Merging Close and Distant Perspectives on Language: Using Linguistic Typology in NLP

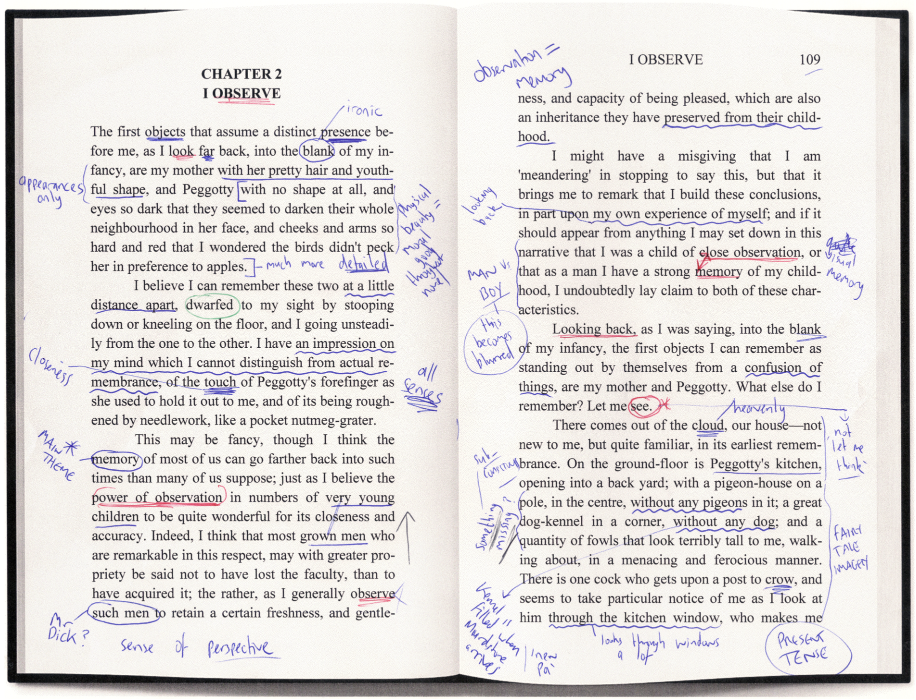
Research Seminar Language Technology
Helsinki University, 18 April 2024

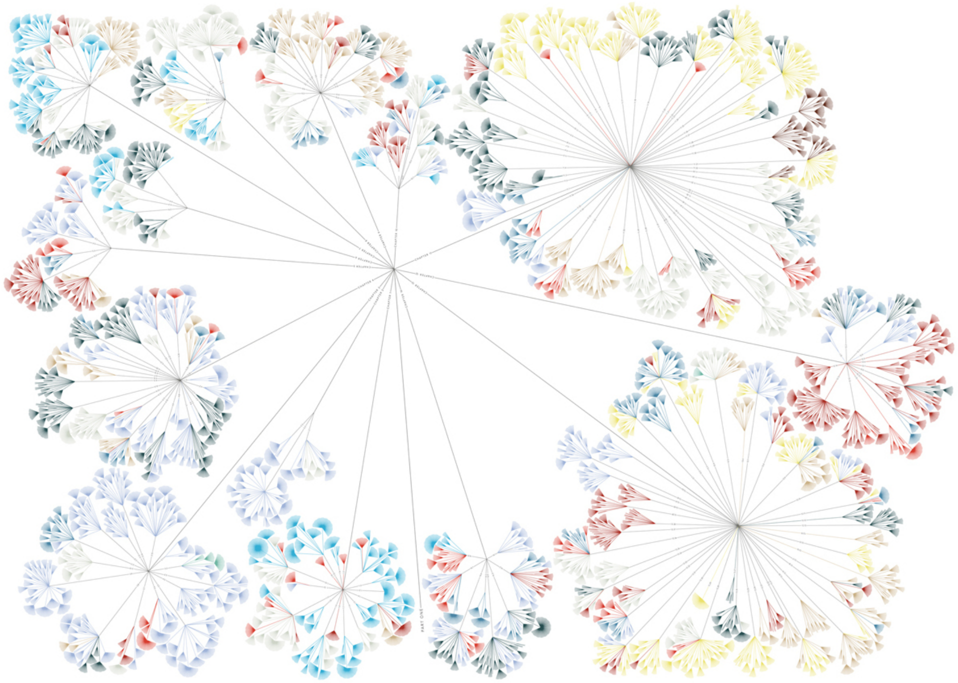
Esther Ploeger (`espl@cs.aau.dk`)

AALBORG UNIVERSITY

Close reading

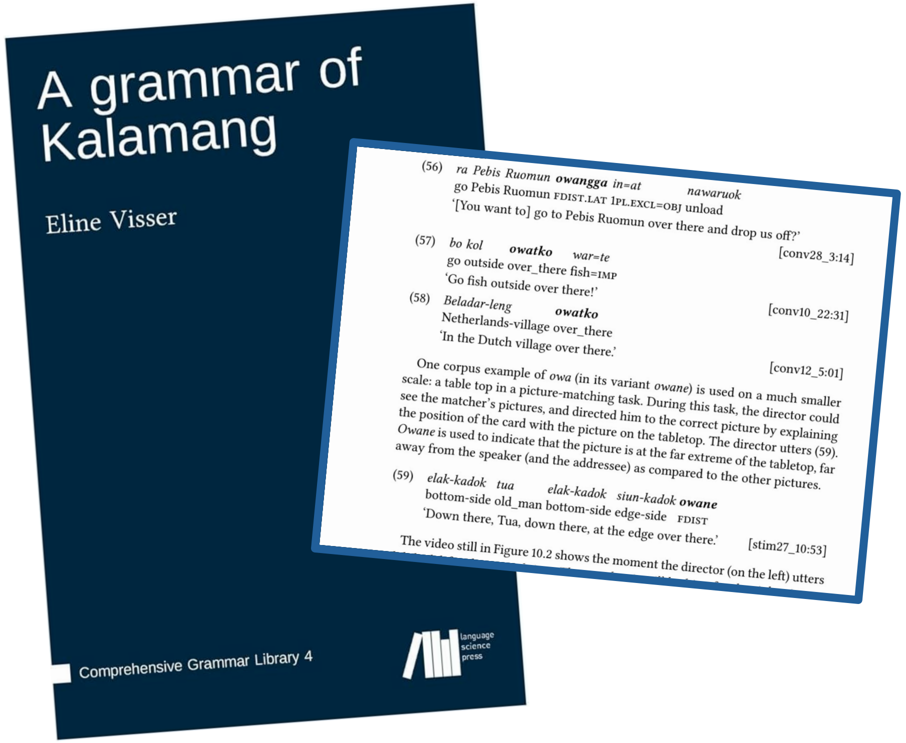
Distant reading



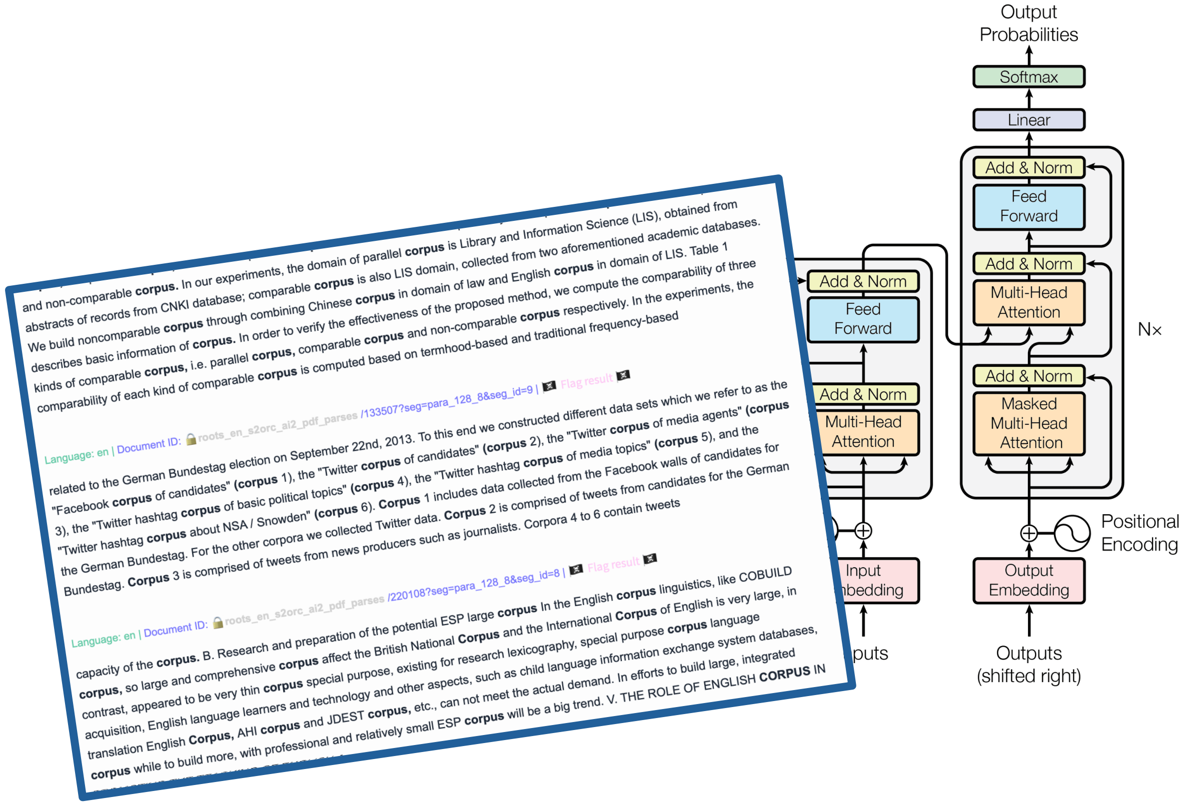


Jänicke et al. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.

Close view of language

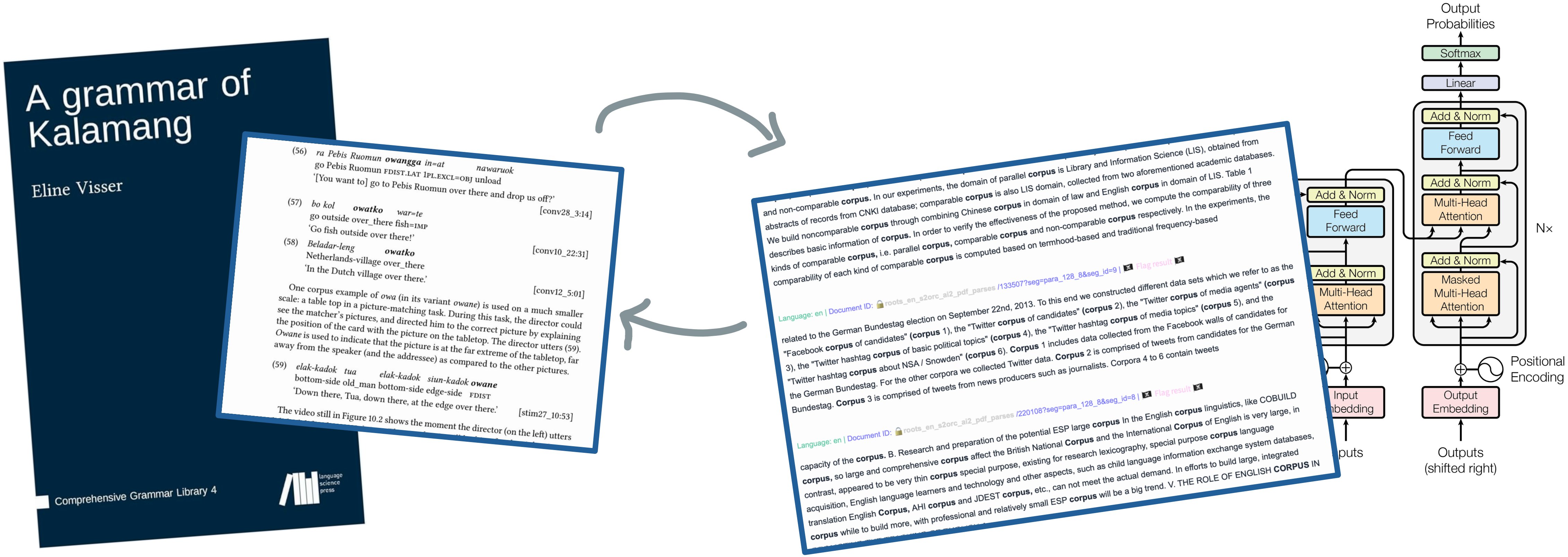


Distant view of language



Close view of language

Distant view of language



This talk

- Some background on linguistic typology
- Using typological information in NLP
  - Interpreting
  - Evaluating
  - Improving

  } Multilingual LMs
- Current issues and future solutions
- Conclusions

Disclaimers

- Indeed, using linguistics in NLP is nothing new…

Disclaimers

- Indeed, using linguistics in NLP is nothing new…
- This talk will not be about computational typology

Disclaimers

- Indeed, using linguistics in NLP is nothing new…
- This talk will not be about computational typology

Uncovering Probabilistic Implications in Typological Knowledge Bases

Johannes Bjerva[a] Yova Kementchedjhieva[a] Ryan Cotterell[?,fi] Isabelle Augenstein[a]
[a]Department of Computer Science, University of Copenhagen
[?]Department of Computer Science, Johns Hopkins University
[fi]Department of Computer Science and Technology, University of Cambridge
bjerva,yova,augenstein@di.ku.dk, rdc42@cam.ac.uk

Abstract

The study of linguistic typology is rooted in the implications we find between linguistic features, such as the fact that languages with object-verb word ordering tend to have post-positions. Uncovering such implications typically amounts to time-consuming manual processing by trained and experienced linguists, which potentially leaves key linguistic universals unexplored. In this paper, we present a

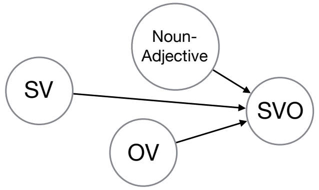


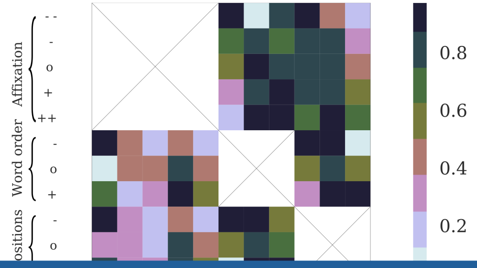
Figure 1: Visualisation of a section of our induced graphical model. Observing the features in the left-most nodes (SV, OV, and Noun-Adjective), can we cor-

A Probabilistic Generative Model of Linguistic Typology

Johannes Bjerva[a] Yova Kementchedjhieva[a] Ryan Cotterell[?,fi] Isabelle Augenstein[a]
[a]Department of Computer Science, University of Copenhagen
[?]Department of Computer Science, Johns Hopkins University
[fi]Department of Computer Science and Technology, University of Cambridge
bjerva,yova,augenstein@di.ku.dk, rdc42@cam.ac.uk

Abstract

In the principles-and-parameters framework, the structural features of languages depend on parameters that may be toggled on or off, with a single parameter often dictating the status of multiple features. The implied covariance between features inspires our probabilisation of this line of linguistic inquiry—we develop a generative model of language



Background

What is linguistic typology?

Selected topics

What is linguistic typology?

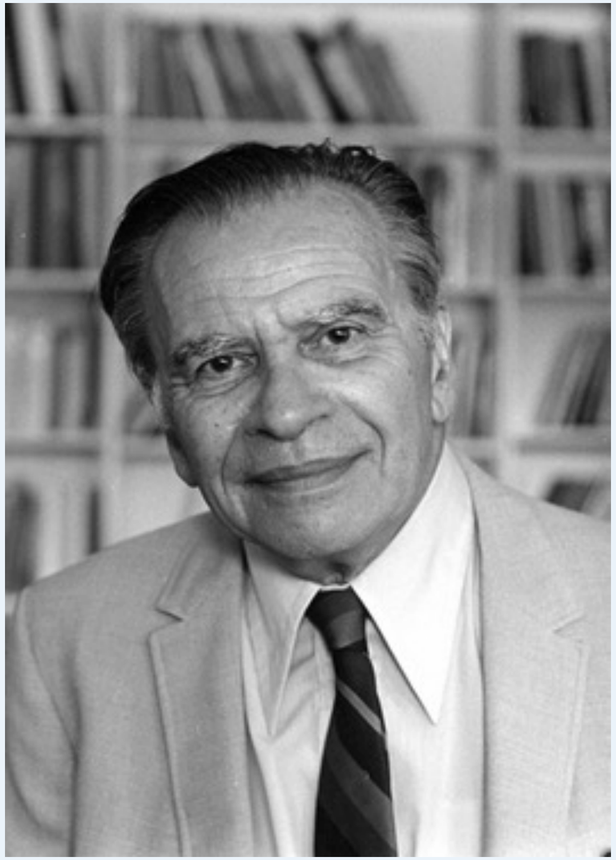
“the classification of the world’s languages according to similarities and differences in their linguistic structures and genetic relationships.”

“Language typology, therefore, is essentially comparative and crosslinguistic.”

Kashyap, A. K. (2019). Language typology. The Cambridge handbook of systemic functional linguistics, 767-792.

‘Universals’

- “Some universals of grammar with particular reference to the order of meaningful elements” (1963)
- 45 linguistic universals
- Universal 3: “Languages with dominant VSO order are always prepositional.”



Joseph Greenberg

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. Universals of language, 2, 73-113.

Language sampling

“**a general theory of grammar must provide a framework for all languages** and not just for, say, Dutch or English. These are just two manifestations of possible languages, and there is no reason to assume a priori that by studying one or two languages we can account for linguistic phenomena in every other language as well.”

Rijkhoff, J., Bakker, D., Hengeveld, K., & Kahrel, P. (1993). A method of language sampling. Studies in Language. 17(1), 169-203.

Language sampling

Three types of sampling methods (Rijkhoff & Bakker, 1998):

- **Random** sampling

Language sampling

Three types of sampling methods (Rijkhoff & Bakker, 1998):

- **Random** sampling
- **Probability** sampling
  - *Languages should be as independent as possible*
  - *Sample from different families, locations, etc.*

Rijkhoff, J., & Bakker, D. (1998). Language sampling. Linguistic Typology, 2(3), 263-314.

Language sampling

Three types of sampling methods (Rijkhoff & Bakker, 1998):

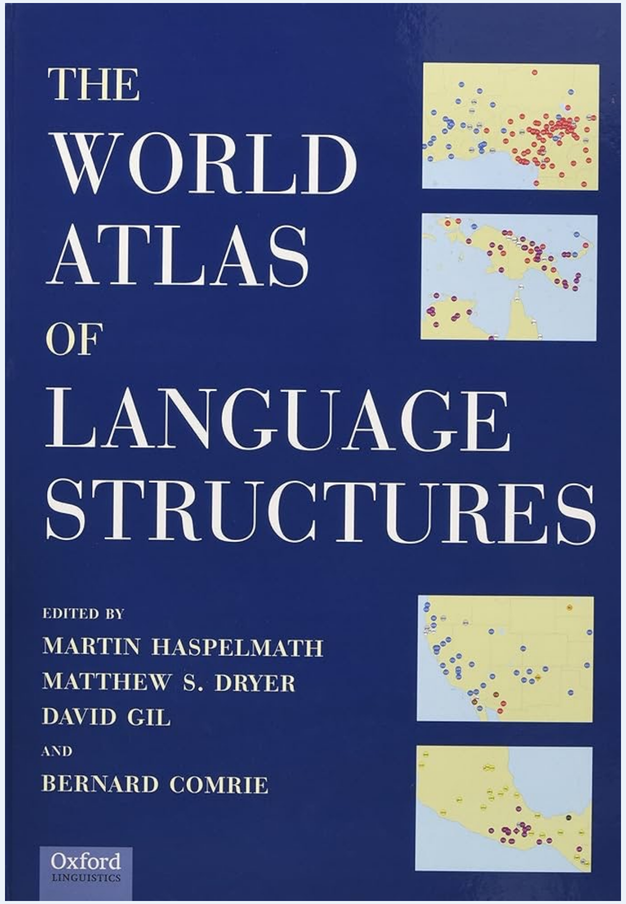
- **Random** sampling
- **Probability** sampling
  - *Languages should be as independent as possible*
  - *Sample from different families, locations, etc.*
- **Variety** sampling
  - *The sample should include the rarest cases*
  - *Exceptional properties should be captured, rule out counterexamples*

Rijkhoff, J., & Bakker, D. (1998). Language sampling. Linguistic Typology, 2(3), 263-314.

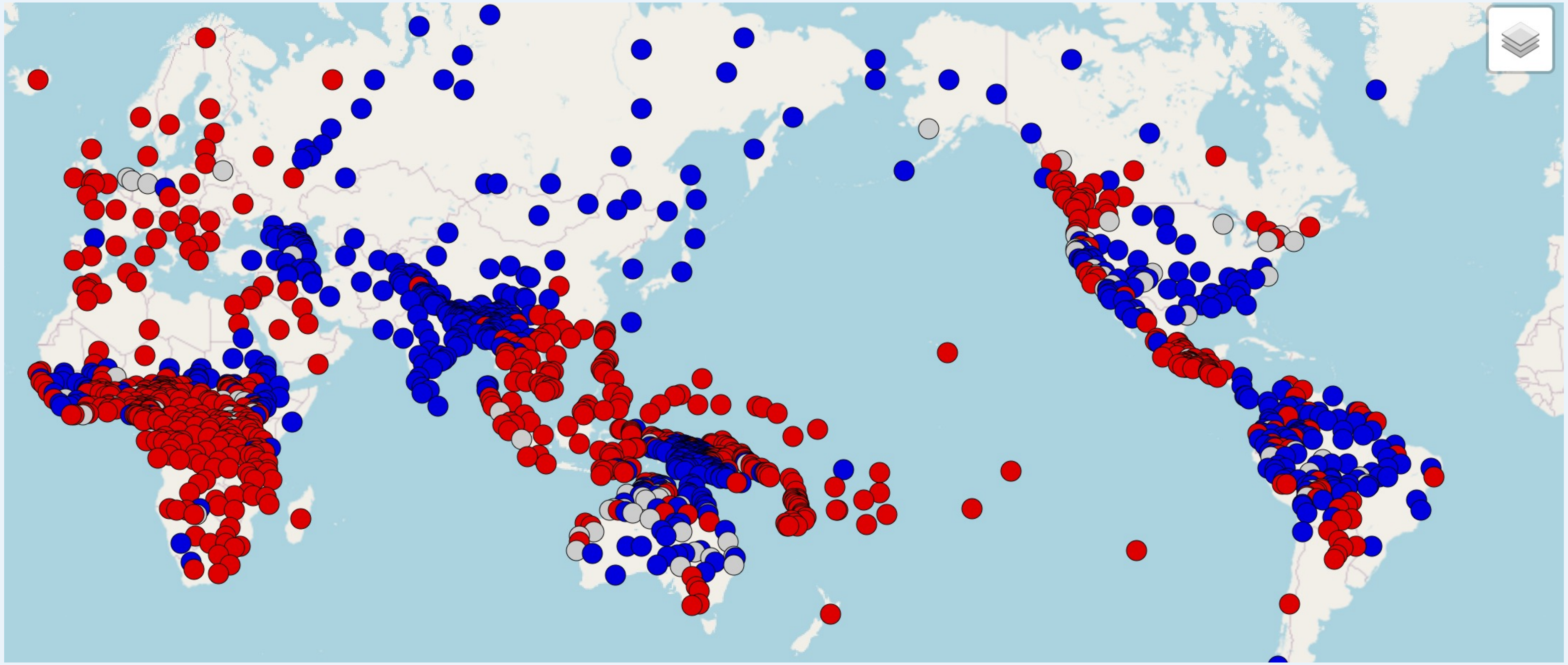
Typological Databases

Visser, E. (2022). A grammar of Kalamang. Language Science Press.

