Keeping Up Appearances—or how to get all Uralic languages included into bleeding edge research and software: generate, convert, and LLM your way into multilingual datasets

Flammie A Pirinen
Divvun
UiT—Norgga árktalaš universitehta
Tromsø, Norway
flammie.pirinen@uit.no

Abstract

The current trends in natural language processing strongly favor large language models and generative AIs as the basis for everything. For Uralic languages that are not largely present in publically available data on the Internet, this can be problematic. In the current computational linguistic scene, it is very important to have representation of your language in popular datasets. Languages that are included in well-known datasets are also included in shared tasks, products by large technology corporations, and so forth. This inclusion will become especially important for under-resourced, under-studied minority, and Indigenous languages, which will otherwise be easily forgotten. In this article, we present the resources that are often deemed necessary for digital presence of a language in the large language model - obsessed world of today. We show that there are methods and tricks available to alleviate the problems with a lack of data and a lack of creators and annotators of the data, some more successful than others.

1 Introduction

In recent years, the landscape of language technology has changed quite rapidly, mainly with the advent large language models, but the overarching shift towards big data has been ongoing for longer. The problem with this shift is, that it is based on the big data for large majority languages, the inclusion of all the smaller languages, including all of the Uralic languages, has come as an afterthought if at all.

The expected solution for the continued sustainability of minority Uralic languages in the landscape of modern languages in the time of large language models is to “generate” more data. Ideally, by ‘generate’, the engineers in large language model contexts mean, that authentic written (or spoken) data needs to be created by native writers who should not make too many spelling or grammar errors and write the most current normative form. This can be an unreachable goal for a language that has fewer than million speakers and writers who are not L1, as while the requirements for large language models are going down over time, they are still orders of magnitude larger that can plausibly be created by limited amount of writers and speakers in limited amount of time.

What we suggest in this paper is to carefully organise the initial work of corpus curation and creation around materials that are of high importance to the contemporary language technology community. We leverage existing resources and language technologies to minimise unnecessary and repetitive work by linguists and language professionals on the language data that is being worked on; automating what can be automated and re-using linguists annotation efforts is a key to efficient development of high-quality human verified gold data.

Our *research question* is, going from existing langauge technology resources: which tools are best suitable for launching and bootstrapping which resources. If language has usable electronical dictionaries, morphological analysers and generators, spell-checkers and so on, what can be used to effectivise the dataset creation and corpus curation. The question is especially interesting now, as there is a possibility to use contemporary multilingual large language models, as well as traditional rule-based, statistical and hybrid language models to perform various pre-processing and processing tasks.

Our *key contributions* from this article are: *the experimental framework* for others to compare and combine methods of gold data annotation for smaller languages, the *pipelines* from traditional rule-based annotations and LLM generations into concrete target formats, and the results of comparing some of the approaches for a low resource Uralic language along with recommendations of what is currently the most effective approach. As a side product we have created, curated and an-

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 123–131
November 28-29, 2024 ©2024 Association for Computational Linguistics

notated beginnings of *several new datasets* for an under-resourced Uralic language.

We have laid out experimental computational linguistics data creation and annotation system that can use both existing rule-based tools as well as large language models to aid the processs. One of the goals of this experiment and the approach is that we want to promote inclusion of more Uralic languages in all of the common language technology datasets. We are considering three separate approaches to help creation of annotated gold data:

1. rule-based generators and generative language models to generate a starting point for a data set, to be proof-read and re-annotated by humans,
2. rule-based analysers creating annotated dataset in legacy and ad hoc formats that are converted and organised into a starting point for human re-annotation, and
3. generative language models providing human annotators with starting points or improvements during annotation process

There are of course other possibilities as well, these are based on our previous experience and iterations with different datasets and projects. It must be noted that the goal here is to generate something comparable to human annotated gold corpus, so we are not planning to automate data generation or annotation. This has to be also contrasted to the reality of limited human resources for working with smaller Uralic languages, we do not necessarily have a possiblity to hire 5 annotators to work on data full hours for several months, but to ask if the language experts who have other main jobs as language experts can use hours or two here and there on the task, this is one of the motivations of our experiment as well.

