BELLA, G.; LEONARD, B.; NICOLAI, G.; GORMAN, K.; GHANGGO ATE, Y.; RYSKINA, M.; MIELKE, S.; BUDIANSKAYA, E.; EL-KHAISSI, C.; PIMENTEL, T.; GASSER, M.; ABBOTT LANE, W.; RAJ, M.; COLER, M.; MONTOYA SAMAME, J. R.; SITICONATZI CAMAITERI, D.; ZUMAETA ROJAS, E.; LÓPEZ FRANCIS, D.; ONCEVAY, A.; LÓPEZ BAUTISTA, J.; SILVA VILLEGAS, G. C.; TORROBA HENNIGEN, L.; EK, A.; GURIEL, D.; DIRIX, P.; BERNARDY, J.-P.; SCHERBAKOV, A.; BAYYR-OOL, A.; ANASTASOPOULOS, A.; ZARIQUIEY, R.; SHEIFER, K.; GANIEVA, S.; CRUZ, H.; KARAHÓǦA, R.; MARKANTONATOU, S.; PAVLIDIS, G.; PLUGARYOV, M.; KLYACHKO, E.; SALEHI, A.; ANGULO, C.; BAXI, J.; KRIZHANOVSKY, A.; KRIZHANOVSKAYA, N.; SALESKY, E.; VANIA, C.; IVANOVA, S.; WHITE, J.; HALL MAUDSLAY, R.; VALVODA, J.; ZMIGROD, R.; CZARNOWSKA, P.; NIKKARINEN, I.; SALCHAK, A.; BHATT, B.; STRAUGHN, C.; LIU, Z.; WASHINGTON, J. N.; PINTER, Y.; ATAMAN, D.; WOLINSKI, M.; SUHARDIJANTO, T.; YABLONSKAYA, A.; STOEHR, N.; DOLATIAN, H.; NURIAH, Z.; RATAN, S.; TYERS, F. M.; PONTI, E. M.; AITON, G.; ARORA, A.; HATCHER, R. J.; KUMAR, R.; YOUNG, J.; RODIONOVA, D.; YEMELINA, A.; ANDRUSHKO,

T.; MARCHENKO, I.; MASHKOVTSEVA, P.; SEROVA, A.; PRUD'HOMMEAUX, E.; NEPOMNIASHCHAYA, M.; GIUNCHIGLIA, F.; CHODROFF, E.; HULDEN, M.; SILFVERBERG, M.; MCCARTHY, A. D.; YAROWSKY, D.; COTTERELL, R.; TSARFATY, R.; VYLOMOVA, E. *UniMorph 4.0: universal morphology*. *In:* Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France, 2022.

BERZAK, Y.; KENNEY, J.; SPADINE, C.; WANG, J. X.; LAM, L.; MORI, K. S.; GARZA, S.; KATZ, B. *Universal dependencies for learner English*. *In:* Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. v. 1, n. Long Papers. Berlin, Germany, 2016.

BONN, J.; BUCHHOLZ, M.; CHUN, J.; COWELL, A.; CROFT, W.; DENK, L.; GE, S.; VAN GYSEL, J. E. L.; HAJIČ, J.; LAI, K.; MARTIN, J. H.; MYERS, S.; PALMER, A.; PALMER, M.; BENET POST, C.; PUSTEJOVSKY, J.; STENZEL, K.; SUN, H.; UREŠOVÁ, Z.; VALLEJOS YOPAN, R.; XUE, N.; ZHAO, J. *Building an infrastructure for uniform meaning representation*. *In:* Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resource and Evaluation (LREC-COLING 2024). Torino, Italy, 2024.

DANILOVA, V.; STYMNE, S. *UD-MULTIGENRE – a UD-based dataset enriched with instance-level genre annotations*. *In:* Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL). Singapore, 2023. p. 253-267.

DE MARNEFFE, M.-C.; MANNING, C.; NIVRE, J.; ZEMAN, D. *Universal dependencies*. *Computational Linguistics*, v. 47, n. 2, 2021. p. 255-308.

DE MARNEFFE, M.-C.; NIVRE, J.; ZEMAN, D. *Function words in universal dependencies*. *Linguistic Analysis*, v. 43, n. 3-4, 2024. p. 549-588.

DI NUOVO, E.; SANGUINETTI, M.; MAZZEI, A.; CORINO, E.; BOSCO, C. *VALICO-UD: treebanking an Italian learner corpus in universal dependencies*. *Italian Journal of Computational Linguistics*, v. 8, n. 1, 2022.

DROGANOVA, K.; ZEMAN, D. *Towards deep universal dependencies*. *In:* Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019). Paris, France, 2019.

DROGANOVA, K.; ZEMAN, D. *Towards a unified taxonomy of deep syntactic relations*. *In:* Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italy, 2024. p. 16412-16421.

FUTRELL, R.; MAHOWALD, K.; GIBSON, E. *Large-scale evidence of dependency length minimization in 37 languages*. *Proceedings of National Academy of Sciences*, v. 112, n. 33, 2015. p. 10336-10341.

GERDES, K.; GUILLAUME, B.; KAHANE, S.; PERRIER, G. *SUD or surface-syntactic universal dependencies: an annotation scheme near-isomorphic to UD*. *In:* Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). Bruxelles, Belgium, 2018.

GERDES, K.; KAHANE, S.; CHEN, X. *Typometrics: from implicational to quantitative universals in word order typology*. *Glossa: a journal of general linguistics*, v. 6, n. 1, p. 17, 2021.

