UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, Alina Wróblewska

LISN, Paris-Saclay University, CNRS, France; ÚFAL MFF, Charles University, Czechia;
Romanian Academy Research Institute for Artificial Intelligence, Romania;
INESC-ID Lisboa, Portugal; Moldova State University, Vladimir Andrunachievici
Institute of Mathematics and Computer Science, Moldova;
FNRS, Université catholique de Louvain, Belgium;
University of Ljubljana and JSI, Slovenia; Istanbul Technical University, Türkiye;
Aristotle University of Thessaloniki and ILSP, ATHENA RC, Greece;
LORIA, Inria, France; ILSP and Archimedes Unit, ATHENA RC, Greece;
Open University of Israel; Uppsala University and RISE, Sweden;
University of Galway, Ireland; LIS, Aix-Marseilles University, CNRS, France;
ADAPT Centre, DCU, Ireland; University of Warsaw, Poland; ICS PAS, Poland
Corresponding author: agata.savary@universite-paris-saclay.fr

Abstract

This paper presents the objectives, organization and activities of the UniDive COST Action, a scientific network dedicated to universality, diversity and idiosyncrasy in language technology. We describe the objectives and organization of this initiative, the people involved, the working groups and the ongoing tasks and activities. This paper is also an open call for participation towards new members and countries.

Keywords: universality, diversity, idiosyncrasy, language technology, scientific network

1. Introduction

Natural language processing (NLP) is currently booming, to the benefit of many end users. However, this technological progress poses an important challenge: accounting for and fostering language diversity. We present UniDive, an initiative which takes two original stands on this challenge. Firstly, it addresses both inter- and intra-language diversity, i.e., diversity understood both in terms of the differences among the existing languages and among the linguistic phenomena exhibited within a language. Phenomena currently under study are: morphological features, syntactic dependencies, multiword expressions and other idiosyncratic constructions, as well as word formation processes and their links with the notion of "wordhood". Secondly, UniDive does not assume that linguistic diversity is to be protected against technological progress but strives for reconciling both of these aims. Its approach is to: (i) pursue NLP-applicable universality of terminologies and methodologies, (ii) quantify inter- and intra-linguistic diversity, (iii) boost and coordinate universality- and diversity-driven development of language resources and tools, for a large variety of linguistic phenomena in a large number of languages, including low-resourced ones.

UniDive is a COST Action[1], i.e. a scientific network funded (for 2022-2026) by the European Union via COST (European Cooperation in Science and Technology). COST Actions connect researchers, from Europe and beyond, via networking instruments such as meetings, conferences, workshops, short-term scientific missions and training schools. UniDive is open to new members throughout its entire duration.

2. State of the Art

The three foundational concepts for UniDive are diversity, universality and idiosyncrasy.

2.1. Universality

The study of language universals has a long-standing tradition (Greenberg, 1996; Chomsky,

[1] For the COST-hosted portal of UniDive see https://www.cost.eu/actions/CA21167/.

1975), prevails in mainstream theoretical linguistics and is a central issue in typology. But the existence of absolute universals is a subject of a major controversy. Evans and Levinson (2009) claim that the existence of a Universal Grammar is a myth, that statistical tendencies ("statistical universals") should be considered instead and that linguistic research should use diversity as a starting point. Others argue that diversity is a surface phenomenon, while universality, conversely, can be captured at the right level of abstractness (Tallerman, 2009). In NLP, researchers are more agnostic towards the theoretical status of language universals, rather emphasizing the usefulness of cross-linguistically consistent and applicable language descriptions. The objective of defining such descriptions is referred to in UniDive as *universality*.

Universality holds a pivotal role in NLP and its practical realization has facilitated the expeditious advancement of this discipline. Widely acknowledged presumptions of universality serve as the foundation for open and cooperative NLP initiatives. UniDive directly builds upon three of them: Universal Dependencies (de Marneffe et al., 2021), which posits standarized guidelines for morphosyntactic annotation in treebanks, PARSEME (Savary et al., 2023a), which advocates for unified directives concerning the annotation of multiword expressions (Sec. 2.3), and UniMorph (Kirov et al., 2018), which proposes universal guidance on annotating morphological properties in inflectional languages. Inspired by these well-established endeavors, new ones emerge, e.g. CorefUD (Nedoluzhko et al., 2022), which establishes a standardized format for coreference resolution, and Universal Anaphora (Poesio et al., 2023), which promotes cross-linguistically universal anaphoric interpretation.

The importance of these universality-driven initiatives is multifaceted. By sharing datasets that are annotated consistently and uniformly across multiple languages, they enable cross-linguistic comparative research and the development of robust and versatile NLP models. By providing a unified foundation for linguistic annotation, they promote shared linguistic understanding. Last but not least, they highlight the importance of linguistic diversity and the need for inclusive approaches in NLP research.

