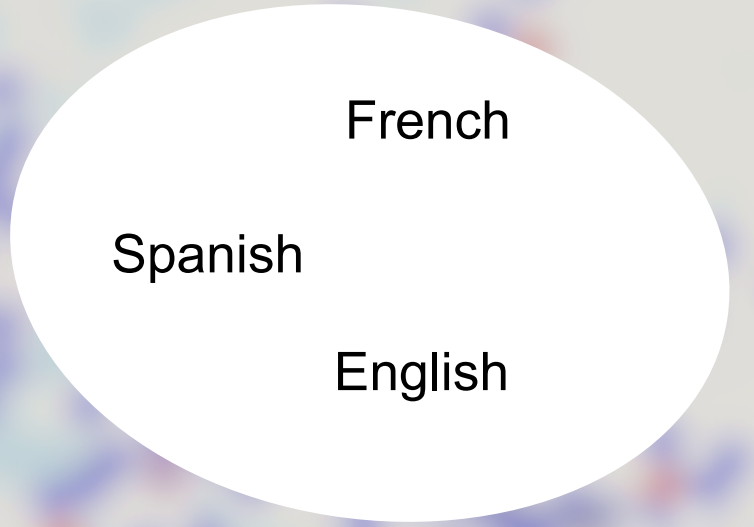
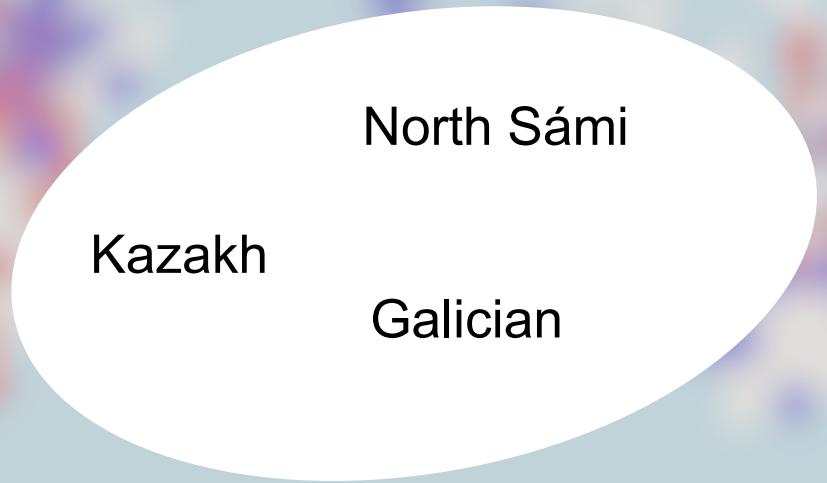
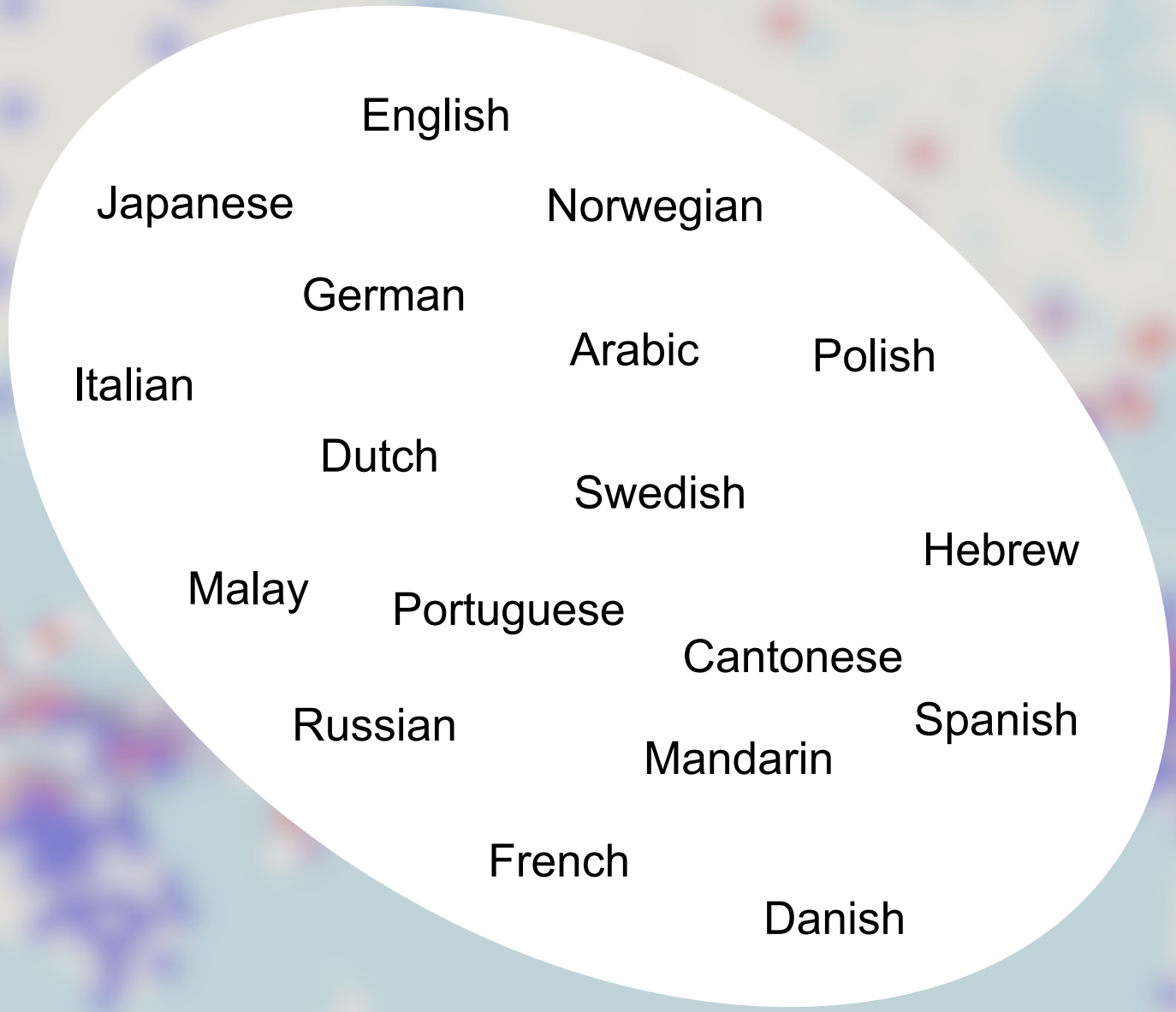
Beyond Post-Hoc Typological Diversity in NLP

Esther Ploeger
Dept. of Computer Science
Aalborg University
AAU-NLP
espl@cs.aau.dk

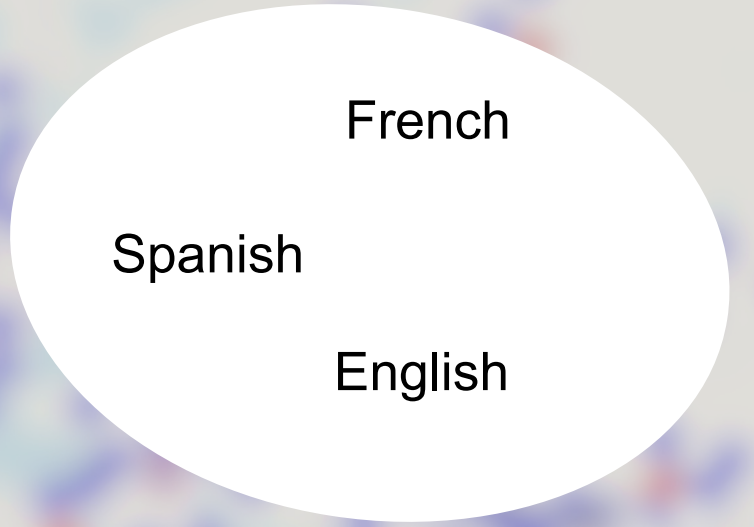
9 January 2025

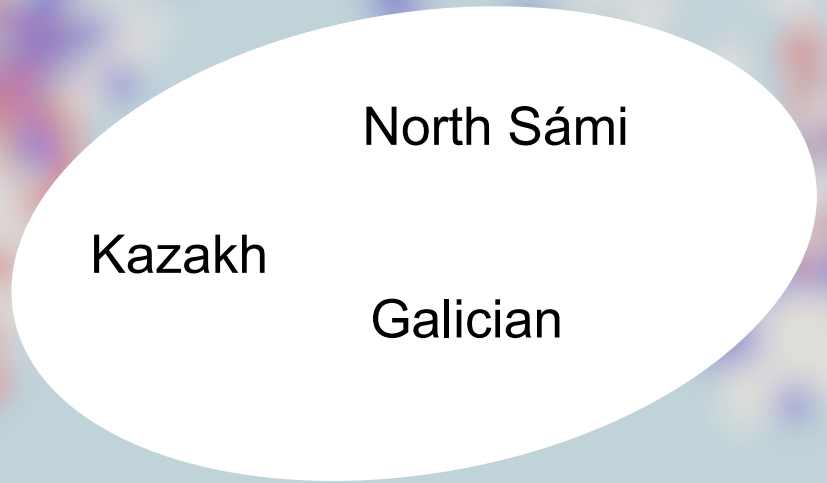

French
Spanish
English

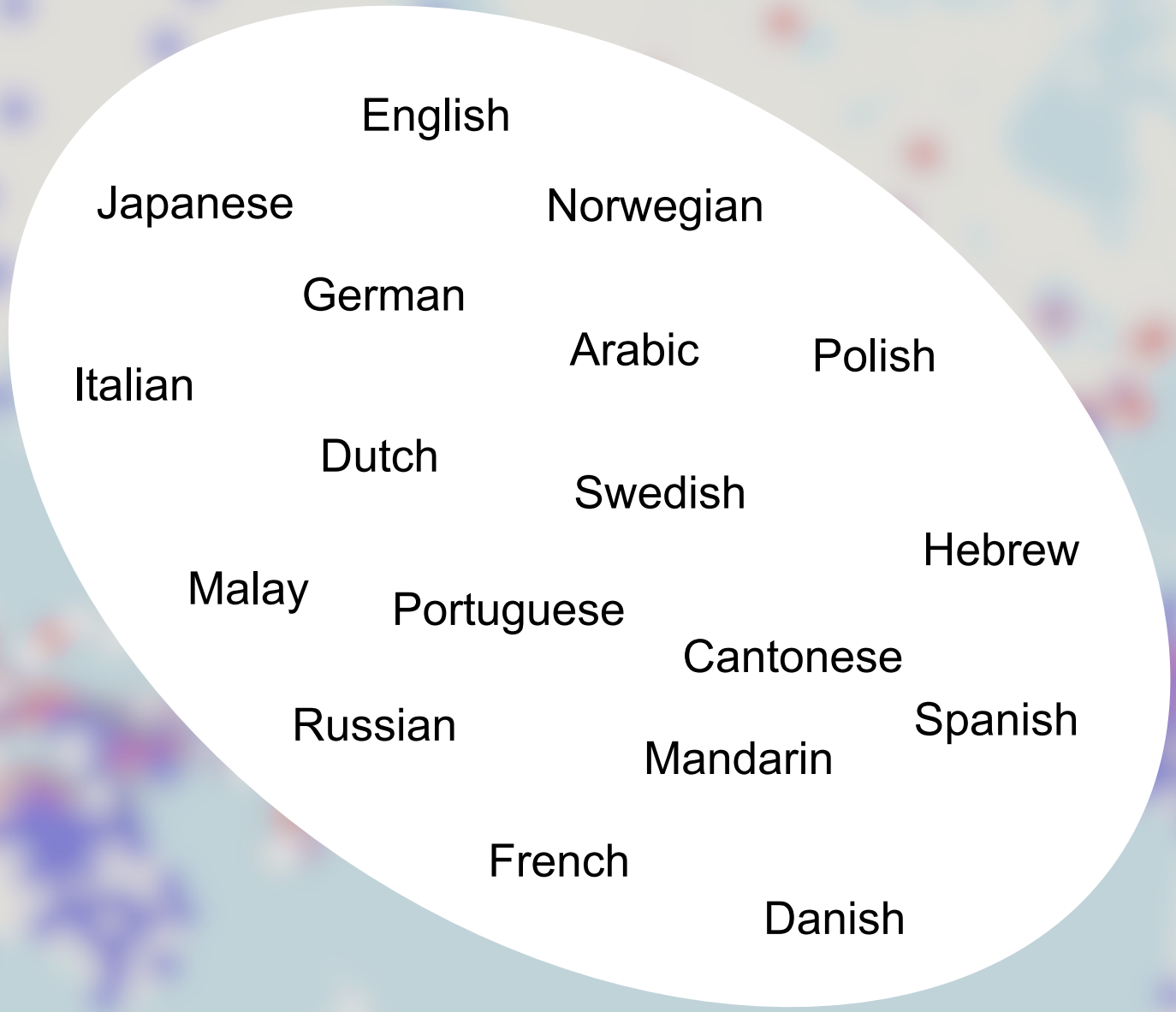

North Sámi
Kazakh
Galician

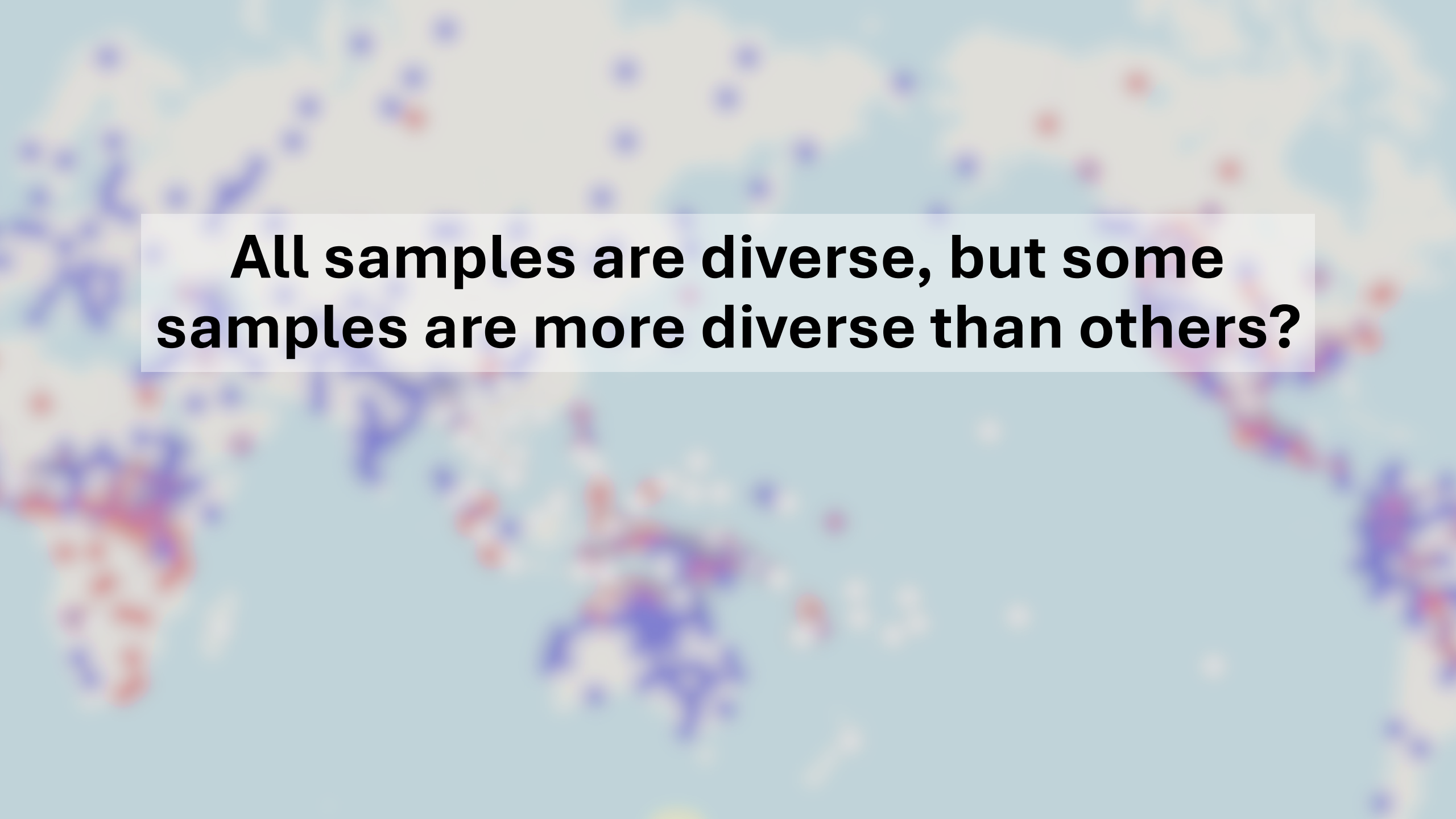

English
Japanese
Norwegian
German
Arabic
Polish
Italian
Dutch
Swedish
Hebrew
Malay
Portuguese
Cantonese
Russian
Spanish
Mandarin
French
Danish

“We evaluate on *N* typologically diverse languages.”