Typological Databases

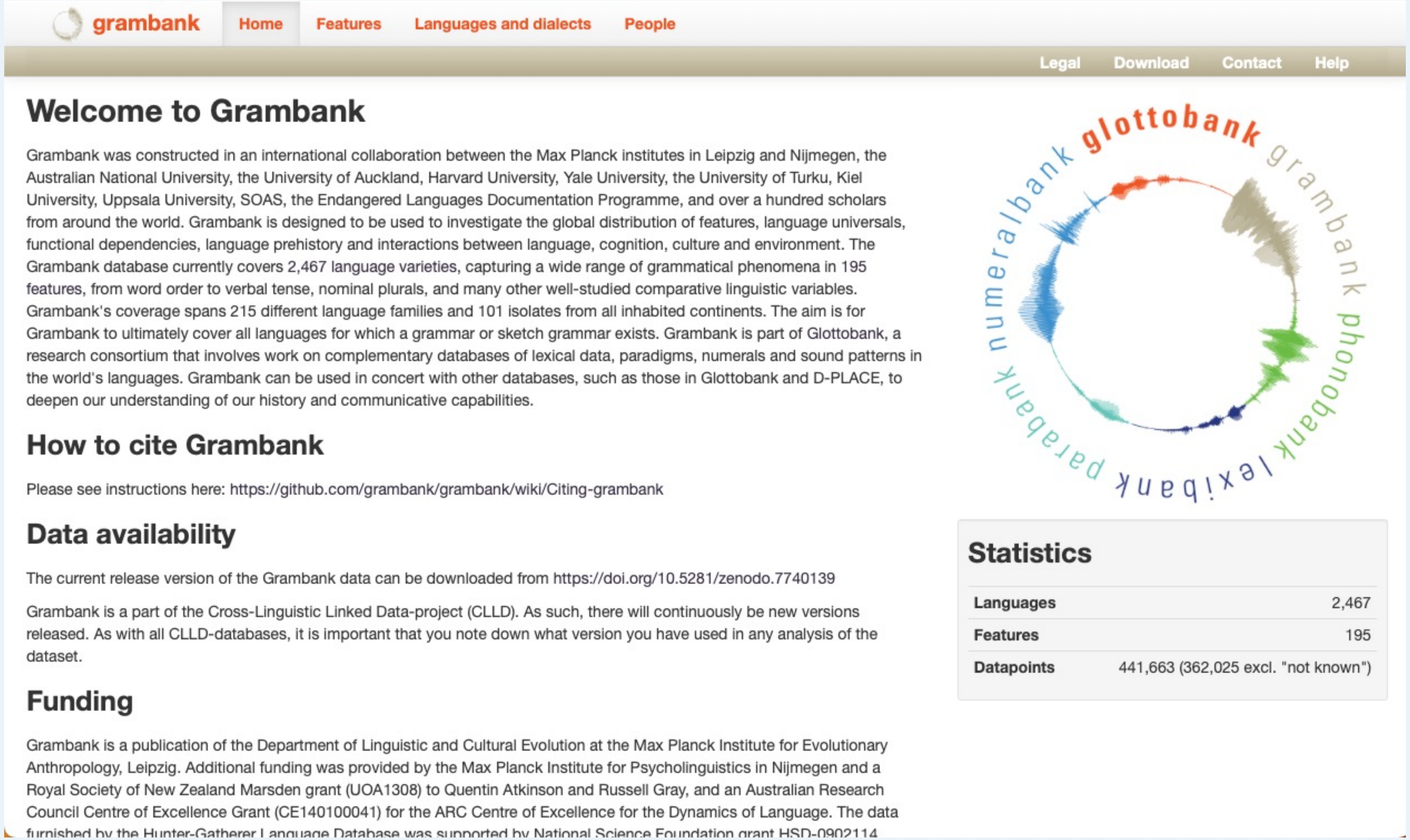


Typological Databases



Matthew S. Dryer. 2013. Order of Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

Typological Databases


grambank Home Features Languages and dialects People

Legal Download Contact Help

Welcome to Grambank

Grambank was constructed in an international collaboration between the Max Planck institutes in Leipzig and Nijmegen, the Australian National University, the University of Auckland, Harvard University, Yale University, the University of Turku, Kiel University, Uppsala University, SOAS, the Endangered Languages Documentation Programme, and over a hundred scholars from around the world. Grambank is designed to be used to investigate the global distribution of features, language universals, functional dependencies, language prehistory and interactions between language, cognition, culture and environment. The Grambank database currently covers 2,467 language varieties, capturing a wide range of grammatical phenomena in 195 features, from word order to verbal tense, nominal plurals, and many other well-studied comparative linguistic variables. Grambank's coverage spans 215 different language families and 101 isolates from all inhabited continents. The aim is for Grambank to ultimately cover all languages for which a grammar or sketch grammar exists. Grambank is part of Glottobank, a research consortium that involves work on complementary databases of lexical data, paradigms, numerals and sound patterns in the world's languages. Grambank can be used in concert with other databases, such as those in Glottobank and D-PLACE, to deepen our understanding of our history and communicative capabilities.

glottobank grambank phonobank lexibank parabank numeralbank

How to cite Grambank

Please see instructions here: https://github.com/grambank/grambank/wiki/Citing-grambank

Data availability

The current release version of the Grambank data can be downloaded from https://doi.org/10.5281/zenodo.7740139

Grambank is a part of the Cross-Linguistic Linked Data-project (CLLD). As such, there will continuously be new versions released. As with all CLLD-databases, it is important that you note down what version you have used in any analysis of the dataset.

Statistics

Languages	2,467
Features	195
Datapoints	441,663 (362,025 excl. "not known")

Funding

Grambank is a publication of the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology, Leipzig. Additional funding was provided by the Max Planck Institute for Psycholinguistics in Nijmegen and a Royal Society of New Zealand Marsden grant (UOA1308) to Quentin Atkinson and Russell Gray, and an Australian Research Council Centre of Excellence Grant (CE140100041) for the ARC Centre of Excellence for the Dynamics of Language. The data furnished by the Hunter-Gatherer Language Database was supported by National Science Foundation grant HSD-0902114

Skirgård, Hedvig et al. (2023). Grambank v1.0 (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7740140

Typological Databases

Features

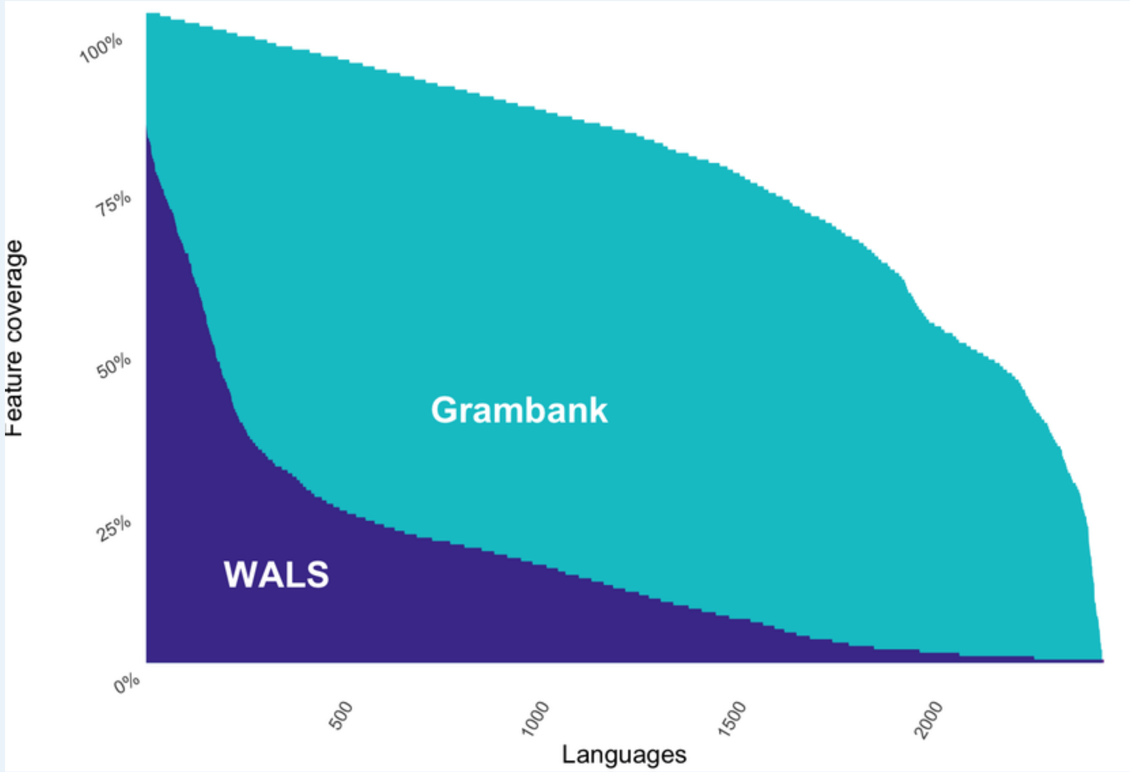
Showing 1 to 100 of 195 entries

← Previous 1 2 Next →

Id	Feature	Patron	Languages and dialects	Details
Search	Search		Search	
GB020	Are there definite or specific articles?	Jay Latarche and Jeremy Collins	2198	Values and description
GB021	Do indefinite nominals commonly have indefinite articles?	Jay Latarche and Jeremy Collins	2221	Values and description
GB022	Are there prenominal articles?	Jay Latarche and Jeremy Collins	2208	Values and description
GB023	Are there postnominal articles?	Jay Latarche and Jeremy Collins	2205	Values and description
GB024	What is the order of numeral and noun in the NP?	Hannah J. Haynie	2199	Values and description
GB025	What is the order of adnominal demonstrative and noun?	Jay Latarche and Jeremy Collins	2259	Values and description
GB026	Can adnominal property words occur discontinuously?	Hannah J. Haynie	1771	Values and description
GB027	Are nominal conjunction and comitative expressed by different elements?	Hedvig Skirgård	1778	Values and description

Skirgård, Hedvig et al. (2023). Grambank v1.0 (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7740140

Typological Databases



Skirgård, Hedvig et al. (2023). Grambank v1.0 (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7740140

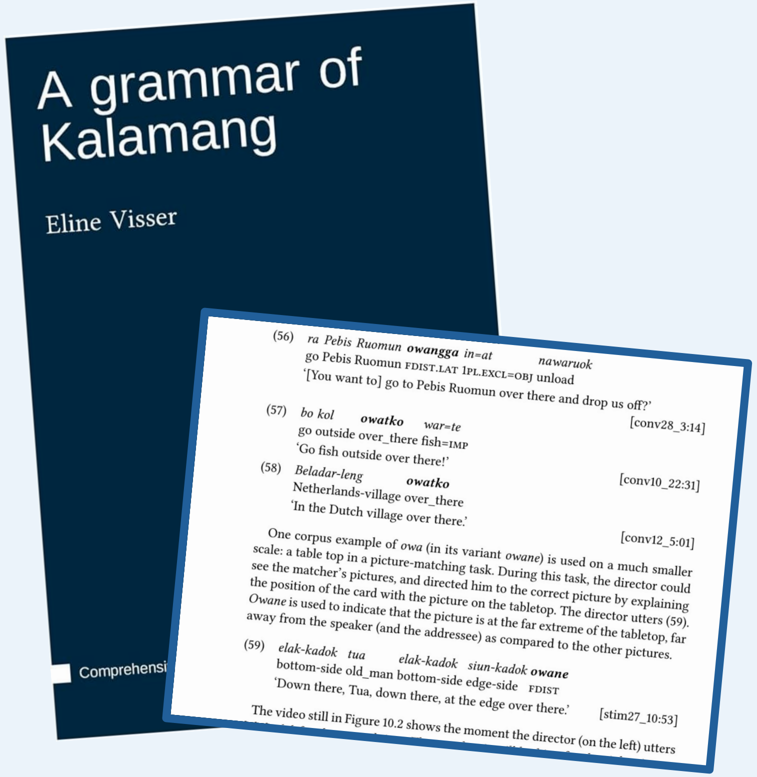
Typological Databases

Grambank	WALS
Comparable number of features and languages	
More datapoints (higher coverage per lang/feat)	Fewer datapoints (lower coverage per lang/feat)
Mostly coded in binary values (“what is possible?”)	Mostly coded in multi-value values (“what is dominant?”)
“Care was taken to avoid strict logical dependencies between features”	
Grammar	Phonology, lexicon, sign languages, ‘other’, …
Actively maintained	No longer maintained

Typological Databases

"The scale, completeness, reliability, format, and documentation of Grambank make it a useful resource for linguistically-informed models, cross-lingual NLP, and research targeting less-resourced languages."

Typological Databases

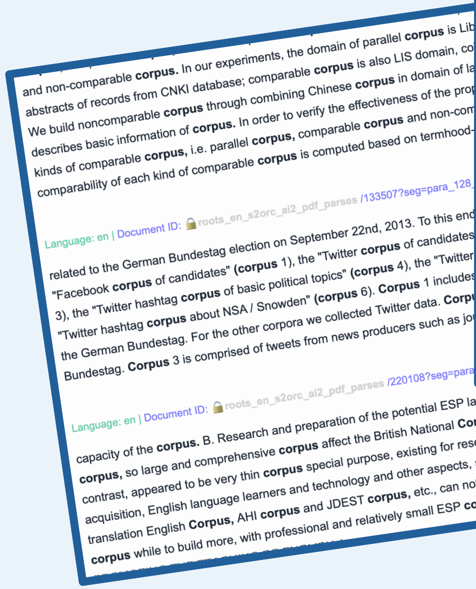


(56) ra Pebis Ruomun **owangga** in=at nawaruok
go Pebis Ruomun FDIST.LAT 1PL.EXCL=OBJ unload
'[You want to] go to Pebis Ruomun over there and drop us off?'
[conv28_3:14]
(57) bo kol **owatko** war=te
go outside over_there fish=IMP
'Go fish outside over there!'
[conv10_22:31]
(58) Beladar-leng **owatko**
Netherlands-village over_there
'In the Dutch village over there.'
[conv12_5:01]

One corpus example of *owa* (in its variant *owane*) is used on a much smaller scale: a table top in a picture-matching task. During this task, the director could see the matcher's pictures, and directed him to the correct picture by explaining the position of the card with the picture on the tabletop. The director utters (59). *Owane* is used to indicate that the picture is at the far extreme of the tabletop, far away from the speaker (and the addressee) as compared to the other pictures.

(59) elak-kadok tua elak-kadok siun-kadok **owane**
bottom-side old_man bottom-side edge-side FDIST
'Down there, Tua, down there, at the edge over there.' [stim27_10:53]

The video still in Figure 10.2 shows the moment the director (on the left) utters



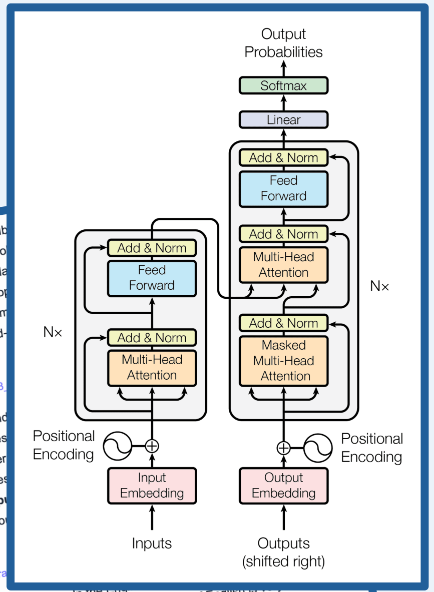
and non-comparable **corpus**. In our experiments, the domain of parallel **corpus** is Lib
abstracts of records from CNKI database; comparable **corpus** is also LIS domain, co
We build noncomparable **corpus** through combining Chinese **corpus** in domain of la
describes basic information of **corpus**. In order to verify the effectiveness of the prop
kinds of comparable **corpus**, i.e. parallel **corpus**, comparable **corpus** and non-com
comparability of each kind of comparable **corpus** is computed based on termhood-

Language: en | Document ID: roots_en_s2orc_ai2_pdf_parses /133507?seg=para_128_

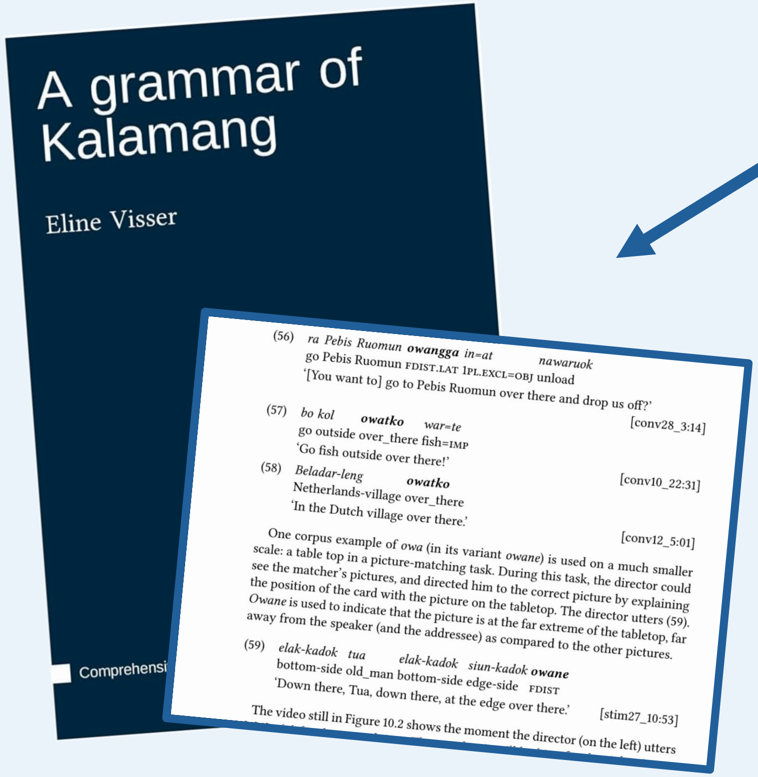
related to the German Bundestag election on September 22nd, 2013. To this end
"Facebook **corpus** of candidates" (**corpus** 1), the "Twitter **corpus** of candidates
3), the "Twitter hashtag **corpus** of basic political topics" (**corpus** 4), the "Twitter
"Twitter hashtag **corpus** about NSA / Snowden" (**corpus** 6). **Corpus** 1 includes
the German Bundestag. For the other corpora we collected Twitter data. **Corpu**
Bundestag. **Corpus** 3 is comprised of tweets from news producers such as jo

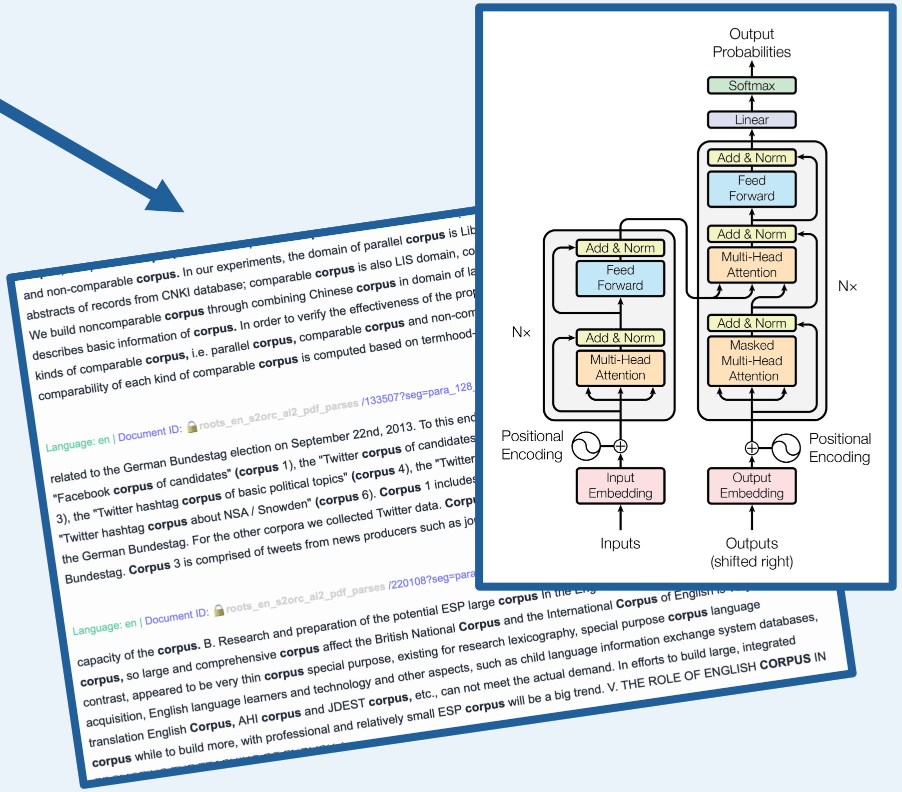
Language: en | Document ID: roots_en_s2orc_ai2_pdf_parses /220108?seg=para

capacity of the **corpus**. B. Research and preparation of the potential ESP large **corpus** in the Eng
corpus, so large and comprehensive **corpus** affect the British National **Corpus** and the International **Corpus** of English
contrast, appeared to be very thin **corpus** special purpose, existing for research lexicography, special purpose **corpus** language
acquisition, English language learners and technology and other aspects, such as child language information exchange system databases,
translation English **Corpus**, AHI **corpus** and JDEST **corpus**, etc., can not meet the actual demand. In efforts to build large, integrated
corpus while to build more, with professional and relatively small ESP **corpus** will be a big trend. V. THE ROLE OF ENGLISH **CORPUS** IN



Typological Databases





Linguistic Typology in NLP

Linguistic Typology in NLP

Using linguistic typology for the

Interpreting

Evaluating

Improving

of multilingual language models

Linguistic Typology in NLP

Using linguistic typology for the

Interpreting

Evaluating

Improving

of multilingual language models

The Past, Present, and Future of Typological Databases in NLP (Baylor et al., Findings 2023)

Model interpretability with typology

Probing classifiers:

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207-219

Model interpretability with typology

Probing classifiers:

Internal Representations → **Classifier** → **External Properties**

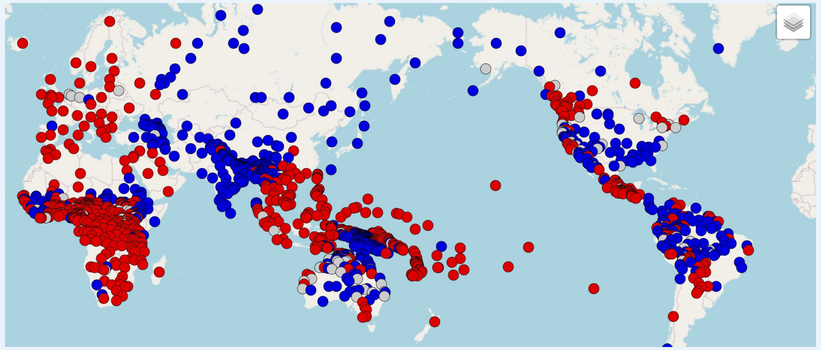
Multilingual model; different representation per language

Model interpretability with typology

Probing classifiers:

Internal Representations → **Classifier** → **External Properties**

Multilingual model; different representation per language



Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207-219

Model interpretability with typology

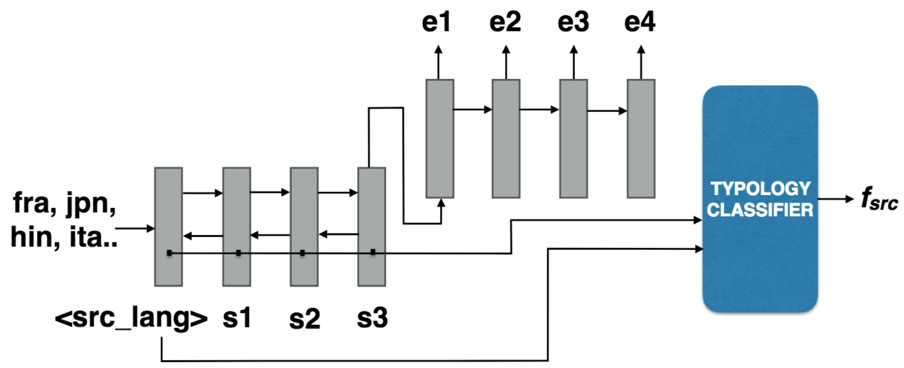


Figure 1: Learning representations from multilingual neural MT for typology classification. (Model MTBOTH)

"This work presents a more holistic analysis of **whether we can discover what neural networks learn about the linguistic concepts of an entire language** by aggregating their representations over a large number of the sentences in the language."