2 Background

The Uralic languages, especially besides the bigger national languages, are relatively under-resourced; the size of freely available texts is measured in millions of tokens or less. However, Uralic languages do have strong traditions of rule-based language technology. Also, lately, the large language model -based language technology has showed itself as a viable option for some use cases. Our approach to resource creation to overcome some of the under-resourcedness problem is thus to see if we can leverage the existing technology to supplement the well-planned tactical selection of language dataset resources. In this article, we suggest curating and creating data that are highly relevant for the large language model building industry and also for the researchers of languages in language technology and linguists as well. While majority of industry and researchers concern themselves with basically English and maybe handful of commercially plausible majority languages of the world, we have discovered some related research both from the industry and the researchers who specialise in minority and under-resourced languages.

As one reference point, we study what technology companies and central research groups in LLM-based language technology have said about support for smaller language in the recent years; One reason for writing this article and its experiments is also inspired by these works: Meta and FAIR research group (Facebook's AI Research) have released resources and studies under the moniker of *No language left behind* (NLLB) (Costa-jussà et al., 2022), also known for datasets and evaluation schemes under *Scaling neural machine translation to next 200 languages* (FLoRES) (Team et al., 2024). Unsurprisingly, this data set has so far included only Finnish, Estonian, and Hungarian when it comes to Uralic language inclusion. Alphabet and Google research have also been active on extending the range of languages supported under the name of *next 1000 languages* (Bapna et al., 2022). They have also published several research papers listing exactly the sources they use to gather information and data on the languages (Ritchie et al., 2024), this is directly useful information to know that, if you want to be included in Google's considerations list of languages that might be supported or relevant, perhaps you want to have data in the resources and datasets they use.

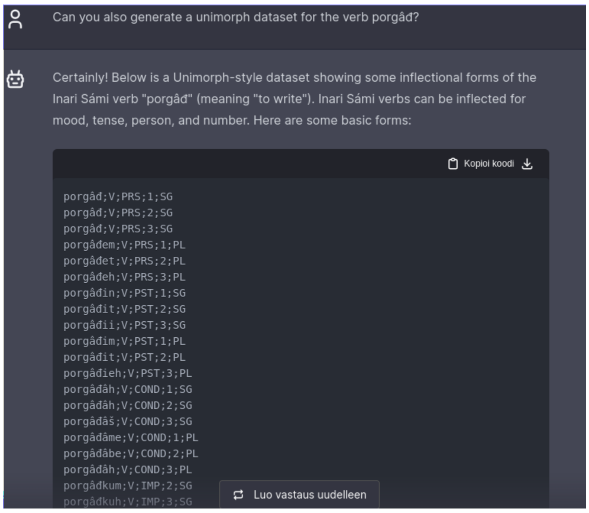
The resources that we use in this articles experiments here have also been used for several years now in the academic community as the go-to resource to measure if your tool works with the given language. For example, the *Universal Dependencies* (UD) treebanks (Zeman et al., 2024), are used in a huge number of papers investigating computational linguistic methods in a large number of languages, including the annual shared tasks in syntactic parsing. It would thus appear that UD as a resource has passed the test of time. Secondly we have seen the *Unimorph* dataset, that concerns morphology of languages, has been used widely in the research and applications. Namely with research of morphophonology and machine learning there have been regular shared tasks. We have explicitly left out parallel corpora and machine translations from this article for two reasons: firstly it is already a main focus of the large corporations and research groups working on the natural language engineering tasks and secondly our corpus selection is based on aiming to have a large subset of professionally human-translated texts as the source texts in these datasets, we find these are much more valuable than machine translated or post-edited texts, for the early phases of big data building we are in.