French
Spanish
English


North Sámi
Kazakh
Galician


English
Japanese
Norwegian
German
Arabic
Polish
Italian
Dutch
Swedish
Hebrew
Malay
Portuguese
Cantonese
Russian
Spanish
Mandarin
French
Danish


All samples are diverse, but some samples are more diverse than others?

All samples are diverse, but some samples are more diverse than others?

Can we approach typological diversity more systematically?

This Talk

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
wessel.poelman@kuleuven.be

Andreas Holck Høeg-Petersen
Aalborg University
Department of Computer Science
ahhp@cs.aau.dk

Anders Schlichtkrull
Aalborg University
Department of Computer Science
andsch@cs.aau.dk

Miryam de Lhoneux
KU Leuven
Department of Computer Science
miryam.delhoneux@kuleuven.be

Johannes Bjerva
Aalborg University
Department of Computer Science
jbjerva@cs.aau.dk

Beyond individual languages, multilingual natural language processing (NLP) research increasingly aims to develop models that perform well across languages generally. 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 highly typologically diverse languages given a sampling frame, informed by language typology. We compare sampling methods with a range of metrics and find that our systematic methods consistently retrieve more typologically diverse language selections than previous methods in NLP. Moreover, we provide evidence that this affects generalizability in multilingual model evaluation, emphasizing the importance of diverse language sampling in NLP evaluation.

1. Introduction

Data-driven approaches to language technology have shifted the realm of possibility in multilingual NLP. Distributed word representations (Mikolov et al. 2013) have lifted

This Talk

- Motivation
- Background
- Methodology and results
- Limitations and next steps
- Conclusion

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
wessel.poelman@kuleuven.be

Andreas Holck Høeg-Petersen
Aalborg University
Department of Computer Science
ahhp@cs.aau.dk

Anders Schlichtkrull
Aalborg University
Department of Computer Science
andsch@cs.aau.dk

Miryam de Lhoneux
KU Leuven
Department of Computer Science
miryam.delhoneux@kuleuven.be

Johannes Bjerva
Aalborg University
Department of Computer Science
jbjerva@cs.aau.dk

Beyond individual languages, multilingual natural language processing (NLP) research increasingly aims to develop models that perform well across languages generally. 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 highly typologically diverse languages given a sampling frame, informed by language typology. We compare sampling methods with a range of metrics and find that our systematic methods consistently retrieve more typologically diverse language selections than previous methods in NLP. Moreover, we provide evidence that this affects generalizability in multilingual model evaluation, emphasizing the importance of diverse language sampling in NLP evaluation.

1. Introduction

Data-driven approaches to language technology have shifted the realm of possibility in multilingual NLP. Distributed word representations (Mikolov et al. 2013) have lifted


Why?

Generalizable multilingual NLP

Generalizable multilingual NLP

→ ***How*** generalizable?

Generalizable multilingual NLP

→ ***How*** generalizable? → We cannot test it on every language ever…

Generalizable multilingual NLP

→ ***How*** generalizable? → We cannot test it on every language ever…

↓

How do we select?


What?

What is typological diversity?

What is typological diversity?

In NLP: (almost) no criteria

What is typological diversity?

What is **typological** diversity?

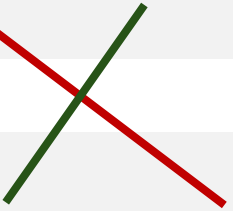
Linguistic Typology

English:	*The red apple*
French:	*La pomme rouge*

Linguistic Typology

English: *The **red** **apple***

French: *La **pomme** **rouge***

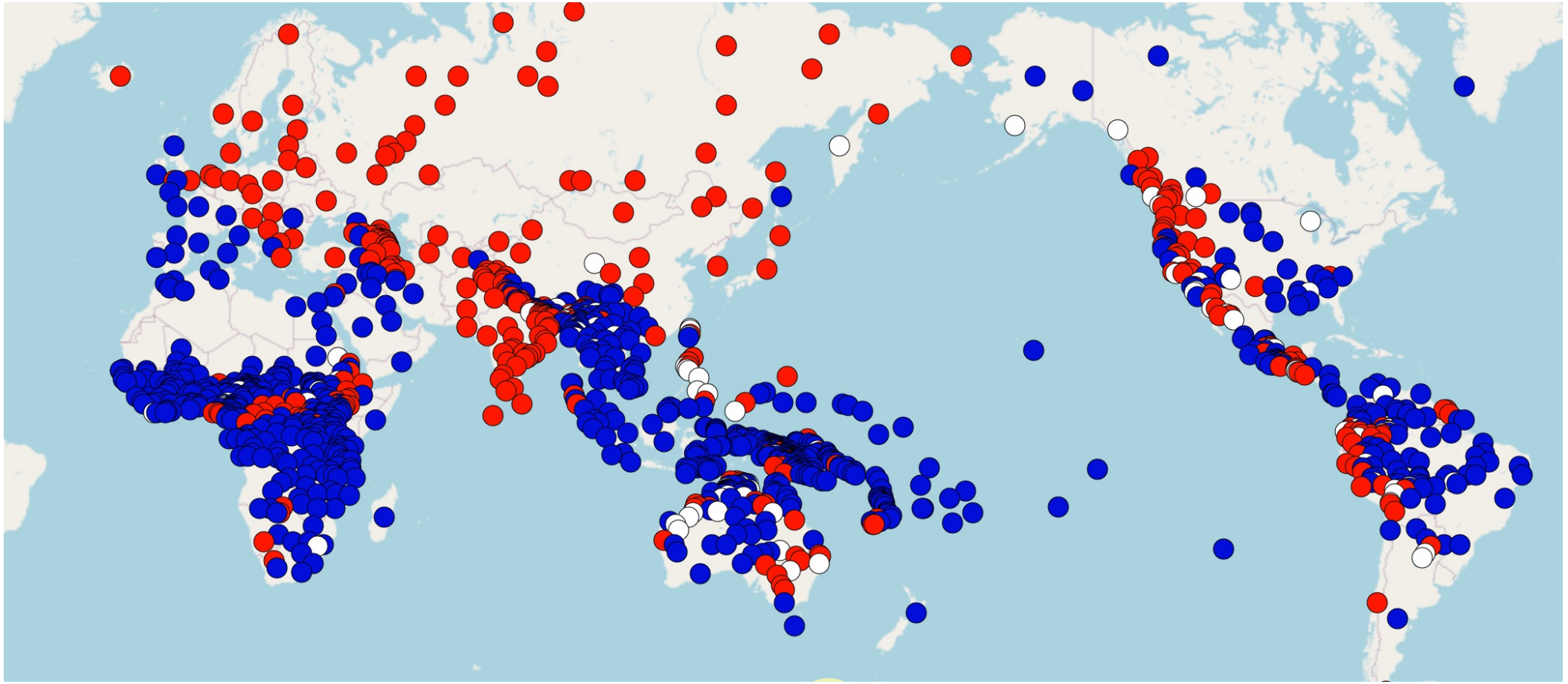


Linguistic Typology

English: *The **red** **apple***

French: *La **pomme** **rouge***

Typological feature: **adjective-noun order**



Matthew S. Dryer. 2013. Order of Adjective and Noun. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.4)

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

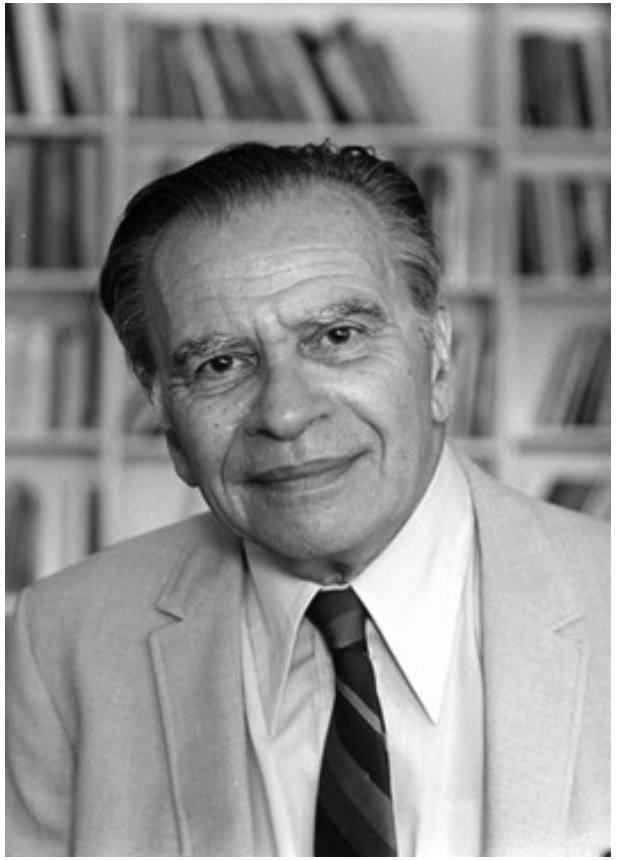
Linguistic Typology

"the classification of the world's languages according to similarities and differences in their linguistic structures..."

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

Greenberg's Universals

- "Some universals of grammar with particular reference to the order of meaningful elements" (1963)



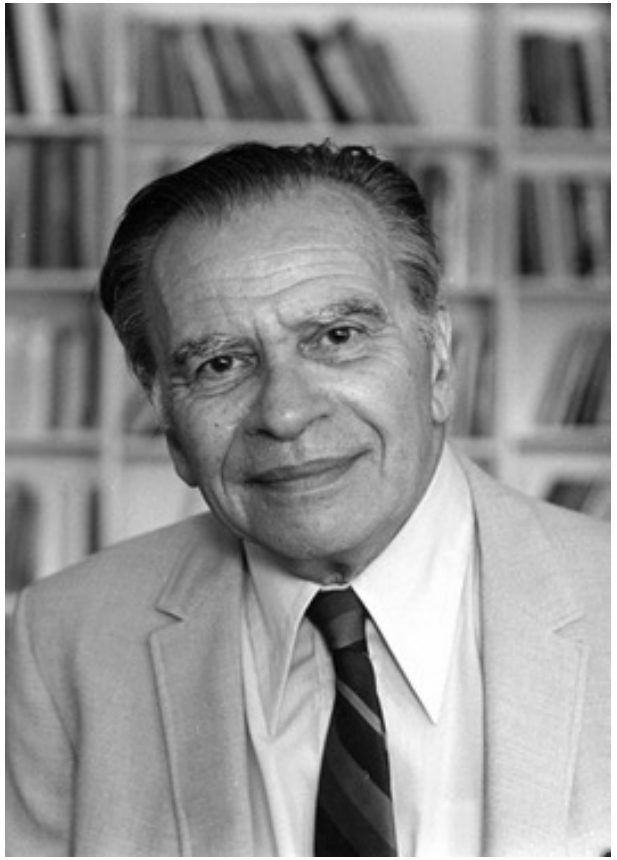
Joseph Greenberg

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

Greenberg's Universals

- "Some universals of grammar with particular reference to the order of meaningful elements" (1963)

- 45 linguistic universals

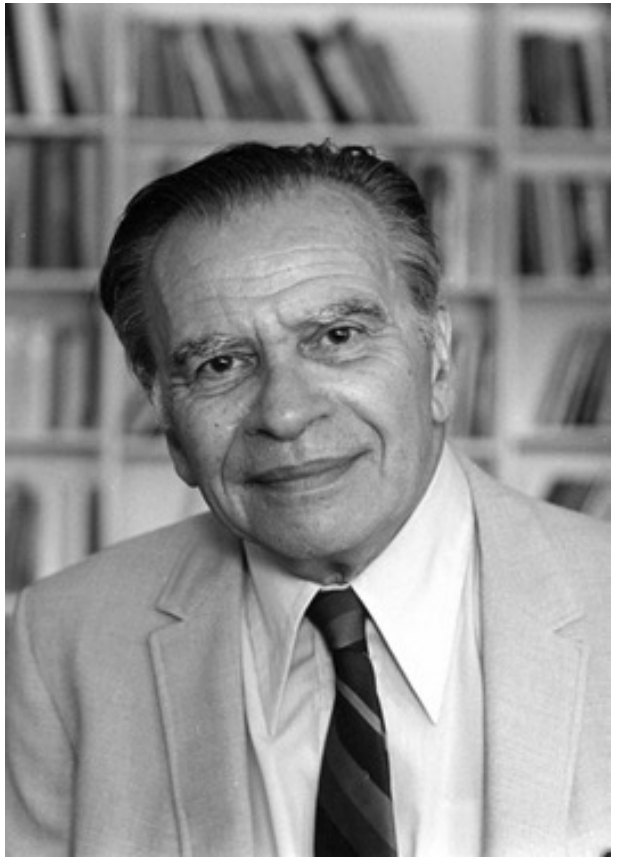


Joseph Greenberg

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

Greenberg's 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."

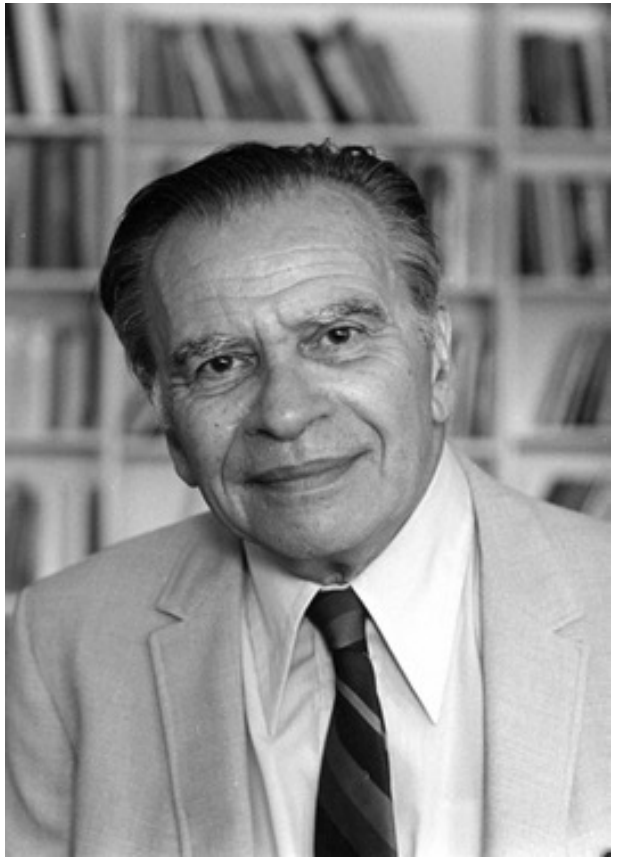


Joseph Greenberg

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

Greenberg's 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.

Greenberg's Universals

"**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.