2.2. Diversity

Diversity has been modelled and measured in many domains, such as as ecology, economy or information theory (Morales et al., 2021). There, formal definitions of diversity often rely on the notions of *items* and *types*. In ecology, *items* are specimens/individuals, while *types* refer to the species these specimens are affiliated to. Given a population of items clustered into types, the concept of diversity is often defined along three distinct dimensions: *variety*, *balance* and *disparity* (Stirling, 1998). *Variety* is the number of types into which items can be classified (sometimes normalized by the number of items). *Balance* is the extent to which the type-item distribution is uniform. *Disparity* is the degree to which types differ from each other, according to a distance metric defined on types.

In linguistics, diversity was mainly addressed in the interlingual sense, e.g. in terms of languages spoken in a given geographical area, different lineages in the phylogenetic tree of languages, or variation among structures within languages (Nettle, 1999), as well as the rate of language extinction (Harmon and Loh, 2010).

In NLP, a growing body of works addresses the need for language technology to cover a larger number of world's languages (Joshi et al., 2020; ImaniGooghari et al., 2023). Some other works stress the need for intra-lingual diversity in training data and its impact on performances in parsing (Narayan and Cohen, 2015), question answering (Yang et al., 2018) and natural language generation (Zhang et al., 2020; Agirre et al., 2016; Zhu et al., 2018; Palumbo et al., 2020; Li et al., 2021; Tevet and Berant, 2021). Lion-Bouton et al. (2022) quantify the intra-linguistic diversity (in terms of variety and balance) of one particular linguistic phenomenon: multiword expressions, which are outstanding representatives of idiosyncrasy, the third major concept addressed by UniDive.

2.3. Idiosyncrasy

Human languages present recurrent patterns that allow humans and computers to deduce generic rules and generalizations from examples. Idiosyncrasy occurs when these patterns are breached, that is, when only a few instances of a larger class present a given characteristic or behaviour. This abstract notion can be applied to any level of linguistic analysis (word senses, syntactic constructions, phonemes, etc.), but in UniDive we focus on idiosyncratic word combinations. Most of the time, these elements are words, and the combinations are called *multiword expressions* (MWEs) (Baldwin and Kim, 2010). When the elements under consideration are under-specified, we speak of *constructions*, in the sense of Construction Grammar (Fillmore et al., 1988; Goldberg, 1995).

The state of the art in MWE modeling encompasses a large body of works. In UniDive, we are notably concerned with MWE lexicons (Losnegaard et al., 2016) and corpora annotated with MWEs (Schneider et al., 2016; Savary et al., 2023a). Of special interest for UniDive is unifying divergent MWE modeling practices in universality-

driven initiatives (Kahane et al., 2017; Savary et al., 2023b) and designing MWE lexicon-corpus interfaces.

In MWE processing, the major tasks include MWE discovery, identification and translation (Constant et al., 2017), as well as semantic compositionality prediction (Cordeiro et al., 2019). One of the challenges lies in the severe difficulty of generalizing beyond the data seen in training (Ramisch et al., 2020). In more generic NLP tasks, recent MWE-related challenges include evaluating neural machine translation (Baziotis et al., 2023), capturing semantic similarity (Tayyar Madabushi et al., 2022) and understanding the behavior of transformer-based language models (Haviv et al., 2023) while explicitly focusing on MWEs.

3. Objectives and Organization

UniDive's main objective is to reconcile language diversity with rapid progress in language technology. To achieve these goals, the Action is focusing on two general efforts: *research coordination* and *capacity building*.

Research coordination objectives include: (i) developing methods for quantifying linguistic diversity, (ii) reaching a common understanding of language universals, (iii) coordinating diversity-driven developments of language resources and NLP tools, (iv) raising awareness regarding the importance of diversity preservation in language technology, and (v) disseminating the outcomes to stakeholders.

Capacity building objectives include: (i) creating a network of experts in a large number of languages working on modelling and processing linguistic phenomena within a common framework, (ii) fostering the capacities of young researchers, (iii) setting up a long-term roadmap for the joint efforts of the universality-driven NLP community.

To achieve its goals, UniDive employs instruments that aim to bring the research community together. Semi-annual management committee (MC) meetings, monthly working group (WG) meetings and meetings of various task groups are held online and provide Action members with the opportunity to discuss research and address managerial issues. Annual in-person general meetings include talks by invited speakers and a workshop where Action members and non-members present peer-reviewed work on the Action's topics. Training events, held annually, either online or in-person, focus on topics that are central to the Action's activities and are especially beneficial to young researchers. In addition, the Action funds short-term scientific missions (STSMs) which enable members to visit institutions located in a country other than their country of affiliation and take advantage

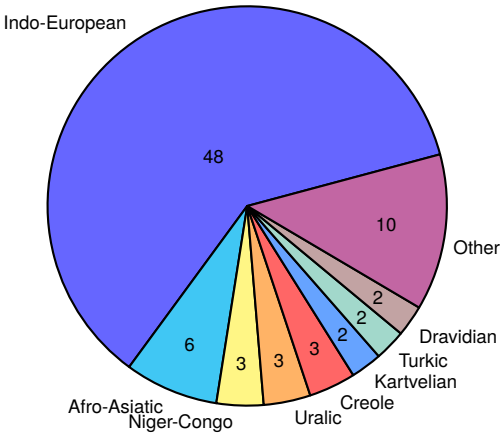

Figure 1: Number of languages in UniDive per language family. *Other* comprises Sumerian, Mongolic, Korean, Sino-Tibetan, Austro-Asiatic, Austronesian, Pama-Nyungan, Uto-Aztecan, Mayan, and Constructed languages.

of knowledge not available in their own institutions. STSMs contribute to the scientific objectives of the Action and foster collaboration between participants.