For the experimentation of this article I have chosen Inari Sámi as a target language; Inari Sámi is a Uralic language, that does not as of now have many of the resources that we are about to create. It is a low-resource Indigenous language with limited amount of speakers and written resources available, but an active speaker community that writes new texts. We have existing tools in rule-based language technology available from the well-known free and open source repository[1]. Furthermore, the most recent versions of large language model -based systems have been seen to support Inari Sámi (instead of just refusing to handle it and deferring to professionals as earlier versions did). Finally, we have a computational linguist who is not a native speaker but is capable of working with the language and has contacts to language experts, we find this is sufficient for initial experimentation, but of course for serious language data building, more expert knowledge is needed.

For some the work on dataset creation there has been previous works, for example in Universal Dependencies and rule-based analyser there are existing methods that have been used for other existing uralic dependencies treebanks, such as the North Sámi (Tyers and Sheyanova, 2017) and Karelian treebanks (Pirinen, 2019). For generation of the UniMorph data, some of the datasets are generated based on rule-based generators (Batsuren et al., 2022), strictly speaking Wiktionary can also be considered as rule-based morphological generation, however, we have not found this mentioned explicitly in existimg articles about unimorph.

3 Methods

Our experimentation concerns the use of existing language technology tools to help the creation of the datasets while following the rules and ideals behind the given datasets. For example, when Universal Dependencies guidelines dictates that the dependency annotation must be manual or human made, we do not use the tools to generate unchecked 1-best annotations that would pollute the dataset. The most common strategy here is to give all plausible hypotheses from the automatic analysis to the linguist to post-edit, but another option is that the post-edited analyses are verified to be plausible analyses of the system (our end goal is to have a gold standard that agrees with the analyser and linguistic expertise).



Can you also generate a unimorph dataset for the verb porgâđ?

Certainly! Below is a Unimorph-style dataset showing some inflectional forms of the Inari Sámi verb "porgâđ" (meaning "to write"). Inari Sámi verbs can be inflected for mood, tense, person, and number. Here are some basic forms:

```
porgâđ;V;PRS;1;SG
porgâđ;V;PRS;2;SG
porgâđ;V;PRS;3;SG
porgâđem;V;PRS;1;PL
porgâđet;V;PRS;2;PL
porgâđeh;V;PRS;3;PL
porgâđin;V;PST;1;SG
porgâđit;V;PST;2;SG
porgâđii;V;PST;3;SG
porgâđim;V;PST;1;PL
porgâđit;V;PST;2;PL
porgâđieh;V;PST;3;PL
porgâđâh;V;COND;1;SG
porgâđâh;V;COND;2;SG
porgâđâš;V;COND;3;SG
porgâđâme;V;COND;1;PL
porgâđâbe;V;COND;2;PL
porgâđâh;V;COND;3;PL
porgâđkum;V;IMP;2;SG
porgâđkuh;V;IMP;3;SG
```

Figure 1: ChatGPT generating data for Inari Sámi UniMorph dataset.

For the existing rule-based systems, we have downloaded and installed well-known GiellaLT softwares, which are freely available from the GitHub with an open source licence (Pirinen et al., 2023).[2] The LLM experimentation is performed using a ChatGPT, the state-of-the-art chatbot interface to a closed-source, commercial neural network.[3] We have chosen ChatGPT since it is the most popular one, it has freely usable version available for most Uralic language researchers even without expensive AI budget. An example of ChatGPT performing UniMorph dataset generation task can be seen in Figure 1.