Model interpretability with typology

	Syntax		Phonology		Inventory	
	-Aux	+Aux	-Aux	+Aux	-Aux	+Aux
NONE	69.91	83.07	77.92	86.59	85.17	90.68
LMVEC	71.32	82.94	80.80	86.74	87.51	89.94
MTVEC	74.90	83.31	82.41	87.64	89.62	90.94
MTCELL	75.91	85.14	84.33	88.80	90.01	90.85
MTBOTH	**77.11**	**86.33**	**85.77**	**89.04**	**90.06**	**91.03**

Table 1: Accuracy of syntactic, phonological, and inventory features using LM language vectors (LMVEC), MT language vectors (MTVEC), MT encoder cell averages (MTCELL) or both MT feature vectors (MTBOTH). Aux indicates auxiliary information of geodesic/genetic nearest neighbors; "NONE -Aux" is the majority class chance rate, while "NONE +Aux" is a 3-NN classification.

Model interpretability with typology

	Syntax		Phonology		Inventory	
	-Aux	+Aux	-Aux	+Aux	-Aux	+Aux
NONE	69.91	83.07	77.92	86.59	85.17	90.68
LMVEC	71.32	82.94	80.80	86.74	87.51	89.94
MTVEC	74.90	83.31	82.41	87.64	89.62	90.94
MTCELL	75.91	85.14	84.33	88.80	90.01	90.85
MTBOTH	**77.11**	**86.33**	**85.77**	**89.04**	**90.06**	**91.03**

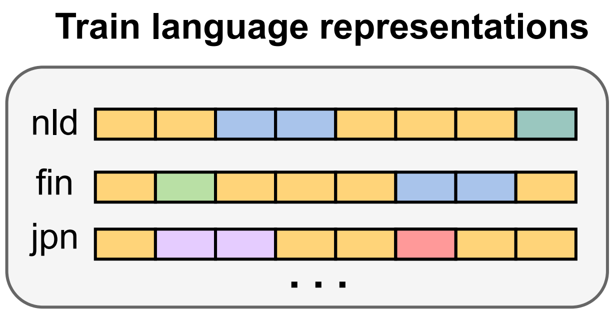
Table 1: Accuracy of syntactic, phonological, and inventory features using LM language vectors (LMVEC), MT language vectors (MTVEC), MT encoder cell averages (MTCELL) or both MT feature vectors (MTBOTH). Aux indicates auxiliary information of geodesic/genetic nearest neighbors; "NONE -Aux" is the majority class chance rate, while "NONE +Aux" is a 3-NN classification.

Model interpretability with typology

“We performed **10-fold cross-validation over the URIEL database**, where we train on 9/10 of the languages to predict 1/10 of the languages for 10 folds over the data.”

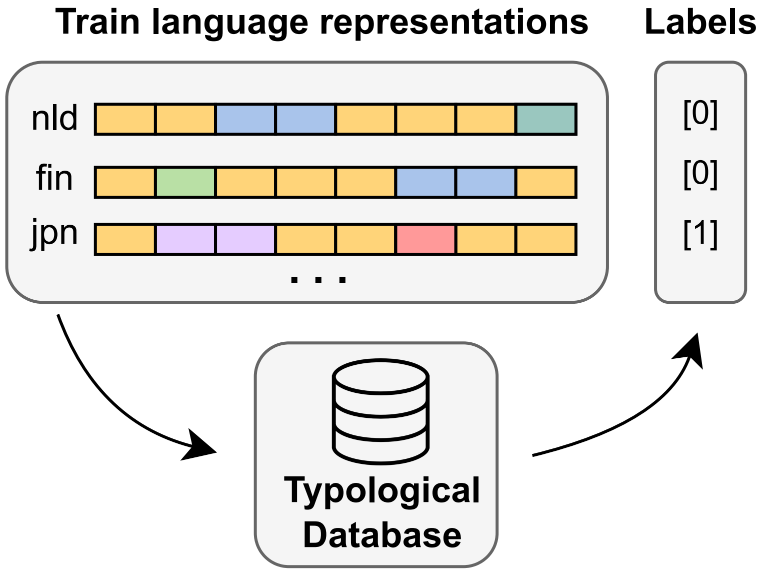
Model interpretability with typology

“We performed **10-fold cross-validation over the URIEL database**, where we train on 9/10 of the languages to predict 1/10 of the languages for 10 folds over the data.”



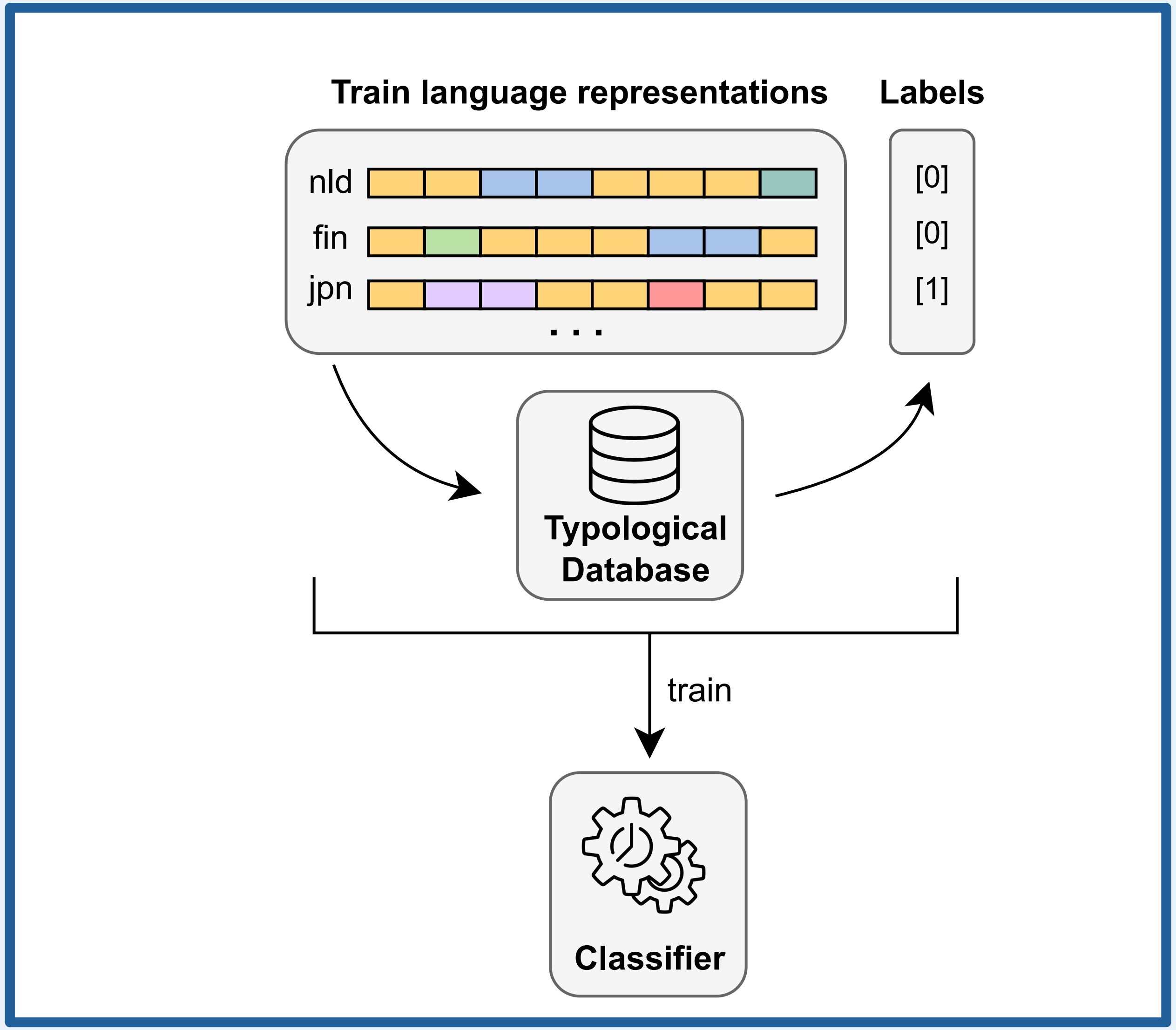
Model interpretability with typology

“We performed **10-fold cross-validation over the URIEL database**, where we train on 9/10 of the languages to predict 1/10 of the languages for 10 folds over the data.”



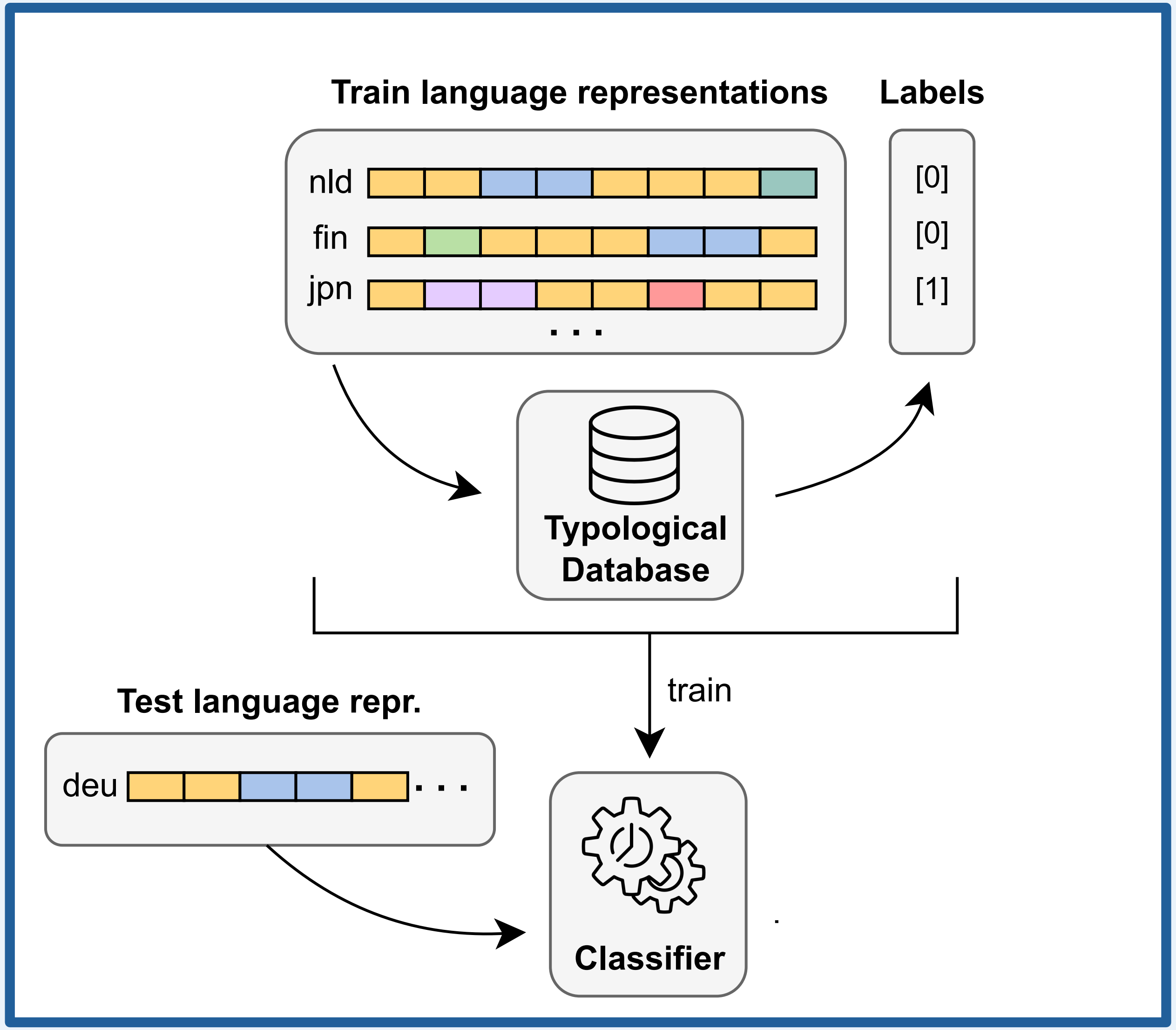
Model interpretability with typology

"We performed **10-fold cross-validation over the URIEL database**, where we train on 9/10 of the languages to predict 1/10 of the languages for 10 folds over the data."



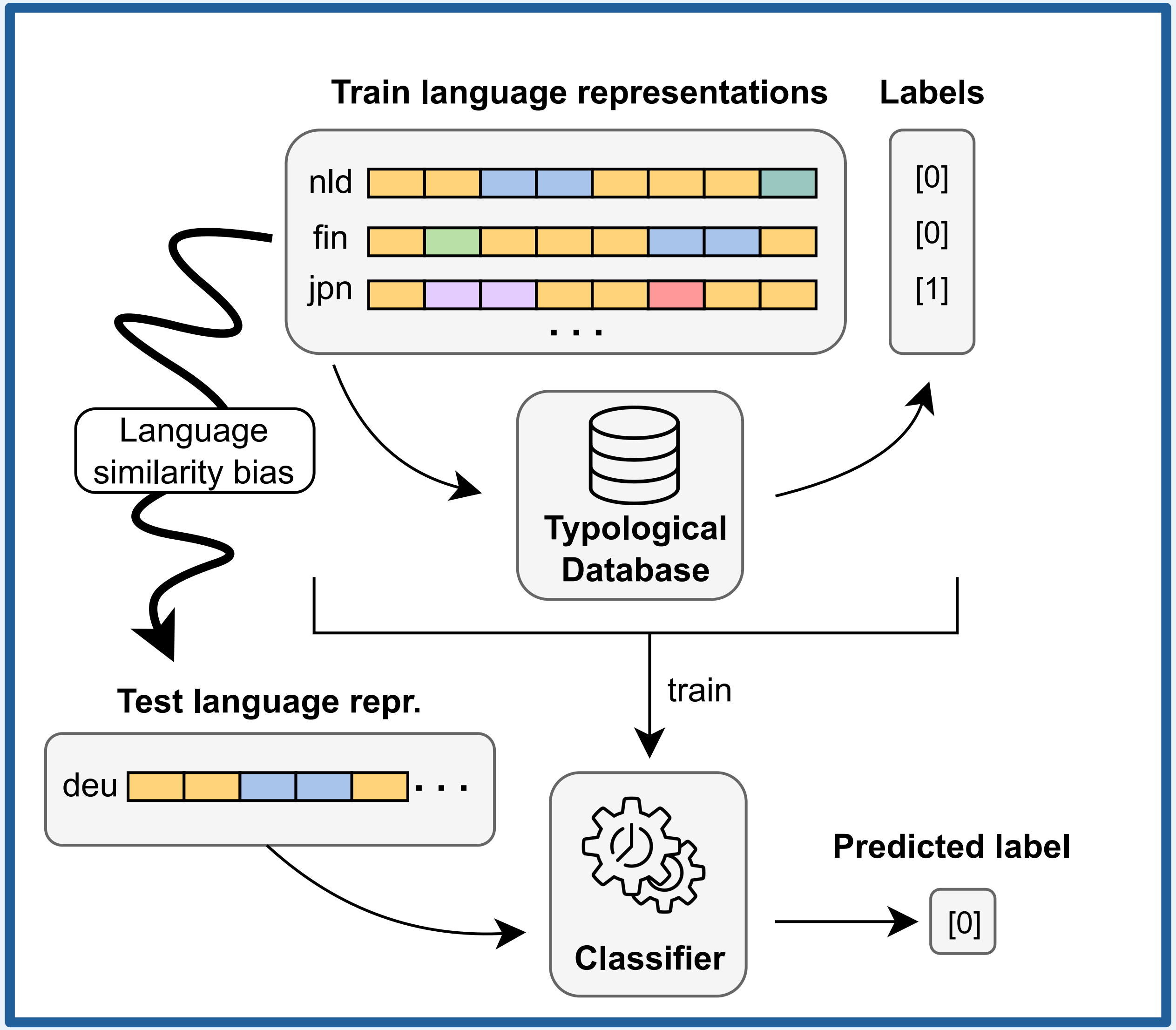
Model interpretability with typology

"We performed **10-fold cross-validation over the URIEL database**, where we train on 9/10 of the languages to predict 1/10 of the languages for 10 folds over the data."



Model interpretability with typology

"We performed **10-fold cross-validation over the URIEL database**, where we train on 9/10 of the languages to predict 1/10 of the languages for 10 folds over the data."



Model interpretability with typology

Östling & Kurfalı (2023): Linguistically Sound Cross-validation

- Do not train and then test on languages from the same family, macroarea and consider long-distance contact

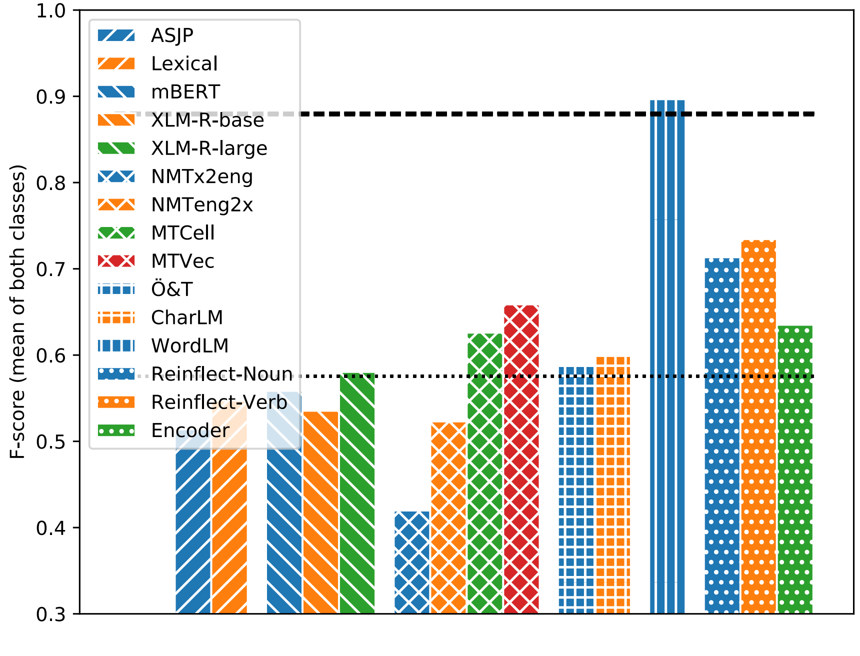
- Minimize the impact of lexical similarity through family-wise Monte Carlo sampling

Östling, R., & Kurfalı, M. (2023). Language embeddings sometimes contain typological generalizations. Computational Linguistics, 49(4), 1003-1051

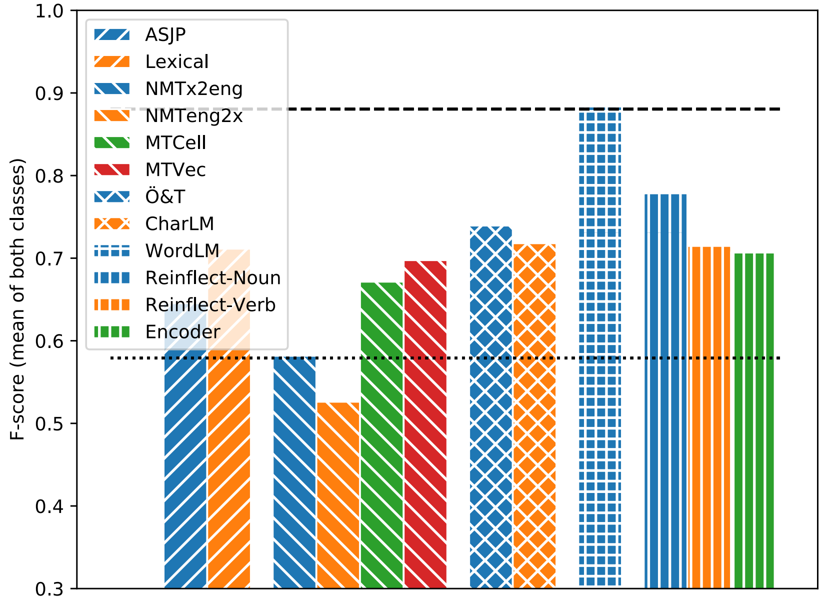
Model interpretability with typology

Östling & Kurfalı (2023): Linguistically Sound Cross-validation

- Do not train and then test on languages from the same family, macroarea and consider long-distance contact
- Minimize the impact of lexical similarity through family-wise Monte Carlo sampling



(a) Order of object and verb, using gold standard labels for training.