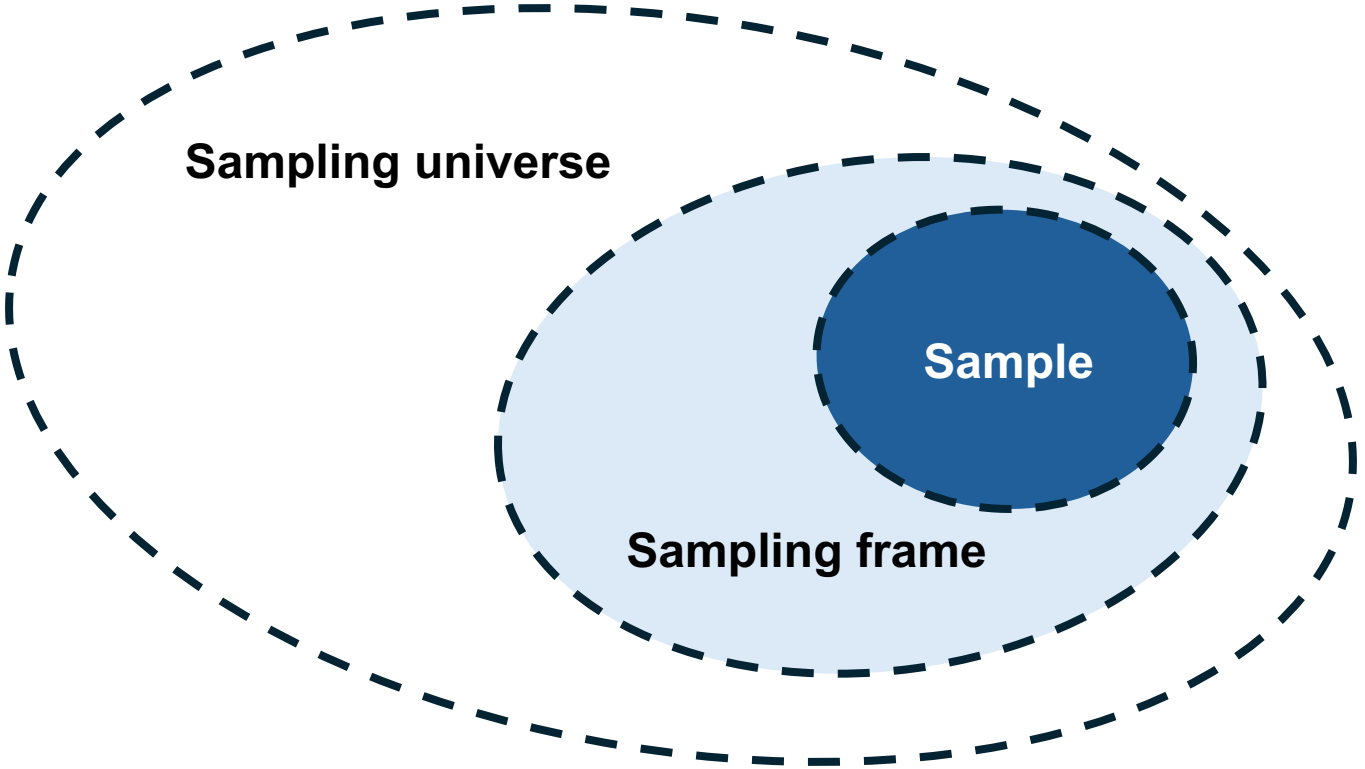
Greenberg's Universals

"**a general th** **e a framework**
for all langu or English.
These are jus languages,
and there is r by studying
one or two la uistic
phenomena i

Generalizable multilingual NLP

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

Language Sampling in Typology



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

Language Sampling in Typology

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

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

Language Sampling in Typology

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

- **Random** sampling

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

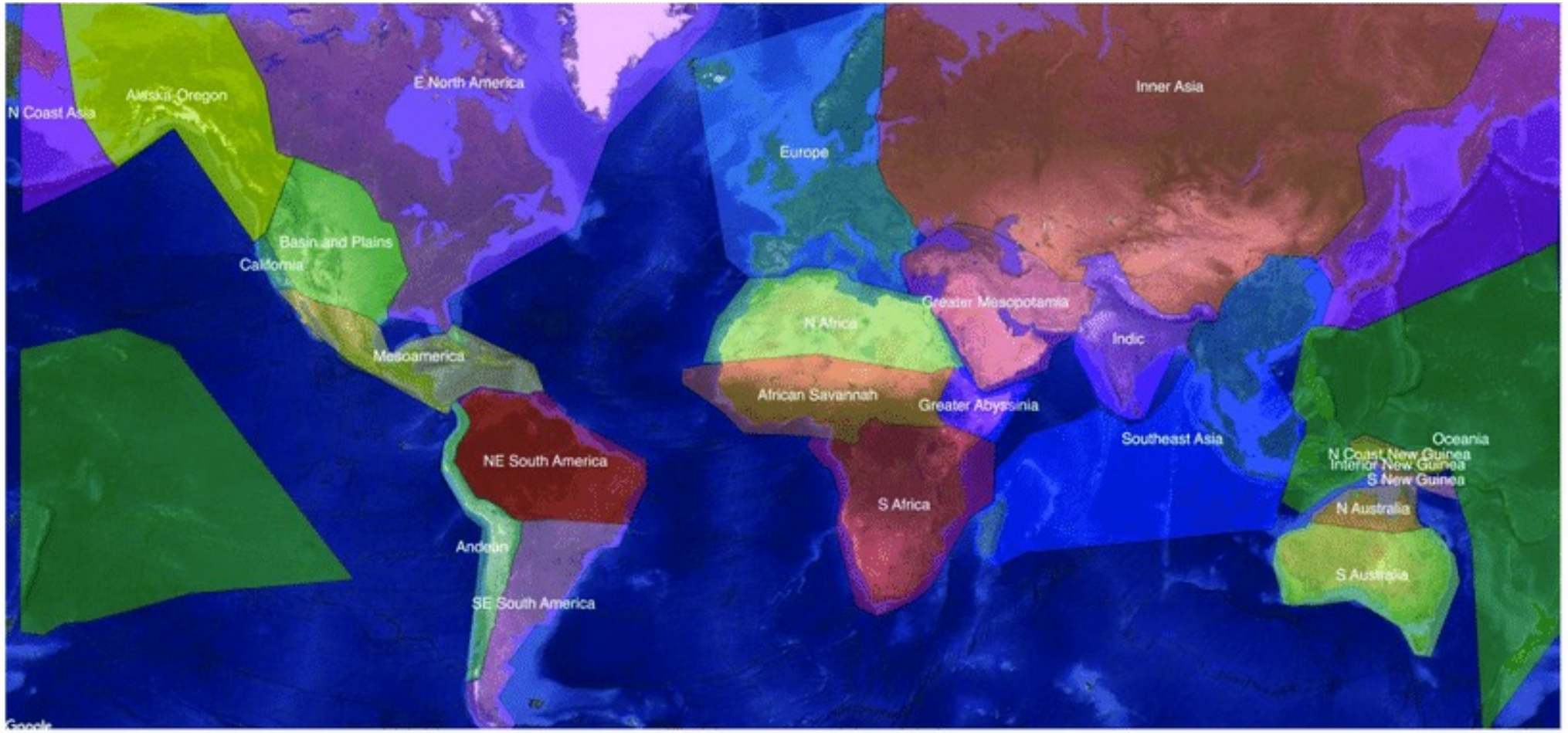
Language Sampling in Typology

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 in Typology



AUTOTYP areas

Bickel et al., 2017

Language Sampling in Typology

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.

Language Sampling in Typology

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*
- **Convenience** sampling (Velupillai, 2012)
  - Based on data availability

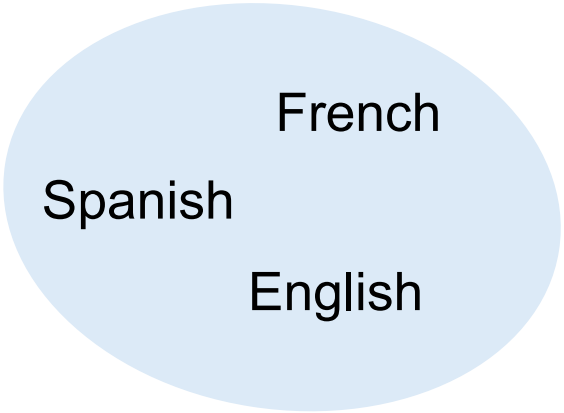
Rijkhoff, J., & Bakker, D. (1998). Language sampling. Linguistic Typology, 2(3), 263-314.

Language Sampling in Typology

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*

- **Convenience** sampling (Velupillai, 2012)
  - Based on data availability

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

What is typological diversity?

What is typological **diversity**?

What is typological **diversity**?

Working definition:
multiple things should be included,

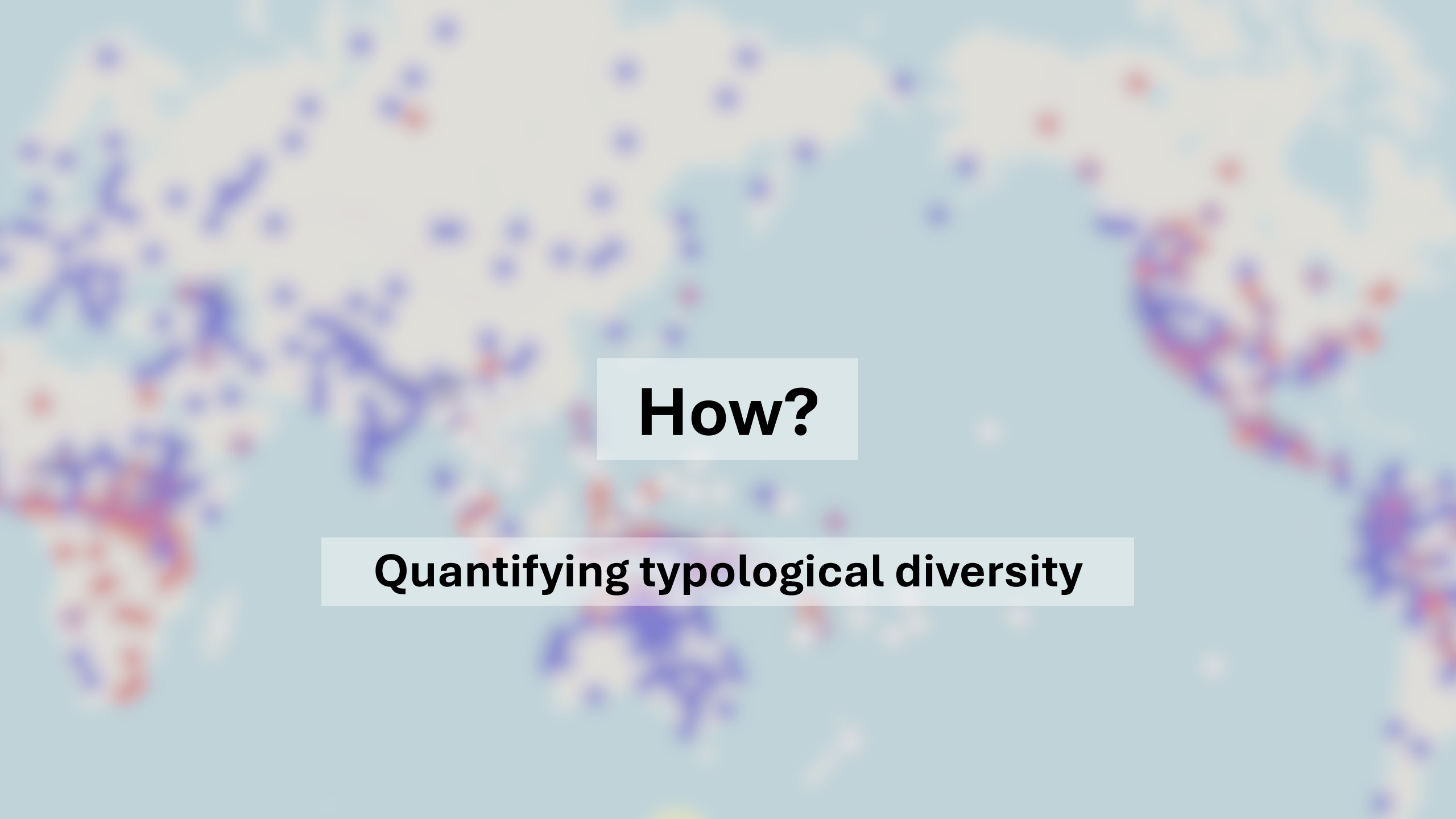
What is typological **diversity**?

Working definition:
multiple things should be included, as much as possible,

What is typological **diversity**?

Working definition:
multiple things should be included, as much as possible, maybe with some balance


How?



How?

Quantifying typological diversity

“multiple things should be included, as much as possible, maybe with some balance”

“multiple ~~things~~ should be included, as much as possible, maybe with some balance”

typological feature values

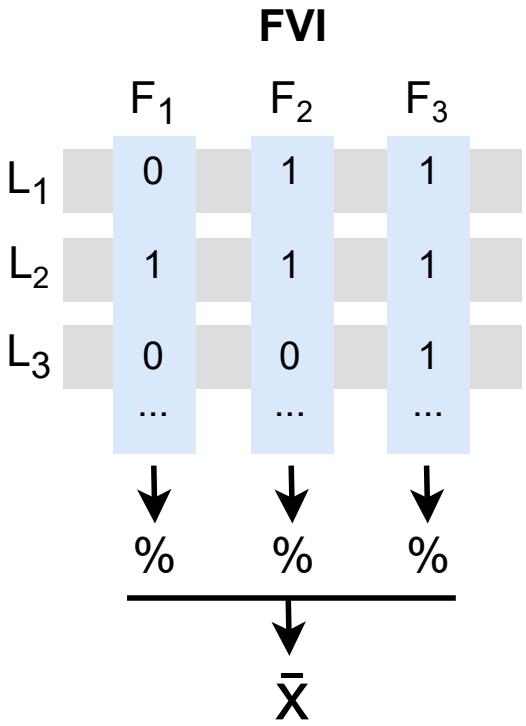
Features

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

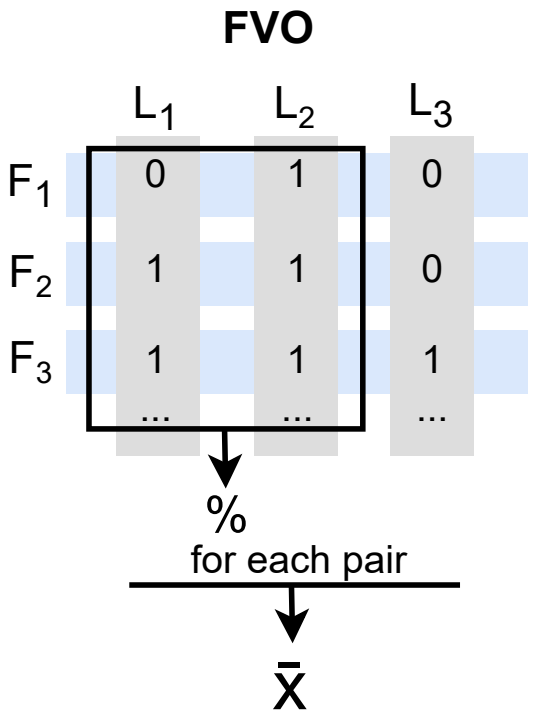
Feature Value Inclusion (FVI)



To what extent are all typological feature values represented in the dataset?

- Typology: *saturation* (Miestamo, Bakker, and Arppe; 2016)
- Likely higher for larger language sets

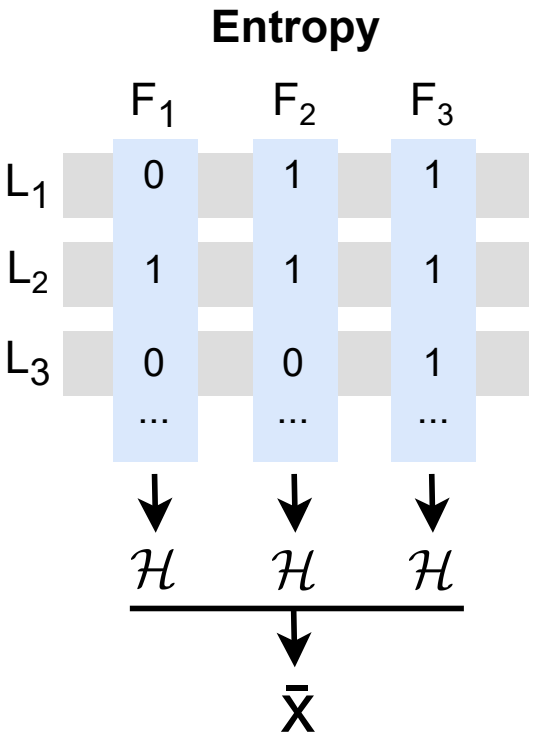
Feature Value Overlap (FVO)



To what extent do the included feature values overlap?

- Typology: inspired by Dahl (2008)
- Likely higher for larger language sets

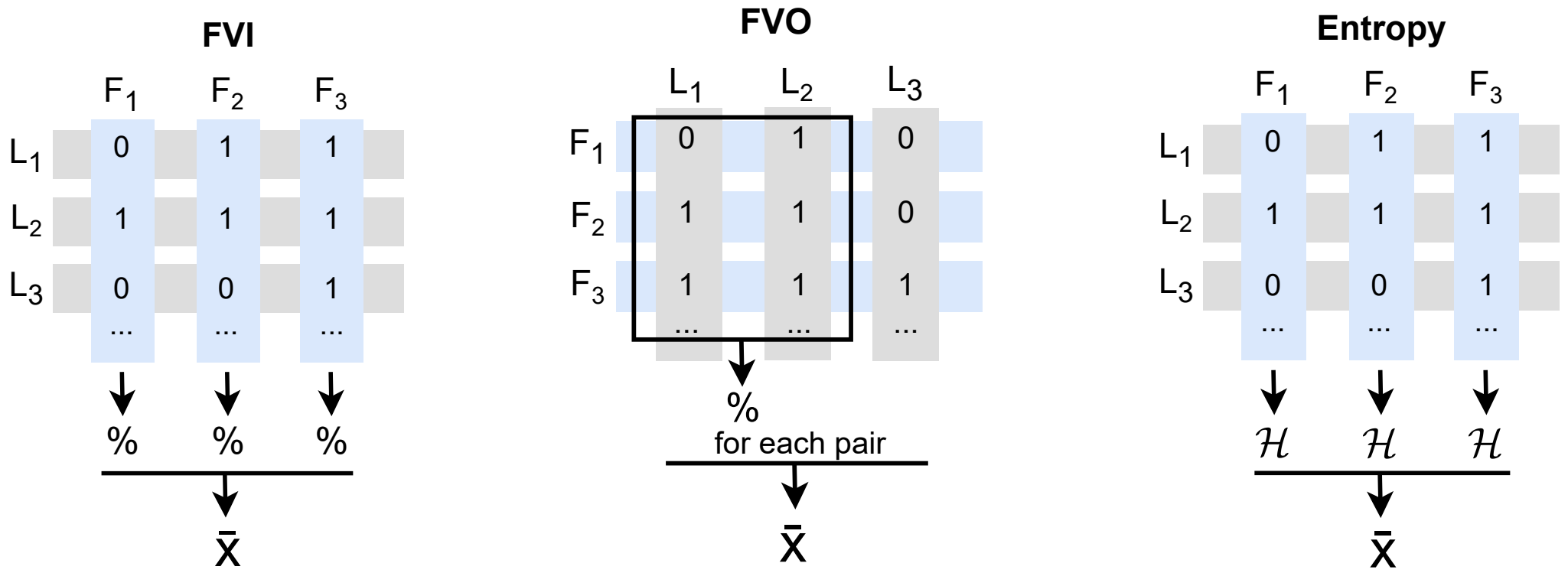
Feature Value Entropy (H)



To what extent is there a skew in the included feature values?