HANA, J.; HLADKÁ, B. *Universal dependencies and non-native Czech*. *In:* Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018). Oslo, Norway, 2018.

JINDAL, I.; RADEMAKER, A.; ULEWICZ, M.; LINH, H.; NGUYEN, H.; TRAN, K.-N.; ZHU, H.; LI, Y. *Universal proposition bank 2.0*. *In:* Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France, 2022.

KYJÁNEK, L.; ŽABOKRTSKÝ, Z.; ŠEVČÍKOVÁ, M.; VIDRA, J. *Universal derivations 1.0, a growing collection of harmonised word-formation resources*. *The Prague Bulletin of Mathematical Linguistics*, v. 115, n. 2, 2020. p. 5-30.

LEE, J.; LI, K.; LEUNG, H. *L1-L2 parallel dependency treebank as learner corpus*. *In:* Proceedings of the 15th International Conference on Parsing Technologies. Pisa, Italy, 2017.

LEVSHINA, N. *Token-based typology and word order entropy: a study based on universal dependencies*. *Linguistic Typology*, v. 23, n. 3, p. 533-572, 2019.

MAYHEW, S.; BLEVINS, T.; LIU, S.; SUPPA, M.; GONEN, H.; IMPERIAL, J. M.; KARLSSON, B.; LIN, P.; LJUBEŠIĆ, N.; MIRANDA, L. J.; PLANK, B.; RIABI, A.; PINTER, Y. *Universal NER: a gold-standard multilingual named entity recognition benchmark*. *In:* Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. v. 1, n. Long Papers. Ciudad de México, Mexico, 2024.

MÜLLER-EBERSTEIN, M.; VAN DER GOOT, R.; PLANK, B. *How universal is genre in universal dependencies? In:* Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021). Sofia, Bulgaria, 2021.

NEDOLUZHKO, A.; NOVÁK, M.; POPEL, M.; ŽABOKRTSKÝ, Z.; ZELDES, A.; ZEMAN, D. *CorefUD 1.0: coreference meets universal dependencies*. *In:* Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France, 2022.

NIVRE, J.; DE MARNEFFE, M.-C.; GINTER, F.; HAJIČ, J.; MANNING, C. D.; PYYSALO, S.; SCHUSTER, S.; TYERS, F.; ZEMAN, D. *Universal dependencies v2: an evergrowing multilingual treebank collection*. *In:* Proceedings of the 12th International Conference on Language Resources and Evaluation. Marseille, France, 2020.

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. *Stanza: a Python natural language processing toolkit for many human languages*. *In:* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online, 2020.

SAVARY, A.; BEN KHELIL, C.; RAMISCH, C.; GIOULI, V.; BARBU MITITELU, V.; HADJ MOHAMED, N.; KRSTEV, C.; LIEBESKIND, C.; XU, H.; STYMNE, S.; GÜNGÖR, T.; PICKARD, T.; GUILLAUME, B.; BEJČEK, E.; BHATIA, A.; CANDITO, M.; GANTAR, P.; IÑURRIETA, U.; GATT, A.; KOVALEVSKAITE, J.; LICHTE, T.; LJUBEŠIĆ, N.; MONTI, J.; PARRA ESCARTÍN, C.; SHAMSFARD, M.; STOYANOVA, I.; VINCZE, V.; WALSH, A. *PARSEME corpus release 1.3*. *In:* Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023). Dubrovnik, Croatia, 2023.

SCHUSTER, S.; MANNING, C. D. *Enhanced English universal dependencies: an improved representation for natural language understanding tasks*. *In:* Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, 2016.

STRAKA, M.; HAJIČ, J.; STRAKOVÁ, J. *UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing*. *In:* Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, 2016.

SUNG, H.; SHIN, G.-H. *Constructing a dependency treebank for second language learners of Korean*. *In:* Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italy, 2024.

ZEMAN, D.; POPEL, M.; STRAKA, M.; HAJIČ, J.; NIVRE, J.; GINTER, F.; LUOTOLAHTI, J.; PYYSALO, S.; PETROV, S.; POTTHAST, M.; TYERS, F.; BADMAEVA, E.; GÖKIRMAK, M.; NEDOLUZHKO, A.; CINKOVÁ, S.; HAJIČ, J. JR.; HLAVÁČOVÁ, J.; KETTNEROVÁ, V.; UREŠOVÁ, Z.; KANERVA, J.; OJALA, S.; MISSILÄ, A.; MANNING, C.; SCHUSTER, S.; REDDY, S.; TAJI, D.; HABASH, N.; LEUNG, H.; DE MARNEFFE, M.-C.; SANGUINETTI, M.; SIMI, M.; KANAYAMA, H.; DE PAIVA, V.; DROGANOVA, K.; MARTÍNEZ ALONSO, H.; ÇÖLTEKIN, Ç.; SULUBACAK, U.; USZKOREIT, H.; MACKETANZ, V.; BURCHARDT, A.; HARRIS, K.; MARHEINECKE, K.; REHM, G.; KAYADELEN, T.; ATTIA, M.; ELKAHKY, A.; YU, Z.; PITLER, E.; LERTPRADIT, S.; MANDL, M.; KIRCHNER, J.; FERNANDEZ ALCALDE, H.; STRNADOVÁ, J.; BANERJEE, E.; MANURUNG, R.; STELLA, A.; SHIMADA, A.; KWAK, S.; MENDONÇA, G.; LANDO, T.; NITISAROJ, R.; LI, J. *CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies*. *In:* Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada, 2017. p. 1-19.