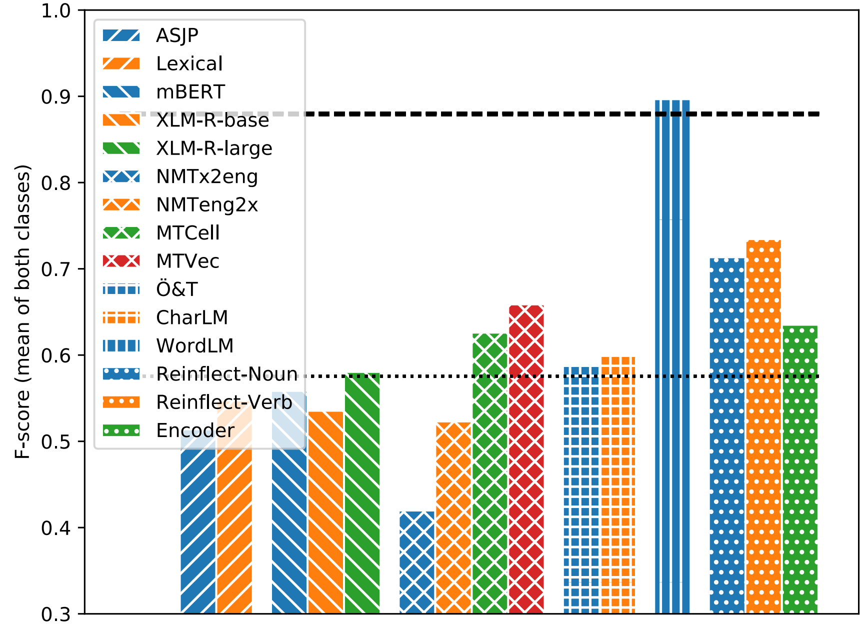
(a) Order of object and verb, using gold standard labels for training and naive cross-validation.

Östling, R., & Kurfalı, M. (2023). Language embeddings sometimes contain typological generalizations. Computational Linguistics, 49(4), 1003-1051

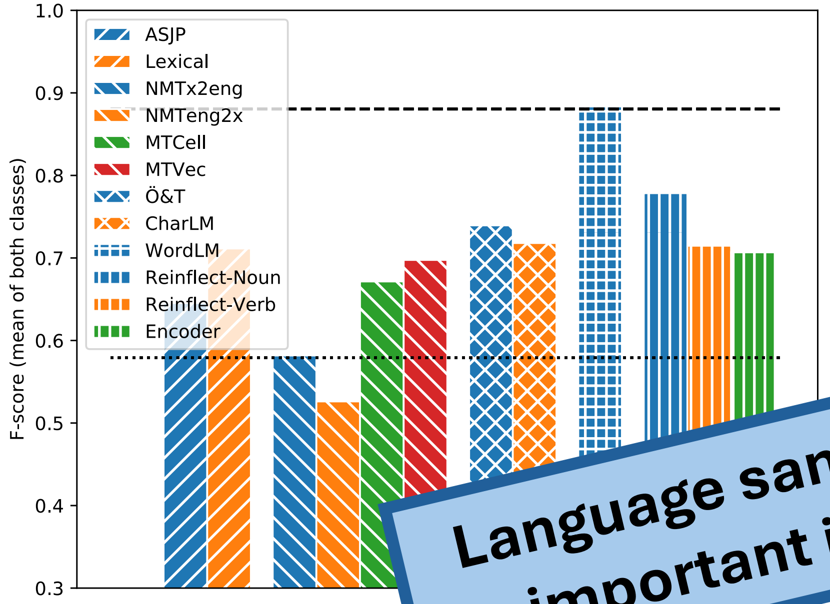
Model interpretability with typology

Östling & Kurfalı (2023): Linguistically Sound Cross-validation

- Do not train and then test on languages from the same family, macroarea and consider long-distance contact
- Minimize the impact of lexical similarity through family-wise Monte Carlo sampling



(a) Order of object and verb, using gold standard labels for training.



(a) Order of object and verb, using gol… training and naive cross-validation.

Language sampling is important in NLP!

Östling, R., & Kurfalı, M. (2023). Language embeddings sometimes contain typological generalizations. Computational Linguistics, 49(4), 1003-1051

Model interpretability with typology

Probing classifiers:

Hewitt & Liang (2019): Designing and Interpreting Probes with Control Tasks

- Do the representations encode linguistic structure or does probe just learn the linguistic task?
- Control tasks
- “A good probe should be *selective*, achieving high linguistic task accuracy and low control task accuracy.”

Designing and Interpreting Probes with Control Tasks (Hewitt & Liang, EMNLP-IJCNLP 2019)

Linguistic Typology in NLP

~~Interpreting~~

Using linguistic typology for the **Evaluating** of multilingual language models

Improving

The Past, Present, and Future of Typological Databases in NLP (Baylor et al., Findings 2023)

Typologically fair multilingual evaluation

A Call for Consistency in Reporting Typological Diversity

Wessel Poelman*♣ Esther Ploeger*♠ Miryam de Lhoneux♣ Johannes Bjerva♠
♣Department of Computer Science, KU Leuven, Belgium
♠Department of Computer Science, Aalborg University, Denmark
{wessel.poelman,miryam.delhoneux}@kuleuven.be {espl,jbjerva}@cs.aau.dk

1 Introduction

In order to draw generalizable conclusions about the performance of multilingual models across languages, it is important to evaluate on a set of languages that captures linguistic diversity. Linguistic typology is increasingly used to justify language selection, inspired by language sampling in linguistics (e.g., Rijkhoff and Bakker, 1998). In other words, more and more papers suggest generalizability by evaluating on 'typologically diverse languages' (see Figure 1). However, justifications for 'typological diversity' exhibit great variation, as there seems to be no set definition, methodology or consistent link to linguistic typology. In this work, we provide a systematic insight into how previous work in the ACL Anthology uses the term 'typological diversity'. Our two main findings are:

1. What is meant by typologically diverse language selection is not consistent.
2. The actual typological diversity of the language sets in these papers varies greatly.

We argue that, when making claims about 'typological diversity', an operationalization of this should be included. A systematic approach that quantifies this claim, also with respect to the number of languages used, would be even better.

2 Systematic Annotation of Claims

We systematically investigate which papers make claims regarding typological diversity, and which languages they actually use. First, we retrieve[1] all papers in the ACL Anthology that contain the following search string in either the title or abstract:

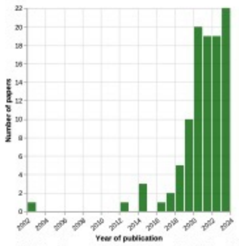


Figure 1: Number of papers in the ACL Anthology claiming a 'typologically diverse' set of languages over the years.

```
typological.+?diverse|
typological.+?diversity|
diverse.+?typological
```

Examples of this are not only *typologically diverse*, but also *typologically maximally diverse language* and *typologically and genetically diverse languages*. In total, this retrieves 140 papers, with the earliest being published in 2002, and the most recent being published in 2023. It contains papers from conferences (e.g., *ACL, EMNLP), journals (e.g., TACL, CL) and workshops (e.g., SIGTYP, SIGMORPHON).

We manually annotate whether these papers contain a claim regarding the typological diversity of their language selection. An example of such a claim is: "*we evaluate on a set of ten typologically diverse languages*" (Pimentel et al., 2020). A paper does not make a claim if it describes related work that claims to use 'a diverse typological test set', for instance. Our annotation is done separately by two annotators (the first two authors). We calculate inter-annotator agreement and retrieve a Cohen's κ of 0.64 ('substantial agreement'). After resolving the disagreements, we are left with 103 papers that

* Equal contribution.

[1] Using the acl-anthology-py package: https://github.com/mbollmann/acl-anthology-py. Papers retrieved on December 11, 2023.

What is 'Typological Diversity' in NLP?

Esther Ploeger*♠ Wessel Poelman*♣ Miryam de Lhoneux♣ Johannes Bjerva♠
♠Department of Computer Science, Aalborg University, Denmark
♣Department of Computer Science, KU Leuven, Belgium
{espl,jbjerva}@cs.aau.dk {wessel.poelman,miryam.delhoneux}@kuleuven.be

Abstract

The NLP research community has devoted increased attention to languages beyond English, resulting in considerable improvements for multilingual NLP. However, these improvements only apply to a small handful of the world's languages. Aiming to extend this, an increasing number of papers aspires to enhance *generalizable* multilingual performance *across languages*. To this end, linguistic typology is commonly used to motivate language selection, on the basis that a broad typological sample ought to imply generalization across a broad range of languages. These selections are often described as being 'typologically diverse'. In this work, we systematically investigate NLP research that includes claims regarding 'typological diversity'. We find that there are no set definitions or criteria for such claims. We introduce metrics to approximate the diversity of language selection along several axes and find that the results vary considerably across papers. Furthermore, we show that skewed language selection can lead to overestimated multilingual performance. We recommend that future work includes an operationalization of 'typological diversity', empirically justifying the diversity of language samples.

github.com/WPoelman/typ-div

1 Introduction

Most research in the field of natural language processing (NLP) is conducted on the English language (Ruder et al., 2022). Competitive monolingual language modelling beyond English remains challenging, as current state-of-the-art methods rely on the availability of large amounts of data, which are not available for most other languages (Joshi et al., 2020). This data sparsity can be mitigated by leveraging cross-lingual transfer through the training of a language model on multilingual data.

* Equal contribution.

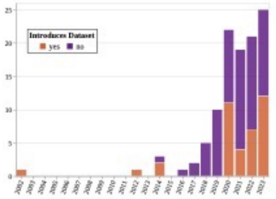


Figure 1: Number of papers with 'typological diversity' claims published by year.

Despite the potential of multilingual language modelling, common methodologies are primarily developed for English. There clearly is no guarantee that an approach that works for one language will work equally well for others (Gerz et al., 2018). For instance, morphologically complex languages can be over-segmented by current widely-used tokenization methods (Rust et al., 2021). Evaluation on a broad range of languages is important for drawing more generalizable conclusions about the performance of multilingual language technology. For instance, including only morphologically simple languages such as English can give an unrealistic image of the effectiveness of a tokenization method, simply because morphologically simple languages are generally easier to tokenize compared to complex ones. Current work increasingly evaluates models on multiple languages, but because of practical and data constraints, it is not realistic to test a model on the thousands of languages in the world.

In order to still ensure a degree of generalizability, previous work recognizes the importance of diverse language sampling. Ponti et al. (2020) suggest that merely evaluating on a small set of similar languages is an unreliable method for estimating a multilingual model's performance, since such

A Principled Framework for Evaluating on Typologically Diverse Languages

Esther Ploeger
Aalborg University
Department of Computer Science
espl@cs.aau.dk

Wessel Poelman
KU Leuven
Department of Computer Science
email@email.com

Andreas Holck Høeg-Petersen
Aalborg University
Department of Computer Science
email@email.com

Anders Schlichtkrull
Aalborg University
Department of Computer Science
email@email.com

Miryam de Lhoneux
KU Leuven
Department of Computer Science
email@email.com

Johannes Bjerva
Aalborg University
Department of Computer Science
email@email.com

Beyond individual languages, multilingual NLP research increasingly aims to develop models that perform well across languages. However, evaluating these systems on all the world's languages is practically infeasible. To attain generalizability, representative language sampling is essential. Previous work argues that generalizable multilingual evaluation sets should contain languages with diverse typological properties. However, 'typologically diverse' language samples have been found to vary considerably in this regard, and popular sampling methods are flawed and inconsistent. We present a language sampling framework for selecting the most typologically diverse languages given a sampling frame. Our approach accommodates multiple sampling objectives from linguistic typology, and is evaluated with a range of metrics. We find that our systematic sampling method consistently retrieves more typologically diverse language selections than previous methods. Moreover, we provide additional evidence that this affects generalizability in multilingual model evaluation, emphasizing the importance of diverse language sampling.

1. Introduction

Multilingual natural language processing (NLP) has seen major improvements in the last decade. Pre-trained language models such as multilingual BERT (Devlin et al. 2019), XLM-R (Conneau et al. 2020) and mT5 () facilitate cross-lingual transfer into languages for which there are limited or no monolingual models available. This has made them increasingly popular in few-shot or zero-shot scenarios. More recently, multilingual

Typologically fair multilingual evaluation

- Multilingual NLP increasingly aims at generalizability across languages
- Recent work implies generalizability by claiming to rely on *linguistic typology*

Typologically fair multilingual evaluation

- Multilingual NLP increasingly aims at generalizability across languages
- Recent work implies generalizability by claiming to rely on *linguistic typology*

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages

Jack FitzGerald* Christopher Hench Charith Peris
Scott Mackie
Aaron Nash
Richa Singh
Misha Britan

TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages

Eunsol Choi Michael Collins Dan Garrette
Vitaly Nikolaev Jennimaria Palomaki
Google Research
tydiqa@google.com

TyDiP: A Dataset for Politeness Classification in Nine Typologically Diverse Languages

Anirudh Srinivasan Eunsol Choi
Department of Computer Science
The University of Texas at Austin
{anirudhs, eunsol}@utexas.edu

Abstract

(Choi et al., 2018), and the Natural Questions (NQ) (Kwiatkowski et al., 2019).

However, many people who might benefit from QA systems do not speak English. The languages of the world exhibit an astonishing breadth of linguistic phenomena used to express meaning; the World Atlas of Language Structures (Comrie and Gil, 2005; Dryer and Haspelmath,

Typologically fair multilingual evaluation

- Multilingual NLP increasingly aims at generalizability across languages
- Recent work implies generalizability by claiming to rely on *linguistic typology*

“We evaluate on 12 typologically diverse languages.”

Typologically fair multilingual evaluation

- Multilingual NLP increasingly aims at generalizability <u>across</u> languages
- Recent work implies generalizability by claiming to rely on *linguistic typology*

“We evaluate on 12 typologically diverse languages.”

What does this mean?

Typologically fair multilingual evaluation

Data collection

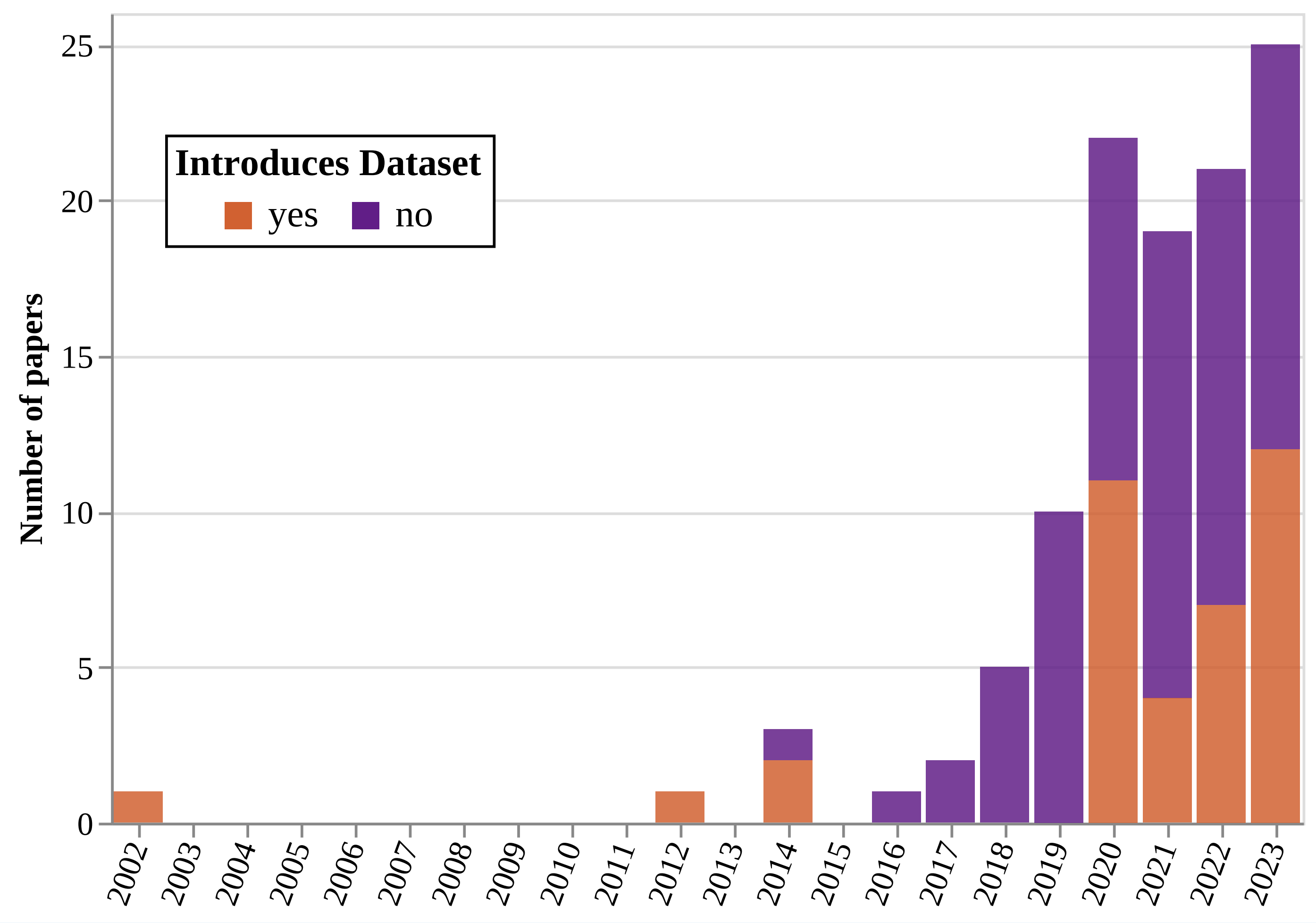
1. Retrieve papers that contain typological diversity* in their title or abstract

2. Annotate whether the paper claims that a language set is typologically diverse. If so:
   - Does it introduce a **new dataset**?
   - **Which languages** does it contain?

We retrieve **194** papers, of which **110** contain a claim of typological diversity.

Typologically fair multilingual evaluation

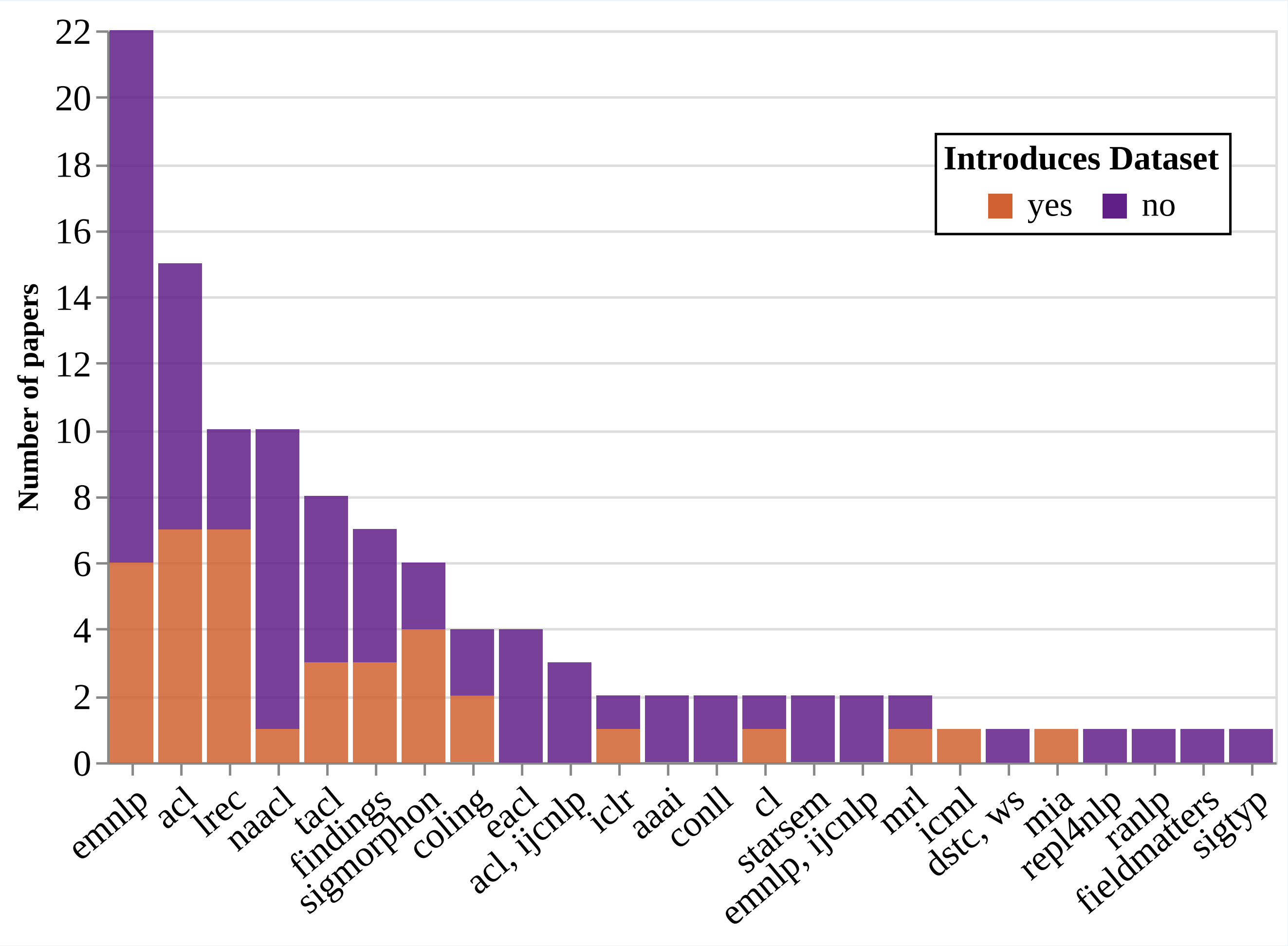
A high-level overview



Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

A high-level overview



Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

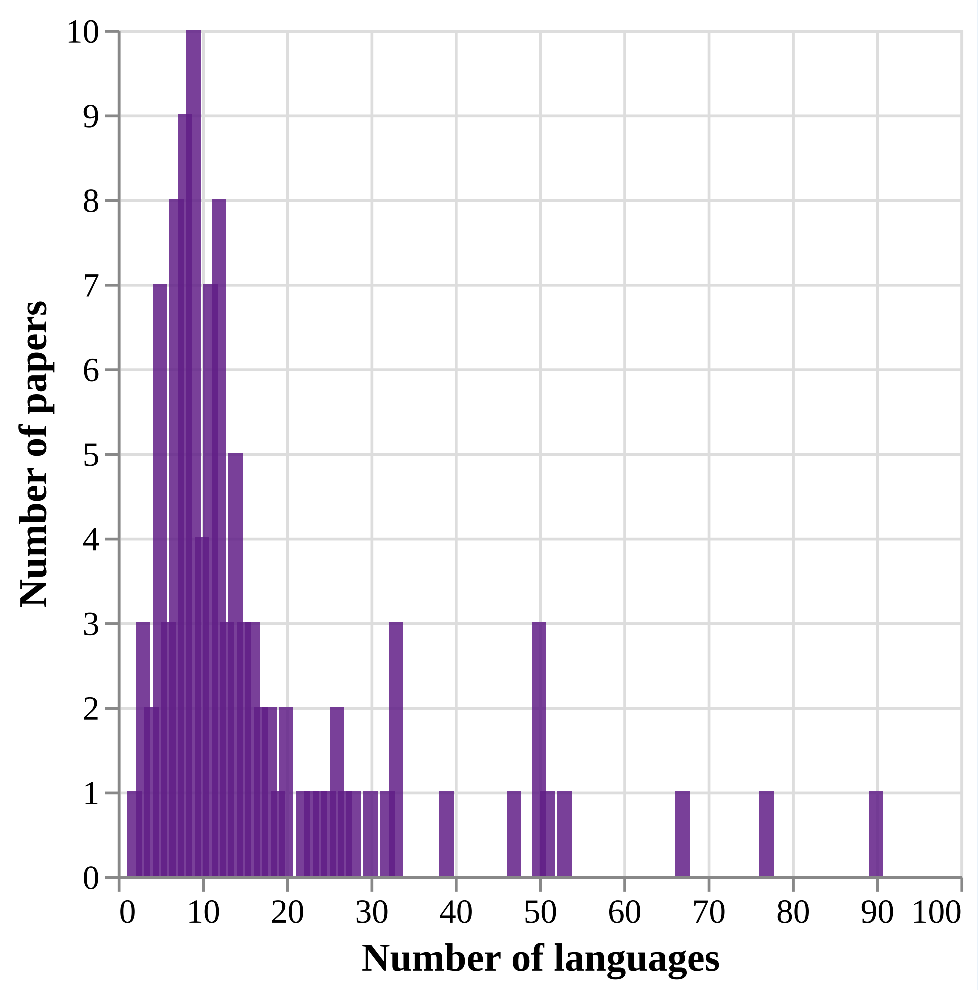
Typologically fair multilingual evaluation