Within a large network like this, efficient communication is needed to share thoughts, ideas, opinions, feedback on research and administration issues. In addition to mailing lists covering various groups and committees, UniDive uses Telegram, selected on the basis of a preference survey, for instantaneous communication. For external communication, we rely on UniDive's website[2], social media platforms, and collaborative platforms for online documentation and meetings.

4. People

Formally, a COST Action consists of countries that send their representatives to the MC. But in practice, obviously, the work is done by *people* who enter one or more WGs; this community reaches far beyond the MC membership. The Action remains open to newcomers throughout its duration.[3]

COST Actions put a lot of weight on balanced representation w.r.t. gender, age, and geography. The latter means that certain countries, mostly from the Eastern half of Europe, are designated 'Inclusiveness Target Countries' (ITC)[4] and a balance between ITC and non-ITC is sought (since historically, researchers from ITCs were underrepresented at international events).

[2] https://unidive.lisn.upsaclay.fr/

[3] See: https://unidive.lisn.upsaclay.fr/doku.php?id=how_to_join_us.

[4] https://www.cost.eu/about/members/

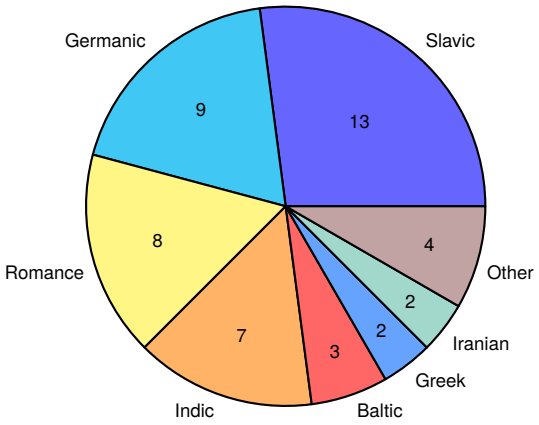

Figure 2: Number of languages in UniDive per Indo-European genus. *Other* comprises Celtic, Italic, Albanian, and Armenian.

At the time of writing, UniDive comprises 37 countries (out of all 43 COST Members, Cooperating and Partner Members); 24 of those are ITCs. The WGs have 330 participants in total (many of them registered in multiple WGs). 58% are female, 42% are young researchers by COST criteria, and 49% are based in ITC.

Given the goals of UniDive, an important factor is also the range of natural languages in which the participants are proficient. We conducted several surveys where we asked members about their native language, the languages they work on and other languages they have expertise in. Not surprisingly, the vast majority of members listed Indo-European languages; nevertheless, there are also languages from 17 other families (Figures 1 and 2). In total, 77 languages were mentioned individually but some members work on language groups and some stated directly that their work is multilingual, not restricted to any particular language or group.

5. Activities

The scientific activity in UniDive is structured in 4 working groups (WGs).

WG1 Corpus Annotation: WG1 focuses on the annotation aspects of corpora development, as annotated corpora constitute one of the Action's fundamental operational tool for fostering and analyzing NLP-applied universality. Current activities are centered around Universal Dependencies (UD) and PARSEME (Sec. 2.1), whose latest corpus releases are 2.13 (Zeman et al., 2023) and 1.3 Savary et al. (2023), respectively. The main aim of WG1 is to maintain and extend this momentum towards large-scale high-quality multilingual linguistic annotation. Diversity is under-represented in the existing universality-driven projects and WG1 aims to support the development of annotated resources for new languages. Another aim is to unify and enhance cross-lingual annotation guidelines for morpho-syntax and MWEs, by also accounting for language typology at various levels of linguistic description. Work is also planned on tools, file formats and related infrastructure supporting corpus development.

WG2 Lexicon-corpus interface: In the quest for diversity, electronic lexicons are complementary to corpora. While the former aim at holistic language modelling, describing possibly many linguistic objects, in the latter many phenomena are rare. In this context, WG2 carries out a survey about segmentation conventions in different UD treebanks and how they coincide with Haspelmath's (2023) definition of a "word". The outcomes will help spot and illustrate segmentation inconsistencies in UD and formulate recommendations for future annotation projects. WG2 also focuses on adding new languages to the ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al., 2021). Provided that an open license sense inventory (a dictionary) is available, any language can join this task of linking words (including MWEs) in the corpus with senses from the dictionary. Finally, WG2 is carrying out a survey on MWE lexicons which would update the previous effort by (Losnegaard et al., 2016), in an attempt to define a proof-of-concept for lexical encoding of idiosyncratic properties in MWEs, with an eye to lexicon-corpus interlinking mentioned above.