When working with a preexisting computational linguistic, rule-based system, one of the main engineering efforts lies on the conversion. Although

[1] https://giellalt.github.io/lang-smn

[2] https://giellalt.github.io

[3] The version tested at the time of writing identifies itself as GPT-4, which was the newest model at the time we began experimenting but has probably been outdated by the time of the publication.

it sounds trivial, there is a lot of linguistic and engineering work to be taken into account here: the actual format of the analyses is rarely exactly the same, so a mapping needs to be devised, for example, converting "noun" analyses from `+N` to `N;` or `NOUN`. The mappings can also be 1:n or m:1, merging and joining 'tags', as well as more involved re-writings. There are a lot of other technical minor details related to such generations and conversions that are beyond the scope of this article, for example, we needed an algorithm that could remove duplicate forms that is aware of Unicode normalisation forms and folding to avoid having the linguist read word forms that look exactly the same several times. The topic of conversions in itself is large enough to deserve its own article,[4] for the purposes of this article we will point the readers to our github repositorium containing freely available scripts.[5] Some examples of conversions are given in the Figure 2.

The experiments with LLMs are based on the currently available free ChatGPT interface prompted in English. We begin prompting with the most straightforward requests, e.g. "can you generate a unimorph annotated list of all word-forms Inari Sámi noun táálu?", "create a CONLL-U annotated version of this sentence", etc.

It might be noteworthy, that since our goal is inclusion of our Uralic languages in the relevant datasets, there is also a component of social engineering involved in all of the dataset creations. Merely producing text files that contain acceptable data is only a first step. The datasets we have selected to experiment with, the selection has been also based on the openness and documentation of the contribution process; all of the given datasets exist on GitHub, and the contribution process is detailed in the documentation and happens largely over GitHub only. This is in contrast to the commercially backed datasets mentioned earlier; while it would be very valuable to have all Uralic languages in the *No Languages Left Behind* and *Next Thousand Languages*, the way to contribute here is not immediately so obvious and available to larger audiences.

4 Corpora and Data Selection

The corpora available for low-resource Uralic languages are scarce and limited. The whole corpora of publically available web crawl data is typically less than the millions of tokens that is often advertised as minimum requirement of large language models. Furthermore, the data that is available is limited by licences, quality, and genres: While some argue that all data that can be crawled is free to use for language technologies, in practice ethical use requires selecting only the data that has explicitly been licenced with a suitable licence, such as Wikipedia or data coming from governmental public domain records—or that has been personally licenced with the author for the specific use. That furthermore limits both quality—wikipedia data is written by language learners—and genres—government's publication are mainly politics, healthcare and such.

In this experiment we have used primarily freely licenced data from Saami international corpora (SIKOR), (SIKOR, 2021) but we have also performed a short experiment on self-created and self-translated data that large language model should not contain from beforehands.

5 Experimental results

The main results of our experiment will be the actual datasets we can produce. To quantify the usefulness of the langauge technology tools we have measured post-edit distances. We have also performed a linguistic error analysis to quantify the errors made, the effect on the time/effort tradeoff is further discussed in the Section 6.

In our experiment in creating datasets for Unimorph, we used both the rule-based system and the LLM to generate the full datasets, that can be read and corrected by a human. The results of generating are shown in the table 1. The expected forms is based on the linguistic grammars we have available (Morottaja and Olthuis, 2023). We have measured the numbers of forms generated, Coverage counted as proportion of generated unique forms out of expected and Accuracy as proportion of fully correct forms and analyses of all generated. In general rule-based approach is close to the gold standard, which is expected from rule-based systems, the LLM has also generated a smaller subset of forms with lower accuracy.

In our experiments in Universal Dependencies annotation, we used the rule-based system to gen-

[4] we have attempted to write one such article, even at very condensed format it easily exceeds 8 pages that is the maximum for average conference article in language technologies.