- Similar to typology index (Ponti et al., 2020)

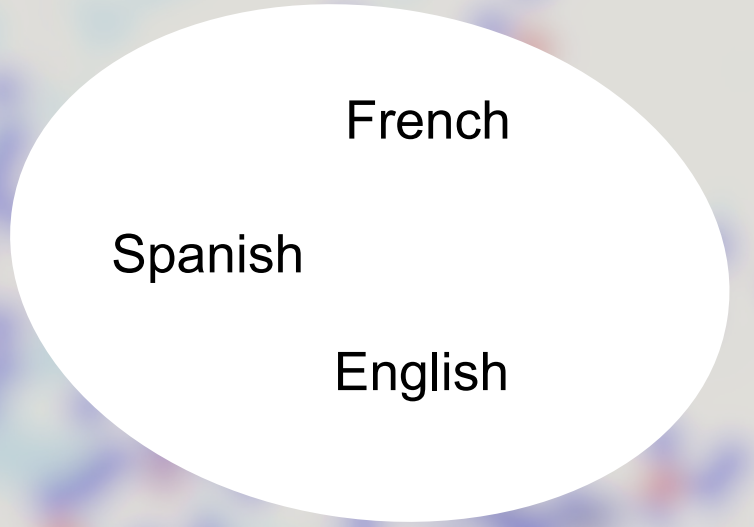
Post-hoc Diversity Measurement

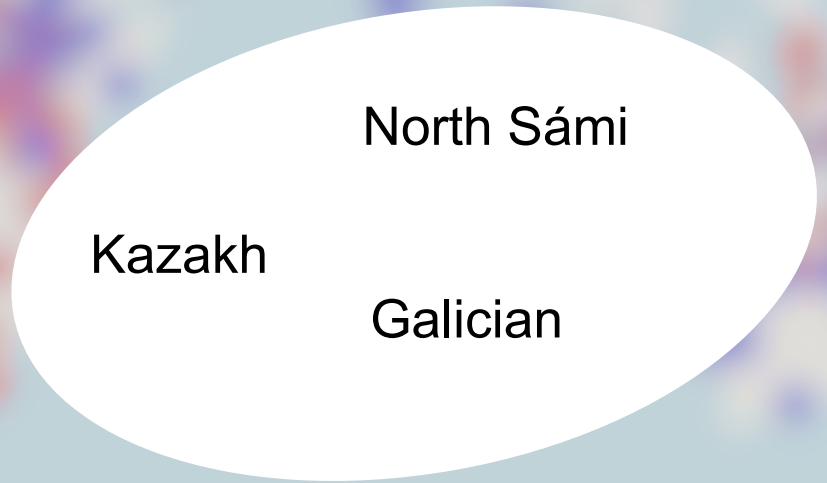


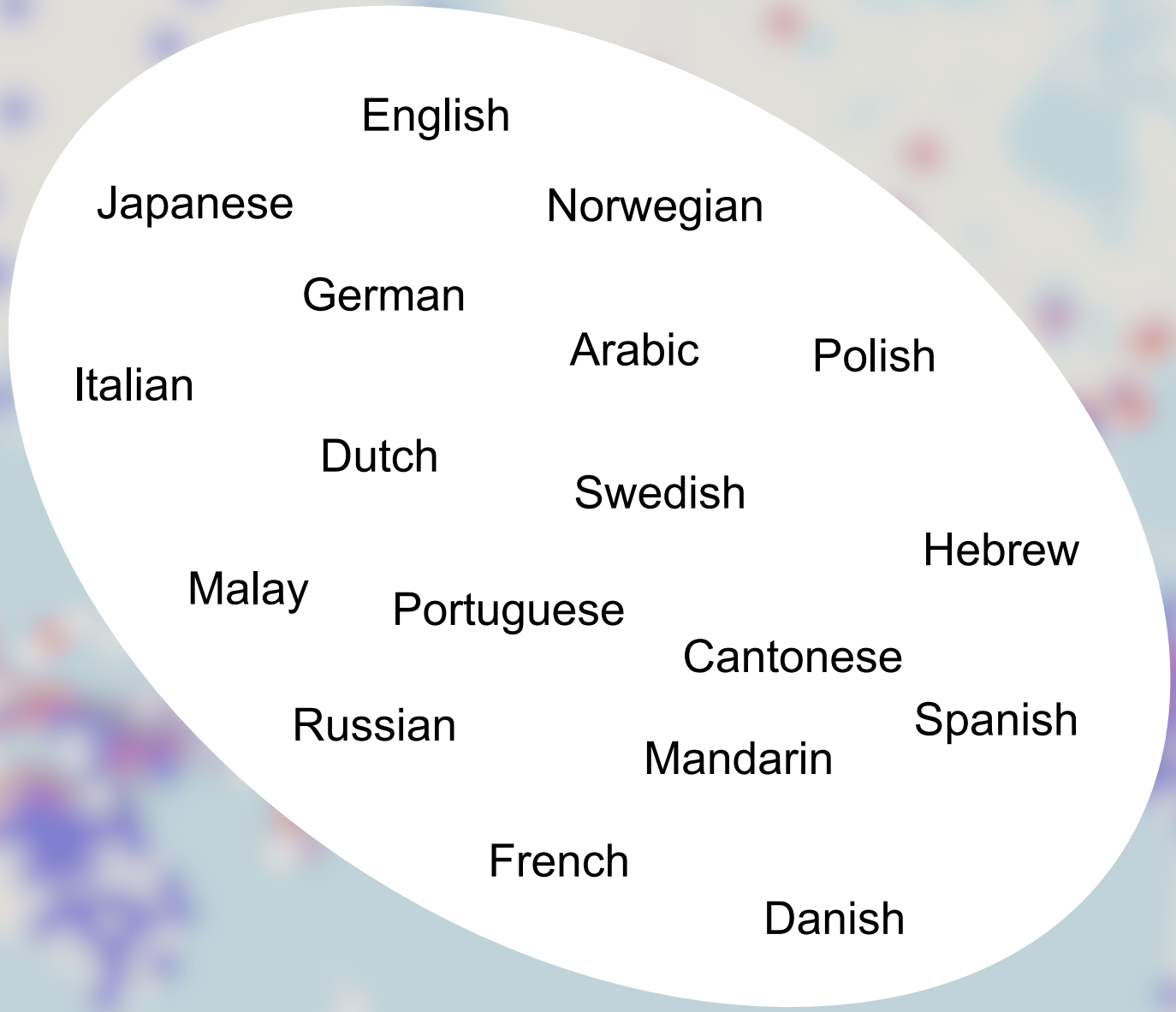
Post-hoc Diversity Measurement

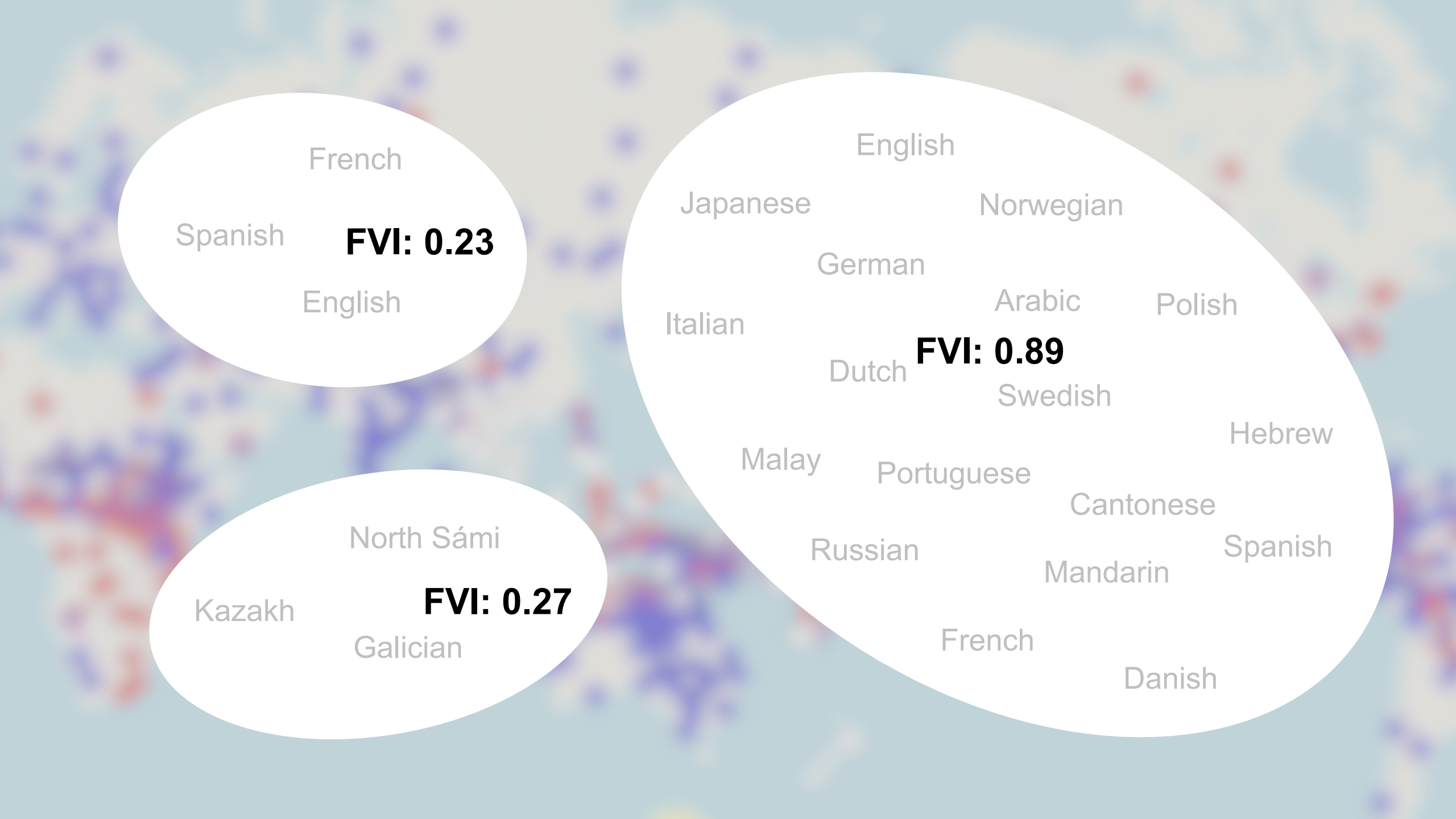
```python
from typdiv_sampling.evaluation import Evaluator

# With default settings.
evaluator = Evaluator()

sample = ['kore1280', 'russ1263', 'stan1290']
evaluator.evaluate_sample(sample)
> Result(
    run=None, # Optional result to keep track of averages across runs, unused here.
    ent_score_with=0.5374,
    ent_score_without=0.4954,
    fvi_score=0.7686,
    mpd_score=0.7836,
    fvo_score=0.6302,
    sample={'russ1263', 'kore1280', 'stan1290'},
)
```

https://github.com/esther2000/typdiv-sampling

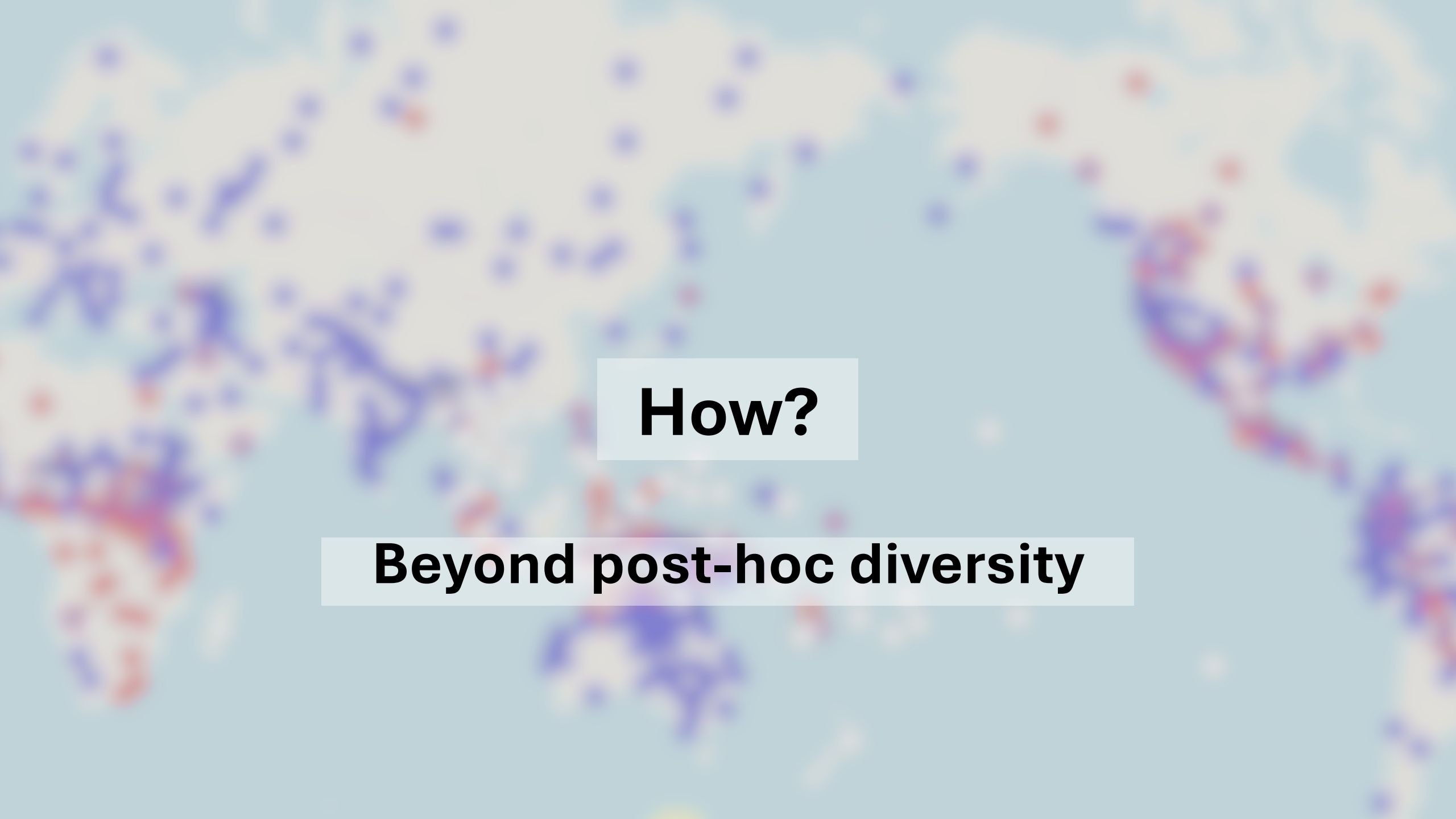

French
Spanish
English


North Sámi
Kazakh
Galician


English
Japanese
Norwegian
German
Arabic
Polish
Italian
Dutch
Swedish
Hebrew
Malay
Portuguese
Cantonese
Russian
Spanish
Mandarin
French
Danish


French
Spanish
FVI: 0.23
English
English
Japanese
Norwegian
German
Arabic
Polish
Italian
FVI: 0.89
Dutch
Swedish
Hebrew
Malay
Portuguese
Cantonese
Russian
Spanish
Mandarin
French
Danish
North Sámi
Kazakh
FVI: 0.27
Galician


How?


How?
Beyond post-hoc diversity

Why can’t we just use language families?

Why can’t we just use language families?

Majewska et al. (2020) ‘sampled languages from 5 different language families to ensure typological diversity’

Why can’t we just use language families?

Majewska et al. (2020) ‘sampled languages from 5 different language families to ensure typological diversity’

Also, it is common in linguistic typology (probability sampling)

Why can’t we just use language families?

Majewska et al. (2020) ‘sampled languages from 5 different language families to ensure typological diversity’

Also, it is common in linguistic typology (probability sampling)

Dahl (2008): “[genealogically] related languages that are no longer in contact with each other can in a few thousand years develop typological profiles that are no longer indicative of a common origin.”