WG3 Multilingual and cross-lingual language technology: The work in WG3 is concerned with multilingual and cross-lingual NLP tools, including but not limited to tools for morphosyntactic and semantic analysis, and for discovery and identification of MWEs. The first ongoing effort focuses on documentation, so as to provide easy access to tools that apply to multiple languages, in particular low-resourced ones, notably through cross-lingual learning. The second current focus is on organizing multilingual evaluation campaigns which would shed new light on how existing language technology tools, ranging from traditional syntactic and semantic analysers to large language models, deal with universality, diversity and idiosyncracy within and across languages. This activity will be informed by the work of WG4 on metrics for intra- and inter-language diversity.

WG4 Quantifying and promoting diversity: The work in WG4 is transversal to the other working groups, aiming at an actionable definition of diversity. The main goal is to propose metrics for intra- and inter-language diversity in resources and tools. Such metrics will be used to (i) assess how diverse multilingual shared-tasks/resources are in terms of spanning a large variety of languages

and language phenomena, (ii) favor tools performing well on rare and diverse phenomena and on low-resourced languages (instead of only reporting scores such as F1, a diversity score would also rank systems submitted to multilingual shared-tasks). To achieve such goals, WG4 will use one of the forces of COST actions: networking. By integrating pre-existing groups dedicated to NLP-applicable universality, with experts of notably low-resourced languages and typologists, WG4 is aiming at promoting diversity in NLP. So far, the effort has focused on documenting existing measures of diversity and collecting multilingual shared-tasks data to test the metrics WG4 will come up with.

6. Conclusions

Despite the apparent contradictions between the notions of universality, diversity and idiosyncrasy, they can in fact be seen as complementary. Universality promotes diversity via inclusiveness. Idiosyncrasy, understood as linguistic behaviors deviating from universals across languages and/or strong generalisations in a language, necessarily contributes to diversity. Finally, what is seen as idiosyncratic in one language can be studied as a potential generalisation across a number of languages or, even, as a universal. UniDive has a huge potential to collectively leverage this complementary nature and thus contribute to reconciling language diversity with rapid progress in language technology.

7. Acknowledgments

This paper is funded by the CA21167 COST Action UniDive, supported by COST (European Cooperation in Science and Technology). The work was also supported in part by Israeli Ministry of Science and Technology grant No. 0002336 (Nurit Melnik, PI),

8. Bibliographical References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of natural language processing*, volume 2, pages 267–292. CRC Press, Boca Raton, USA.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Noam Chomsky. 1975. *Reflections on Language*. Temple Smith, London.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

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

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Charles J Fillmore, Paul Kay, and Mary Catherine O'connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Joseph H. Greenberg, editor. 1996. *Universals of language*. MIT Press.

David Harmon and Jonathan Loh. 2010. The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language Documentation and Conservation*, 4.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall

through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - Propositions for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 181–189, Prague, Czech Republic.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Junyi Li, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2021. TextBox: A unified, modularized, and extensible framework for text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 30–39, Online. Association for Computational Linguistics.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia. European Language Resources Association (ELRA).

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. 2021. Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Electronic lexicography in the 21st century: post-editing lexicography*, pages 377–395, Brno.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115.

Shashi Narayan and Shay B. Cohen. 2015. Diversity in spectral learning for natural language parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1868–1878, Lisbon, Portugal. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Daniel Nettle. 1999. *Linguistic diversity*. Oxford University Press, Oxford.

Enrico Palumbo, Andrea Mezzalira, Cristina Marco, Alessandro Manzotti, and Daniele Amberti. 2020. Semantic diversity for natural language understanding evaluation in dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 44–49, Online. International Committee on Computational Linguistics.

Massimo Poesio, Amir Zeldes, Anna Nedoluzhko, Sopan Khosla, Ramesh Manuvinakurike, Nafise Moosavi, Vincent Ng, Maciej Ogrodniczuk, Sameer Pradhan, Carolyn Rose, Michael Strube, Juntao Yu, Yulia Grishina, Yufang Hou, and Fred Landragin. 2023. Universal Anaphora 1.0 – Proposal for Discussion. Work in progress.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023a. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023b. Parseme meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Andrew Stirling. 1998. On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper*, 28:1–156.

Maggie Tallerman. 2009. If language is a jungle, why are we all cultivating the same plot? *Behavioral and Brain Sciences*, 32:469 – 470.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020. Trading off diversity and quality in natural language generation. *arXiv preprint arXiv:2004.10450*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

9. Language Resource References

Savary, Agata and Ramisch, Carlos and Guillaume, Bruno and Hawwari, Abdelati and Walsh, Abigail and Fotopoulou, Aggeliki and Bielinskienė, Agnė and Estarrona, Ainara and Gatt, Albert and Butler, Alexandra and Rademaker, Alexandre and Maldonado, Alfredo and Villavicencio, Aline and Farrugia, Alison and Muscat, Amanda and Gatt, Anabelle and Antić, Anđela and De Santis, Anna and Raffone, Annalisa and Riccio, Anna and Pascucci, Antonio and Gurrutxaga, Antton and Bhatia, Archna and Vaidya,