A high-level overview

Median: 11
Q1: 8
Q3: 18

Minimum: 2
Maximum: 90

Total languages: 315



Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

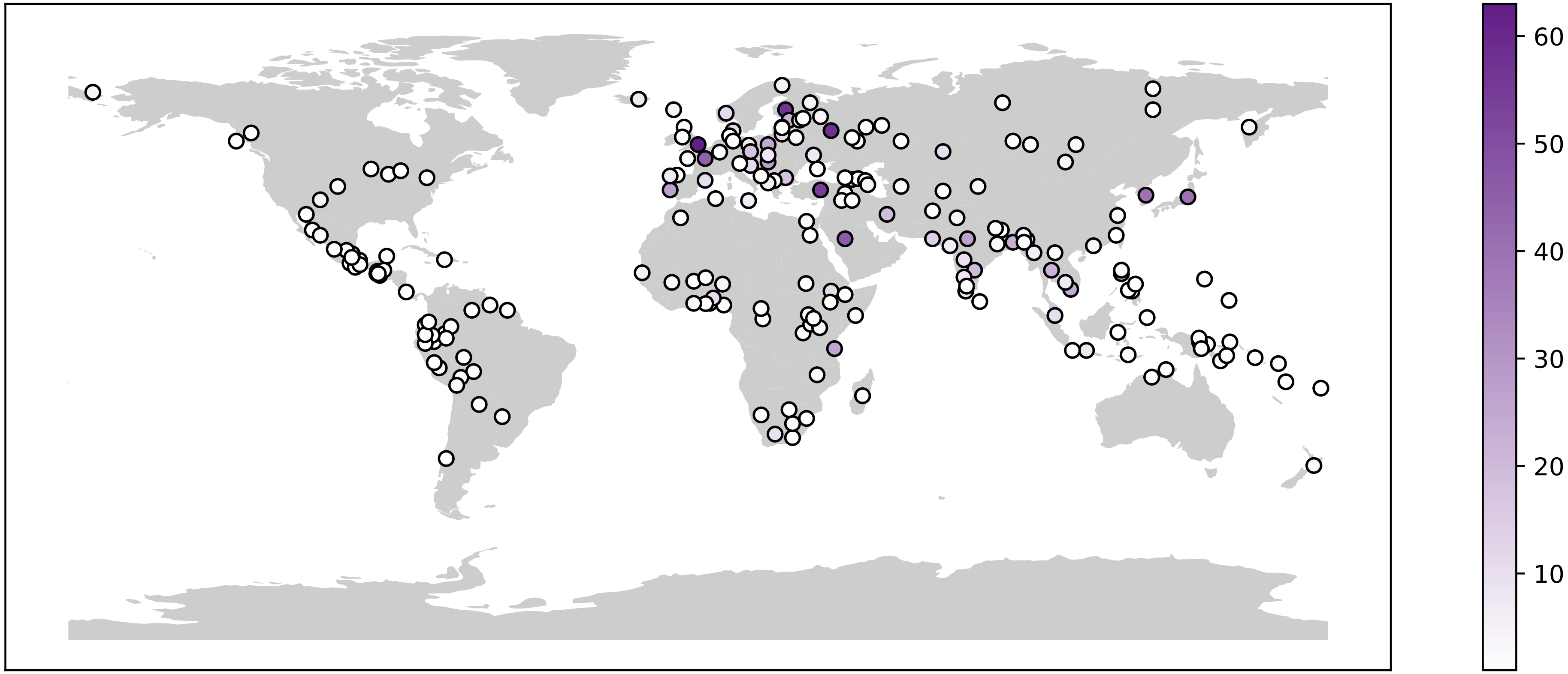


Figure 4: Map of languages in all papers claiming ‘typological diversity’, where the hue corresponds number of papers that uses a language. Coordinates are taken from WALS.

Typologically fair multilingual evaluation

Justifications

No justification

No information on the sampling criteria or method

Genealogical groupings as a proxy

Xu et al. (2022) aim to cover “a reasonable variety of language families”

Often post-hoc

(Some) typological features

Jancso et al. (2020): clustering with typological databases

Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

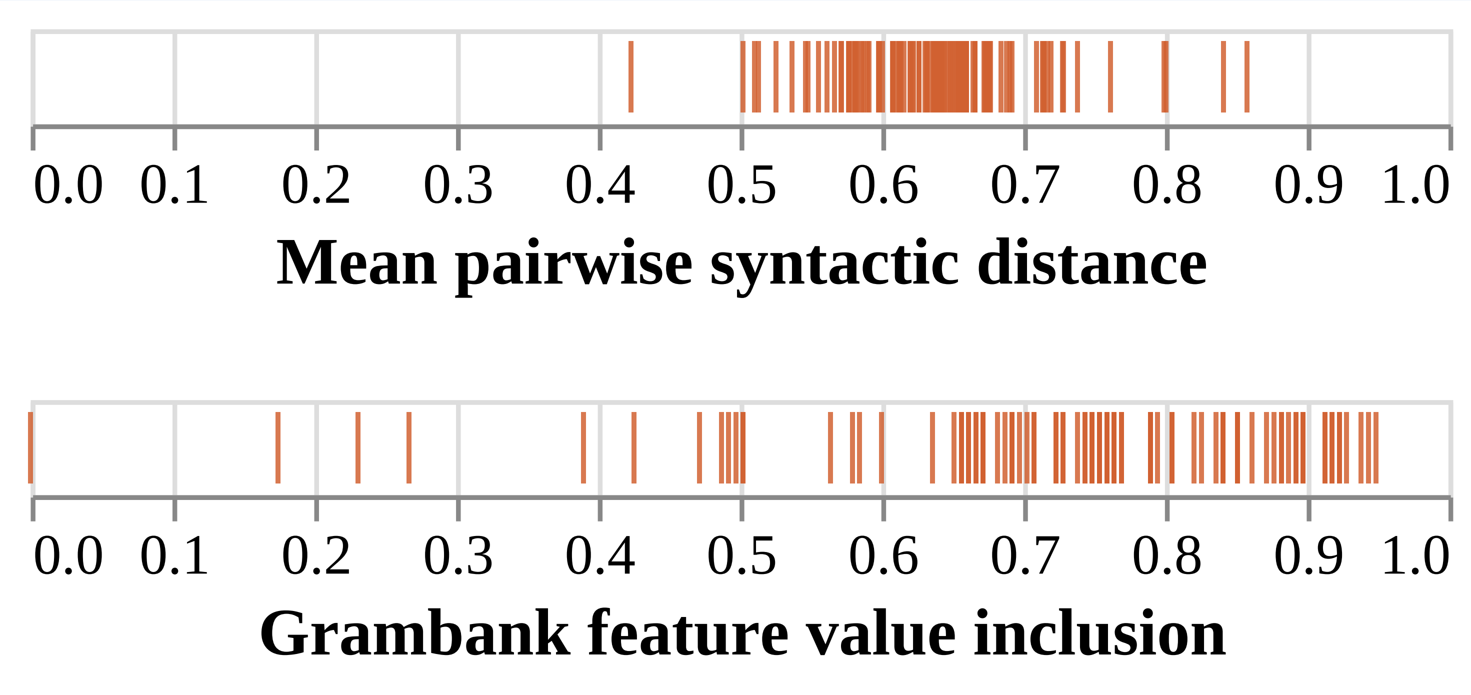
Typologically fair multilingual evaluation

What about the actual ‘typological diversity’?

Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

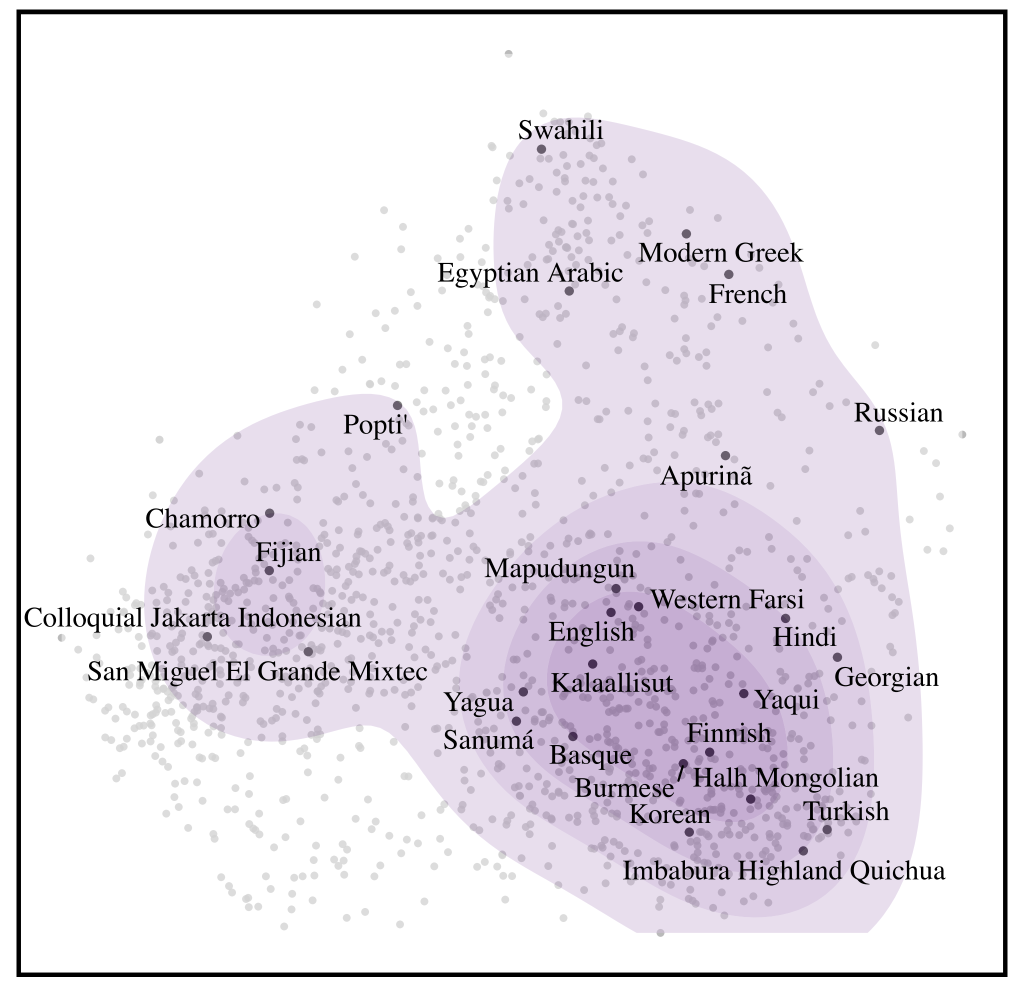
Approximations of typological diversity



Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

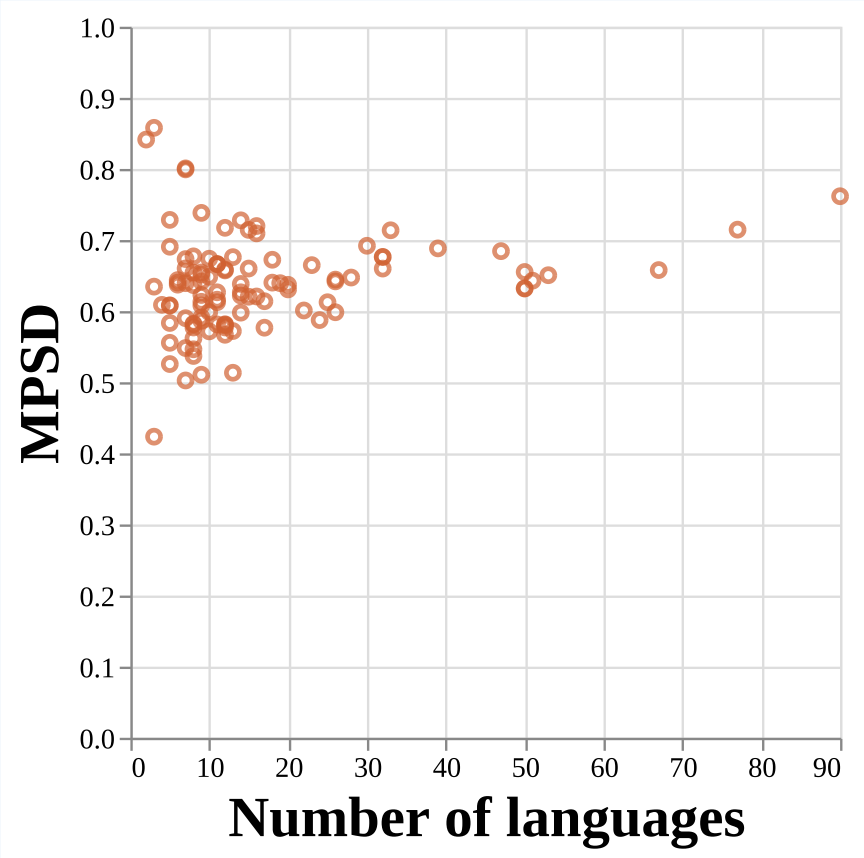
In the best* case

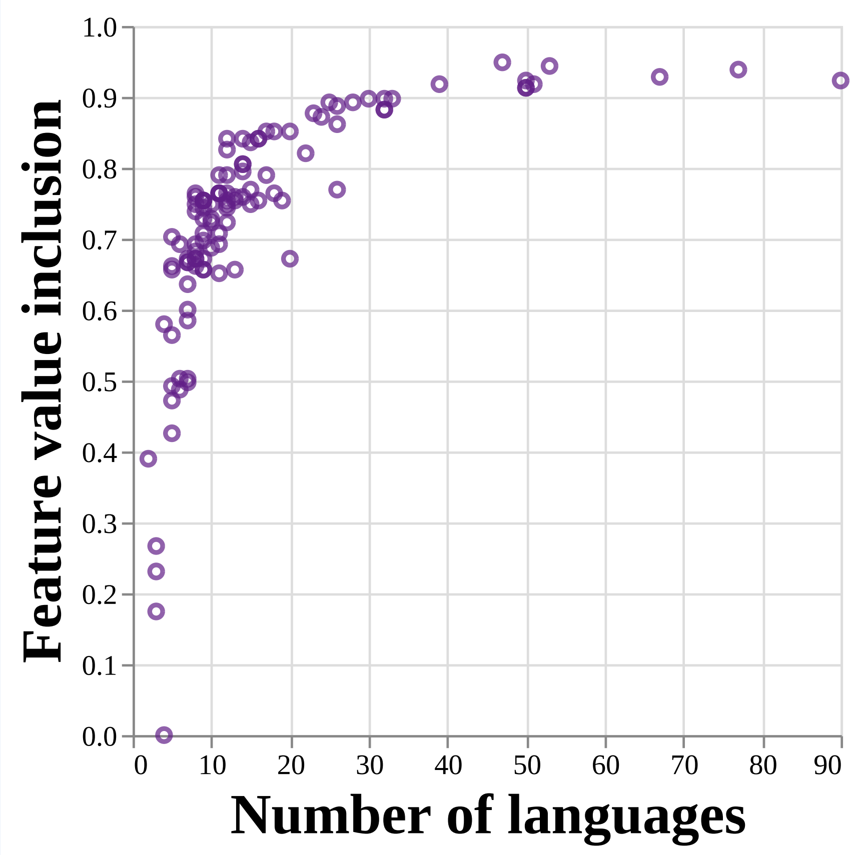


Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

The more languages, the better?

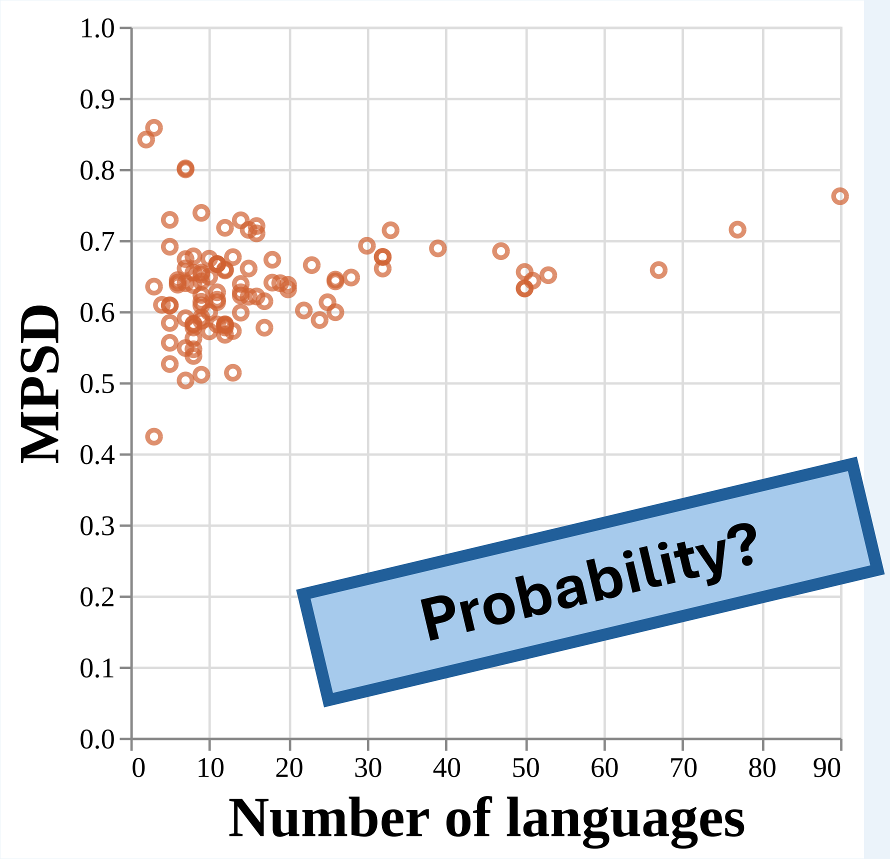


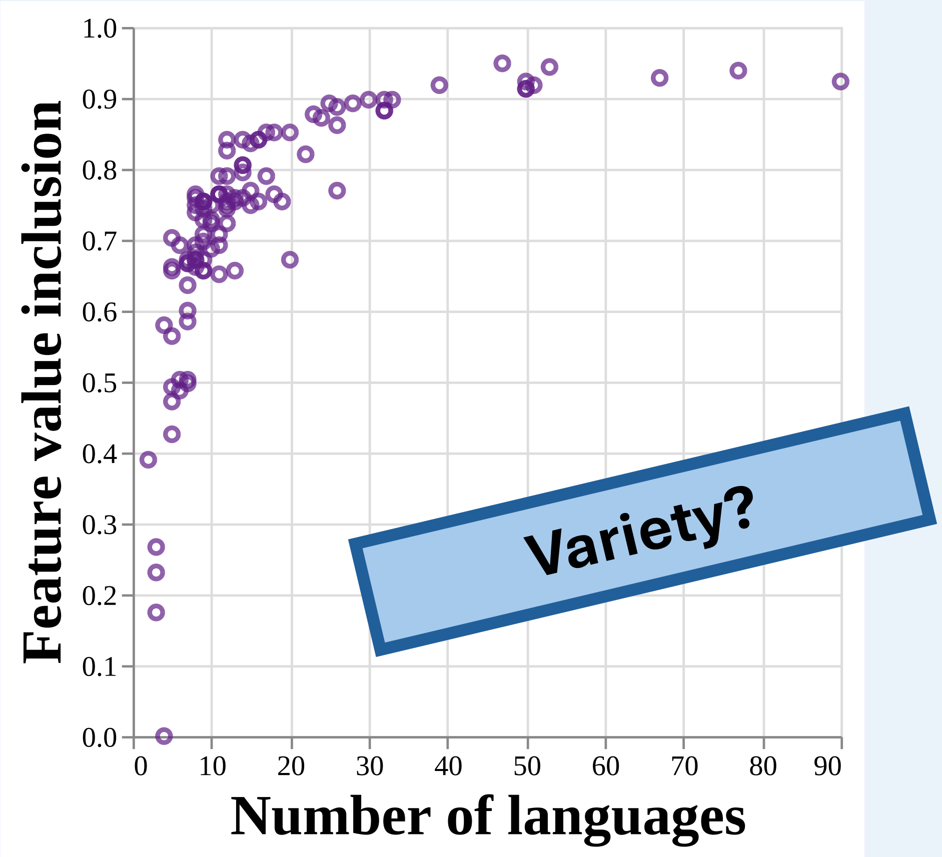


Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

The more languages, the better?





Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

What does this mean for evaluation?

Subtask	Model	**Overall**	**By F**	△	Strong Pre	Weak Pre	Equal Pre & Suf	Strong Suf	Weak Suf	Little Aff	NA
Mewsli-X★	XLM-R-L	45.75 (*11*)	36.23 (*11*)	-9.52	- (*0*)	- (*0*)	- (*0*)	47.86 (*10*)	24.60 (*1*)	- (*0*)	- (*0*)
	mBERT	38.58 (*11*)	27.29 (*11*)	-11.29	- (*0*)	- (*0*)	- (*0*)	41.09 (*10*)	13.50 (*1*)	- (*0*)	- (*0*)
XNLI♦	XLM-R	79.24 (*15*)	76.54 (*15*)	-2.70	- (*0*)	71.20 (*1*)	- (*0*)	80.06 (*12*)	- (*0*)	78.35 (2)	- (*0*)
	mBERT	66.51 (*15*)	60.17 (*15*)	-6.35	- (*0*)	49.30 (*1*)	- (*0*)	68.60 (*12*)	- (*0*)	62.60 (2)	- (*0*)
	mT5	84.85 (*15*)	82.92 (*15*)	-1.92	- (*0*)	80.60 (*1*)	- (*0*)	85.57 (*12*)	- (*0*)	82.60 (2)	- (*0*)

Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

When it comes to ‘typological diversity’ in NLP …

- There are no set definitions or criteria
- There is no consistent link with linguistic typology
- According to our approximations, the actual typological diversity varies considerably
- This can affect downstream evaluation

Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is' Typological Diversity'in NLP?. arXiv preprint arXiv:2402.04222.

Typologically fair multilingual evaluation

Ok... But how could we actually improve upon this?

Typologically fair multilingual evaluation

Ok... But how could we actually improve upon this?