[5] `anonymised`

E.g. *Finite State Morphology* to *Unimorph*

```
táálu      táálu+N+Sg+Nom              <-> táálu      táálu   N;SG;NOM
táálust    táálu+N+Sg+Loc              <-> táálust    táálu   N;SG;LOC
tálustân   táálu+N+Sg+Loc+PxSg1        <-> tálustân   táálu   N;SG;LOC;PSS1S
```

E.g. *VISL CG 3* to *Universal Dependencies*

```
"<mun>"
    "mun" Pron Pers Sg1 Nom @SUBJ> #1->2
:
"<juuhim>"
     "juuhâđ" <mv> V TV Ind Prt Sg1 @FMV #2->0
:
"<vuolâ>"
    "vuolâ" N Sem/Drink Sg Acc @<OBJ #3->2
                                        ^^^
                                        |||
                                        vvv
# textid = example.1
# text = mun juuhim vuolâ
1 mun mun PRON Pron Pers Case=Nom|Number=Sing|Person=1|PronType=Pers 2 nsubj _ _
2 juuhim juuhâđ VERB V TV Mood=Ind|Number=Sing|Person=1|Tense=Past 0 root _ _
3 vuolâ vuolâ NOUN N Sem/Drink Case=Acc|Number=Sing 2 obj _ _
```

Figure 2: Conversions between traditional rule-based analyses and target dataset formats

POS	**Expected forms**	**RB forms**	**RB Cov %**	**RB Acc %**	**LLM forms**	**LLM Cov %**	**LLM Acc %**
Nouns	58	100*	100 %	99 %	14	15 %	21 %
Verbs	57	55	96 %		22	39 %	0 %
Adjectives	51	61	100 %		14	20 %	10 %

Table 1: Unimorph dataset creation statistics. Expected forms is number of forms based on the grammar, RB from rule-basd generator and LLM from large language model, Coverage and Accuracy measured in % units. * Some extra forms in rule-based model are due to allomorphy which was not accounted for expected forms.

System	Full WER	Dep WER
Rule-Based	0.47	0.22
LLM	1.00	0.52

Table 2: Caption

erate ambiguous listing of all potential readings of the sentence with annotations, according to the guidelines in previous works by Pirinen (2019), and asked LLM to generate similar hypotheses likewise. In Table 2 we measure the post edit distance of the sentences fixed and re-annotated, the error rates are calculated as $E = \frac{S+I}{N}$, where E is the error rate, S is number of substitutions made, I is the insertions made, and N number of readings (i.e. N is number of CONLL-U lines with an index). We do not have D for deletions since both methods generated correctly generated one token per token in the input and there are so far no retokenisation requirements (multi-word tokens, multi-token words etc.), however LLM missed some punctuation tokens causing an insertion to be required. The full error rate basically counts whole lines of CONLL-U when making matches and dep error rate just the dep field.

6 Discussion

We have tested rule-based and LLM-based annotations as a help in linguistic work. Currently, for morphology we get clearly better results with the rule-based tools and the results are good enough that it makes work on dataset creation more effective. If we analyse the errors that the systems make, we see that rule-based system includes some results with linguistically motivated potential errors, like wrong stem alternation or missing accent in a suffix. The errors in LLM generated version are that it just uses seemingly random suffixes with unchanged stem, it also uses some forms like cases that do not exist in Inari Sámi (but for example exist in Finnish), all in all cleaning this data would possible even be slower than writing the data by hand. When we error-analyse the dependency analysis the results get more interesting, like both starting points require quite a significant amount of work to get to gold-standard state, but this is also to be expected if reference the past experiences of UD annotation from converted or machine analysed starting point. What is interesting is that the LLM can sometimes generate quite accurate depdendency subgraphs of certain expressions, for example personal names, we assume this is due to them appearing in very similar form in existing English documentations, where high level dependency structure is the same even if there are slight variations in the morphological level.

There are a large number of different large language models and generative artificial intelligence that could possibly be used to experiment this and that is a common feedback we get. We are using a version of a popular LLM that is available to us, without excessive extra costs. This is also available to most researchers who are the target audience of this paper.