Why can’t we just use language families?

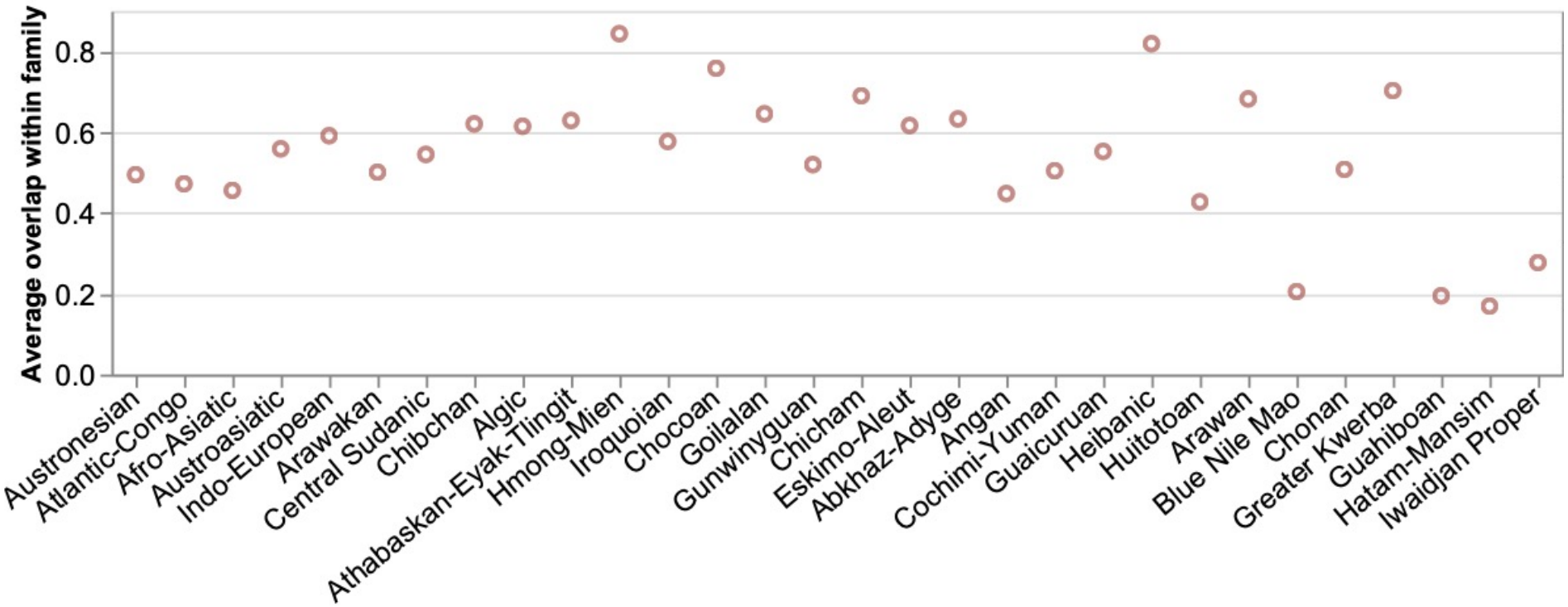
Majewska et al. (2020) ‘sampled languages from 5 different language families to ensure typological diversity’

Also, it is common in linguistic typology (probability sampling)

Dahl (2008): “[genealogically] related languages that are no longer in contact with each other can in a few thousand years develop typological profiles that are no longer indicative of a common origin.”

German and Dutch <> Dutch and Hindi?

Why can’t we just use language families?

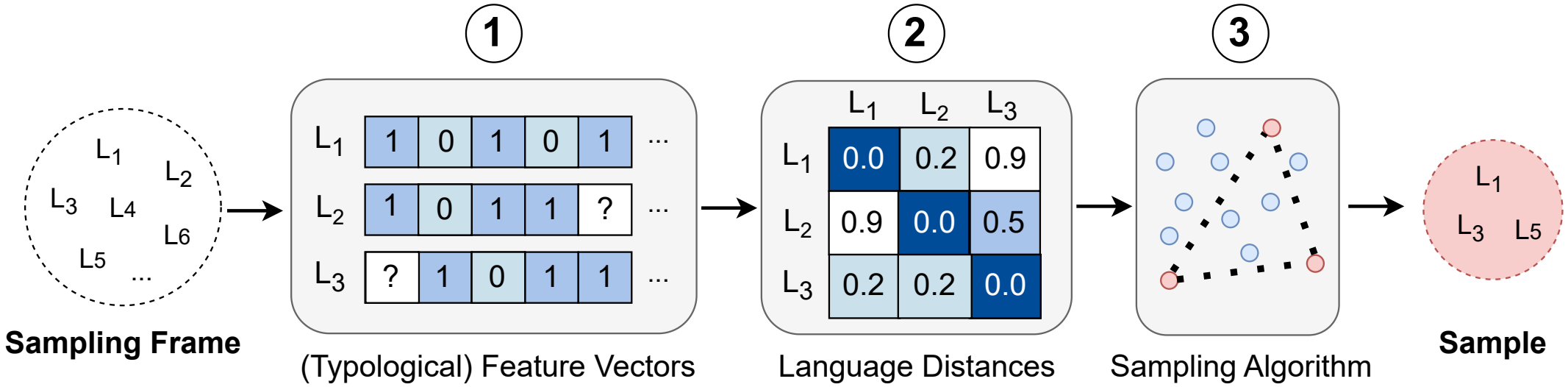


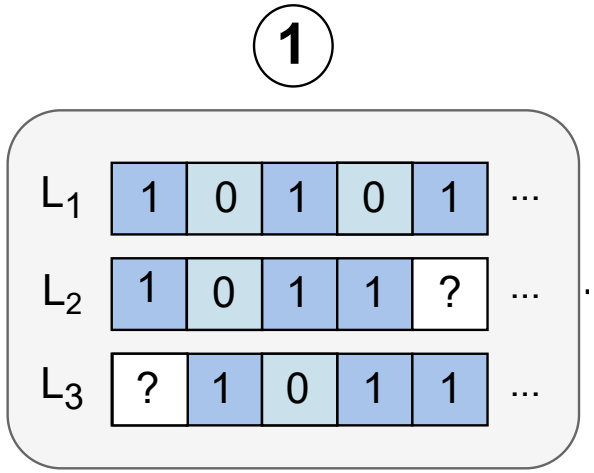
A Framework for Systematic Language Sampling

In NLP, the variables under study are often not typological variables

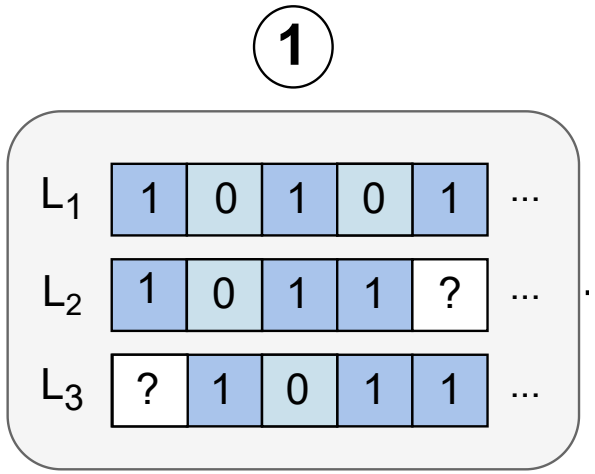
There is no circularity in using them directly!

A Framework for Systematic Language Sampling





(Typological) Feature Vectors



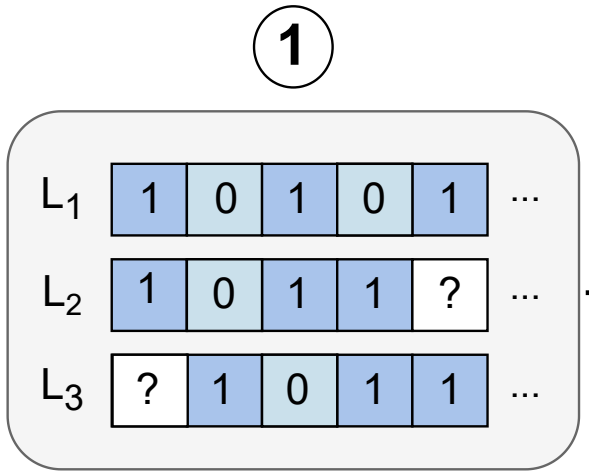
(Typological) Feature Vectors

Features

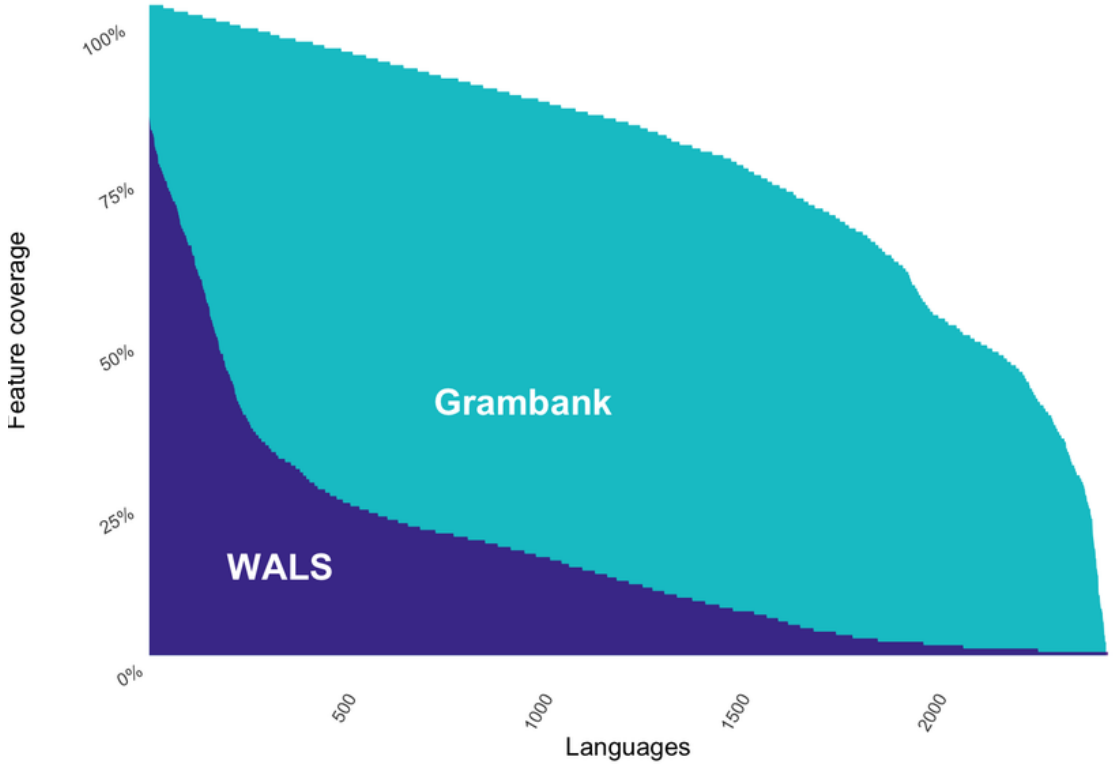
Showing 1 to 100 of 195 entries

← Previous 1 2 Next →

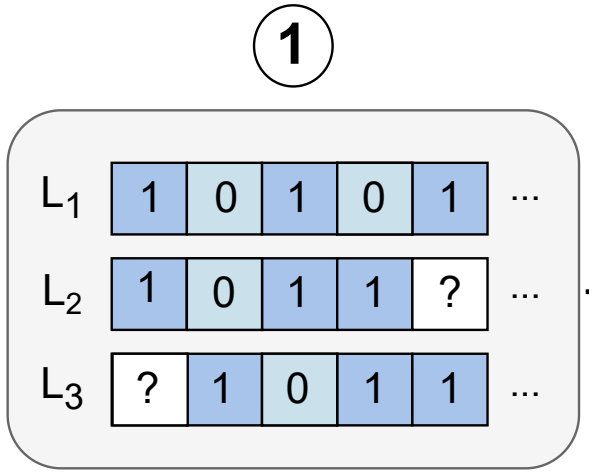
Id	Feature	Patron	Languages and dialects	Details
Search	Search		Search	
GB020	Are there definite or specific articles?	Jay Latarche and Jeremy Collins	2198	Values and description
GB021	Do indefinite nominals commonly have indefinite articles?	Jay Latarche and Jeremy Collins	2221	Values and description
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



(Typological) Feature Vectors



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

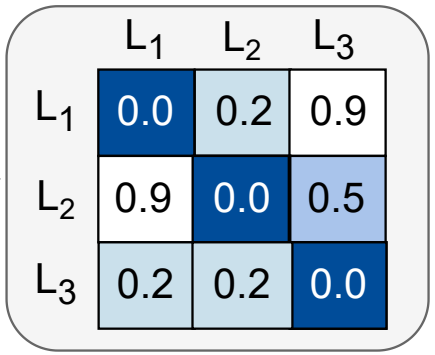


(Typological) Feature Vectors

"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."

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

2



Language Distances

$$dist(l, l') = \sqrt{w(l, l') \cdot \sum_{f \in s(l,l')} (V(l)_f - V(l')_f)^2}$$

2

	L_1	L_2	L_3
L_1	0.0	0.2	0.9
L_2	0.9	0.0	0.5
L_3	0.2	0.2	0.0

Language Distances

2

	L_1	L_2	L_3
L_1	0.0	0.2	0.9
L_2	0.9	0.0	0.5
L_3	0.2	0.2	0.0

Language Distances

$$dist(l, l') = \sqrt{w(l, l') \cdot \sum_{f \in s(l,l')} (V(l)_f - V(l')_f)^2}$$

$$s(l, l') = \{f \in \{1 .. d\} \mid V(l)_f \neq \texttt{NaN} \text{ and } V(l')_f \neq \texttt{NaN}\}$$

2

	L_1	L_2	L_3
L_1	0.0	0.2	0.9
L_2	0.9	0.0	0.5
L_3	0.2	0.2	0.0

Language Distances

$$dist(l, l') = \sqrt{w(l, l') \cdot \sum_{f \in s(l,l')} (V(l)_f - V(l')_f)^2}$$

$$s(l, l') = \{f \in \{1 .. d\} \mid V(l)_f \neq \texttt{NaN} \text{ and } V(l')_f \neq \texttt{NaN}\}$$

$$w(l, l') = \frac{d}{|s(l, l')|}$$

2

	L_1	L_2	L_3
L_1	0.0	0.2	0.9
L_2	0.9	0.0	0.5
L_3	0.2	0.2	0.0

Language Distances

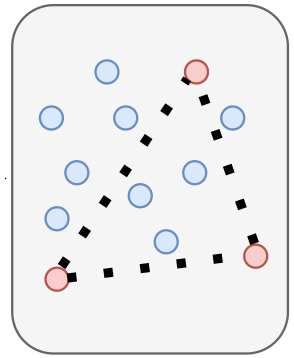
$$dist(l, l') = \sqrt{w(l, l') \cdot \sum_{f \in s(l,l')} (V(l)_f - V(l')_f)^2}$$

$$s(l, l') = \{f \in \{1 .. d\} \mid V(l)_f \neq \texttt{NaN} \text{ and } V(l')_f \neq \texttt{NaN}\}$$

$$w(l, l') = \frac{d}{|s(l, l')|}$$

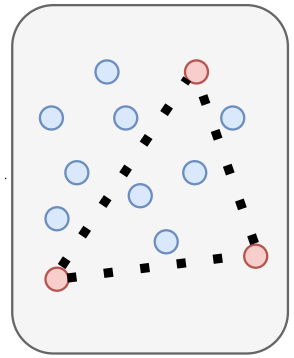
`sklearn.metrics.pairwise.nan_euclidean_distances()`

3



Sampling Algorithm

3

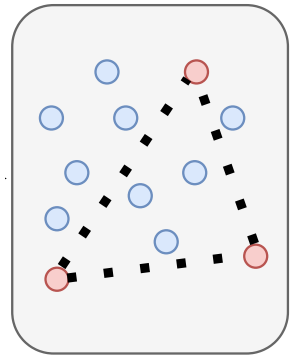


Sampling Algorithm

Maximum Diversity Problem

Finding a size k set of points where the sum of distances between all points in the set is maximal

3



Sampling Algorithm

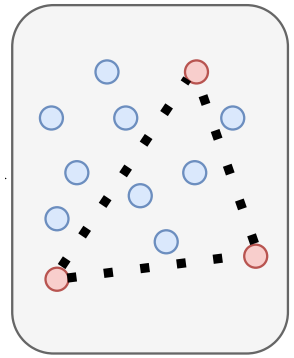
Maximum Diversity Problem

Finding a size k set of points where the sum of distances between all points in the set is maximal

MaxMin Diversity Problem

Finding a size k set of points where the closest two points are maximally distant