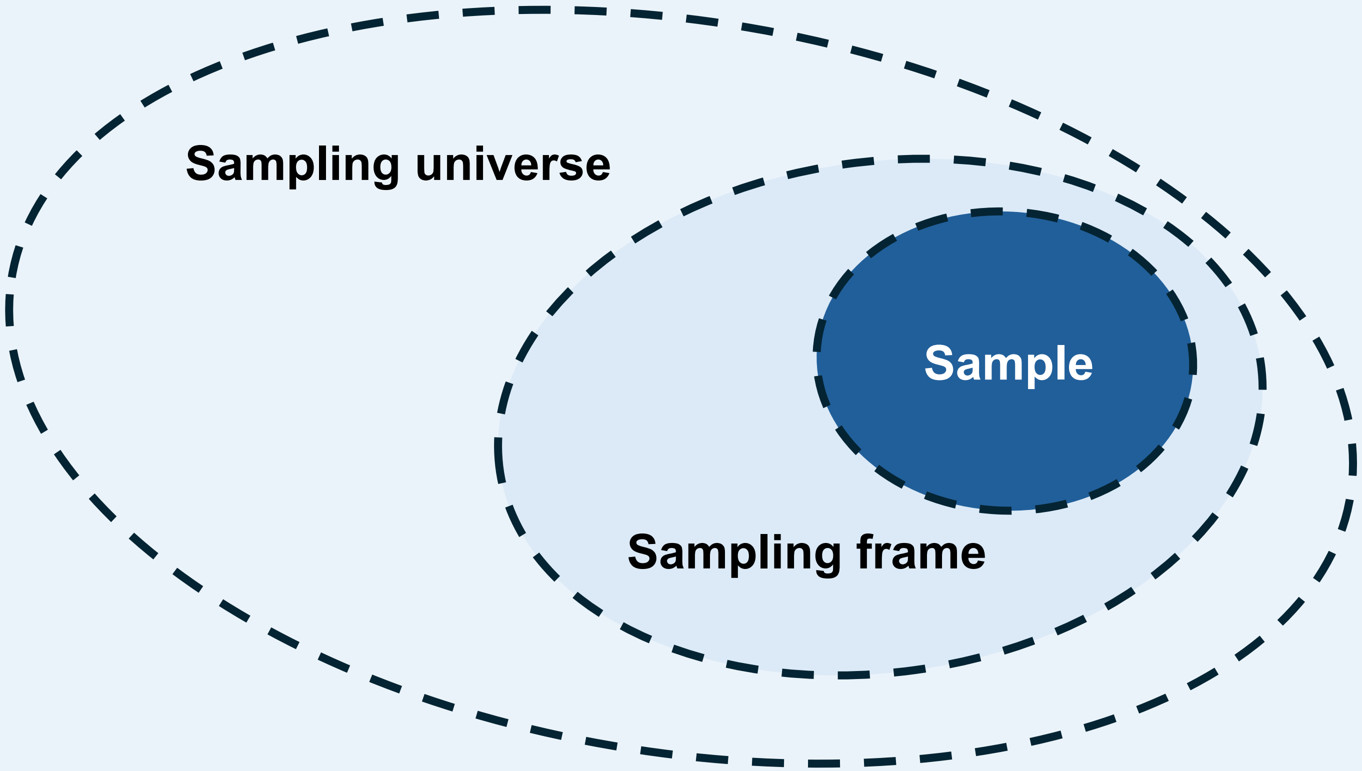
“A Principled Framework for Evaluating on Typologically Diverse Languages”

Typologically fair multilingual evaluation

Task: Select a given number of languages from a sampling frame, such that we maximize typological diversity

Typologically fair multilingual evaluation

Task: Select a given number of languages from a sampling frame, such that we maximize typological diversity



Bell, A. (1978). Language samples. Universals of human language, 1, 123-156.

Typologically fair multilingual evaluation

Linguistic typology (example):

- **Goal:** investigate relations between typological properties
- **Resources:** sample from diverse families and areas
- **Sampling methods:** random, variety or probability sampling

Typologically fair multilingual evaluation

Linguistic typology (example):

- **Goal:** investigate relations between typological properties
- **Resources:** sample from diverse families and areas
- **Sampling methods:** random, variety or probability sampling

Multilingual NLP (example):

- **Goal**: see how well a language model performs on typologically diverse languages
- **Resources:** sample from diverse families and areas

Typologically fair multilingual evaluation

Linguistic typology (example):

- **Goal:** investigate relations between typological properties
- **Resources:** sample from diverse families and areas
- **Sampling methods:** random, variety or probability sampling

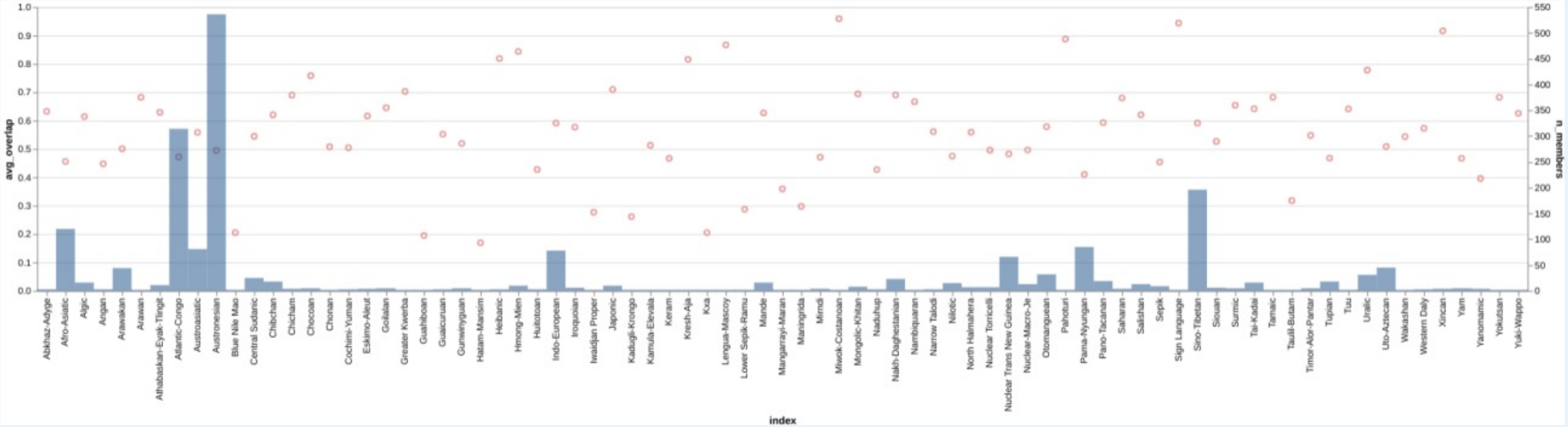
Multilingual NLP (example):

- **Goal**: see how well a language model performs on typologically diverse languages
- ***Resources:*** *sample from diverse families and areas*

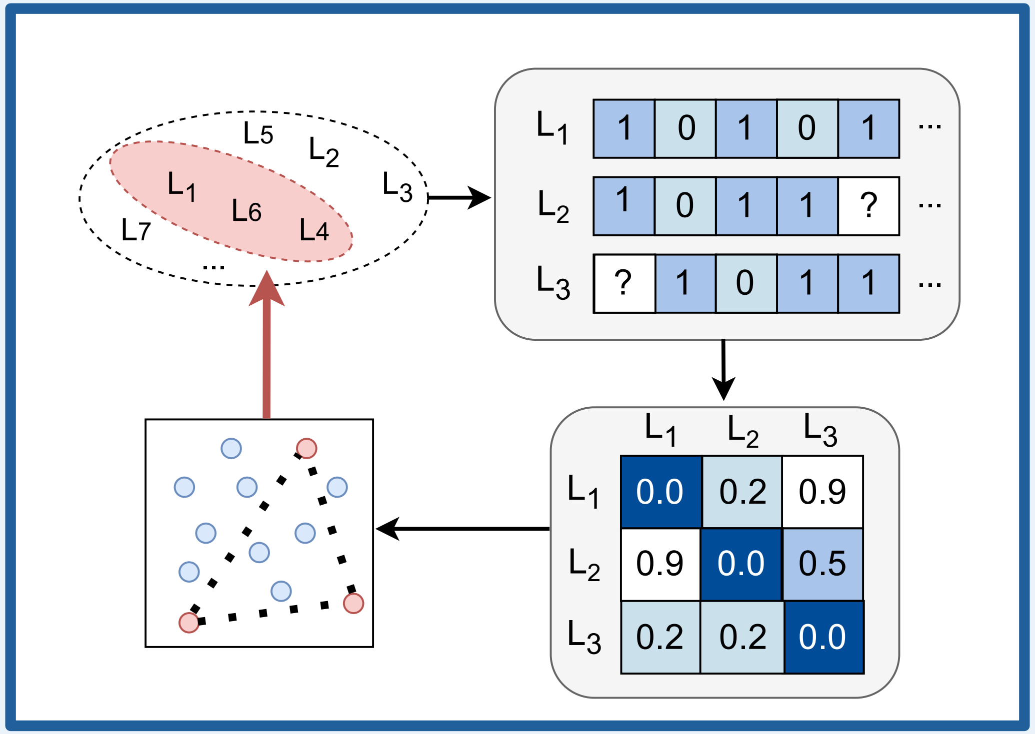
Actually… there is no circularity if we do not investigate typological features directly!

Typologically fair multilingual evaluation

Sampling with families is not ideal in many NLP scenarios!



Typologically fair multilingual evaluation



Typologically fair multilingual evaluation

Sampling algorithm objectives

MaxSum

MaxMin

Typologically fair multilingual evaluation

Sampling algorithm objectives

MaxSum

Sample k languages from N, where we iteratively add the next point that yields **the largest summed distance.**

MaxMin

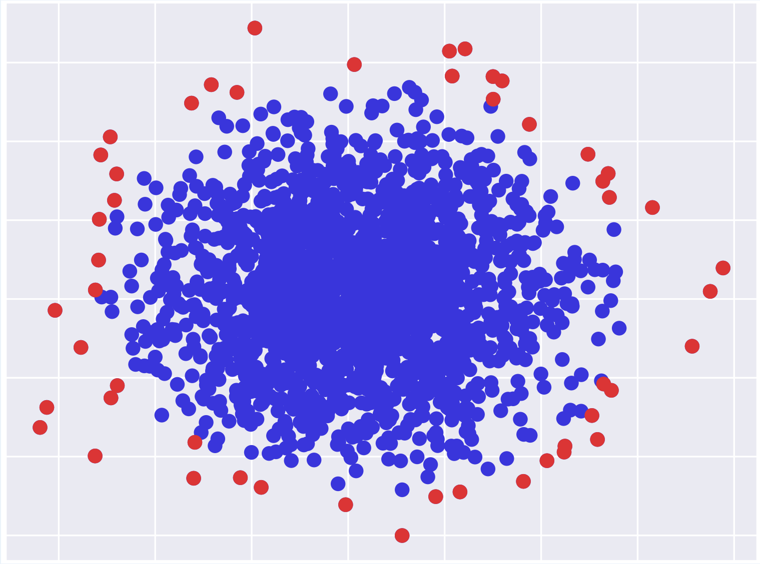
Sample k languages from N, where we iteratively add the next point that yields the **maximum minimum distance between any two points** in k.

Typologically fair multilingual evaluation

Sampling algorithm objectives

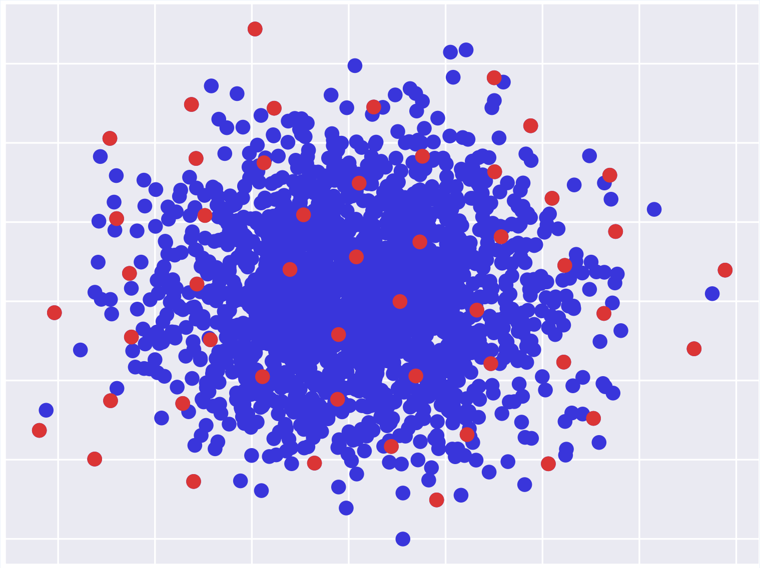
MaxSum

Sample k languages from N, where we iteratively add the next point that yields **the largest summed distance.**



MaxMin

Sample k languages from N, where we iteratively add the next point that yields the **maximum minimum distance between any two points** in k.

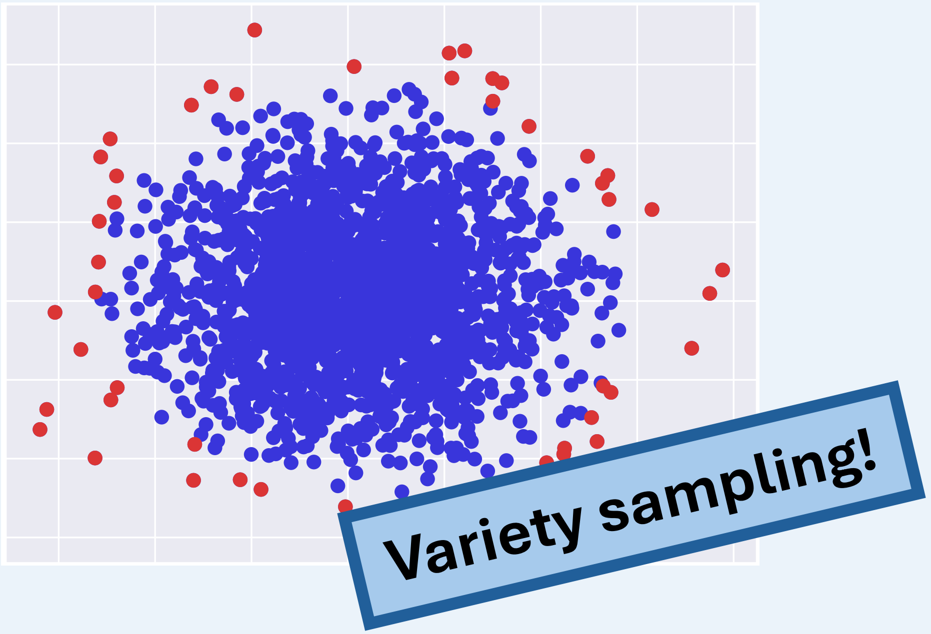


Typologically fair multilingual evaluation

Sampling algorithm objectives

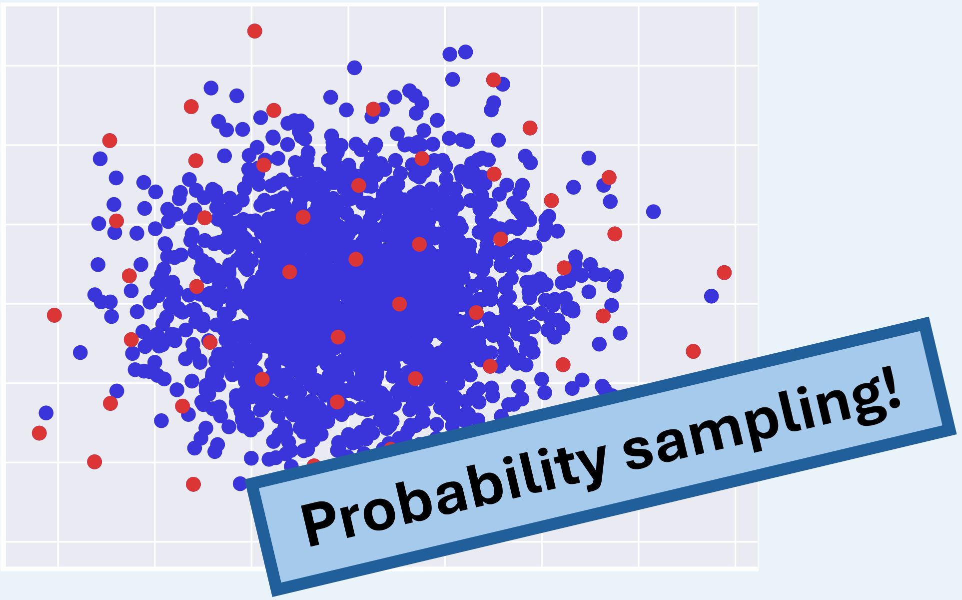
MaxSum

Sample k languages from N, where we iteratively add the next point that yields **the largest summed distance.**



MaxMin

Sample k languages from N, where we iteratively add the next point that yields the **maximum minimum distance between any two points** in k.

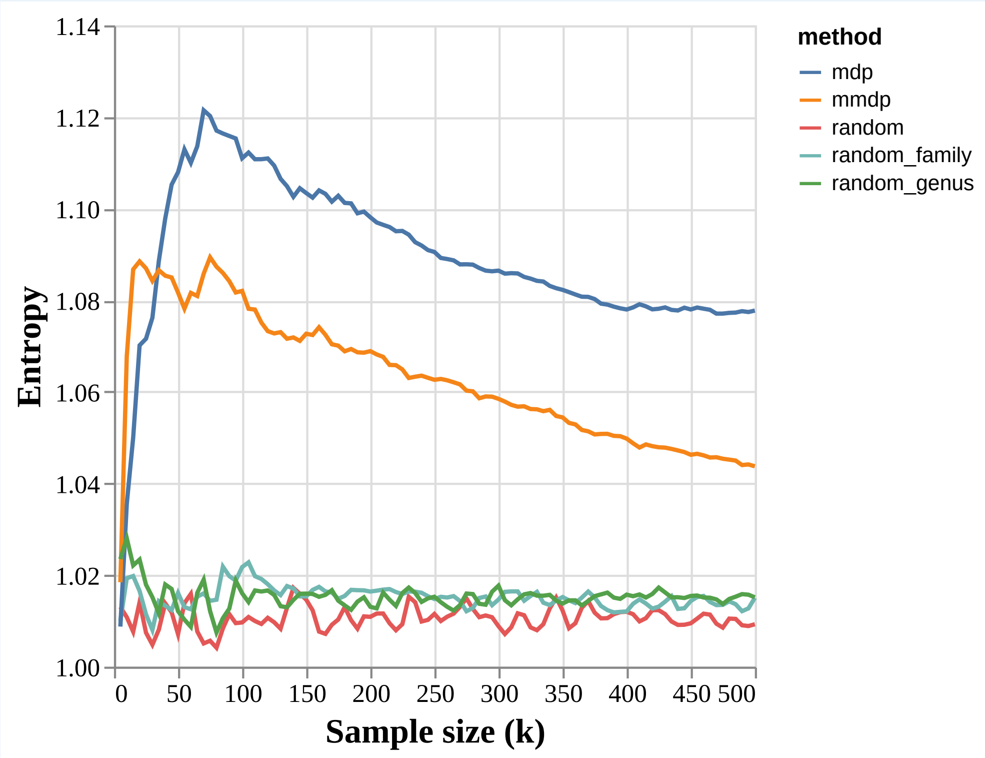


Typologically fair multilingual evaluation

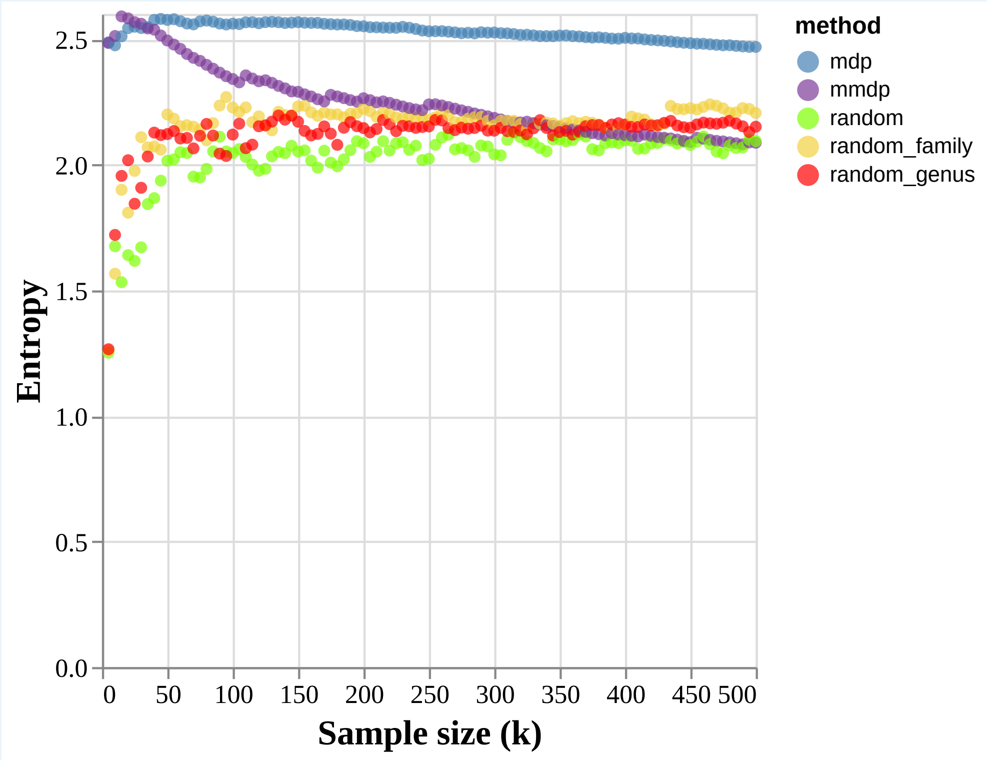
How do our typology-based sampling methods compare to genealogical baselines?

Typologically fair multilingual evaluation

How do our typology-based sampling methods compare to genealogical baselines?



Average per language



Average per feature

Linguistic Typology in NLP

Using linguistic typology for the

~~Interpreting~~

~~Evaluating~~

Improving

of multilingual language models

Improving multilingual NLP with typology?

What are relevant improvements given the current state of multilingual NLP?