A common feedback we get, that there are various techniques that should be used for low resource setup, like fine-tunings, transfer learnings, in context learnings, prompting techniques and so on. We are experimenting in a situation where we start with zero data for the fine-tuning task, we are the ones who will create these data initially, so the use of such data will generally be a future research topic, after we have done the initial data creation. As the methodology here is extremely fast moving and outdates itself in matter of months, we try to begin by only importing either approaches that have been proven and stabilised, perhaps in majority language context, into out lesser resourced languages, or we can perform experimentation that does not tie up too much valuable and scarce resources. Another interesting future research question would be whether it is more beneficial and time-effective to fine-tune early or on-goingly, given the constraints in data and human resources we face in the processing of smaller Uralic languages. We have not found an easy enough recipe to do transfer learning that would not take us more time than actually working on the data creation as described by the approach of this article. Our impression is furthermore that there is currently ongoing research on this topic that we hope will yield some answers that are relevant to us as well.

It is exciting to see that, even if the large language models have rathar disappointing accuracy in generating and annotation of smaller Uralic languages, they are able to generate something that is relevant to the task and occasionally some wordforms or annotations are even correct. This suggests that maybe with further fine-tuning, prompting, in-context learning, transfer learning, and so forth, there could be a usable version of LLM-aided language data annotation and generation in the future.

One question for future work is of course how to integrate these findings to a workflow and softwares for annotation. In this experiment we used normal text editors and raw data formats for data annotation, which is suitable for programmers and short experiments, for the full scale linguistic annotation this would be integrated to a specific editor. And that raises the question of if the ideal way to help linguistic jobs would bear a user interface similar to what we get in the email post writing programs, office tools and programming editors today with a so-called *co-pilot*?

7 Conclusion

We performed several experiments to find out an efficient way of creating NLP datasets for smaller Uralic languages. We have found that using both existing rule-based technology and large language models can help rapid creation of the data, but neither approach is without its caveats. The gold standard remains fully human annotated data, but in lack of that it should be considered if we can achieve reasonable amounts of resources with computer-aided annotation modes.

Limitations

The experimentation on large language models is done using one closed source commercial system and is not reproducible at all, however, this is a common practice in the science of natural language processing in 2024.

The experiments were performed by language learner instead of native speaker or expert, the qualitative results may differ when language experts are working on the same pre-processed data.

Ethics

The large language models used in this experimentation have wasted an estimated several hundreds of litres of drinking water [6] and not insignificant amount of energy (Strubell et al., 2019).[7] If LLM method is taken in to use in the development of annotated gold corpora and data sets, this needs to be taken into consideration until the providers of LLMs resolve the excessive use of natural resources.

No underpaid crowd-sourcers were involved in performing the linguistic tasks, all annotations and evaluations were made by fully paid colleagues.

References

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *Preprint*, arXiv:2205.03983.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. Unimorph 4.0: Universal morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Petter Morottaja and Marja-Liisa Olthuis. 2023. *Inarinsaamen taivutusoppi*. Sámediggi.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. Giellalt—a stable infrastructure for nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649.

Tommi A Pirinen. 2019. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. LinguaMeta: Unified metadata for thousands of languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.

SIKOR. 2021. SIKOR uit norgga árktalaš universitehta ja norgga sámedikki sámi teakstačoakkáldat, veršuvdna 06.11.2018. http://gtweb.uit.no/korp. Accessed: 2024-10-01.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep

[6] https://www.thetimes.com/uk/technology-uk/article/thirsty-chatgpt-uses-four-times-more-water-than-previously-thought-bc0pqswdr

[7] https://disconnect.blog/silicon-valley-is-sacrificing-the-climate-for-ai/

learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Francis M. Tyers and Mariya Sheyanova. 2017. Annotation schemes in North Sámi dependency parsing. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 66–75, St. Petersburg, Russia. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, H͡órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Ansu Berg, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Johnatan Bonilla, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André

Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Emilia Milano, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinh͡ór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Tarık Emre Tıraş, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn H͡órðarson, Vilhjálmur H͡orsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2024. Universal dependencies 2.14. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.