3



Sampling Algorithm

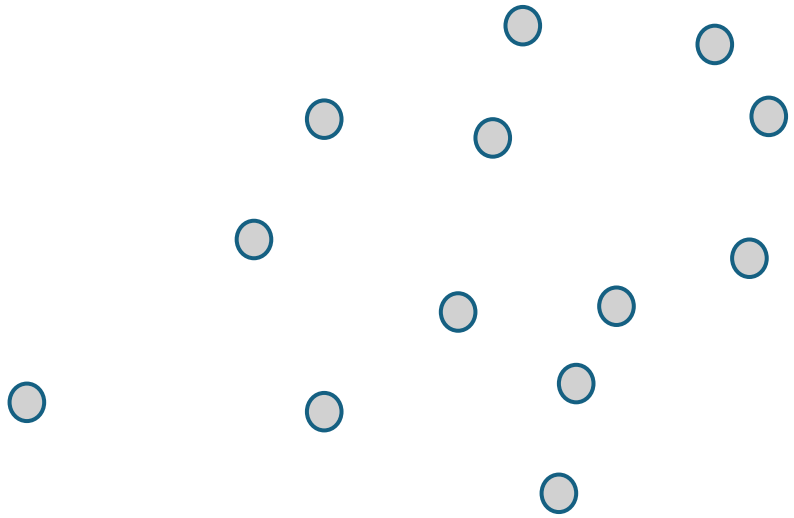
Maximum Diversity Problem
Finding a size k set of points where the sum of distances between all points in the set is maximal

MaxMin Diversity Problem
Finding a size k set of points where the closest two points are maximally distant

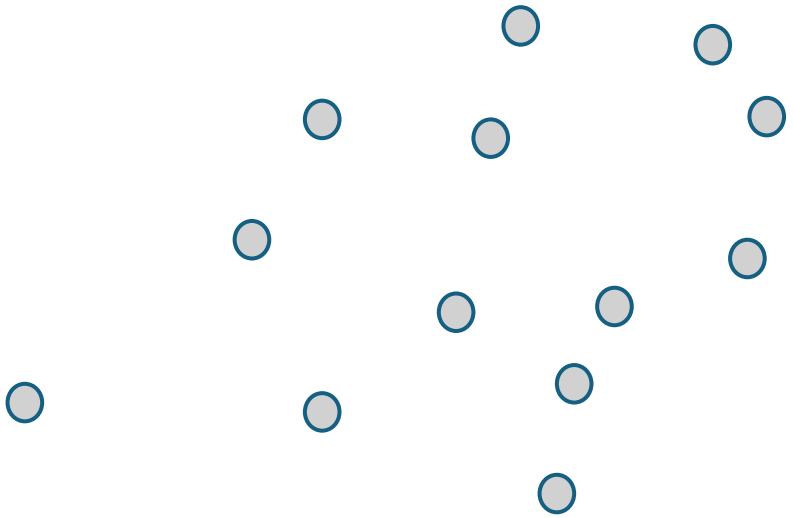
- Both problems are NP-hard (Kuo, Glover, and Dhir; 1993)
- We implement a greedy algorithm