Ashwini and Miral, Ayşenur and QasemiZadeh, Behrang and Priego Sanchez, Belem and Griciūtė, Bernadeta and Erden, Berna and Parra Escartín, Carla and Herrero, Carlos and Carlino, Carola and Pasquer, Caroline and Liebeskind, Chaya and Wang, Chenweng and Ben Khelil, Chérifa and Bonial, Claire and Somers, Clarissa and Aceta, Cristina and Krstev, Cvetana and Bejček, Eduard and Lindqvist, Ellinor and Erenmalm, Elsa and Palka-Binkiewicz, Emilia and Rimkute, Erika and Petterson, Eva and Cap, Fabienne and Hu, Fangyuan and Sangati, Federico and Wick Pedro, Gabriela and Speranza, Giulia and Jagfeld, Glorianna and Blagus, Goranka and Berk, Gözde and Attard, Greta and Eryiğit, Gülşen and Finnveden, Gustav and Martínez Alonso, Héctor and de Medeiros Caseli, Helena and Elyovich, Hevi and Xu, Hongzhi and Xiao, Huangyang and Miranda, Isaac and Jaknić, Isidora and El Maarouf, Ismail and Aduriz, Itziar and Gonzalez, Itziar and Matas, Ivana and Stoyanova, Ivelina and Jazbec, Ivo-Pavao and Busuttil, Jael and Waszczuk, Jakub and Findlay, Jamie and Bonnici, Janice and Šnajder, Jan and Antoine, Jean-Yves and Foster, Jennifer and Chen, Jia and Nivre, Joakim and Monti, Johanna and McCrae, John and Kovalevskaitė, Jolanta and Jain, Kanishka and Simkó, Katalin and Yu, Ke and Azzopardi, Kirsty and Adalı, Kübra and Uria, Larraitz and Zilio, Leonardo and Boizou, Loïc and van der Plas, Lonneke and Galea, Luke and Sarlak, Mahtab and Buljan, Maja and Cherchi, Manuela and Tanti, Marc and Di Buono, Maria Pia and Todorova, Maria and Candito, Marie and Constant, Matthieu and Shamsfard, Mehrnoush and Jiang, Menghan and Boz, Mert and Spagnol, Michael and Onofrei, Mihaela and Li, Minli and Elbadrashiny, Mohamed and Diab, Mona and Rizea, Monica-Mihaela and Hadj Mohamed, Najet and Theoxari, Natasa and Schneider, Nathan and Tabone, Nicole and Ljubešić, Nikola and Vale, Oto and Cook, Paul and Yan, Peiyi and Gantar, Polona and Ehren, Rafael and Fabri, Ray and Ibrahim, Rehab and Ramisch, Renata and Walles, Rinat and Wilkens, Rodrigo and Urizar, Ruben and Sun, Ruilong and Malka, Ruth and Galea, Sara Anne and Stymne, Sara and Louizou, Sevasti and Hu, Sha and Taslimipoor, Shiva and Ratori, Shraddha and Srivastava, Shubham and Cordeiro, Silvio Ricardo and Krek, Simon and Liu, Siyuan and Zeng, Si and Yu, Songping and Arhar Holdt, Špela and Markantonatou, Stella and Papadelli, Stella and Leseva, Svetlozara and Kuzman, Taja and Kavčič, Teja and Lynn, Teresa and Lichte, Timm and Pickard, Thomas and Dimitrova, Tsvetana and Yih, Tsy and Güngör, Tunga and Dinç, Tutkum and Iñurrieta, Uxoa and Tajalli, Vahide and Stefanova, Valentina and Caruso, Valeria and Puri, Vandana and Foufi, Vassiliki and Barbu Mititelu, Verginica and Vincze, Veronika and Kovács, Viktória and Shukla, Vishakha and Giouli, Voula and Ge, Xiaomin and Ha-Cohen Kerner, Yaakov and Öztürk, Yağmur and Yarandi, Yalda and Parmentier, Yannick and Zhang, Yongchen and Zhao, Yun and Urešová, Zdeňka and Yirmibeşoğlu, Zeynep and Qin, Zhenzhen and Stank and Cristescu, Mihaela and Zgreabăn, Bianca-Mădălina and Bărbulescu, Elena-Andreea and Stanković, Ranka. 2023. *PARSEME corpora annotated for verbal multiword expressions (version 1.3)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell and Ackermann, Elia and Aepli, Noëmi and Aghaei, Hamid and Agić, Željko and Ahmadi, Amir and Ahrenberg, Lars and Ajede, Chika Kennedy and Akkurt, Salih Furkan and Aleksandravičiūtė, Gabrielė and Alfina, Ika and Algom, Avner and Alnajjar, Khalid and Alzetta, Chiara and Andersen, Erik and Antonsen, Lene and Aoyama, Tatsuya and Aplonova, Katya and Aquino, Angelina and Aragon, Carolina and Aranes, Glyd and Aranzabe, Maria Jesus and Arıcan, Bilge Nas and Arnardóttir, Þórunn and Arutie, Gashaw and Arwidarasti, Jessica Naraiswari and Asahara, Masayuki and Ásgeirsdóttir, Katla and Aslan, Deniz Baran and Asmazoğlu, Cengiz and Ateyah, Luma and Atmaca, Furkan and Attia, Mohammed and Atutxa, Aitziber and Augustinus, Liesbeth and Avelãs, Mariana and Badmaeva, Elena and Balasubramani, Keerthana and Ballesteros, Miguel and Banerjee, Esha and Bank, Sebastian and Barbu Mititelu, Verginica and Barkarson, Starkaður and Basile, Rodolfo and Basmov, Victoria and Batchelor, Colin and Bauer, John and Bedir, Seyyit Talha and Behzad, Shabnam and Belieni, Juan and Bengoetxea, Kepa and Benli, İbrahim and Ben Moshe, Yifat and Berk, Gözde and Bhat, Riyaz Ahmad and Biagetti, Erica and Bick, Eckhard and Bielinskienė, Agnė and Bjarnadóttir, Kristín and Blokland, Rogier and Bobicev, Victoria and Boizou, Loïc and Borges Völker, Emanuel and Börstell, Carl and Bosco, Cristina and Bouma, Gosse and Bowman, Sam and Boyd, Adriane and Braggaar, Anouck and Branco, António and Brokaitė, Kristina and Burchardt, Aljoscha and Campos,