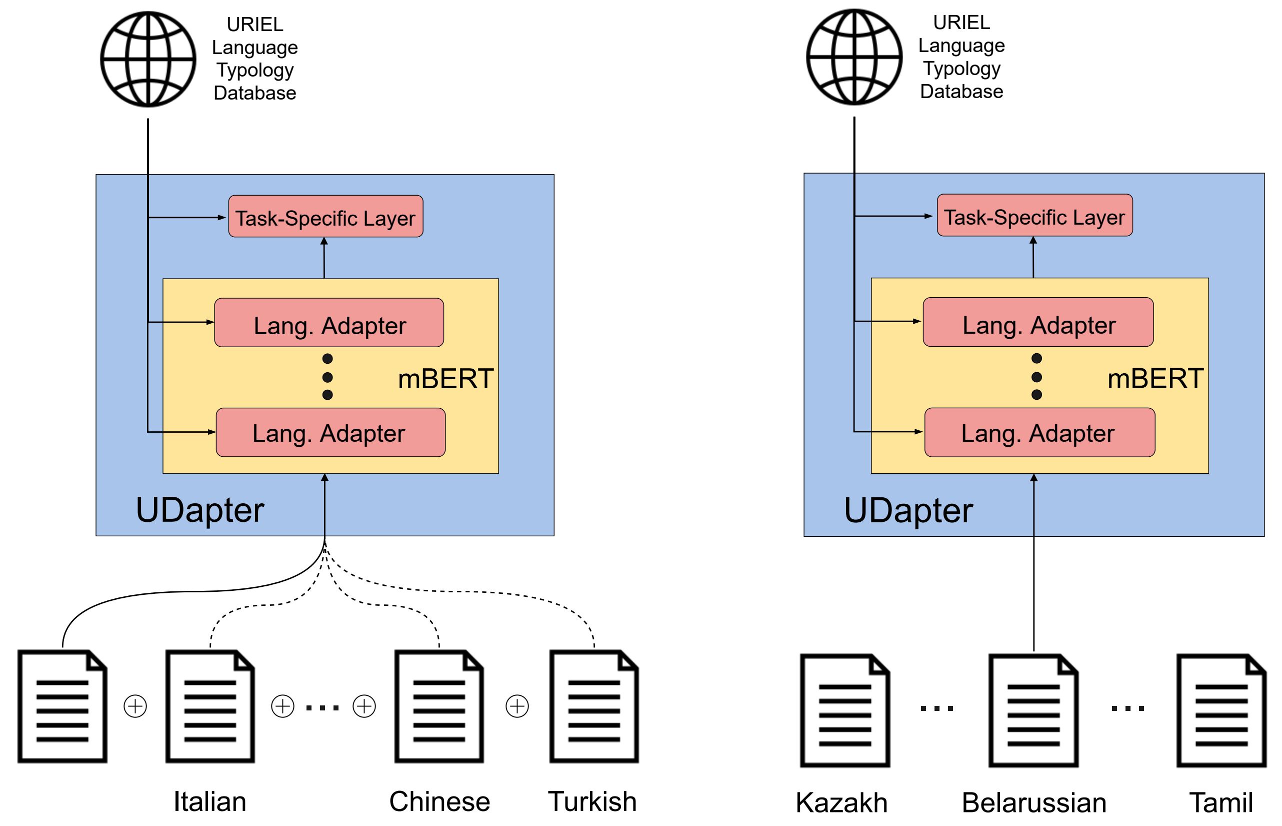
Improving multilingual NLP with typology?

What are relevant improvements given the current state of multilingual NLP?

- Low-resource scenarios
- Efficiency

Improving multilingual NLP with typology?

Üstun et al. (2022): UDapter

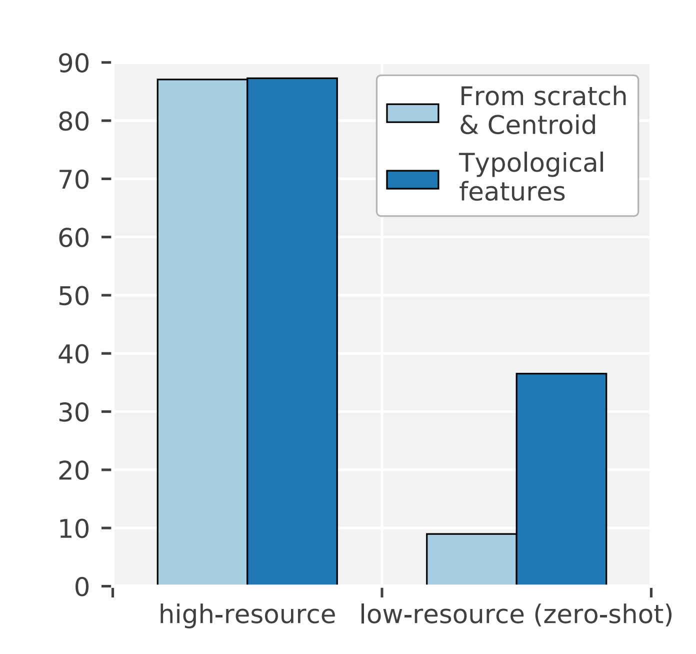


(a) Multilingual Fine-tuning on HR (b) Zero-shot Evaluation on LR

Üstün, A., Bisazza, A., Bouma, G., & Noord, G. V. (2022). UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. Computational Linguistics, 48(3).

Improving multilingual NLP with typology?

Üstun et al. (2022): UDapter

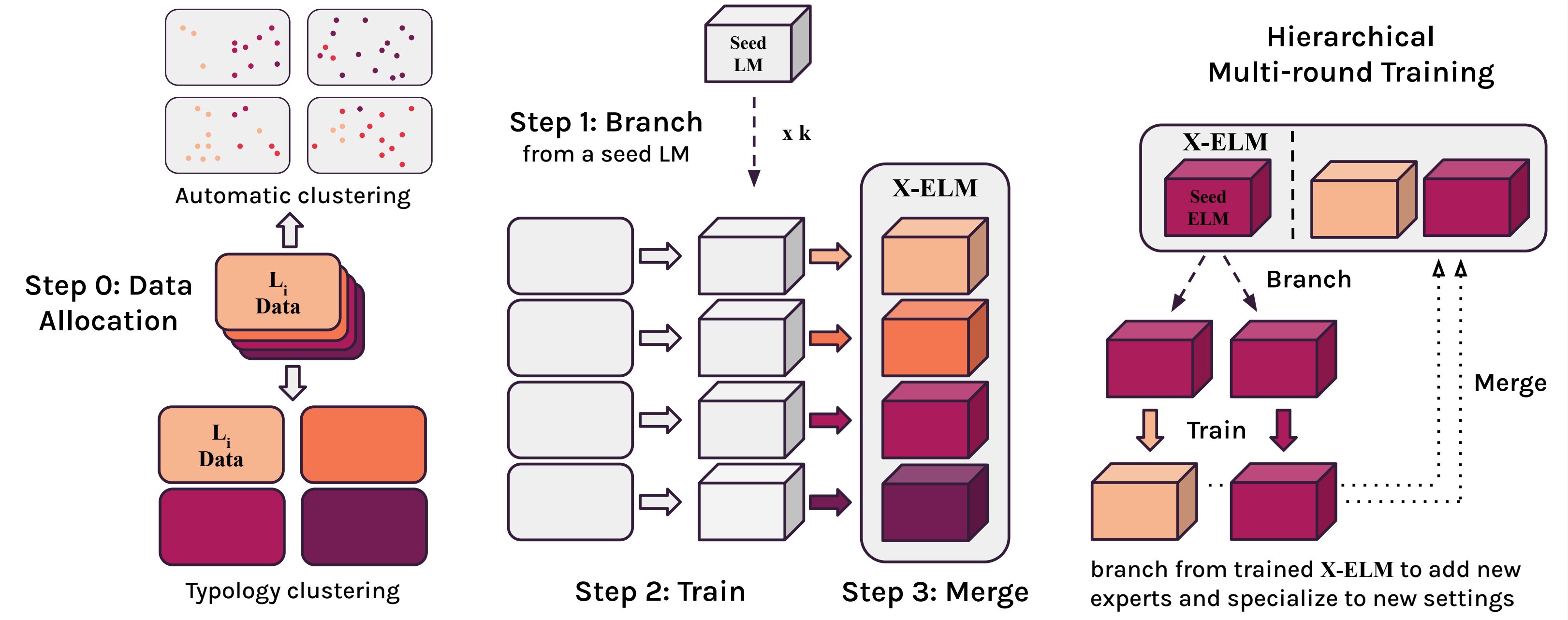


"The main limitation in our approach remains the low representation quality for languages with zero or little data in the pre-trained encoder (multilingual pre-training)."

Üstün, A., Bisazza, A., Bouma, G., & Noord, G. V. (2022). UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. Computational Linguistics, 48(3).

Improving multilingual NLP with typology?

Blevins et al. (2024): X-ELM



Improving multilingual NLP with typology?

Blevins et al. (2024): X-ELM

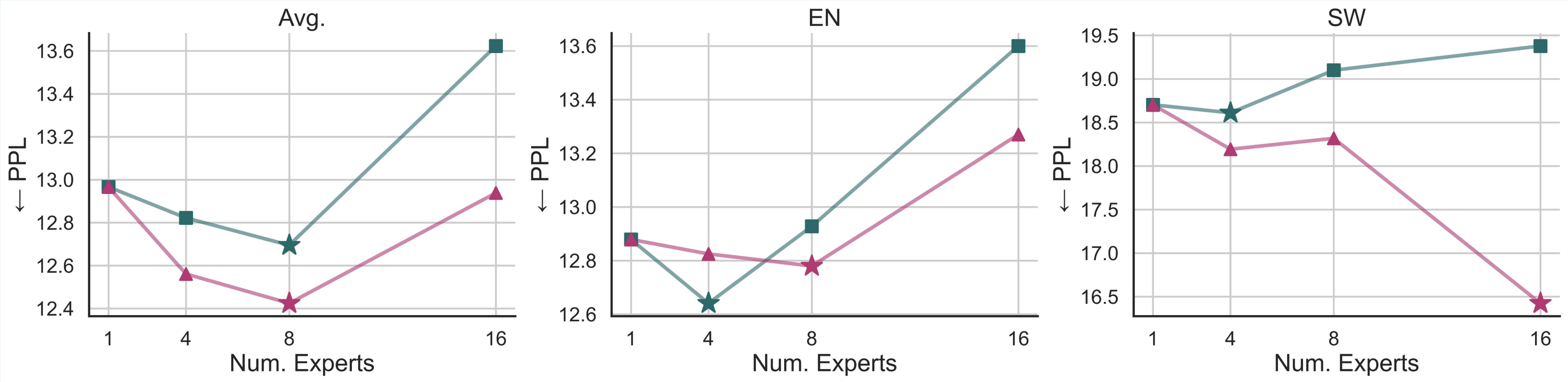


Figure 4: Average and language-specific (EN and SW) perplexities across expert counts (k) when clustering with TF-IDF$_{top1}$ (square) and Linguistic Typology (triangle). The best k for each setting is marked with a star.

Blevins, T., et al. (2024). Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models. arXiv preprint arXiv:2401.10440.

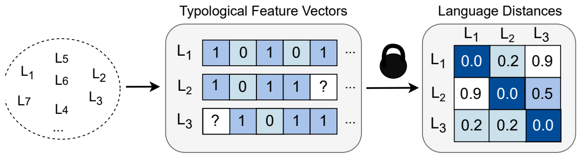
Improving multilingual NLP with typology?

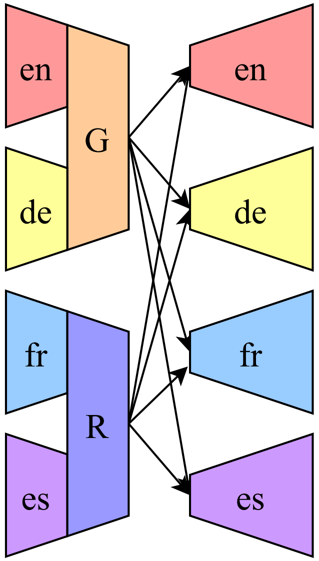
My project here so far

Main RQ:
(How) can we leverage typological groupings in MNMT to maximize transfer between related languages and minimize negative inference (to improve performance)?

Subquestions:
1. Are typological groupings more useful than genealogical groupings?
2. What is the effect of typologically-informed parameter sharing on source language interference?

Approach:





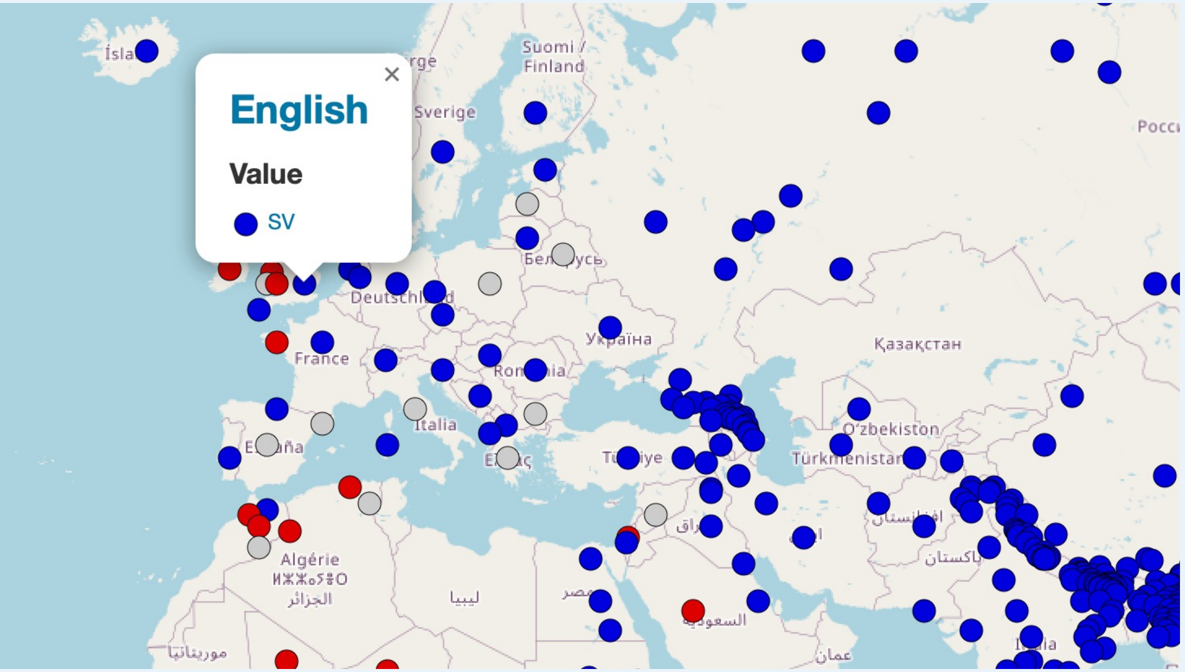
(a) Language group sharing

Current Issues and Solutions

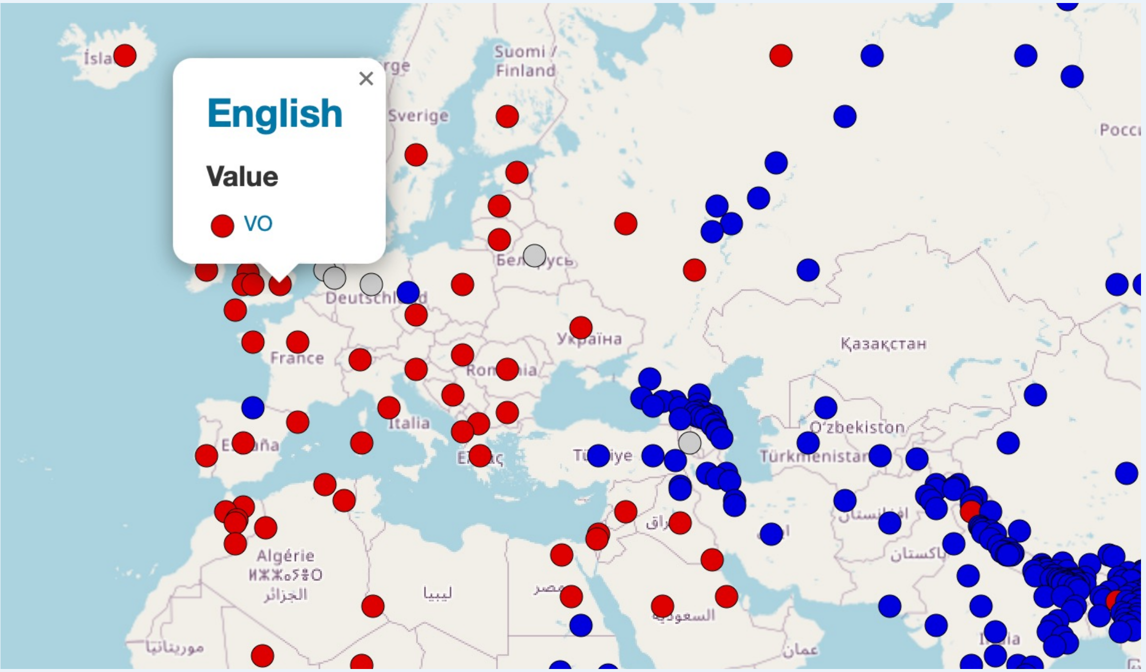
(Adapted from EACL 2024 talk)

Current issues

In typological databases, language are described with concrete datapoints



Feature 82A: Order of Subject and Verb



Feature 83A: Order of Object and Verb

Matthew S. Dryer. 2013. Order of Subject and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3)
Matthew S. Dryer. 2013. Order of Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3)

Current issues

In typological databases, language are described with concrete datapoints

SV:	"Multilingual NLP is challenging."
VS:	"Can we leverage this information for NLP?"

Current issues

In typological databases, language are described with concrete datapoints

“Word order variability should be regarded as a basic assumption, rather than as something exceptional.”

“Gradient approaches follow naturally from the emergentist usage-based view of languages …”

DE GRUYTER MOUTON Linguistics 2023; 61(4): 825–883

Review

Natalia Levshina*, Savithry Namboodiripad*, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova and Natalia Stoynova

Why we need a gradient approach to word order

https://doi.org/10.1515/ling-2021-0098
Received May 13, 2021; accepted April 9, 2022; published online April 25, 2023

***Corresponding authors: Natalia Levshina**, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands, E-mail: natalevs@gmail.com; and **Savithry Namboodiripad**,

Levshina, Natalia, et al. "Why we need a gradient approach to word order." *Linguistics* (2023).

Current issues

Does this matter for NLP?

Current issues

Does this matter for NLP?

- Language models are trained on text

- Ponti et al. (2019):

”this sort of gradient representation is also **more compatible with machine learning algorithms** and particularly with deep neural models that naturally operate with real-valued multi-dimensional word embeddings and hidden states.”

Ponti, E. M. et al. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, *45*(3), 559-601.

A Solution?

Contributions:

- A method for **retrieving gradient word order typology** from UD treebanks
- A **dataset** with continuous word order values
- A new typological **feature prediction task** with baseline results

A Solution?

Five word order features:

- Ordering of adjectives and their nouns
- Ordering of numerals and their nouns
- Ordering of subjects and verbs
- Ordering of objects and verbs
- Ordering of objects and subjects

A Solution?

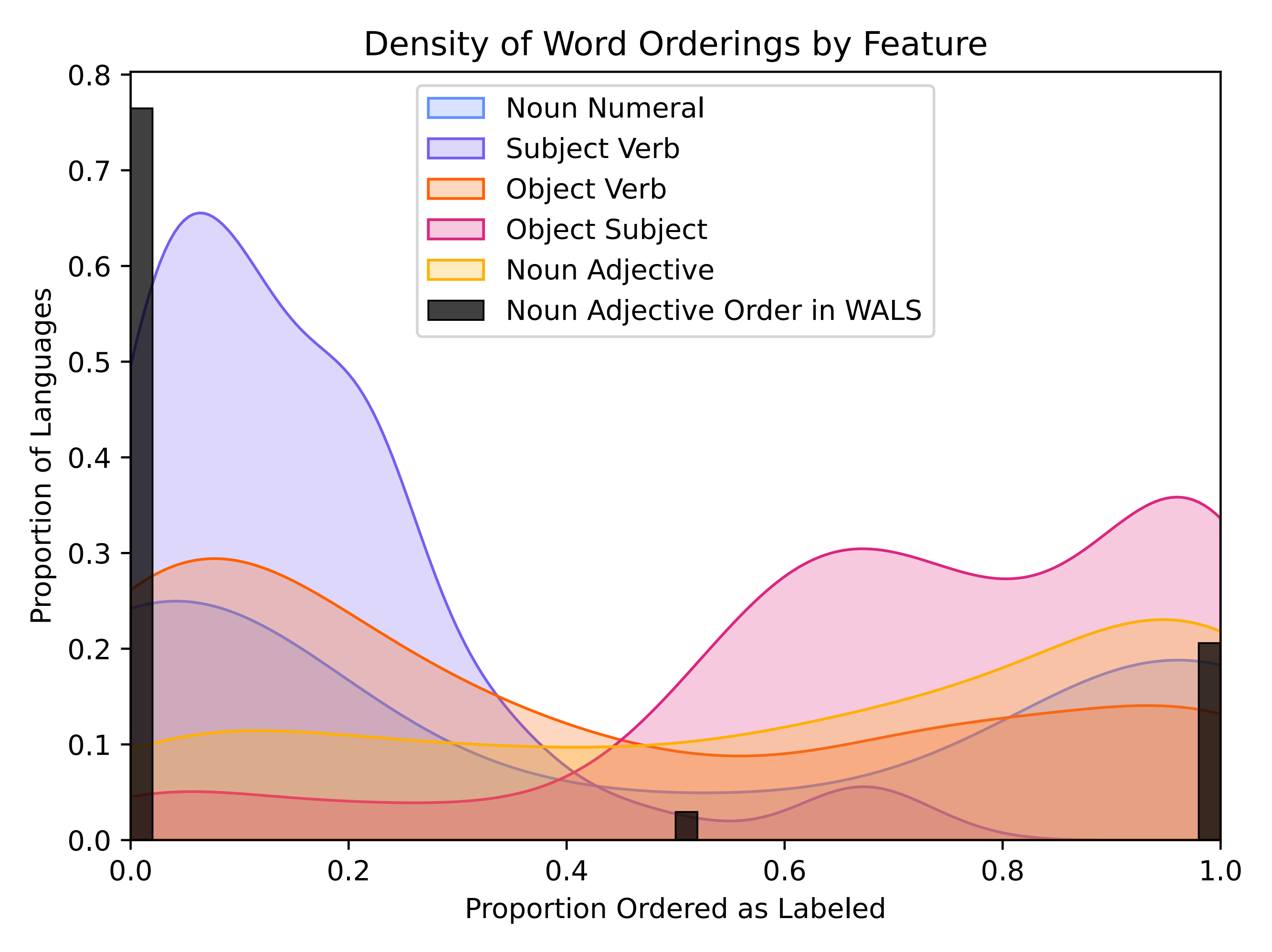
for all $d \in$ UD Datasets **do**
  $na \leftarrow 0$ $\triangleright$ na is the Noun-Adj count
  $an \leftarrow 0$ $\triangleright$ an is the Adj-Noun count
  for all sentence $s \in d$ **do**
    $na \leftarrow na+$ **count** Noun-Adj in s
    $an \leftarrow an+$ **count** Adj-Noun in s
  end for
  $na\_proportion \leftarrow \frac{na}{na+an}$
end for

POS	UD upos value	UD deprels value
Noun	NOUN	–
Adjective	ADJ	amod
Numeral	NUM	nummod
Subject	–	nsubj
Object	–	obj
Verb	VERB	–

Table 4: Tags used to extract the necessary parts of speech from the Universal Dependencies treebank (Nivre et al., 2020). Dashes indicate that that value did not need to be specified.