Sampling Algorithm

T = 0



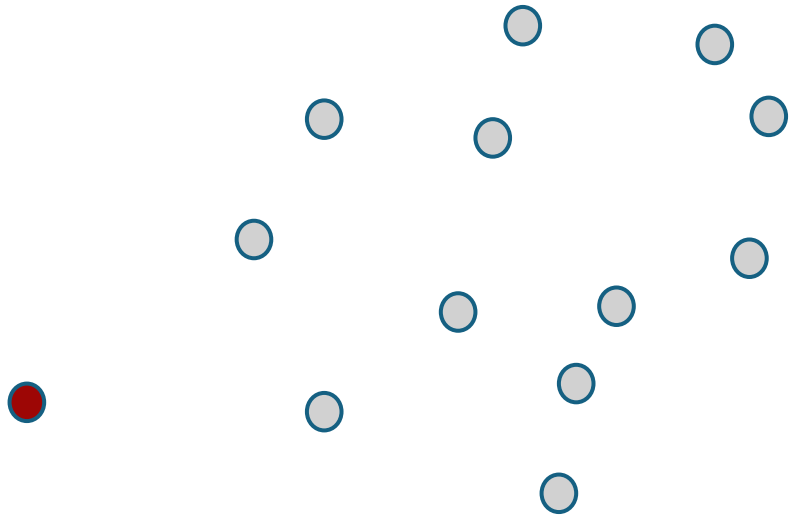
MaxSum



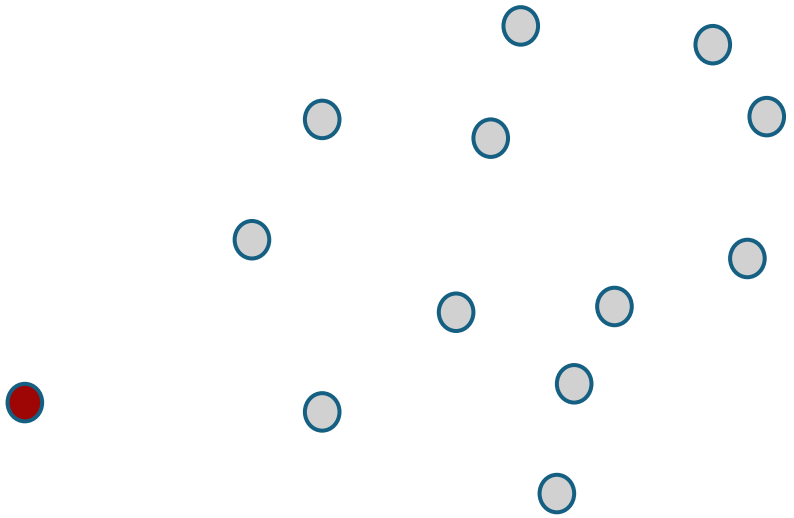
MaxMin

Sampling Algorithm

T = 1



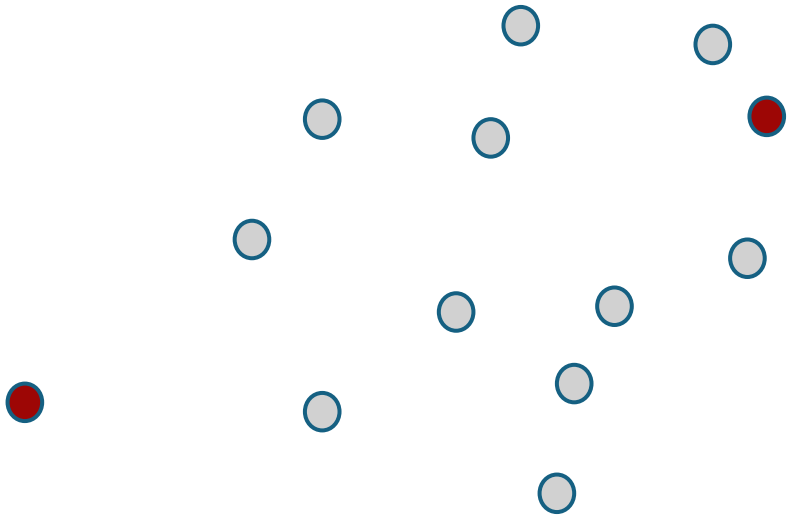
MaxSum



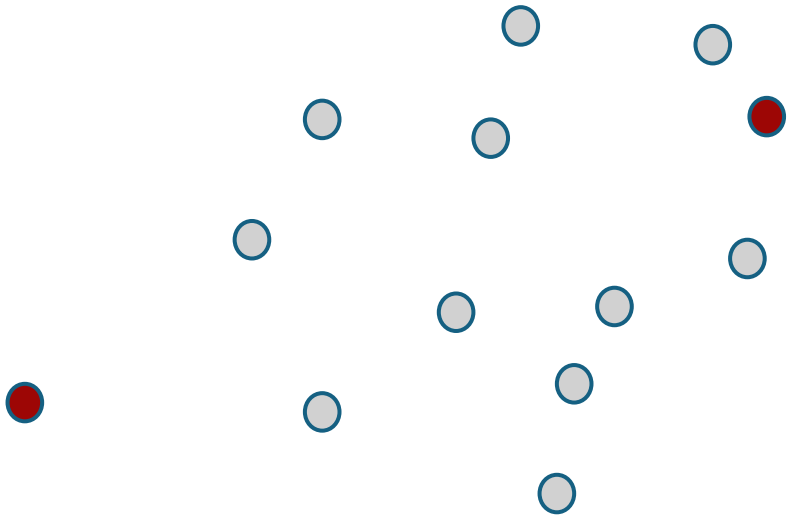
MaxMin

Sampling Algorithm

T = 1



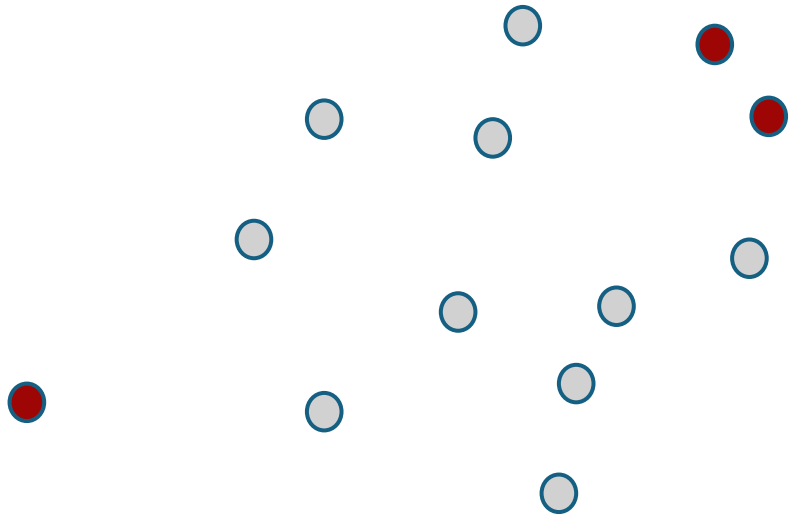
MaxSum



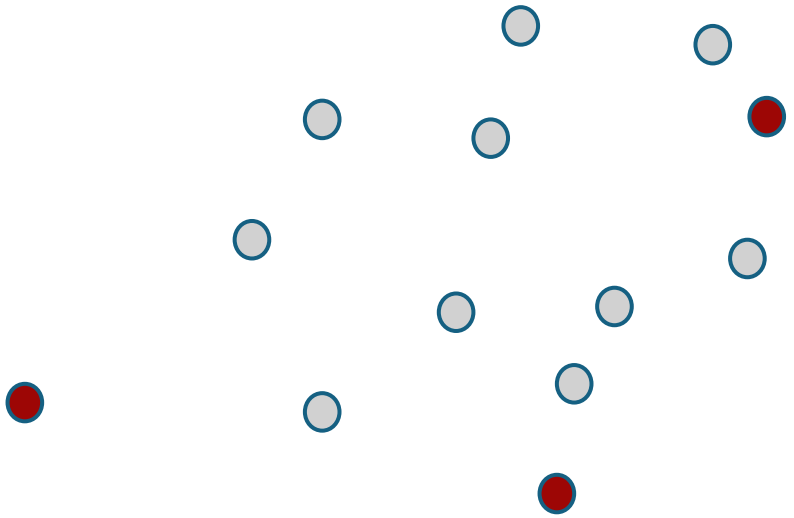
MaxMin

Sampling Algorithm

$T = 2$

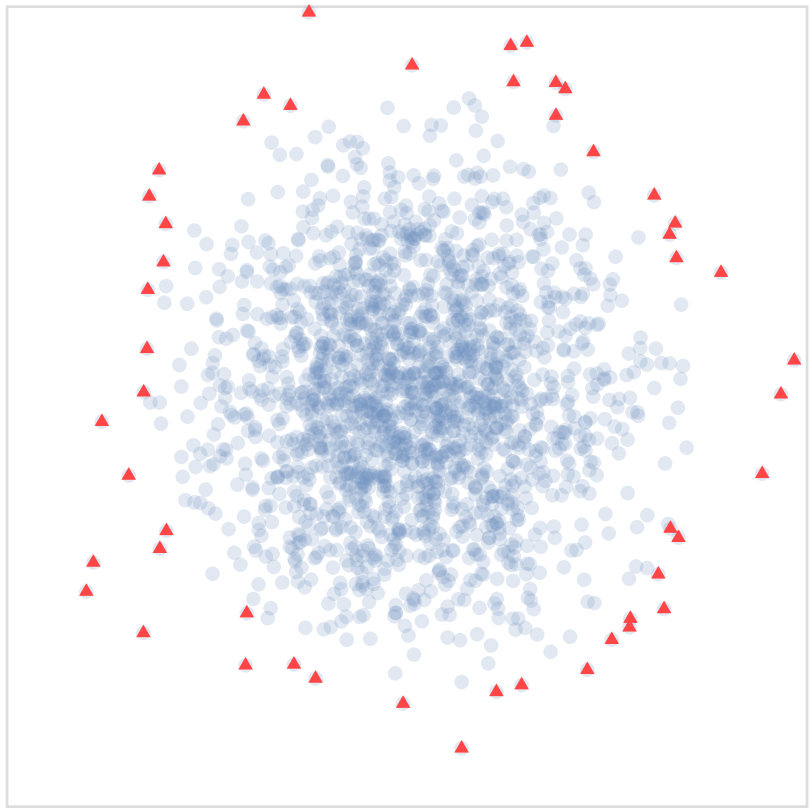


MaxSum

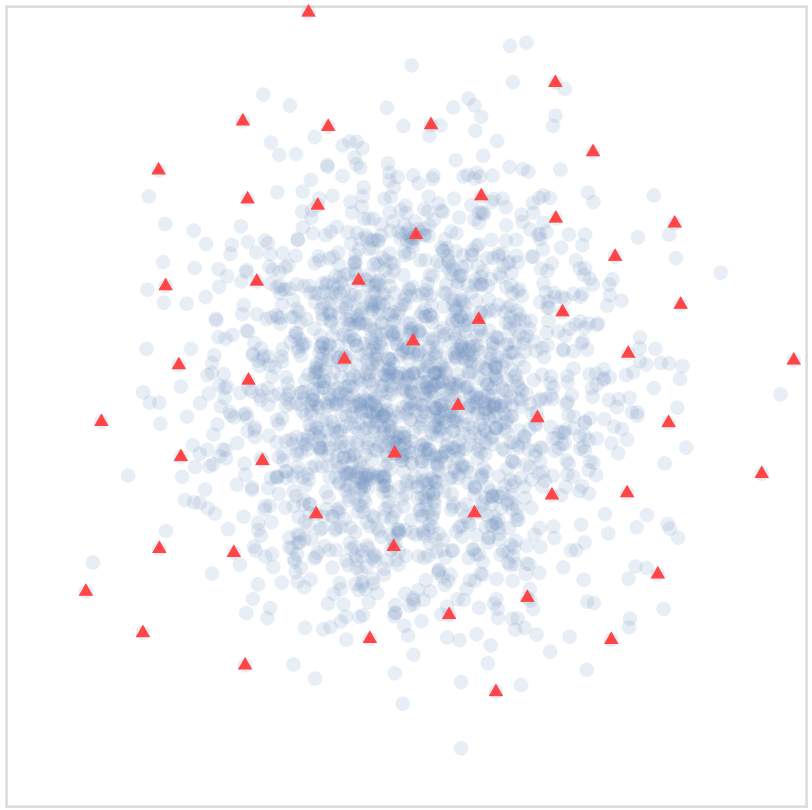


MaxMin

Sampling Algorithm



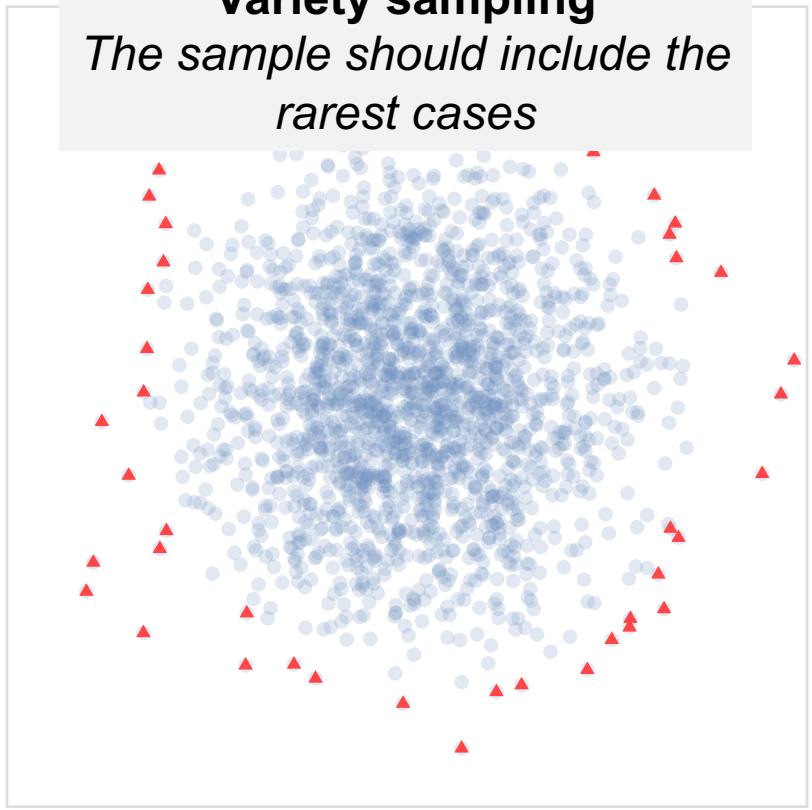
MaxSum



MaxMin

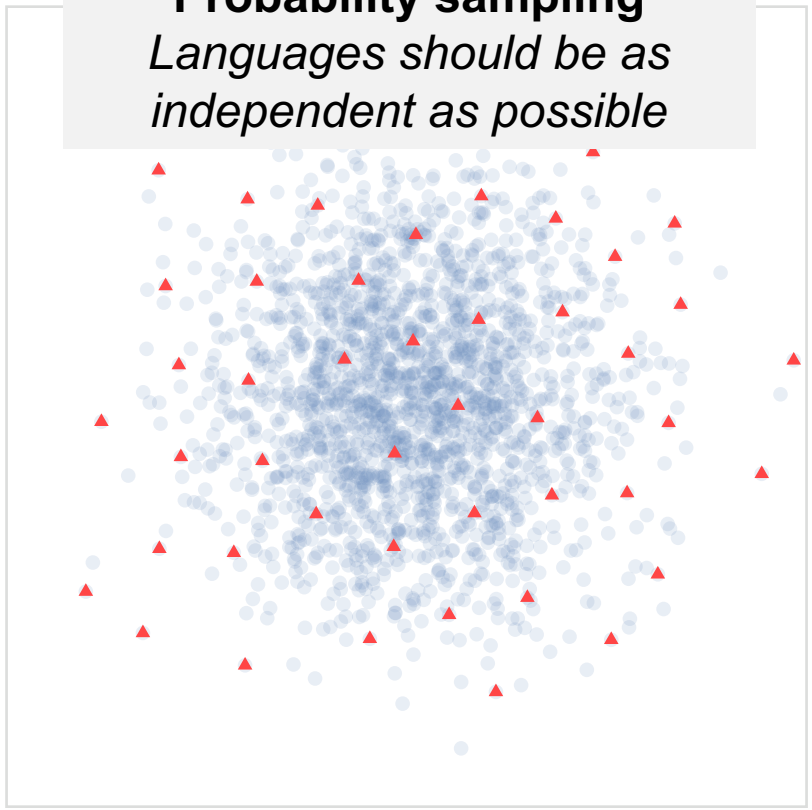
Sampling Algorithm

Variety sampling
The sample should include the rarest cases



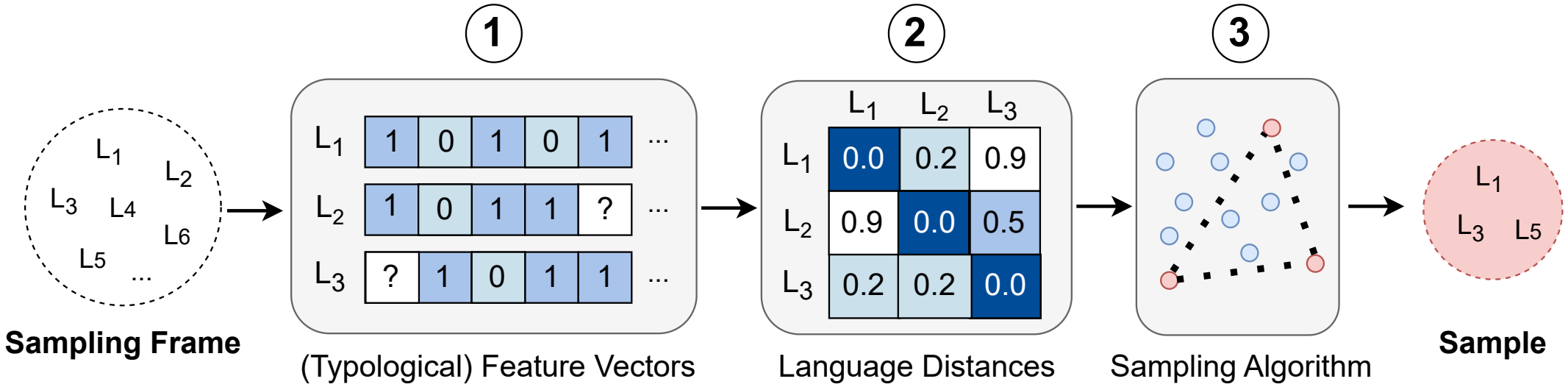
MaxSum

Probability sampling
Languages should be as independent as possible



MaxMin

A Framework for Systematic Language Sampling

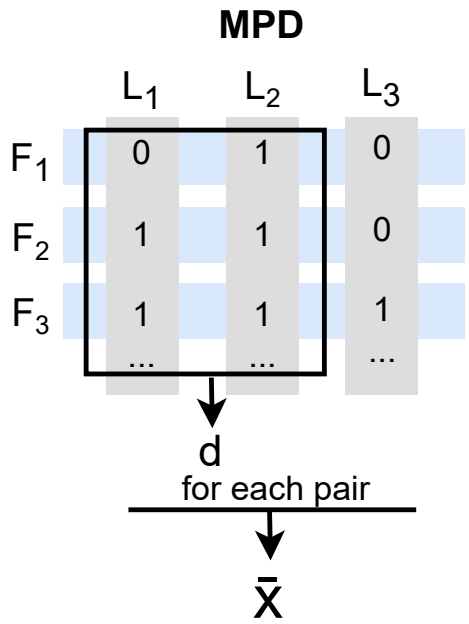


A Framework for Systematic Language Sampling

```python
from typdiv_sampling import Sampler
from pathlib import Path

# A list of glottocodes to sample from.
frame = ['stan1293', 'russ1263', 'finn1318', 'nucl1301', 'stan1290', 'kore1280']
k = 3  # The number of languages to sample.
seed = 1 # A random seed for the non-deterministic methods.

# Initialize with default setup.
sampler = Sampler()
sampler.sample_maxsum(frame, k)
> ['kore1280', 'russ1263', 'stan1290']
```

https://github.com/esther2000/typdiv-sampling

A Sanity Check: Mean Pairwise Distance (MPD)



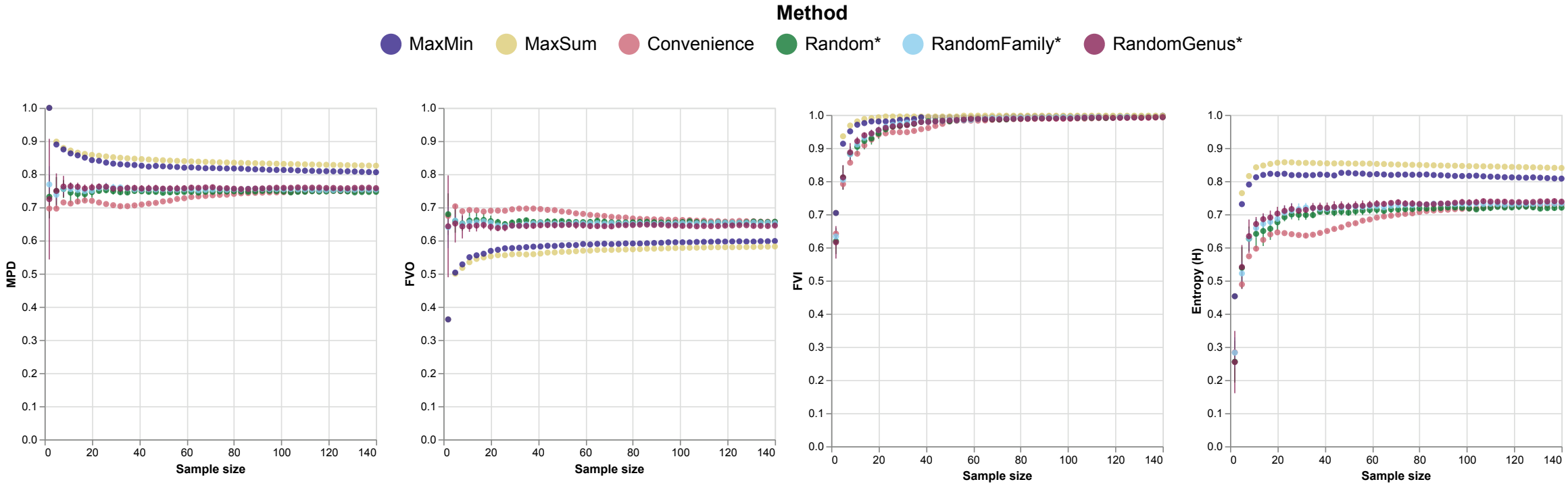
Do we (attempt to) maximize what we think we (attempt to) maxmimize?

Does it work?

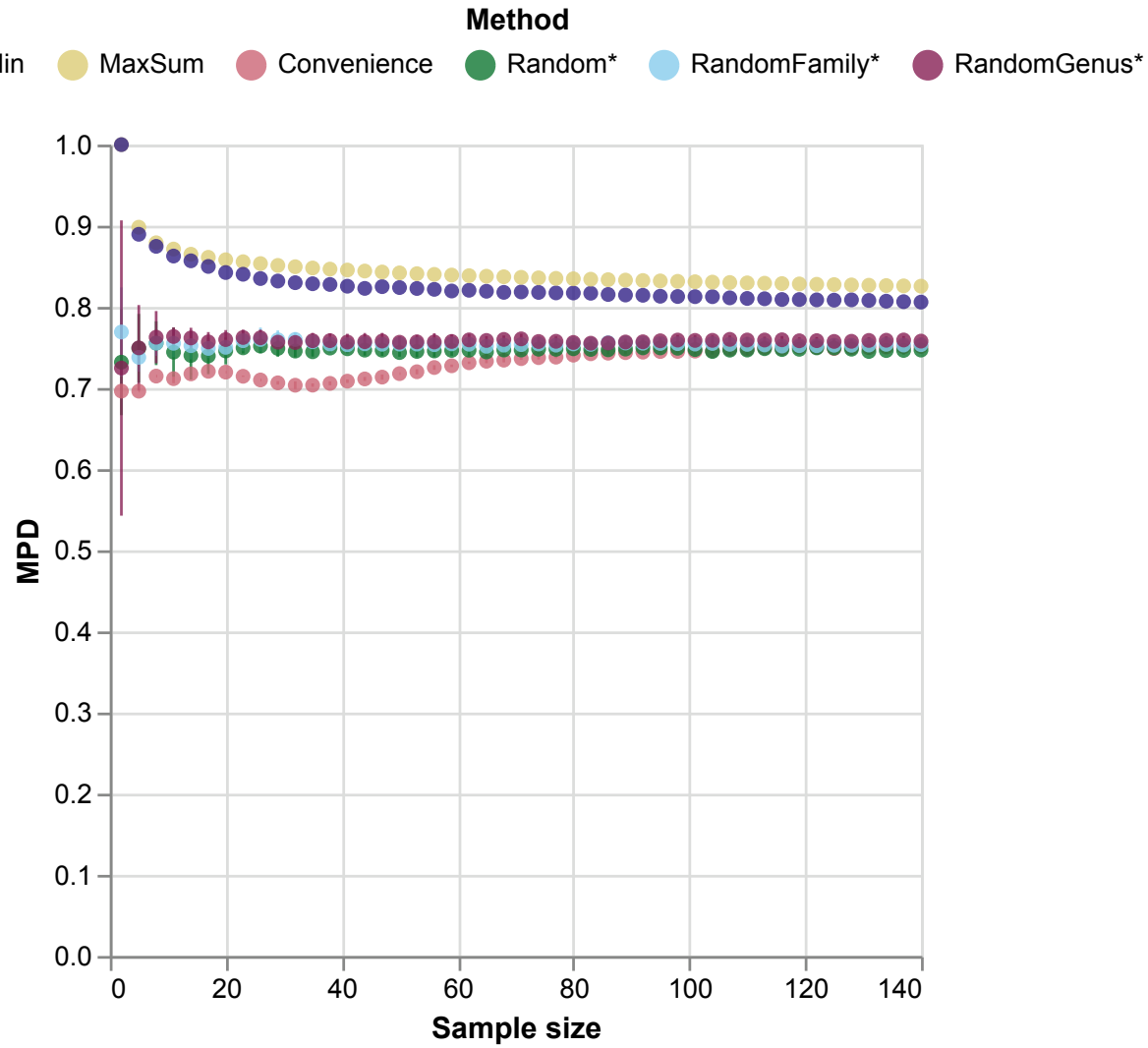
Does it work?


Method
MaxMin
MaxSum
Convenience
Random*
RandomFamily*
RandomGenus*

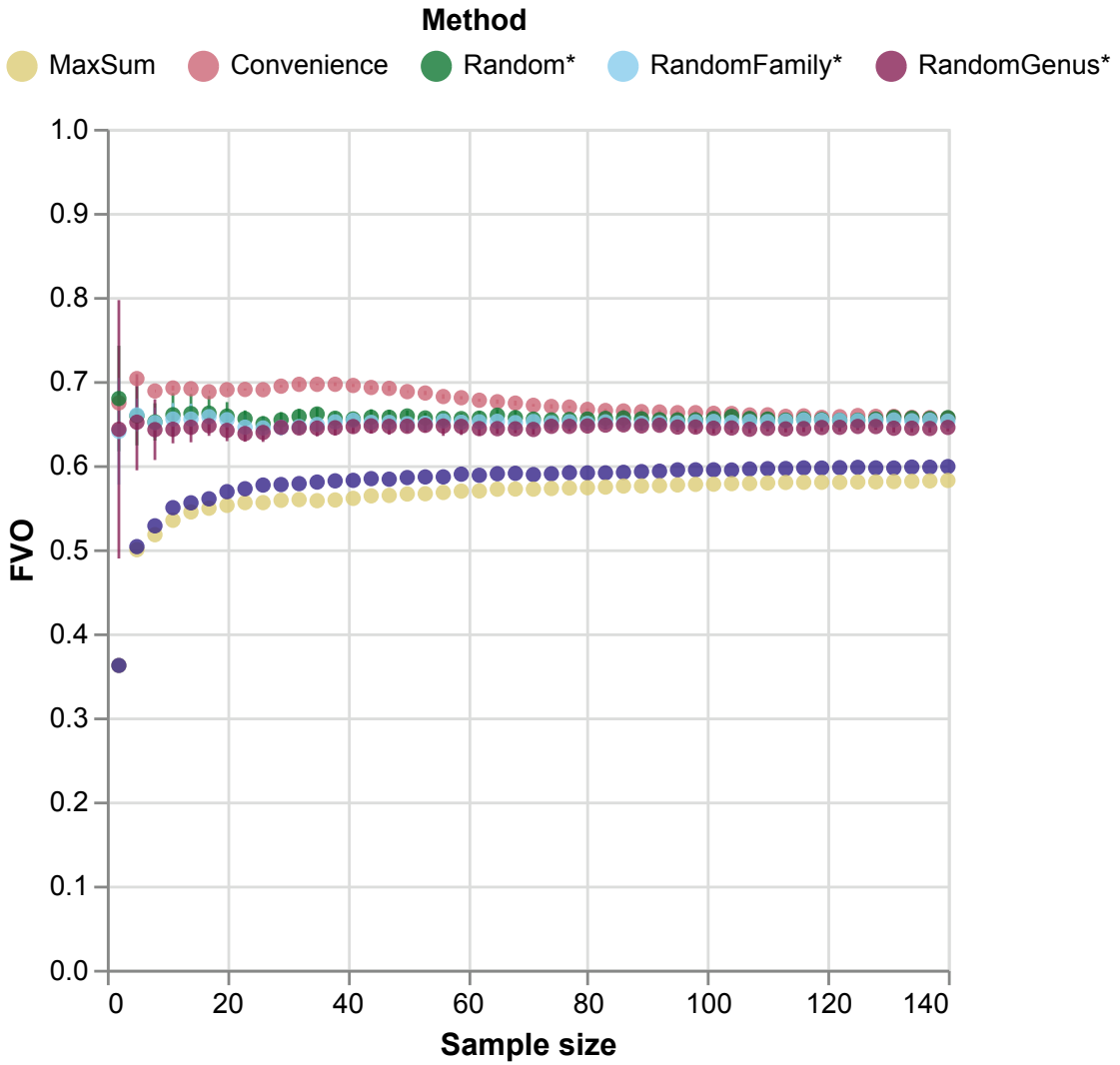
Does it work?