Marisa and Candito, Marie and Caron, Bernard and Caron, Gauthier and Carvalheiro, Catarina and Carvalho, Rita and Cassidy, Lauren and Castro, Maria Clara and Castro, Sérgio and Cavalcanti, Tatiana and Cebiroğlu Eryiğit, Gülşen and Cecchini, Flavio Massimiliano and Celano, Giuseppe G. A. and Čéplö, Slavomír and Cesur, Neslihan and Cetin, Savas and Çetinoğlu, Özlem and Chalub, Fabricio and Chamila, Liyanage and Chauhan, Shweta and Chi, Ethan and Chika, Taishi and Cho, Yongseok and Choi, Jinho and Chun, Jayeol and Chung, Juyeon and Cignarella, Alessandra T. and Cinková, Silvie and Collomb, Aurélie and Çöltekin, Çağrı and Connor, Miriam and Corbetta, Claudia and Corbetta, Daniela and Costa, Francisco and Courtin, Marine and Crabbé, Benoît and Cristescu, Mihaela and Cvetkoski, Vladimir and Dale, Ingerid Løyning and Daniel, Philemon and Davidson, Elizabeth and de Alencar, Leonel Figueiredo and Dehouck, Mathieu and de Laurentiis, Martina and de Marneffe, Marie-Catherine and de Paiva, Valeria and Derin, Mehmet Oguz and de Souza, Elvis and Diaz de Ilarraza, Arantza and Dickerson, Carly and Dinakaramani, Arawinda and Di Nuovo, Elisa and Dione, Bamba and Dirix, Peter and Dobrovoljc, Kaja and Doyle, Adrian and Dozat, Timothy and Droganova, Kira and Duran, Magali Sanches and Dwivedi, Puneet and Ebert, Christian and Eckhoff, Hanne and Eguchi, Masaki and Eiche, Sandra and Eli, Marhaba and Elkahky, Ali and Ephrem, Binyam and Erina, Olga and Erjavec, Tomaž and Essaidi, Farah and Etienne, Aline and Evelyn, Wograine and Facundes, Sidney and Farkas, Richárd and Favero, Federica and Ferdaousi, Jannatul and Fernanda, Marília and Fernandez Alcalde, Hector and Fethi, Amal and Foster, Jennifer and Fransen, Theodorus and Freitas, Cláudia and Fujita, Kazunori and Gajdošová, Katarína and Galbraith, Daniel and Gamba, Federica and Garcia, Marcos and Gärdenfors, Moa and Gerardi, Fabrício Ferraz and Gerdes, Kim and Gessler, Luke and Ginter, Filip and Godoy, Gustavo and Goenaga, Iakes and Gojenola, Koldo and Gökırmak, Memduh and Goldberg, Yoav and Gómez Guinovart, Xavier and González Saavedra, Berta and Griciūtė, Bernadeta and Grioni, Matias and Grobol, Loïc and Grūzītis, Normunds and Guillaume, Bruno and Guiller, Kirian and Guillot-Barbance, Céline and Güngör, Tunga and Habash, Nizar and Hafsteinsson, Hinrik and Hajič, Jan and Hajič jr., Jan and Hämäläinen, Mika and Hà Mỹ, Linh and Han, Na-Rae and Hanifmuti, Muhammad Yudistira and Harada, Takahiro and Hardwick, Sam and Harris, Kim and Haug, Dag and Heinecke, Johannes and Hellwig, Oliver and Hennig, Felix and Hladká, Barbora and Hlaváčová, Jaroslava and Hociung, Florinel and Hohle, Petter and Huang, Yidi and Huerta Mendez, Marivel and Hwang, Jena and Ikeda, Takumi and Ingason, Anton Karl and Ion, Radu and Irimia, Elena and Ishola, Ọlájídé and Islamaj, Artan and Ito, Kaoru and Jagodzińska, Sandra and Jannat, Siratun and Jelínek, Tomáš and Jha, Apoorva and Jiang, Katharine and Johannsen, Anders and Jónsdóttir, Hildur and Jørgensen, Fredrik and Juutinen, Markus and Kaşıkara, Hüner and Kabaeva, Nadezhda and Kahane, Sylvain and Kanayama, Hiroshi and Kanerva, Jenna and Kara, Neslihan and Karahóǧa, Ritván and Kåsen, Andre and Kayadelen, Tolga and Kengatharaiyer, Sarveswaran and Kettnerová, Václava and Kharatyan, Lilit and Kirchner, Jesse and Klementieva, Elena and Klyachko, Elena and Kocharov, Petr and Köhn, Arne and Köksal, Abdullatif and Kopacewicz, Kamil and Korkiakangas, Timo and Köse, Mehmet and Koshevoy, Alexey and Kotsyba, Natalia and Kovalevskaitė, Jolanta and Krek, Simon and Krishnamurthy, Parameswari and Kübler, Sandra and Kuqi, Adrian and Kuyrukçu, Oğuzhan and Kuzgun, Aslı and Kwak, Sookyoung and Kyle, Kris and Laan, Käbi and Laippala, Veronika and Lambertino, Lorenzo and Lando, Tatiana and Larasati, Septina Dian and Lavrentiev, Alexei and Lee, John and Lê Hồng, Phương and Lenci, Alessandro and Lertpradit, Saran and Leung, Herman and Levina, Maria and Levine, Lauren and Li, Cheuk Ying and Li, Josie and Li, Keying and Li, Yixuan and Li, Yuan and Lim, KyungTae and Lima Padovani, Bruna and Lin, Yi-Ju Jessica and Lindén, Krister and Liu, Yang Janet and Ljubešić, Nikola and Lobzhanidze, Irina and Loginova, Olga and Lopes, Lucelene and Lusito, Stefano and Luthfi, Andry and Luukko, Mikko and Lyashevskaya, Olga and Lynn, Teresa and Macketanz, Vivien and Mahamdi, Menel and Maillard, Jean and Makarchuk, Ilya and Makazhanov, Aibek and Mandl, Michael and Manning, Christopher and Manurung, Ruli and Marşan, Büşra and Mărănduc, Cătălina and Mareček, David and Marheinecke, Katrin and Markantonatou, Stella and Martínez Alonso, Héctor and Martín Rodríguez, Lorena and Martins, André and Martins, Cláudia and Mašek, Jan and Matsuda, Hiroshi and Matsumoto, Yuji and Mazzei, Alessandro and McDonald, Ryan and McGuinness, Sarah and Mendonça, Gustavo and Merzhevich, Tatiana and Miekka, Niko and Miller, Aaron and Mischenkova, Karina and Missilä, Anna and Mititelu, Cătălin and Mitrofan, Maria and Miyao, Yusuke and Mojiri Foroushani, AmirHossein and Molnár, Judit and Moloodi, Amirsaeid and Montemagni,