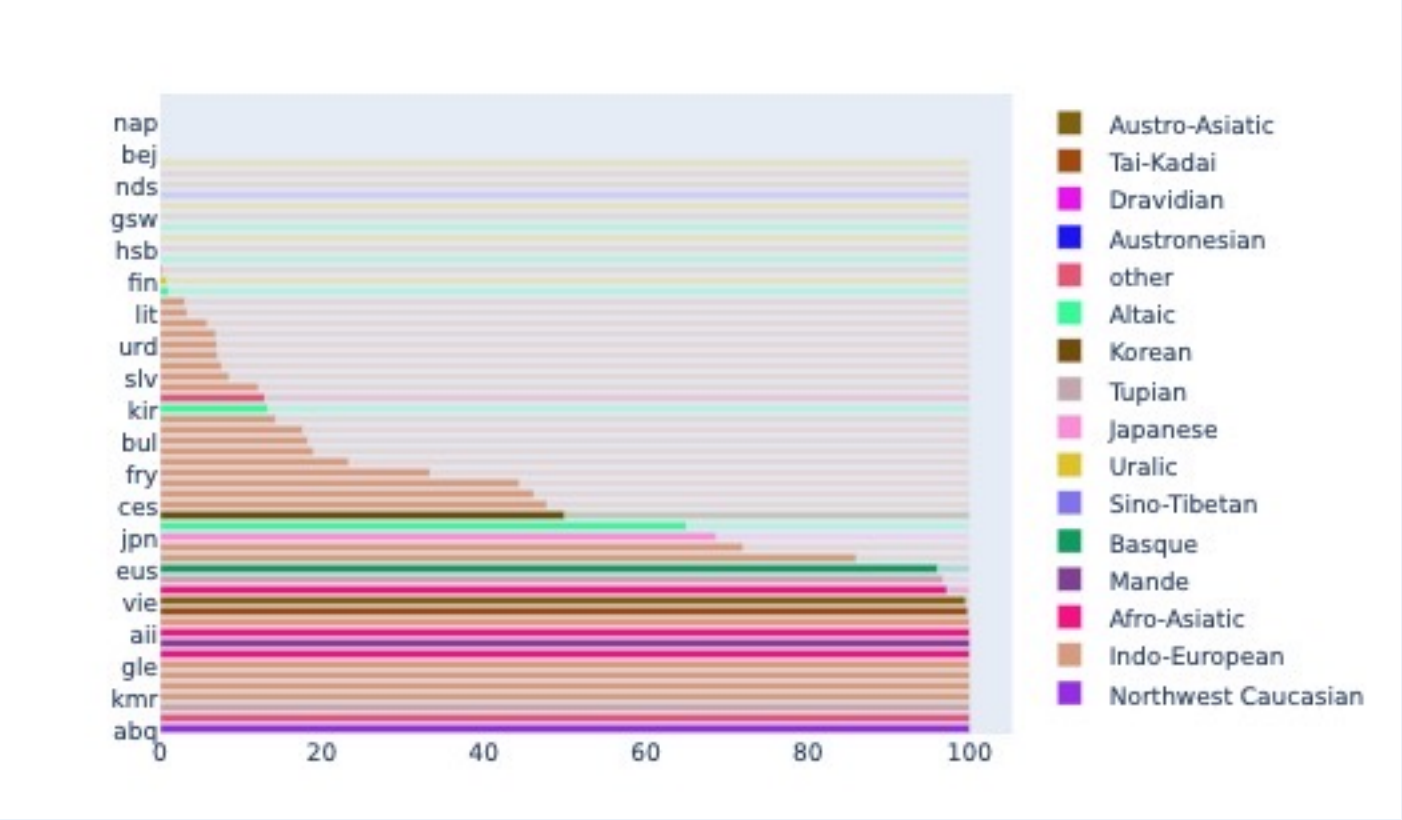
132 UD treebanks, within which there are **91** unique languages

A Solution?



Baylor, E., Ploeger, E., & Bjerva, J. (2024). Multilingual Gradient Word-Order Typology from Universal Dependencies. In Proceedings EACL

A Solution?



Order of adjective and noun

A Solution?

So far:

- Predict typological features based on for instance language embeddings
- Language-level probing of multilingual models
- Logistic regression (e.g. Malaviya et al., 2017; Östling and Kurfalı, 2023)

A Solution?

A new baseline:

- Probing with <u>linear regression</u>

- Language embeddings:
  - Östling and Tiedemann (2017)
  - Malaviya et al. (2017)

- Compare with logistic regression by rounding the values in our dataset

A Solution?

	Östling Linear Regr.	Östling Logistic Regr.	Malaviya Linear Regr.	Malaviya Logistic Regr.
Noun-adjective	0.146	0.261	0.141	0.378
Noun-numeral	0.140	0.132	0.129	0.399
Subject-verb	0.0781	0.306	0.101	0.156
Object-verb	0.169	0.237	0.0757	0.122
Object-subject	0.0127	–	0.0349	0.00940

Table 2: Mean squared error scores for linear regression and logistic regression models for each feature, using language vectors from Östling and Tiedemann (2017) and Malaviya et al. (2017). Better scores are closer to 0.

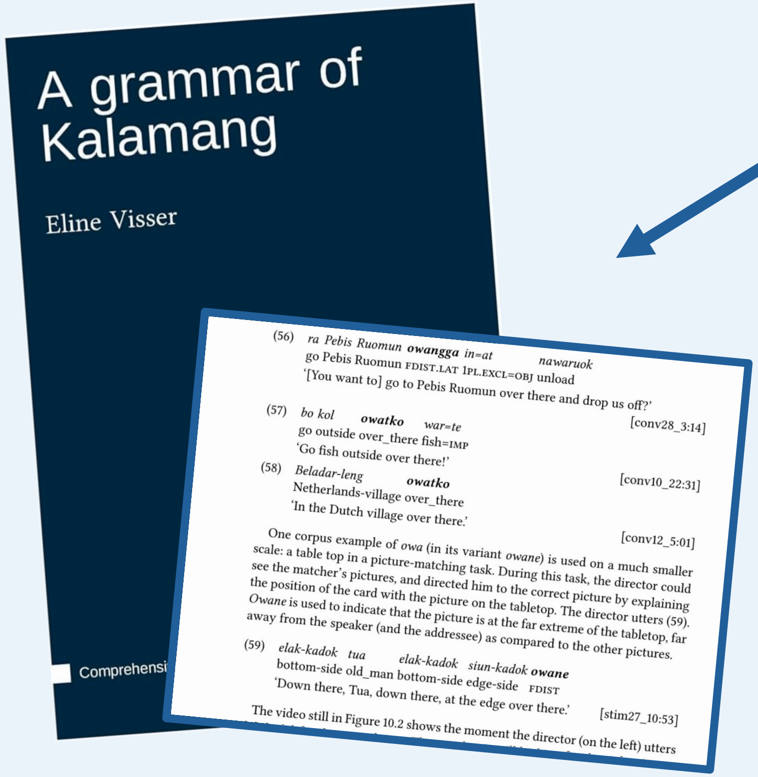
A Solution?

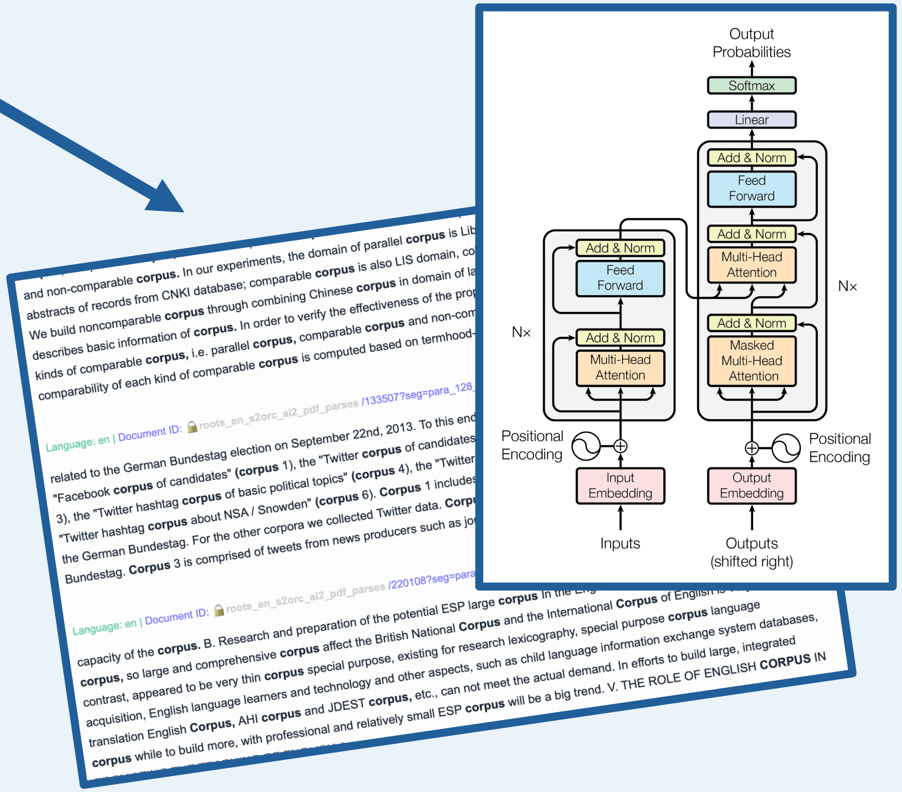
Limitations:

- Text-based typology is heavily influenced by the corpus
- Extension to more features is not trivial
- Extenion to more languages is not trivial

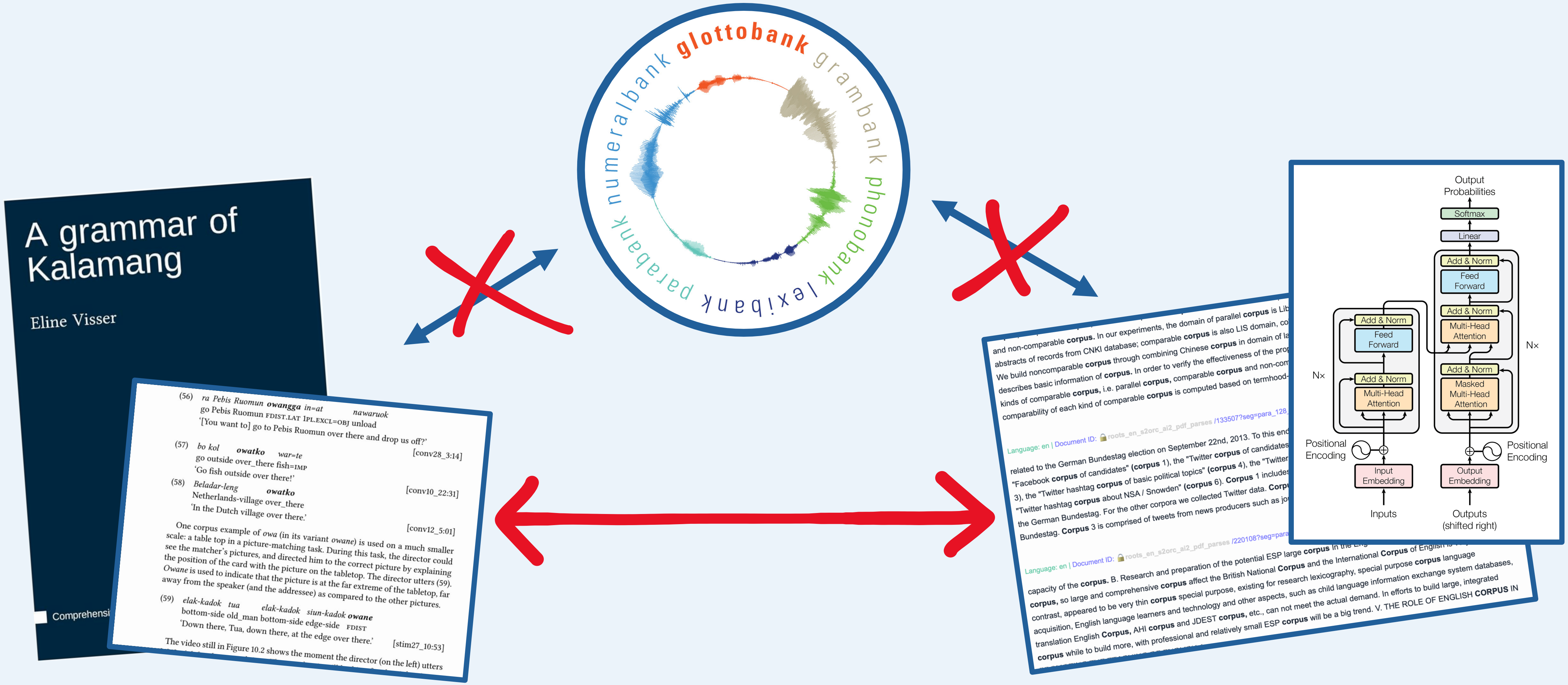
Alternative Solutions

Alternative Solutions





Alternative Solutions



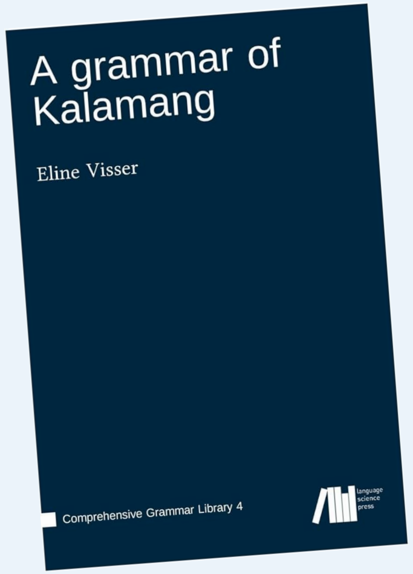
Alternative Solutions

Can we do without typological databases as an intermediate step?

Alternative Solutions

Can we do without typological databases as an intermediate step?

How well can a model "learn a language from a single human-readable book of grammar explanations, rather than a large mined corpus of in-domain data?"

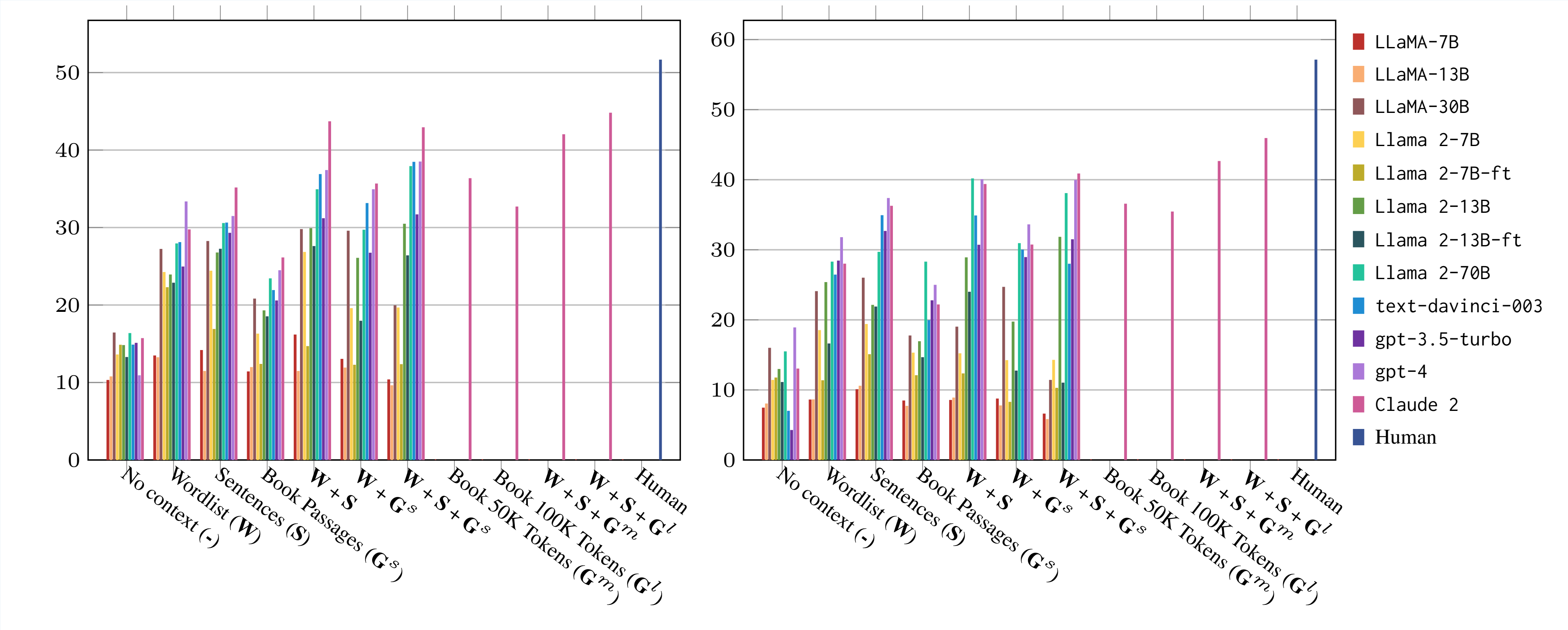


Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., & Melas-Kyriazi, L. (2023, October). A Benchmark for Learning to Translate a New Language from One Grammar Book. In The Twelfth ICLR.

Alternative Solutions

Can we do without typological databases as an intermediate step?

How well can a model "learn a language from a single human-readable book of grammar explanations, rather than a large mined corpus of in-domain data?"



Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., & Melas-Kyriazi, L. (2023, October). A Benchmark for Learning to Translate a New Language from One Grammar Book. In The Twelfth ICLR.

Alternative Solutions

Can we do without typological databases as an intermediate step?

- Beyond machine translation
- More languages

Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions

Kexun Zhang[1] **Yee Man Choi**[1] **Zhenqiao Song**[1] **Taiqi He**[1]
William Yang Wang[2] **Lei Li**[1]
[1]Carnegie Mellon University [2]UC Santa Barbara
{kexunz,yeemanc,zhenqias,taiqih}@andrew.cmu.edu
william@ucsb.edu leili@cs.cmu.edu

Abstract

How can large language models (LLMs) process and translate endangered languages? Many languages lack a large corpus to train a decent LLM; therefore existing LLMs rarely perform well in unseen, endangered languages. On the contrary, we observe that 2000 endangered languages, though without a large corpus, have a grammar book or a dictionary. We propose LINGOLLM, a training-free approach to enable an LLM to process unseen languages that hardly occur in its pre-training. Our key insight is to demonstrate linguistic knowledge of an unseen language in an LLM's prompt, including a dictionary, a grammar book, and morphologically analyzed input text. We implement LINGOLLM on top of two models, GPT-4 and Mixtral, and evaluate their performance on 5 tasks across 8 endangered or low-resource languages. Our results show that LINGOLLM elevates translation capability from GPT-4's 0 to 10.5 BLEU for 10 language directions. Our findings demonstrate the tremendous value of

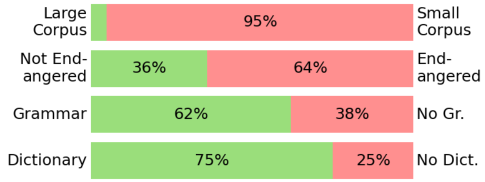


Figure 1: Among the world's ∼7000 languages, 95% don't have enough data (>100K sentences) for training LLMs (Bapna et al., 2022), while most have a grammar book (60%) or dictionary (75%) (Nordhoff and Hammarström, 2011), including many endangered languages (Moseley, 2010). Therefore, we utilize these linguistic descriptions to bring LLMs to endangered languages.

on languages that may not occur in pre-training (Robinson et al., 2023). We believe that speakers of endangered languages deserve equitable access to NLP technologies including LLMs. How can we enable an LLM with language processing capabilities on unseen and endangered languages?

Zhang, K., Choi, Y. M., Song, Z., He, T., Wang, W. Y., & Li, L. (2024). Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions. arXiv preprint arXiv:2402.18025

Alternative Solutions

Grammar is not the only way to take a closer perspective on language

Conclusions

Conclusions

- **Language sampling** seems highly relevant in multilingual NLP

Conclusions

- **Language sampling** seems highly relevant in multilingual NLP
- **Typological features** are potentially useful for interpreting, evaluating and improving multilingual language models

Conclusions

- **Language sampling** seems highly relevant in multilingual NLP
- **Typological features** are potentially useful for interpreting, evaluating and improving multilingual language models
- There are many **open questions** for incorporating linguistic typology in NLP
  - How can questions of language sampling in NLP best be addressed?
  - Can we automatically infer corpus typology? Does this help NLP?
  - Can we leverage linguistic grammars directly?

Funding Acknowledgements

This work was supported by a *Semper Ardens: Accelerate* research grant (CF21-0454) from the Carlsberg Foundation.

The current research visit is co-funded by the Otto Mønsteds Fond.

References (1/3)

Baylor, E., Ploeger, E., & Bjerva, J. (2023, December). The Past, Present, and Future of Typological Databases in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1163-1169).

Baylor, E., Ploeger, E., & Bjerva, J. (2024, January). Multilingual Gradient Word-Order Typology from Universal Dependencies. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207-219.

Bell, A. (1978). Language samples. Universals of human language, 1, 123-156.

Blevins, T., Limisiewicz, T., Gururangan, S., Li, M., Gonen, H., Smith, N. A., & Zettlemoyer, L. (2024). Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models. *arXiv preprint arXiv:2401.10440*.

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. WALS Online (v2020.3)

Grambank's Typological Advances Support Computational Research on Diverse Languages (Haynie et al., SIGTYP 2023)

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. Universals of language, 2, 73-113.

Hewitt, J., & Liang, P. (2019, November). Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2733-2743).

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *EuroVis (STARs)*, *2015*, 83-103.

Kashyap, A. K. (2019). Language typology. The Cambridge handbook of systemic functional linguistics, 767-792.

References (2/3)

Levshina, Natalia, et al. "Why we need a gradient approach to word order." *Linguistics* (2023).

Malaviya, C., Neubig, G., & Littell, P. (2017, September). Learning Language Representations for Typology Prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2529-2535).

Matthew S. Dryer. 2013. Order of Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

Matthew S. Dryer. 2013. Order of Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3)

Matthew S. Dryer. 2013. Order of Subject and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3)

Östling, R., & Kurfalı, M. (2023). Language embeddings sometimes contain typological generalizations. Computational Linguistics, 49(4), 1003-1051.

Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024). What is 'Typological Diversity' in NLP?. arXiv preprint arXiv:2402.04222.

Ponti, E. M. et al. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, *45*(3), 559-601.

Purason, T., & Tättar, A. (2022, June). Multilingual neural machine translation with the right amount of sharing. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 91-100).

Rijkhoff, J., & Bakker, D. (1998). Language sampling. Linguistic Typology, 2(3), 263-314.

Rijkhoff, J., Bakker, D., Hengeveld, K., & Kahrel, P. (1993). A method of language sampling. Studies in Language. 17(1), 169-203.

Skirgård, Hedvig et al. (2023). Grambank v1.0 (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7740140

References (3/3)

Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., & Melas-Kyriazi, L. (2023, October). A Benchmark for Learning to Translate a New Language from One Grammar Book. In The Twelfth International Conference on Learning Representations.

Üstün, A., Bisazza, A., Bouma, G., & Noord, G. V. (2022). UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. Computational Linguistics, 48(3).

Visser, E. (2022). A grammar of Kalamang. Language Science Press.

Zhang, K., Choi, Y. M., Song, Z., He, T., Wang, W. Y., & Li, L. (2024). Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions. *arXiv preprint arXiv:2402.18025*.