Simonetta and More, Amir and Moreno Romero, Laura and Moretti, Giovanni and Mori, Shinsuke and Morioka, Tomohiko and Moro, Shigeki and Mortensen, Bjartur and Moskalevskyi, Bohdan and Muischnek, Kadri and Munro, Robert and Murawaki, Yugo and Müürisep, Kaili and Nainwani, Pinkey and Nakhlé, Mariam and Navarro Horñiacek, Juan Ignacio and Nedoluzhko, Anna and Nešpore-Bērzkalne, Gunta and Nevaci, Manuela and Nguyễn Thị, Lương and Nguyễn Thị Minh, Huyền and Nikaido, Yoshihiro and Nikolaev, Vitaly and Nitisaroj, Rattima and Nourian, Alireza and Nunes, Maria das Graças Volpe and Nurmi, Hanna and Ojala, Stina and Ojha, Atul Kr. and Óladóttir, Hulda and Olúòkun, Adédayọ̀ and Omura, Mai and Onwuegbuzia, Emeka and Ordan, Noam and Osenova, Petya and Östling, Robert and Øvrelid, Lilja and Özateş, Şaziye Betül and Özçelik, Merve and Özgür, Arzucan and Öztürk Başaran, Balkız and Paccosi, Teresa and Palmero Aprosio, Alessio and Panova, Anastasia and Pardo, Thiago Alexandre Salgueiro and Park, Hyunji Hayley and Partanen, Niko and Pascual, Elena and Passarotti, Marco and Patejuk, Agnieszka and Paulino-Passos, Guilherme and Pedonese, Giulia and Peljak-Łapińska, Angelika and Peng, Siyao and Peng, Siyao Logan and Pereira, Rita and Pereira, Sílvia and Perez, Cenel-Augusto and Perkova, Natalia and Perrier, Guy and Petrov, Slav and Petrova, Daria and Peverelli, Andrea and Phelan, Jason and Pierre-Louis, Claudel and Piitulainen, Jussi and Pinter, Yuval and Pinto, Clara and Pintucci, Rodrigo and Pirinen, Tommi A and Pitler, Emily and Plamada, Magdalena and Plank, Barbara and Poibeau, Thierry and Ponomareva, Larisa and Popel, Martin and Pretkalniņa, Lauma and Prévost, Sophie and Prokopidis, Prokopis and Przepiórkowski, Adam and Pugh, Robert and Puolakainen, Tiina and Pyysalo, Sampo and Qi, Peng and Querido, Andreia and Rääbis, Andriela and Rademaker, Alexandre and Rahoman, Mizanur and Rama, Taraka and Ramasamy, Loganathan and Ramisch, Carlos and Ramos, Joana and Rashel, Fam and Rasooli, Mohammad Sadegh and Ravishankar, Vinit and Real, Livy and Rebeja, Petru and Reddy, Siva and Regnault, Mathilde and Rehm, Georg and Riabi, Arij and Riabov, Ivan and Rießler, Michael and Rimkutė, Erika and Rinaldi, Larissa and Rituma, Laura and Rizqiyah, Putri and Rocha, Luisa and Rögnvaldsson, Eiríkur and Roksandic, Ivan and Romanenko, Mykhailo and Rosa, Rudolf and Roșca, Valentin and Rovati, Davide and Rozonoyer, Ben and Rudina, Olga and Rueter, Jack and Rúnarsson, Kristján and Sadde, Shoval and Safari, Pegah and Sahala, Aleksi and Saleh, Shadi and Salomoni, Alessio and Samardžić, Tanja and Samson, Stephanie and Sanguinetti, Manuela and Sanıyar, Ezgi and Särg, Dage and Sartor, Marta and Sasaki, Mitsuya and Saulīte, Baiba and Savary, Agata and Sawanakunanon, Yanin and Saxena, Shefali and Scannell, Kevin and Scarlata, Salvatore and Schang, Emmanuel and Schneider, Nathan and Schuster, Sebastian and Schwartz, Lane and Seddah, Djamé and Seeker, Wolfgang and Seraji, Mojgan and Shahzadi, Syeda and Shen, Mo and Shimada, Atsuko and Shirasu, Hiroyuki and Shishkina, Yana and Shohibussirri, Muh and Shvedova, Maria and Siewert, Janine and Sigurðsson, Einar Freyr and Silva, João and Silveira, Aline and Silveira, Natalia and Silveira, Sara and Simi, Maria and Simionescu, Radu and Simkó, Katalin and Šimková, Mária and Símonarson, Haukur Barri and Simov, Kiril and Sitchinava, Dmitri and Sither, Ted and Skachedubova, Maria and Smith, Aaron and Soares-Bastos, Isabela and Solberg, Per Erik and Sonnenhauser, Barbara and Sourov, Shafi and Sprugnoli, Rachele and Stamou, Vivian and Steingrímsson, Steinþór and Stella, Antonio and Stephen, Abishek and Straka, Milan and Strickland, Emmett and Strnadová, Jana and Suhr, Alane and Sulestio, Yogi Lesmana and Sulubacak, Umut and Suzuki, Shingo and Swanson, Daniel and Szántó, Zsolt and Taguchi, Chihiro and Taji, Dima and Tamburini, Fabio and Tan, Mary Ann C. and Tanaka, Takaaki and Tanaya, Dipta and Tavoni, Mirko and Tella, Samson and Tellier, Isabelle and Testori, Marinella and Thomas, Guillaume and Tonelli, Sara and Torga, Liisi and Toska, Marsida and Trosterud, Trond and Trukhina, Anna and Tsarfaty, Reut and Türk, Utku and Tyers, Francis and Þórðarson, Sveinbjörn and Þorsteinsson, Vilhjálmur and Uematsu, Sumire and Untilov, Roman and Urešová, Zdeňka and Uria, Larraitz and Uszkoreit, Hans and Utka, Andrius and Vagnoni, Elena and Vajjala, Sowmya and Vak, Socrates and van der Goot, Rob and Vanhove, Martine and van Niekerk, Daniel and van Noord, Gertjan and Varga, Viktor and Vedenina, Uliana and Venturi, Giulia and Villemonte de la Clergerie, Eric and Vincze, Veronika and Vlasova, Natalia and Wakasa, Aya and Wallenberg, Joel C. and Wallin, Lars and Walsh, Abigail and Washington, Jonathan North and Wendt, Maximilan and Widmer, Paul and Wigderson, Shira and Wijono, Sri Hartati and Wille, Vanessa Berwanger and Williams, Seyi and Wirén, Mats and Wittern, Christian and Woldemariam, Tsegay and Wong, Tak-sum and Wróblewska, Alina and Wu, Qishen and Yako, Mary and Yamashita, Kayo and Yamazaki, Naoki and Yan, Chunxiao and Yasuoka, Koichi

and Yavrumyan, Marat M. and Yenice, Arife Betül and Yıldız, Olcay Taner and Yu, Zhuoran and Yuliawati, Arlisa and Žabokrtský, Zdeněk and Zahra, Shorouq and Zeldes, Amir and Zhou, He and Zhu, Hanzhi and Zhu, Yilun and Zhuravleva, Anna and Ziane, Rayan. 2023. *Universal Dependencies 2.13*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. PID http://hdl.handle.net/11234/1-5287.