Method
MaxMin
MaxSum
Convenience
Random*
RandomFamily*
RandomGenus*
MPD
FVO
FVI
Entropy (H)
Sample size

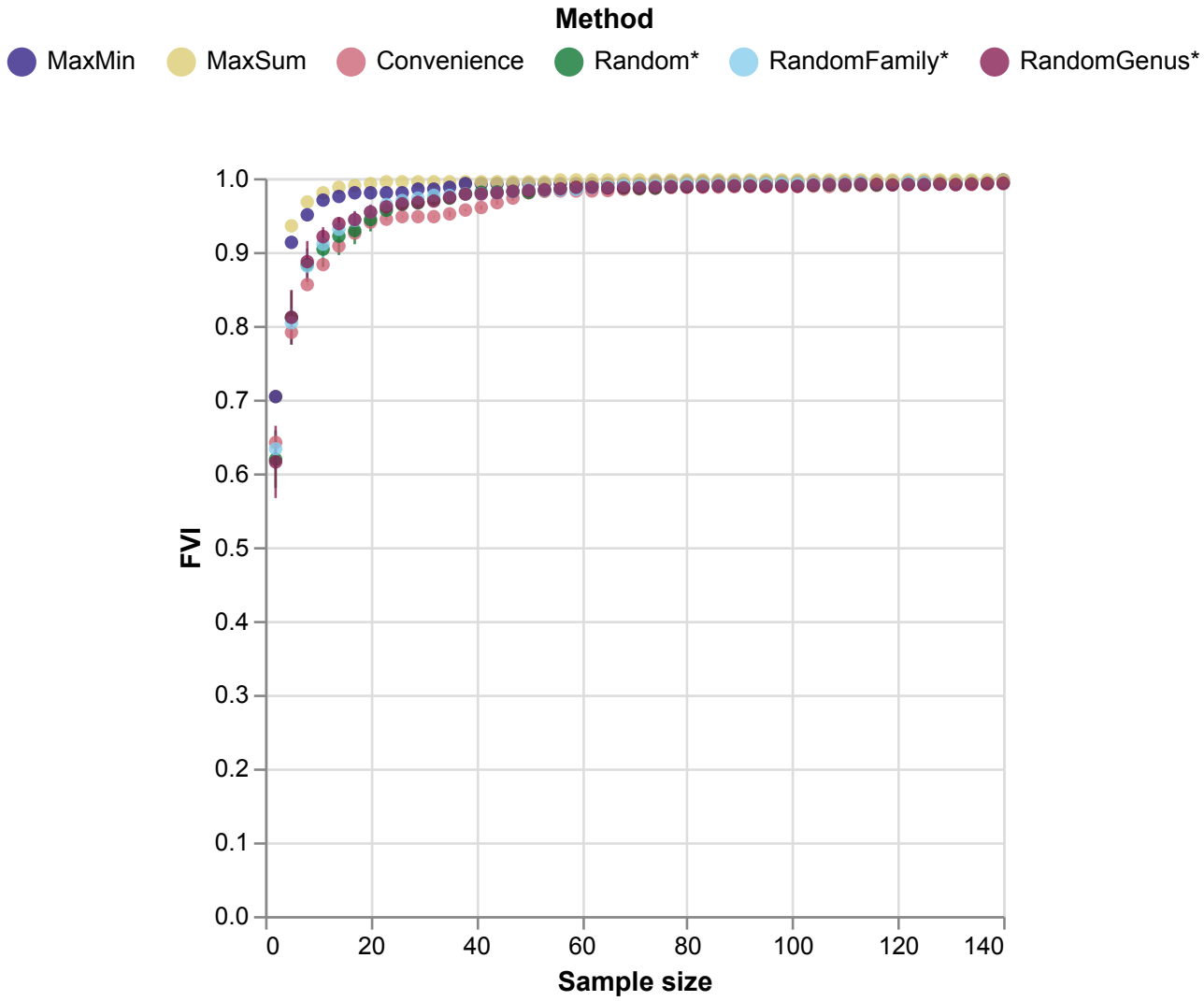
Does it work?


Method
MaxMin
MaxSum
Convenience
Random*
RandomFamily*
RandomGenus*
MPD
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80
100
120
140
Sample size

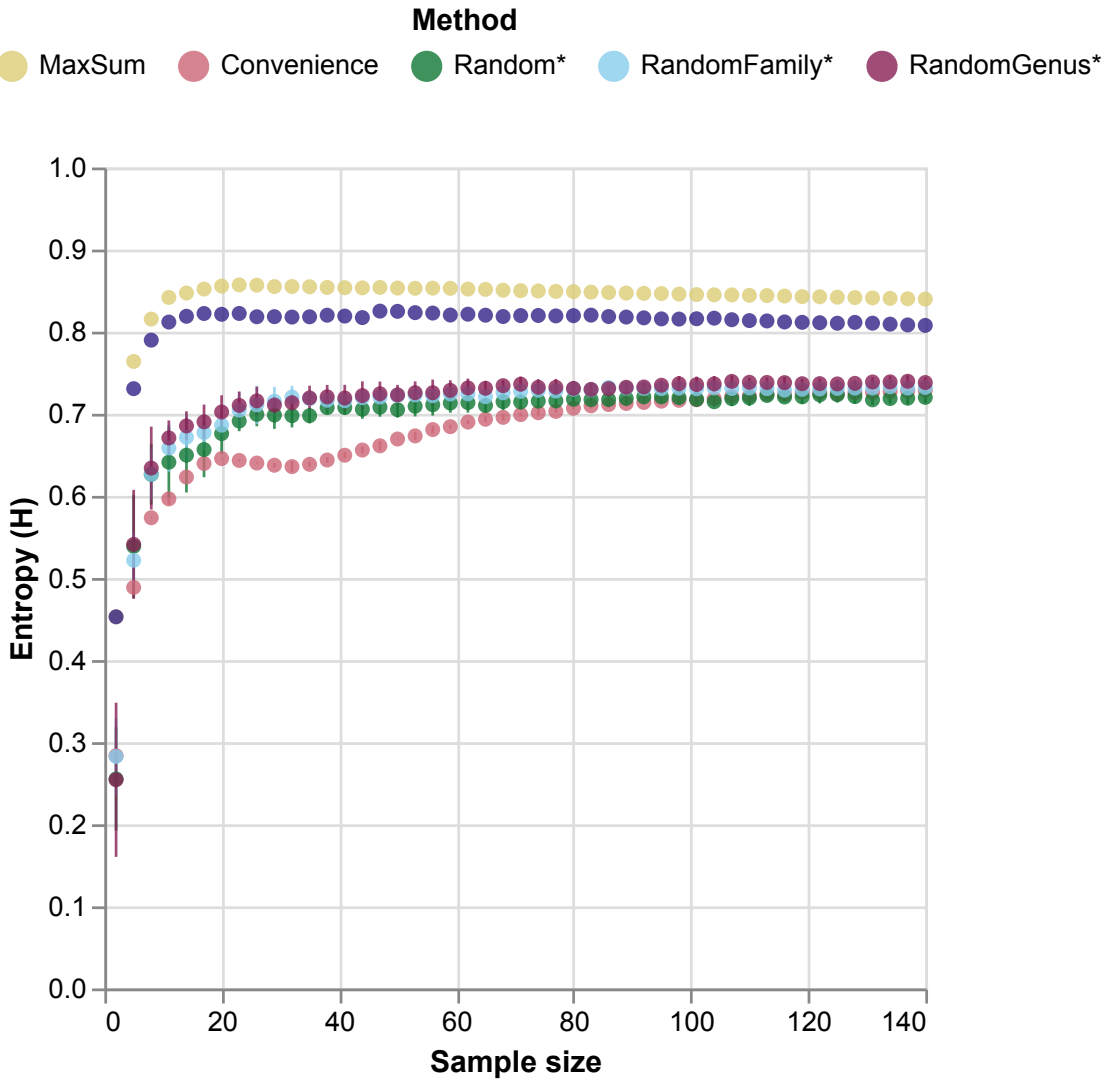
Does it work?


Method
MaxMin
MaxSum
Convenience
Random*
RandomFamily*
RandomGenus*
FVO
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80
100
120
140
Sample size

Does it work?


Method
MaxMin
MaxSum
Convenience
Random*
RandomFamily*
RandomGenus*
FVI
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80
100
120
140
Sample size

Does it work?


Method
MaxMin
MaxSum
Convenience
Random*
RandomFamily*
RandomGenus*
Entropy (H)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80
100
120
140
Sample size

Dataset expansion

Dataset expansion

Dataset	$\|L\|$ ($\|L \cap \text{GB}\|$)	$\mathcal{H}$	FVI
FloRes-200 (Team NLLB et al. 2022)	195 (105)	0.717	0.988
UD v2.14 (Zeman et al. 2024)	158 (80)	0.681	0.985
TyDiQA (Clark et al. 2020)	11 (7)	0.627	0.883
XCOPA (Ponti et al. 2020)	11 (7)	0.599	0.873
Aya Evaluation Suite (human-annotated) (Singh et al. 2024)	7 (5)	0.571	0.841

Dataset expansion

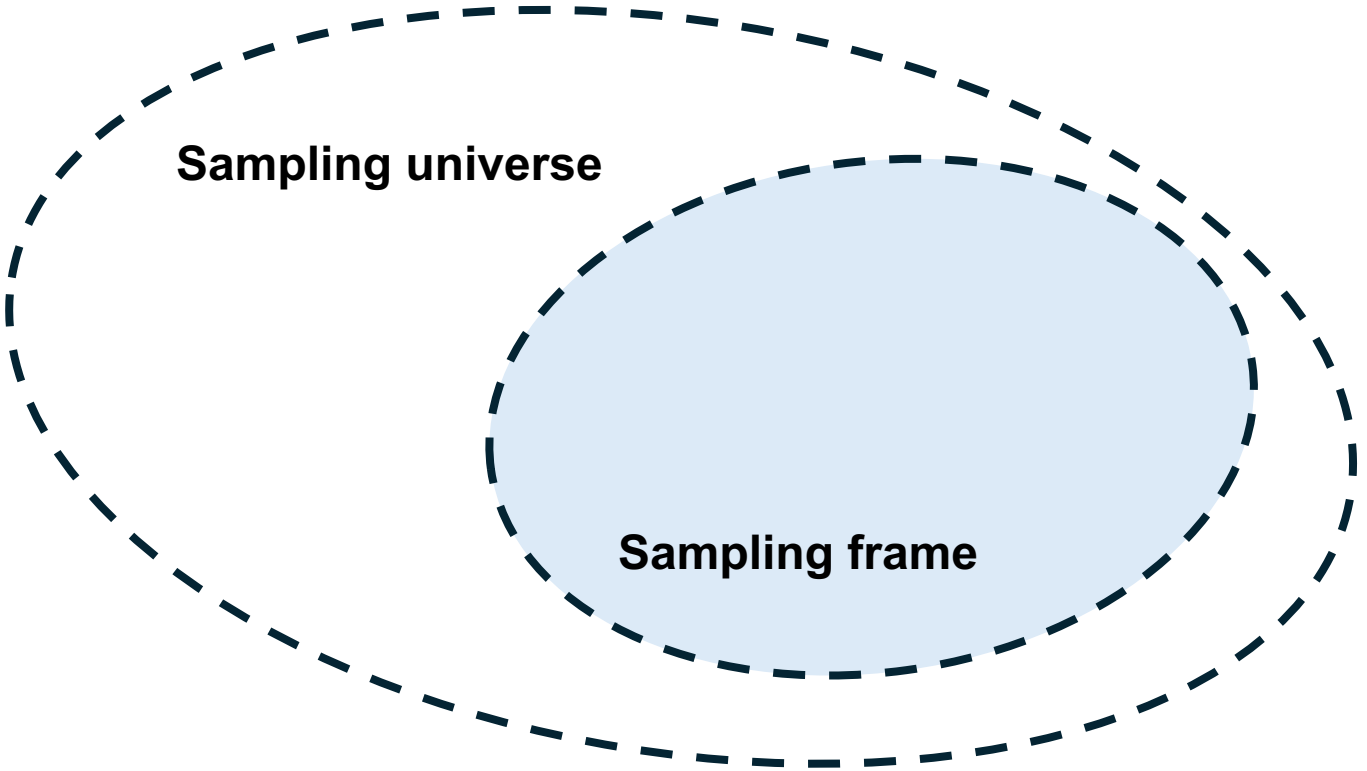
Dataset	$\lvert L\rvert$ ($\lvert L \cap \mathbf{GB}\rvert$)	$\mathcal{H}$	FVI	+ Language
FloRes-200 (Team NLLB et al. 2022)	195 (105)	0.717	0.988	Tariana
UD v2.14 (Zeman et al. 2024)	158 (80)	0.681	0.985	Tariana
TyDiQA (Clark et al. 2020)	11 (7)	0.627	0.883	Movima
XCOPA (Ponti et al. 2020)	11 (7)	0.599	0.873	Tariana
Aya Evaluation Suite (human-annotated) (Singh et al. 2024)	7 (5)	0.571	0.841	Yele

Dataset expansion

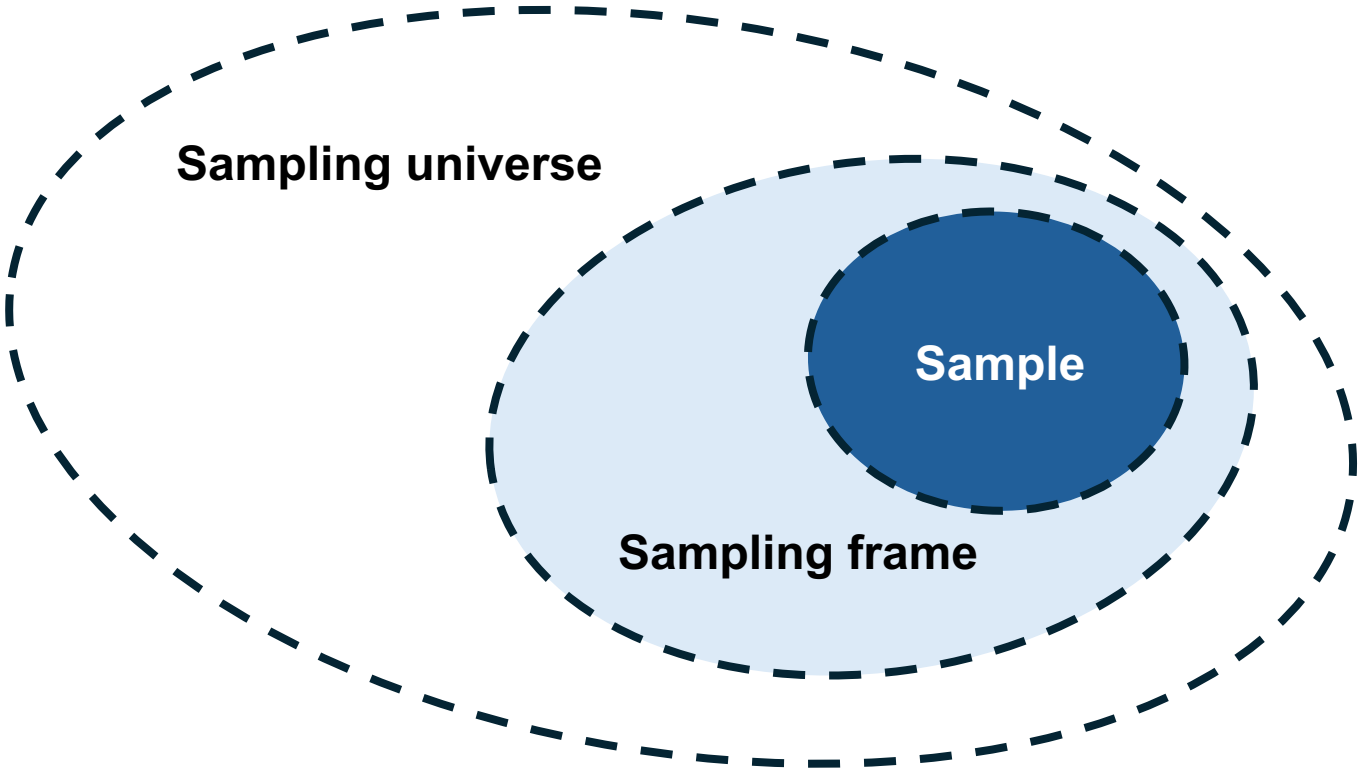
Dataset	$\lvert L\rvert$ ($\lvert L \cap \mathbf{GB}\rvert$)	$\mathcal{H}$	FVI	+ Language	$\mathcal{H}'$	FVI′
FloRes-200 (Team NLLB et al. 2022)	195 (105)	0.717	0.988	Tariana	0.721	0.988
UD v2.14 (Zeman et al. 2024)	158 (80)	0.681	0.985	Tariana	0.687	0.985
TyDiQA (Clark et al. 2020)	11 (7)	0.627	0.883	Movima	0.693	0.928
XCOPA (Ponti et al. 2020)	11 (7)	0.599	0.873	Tariana	0.667	0.913
Aya Evaluation Suite (human-annotated) (Singh et al. 2024)	7 (5)	0.571	0.841	Yele	0.660	0.898

What about more realistic scenarios?

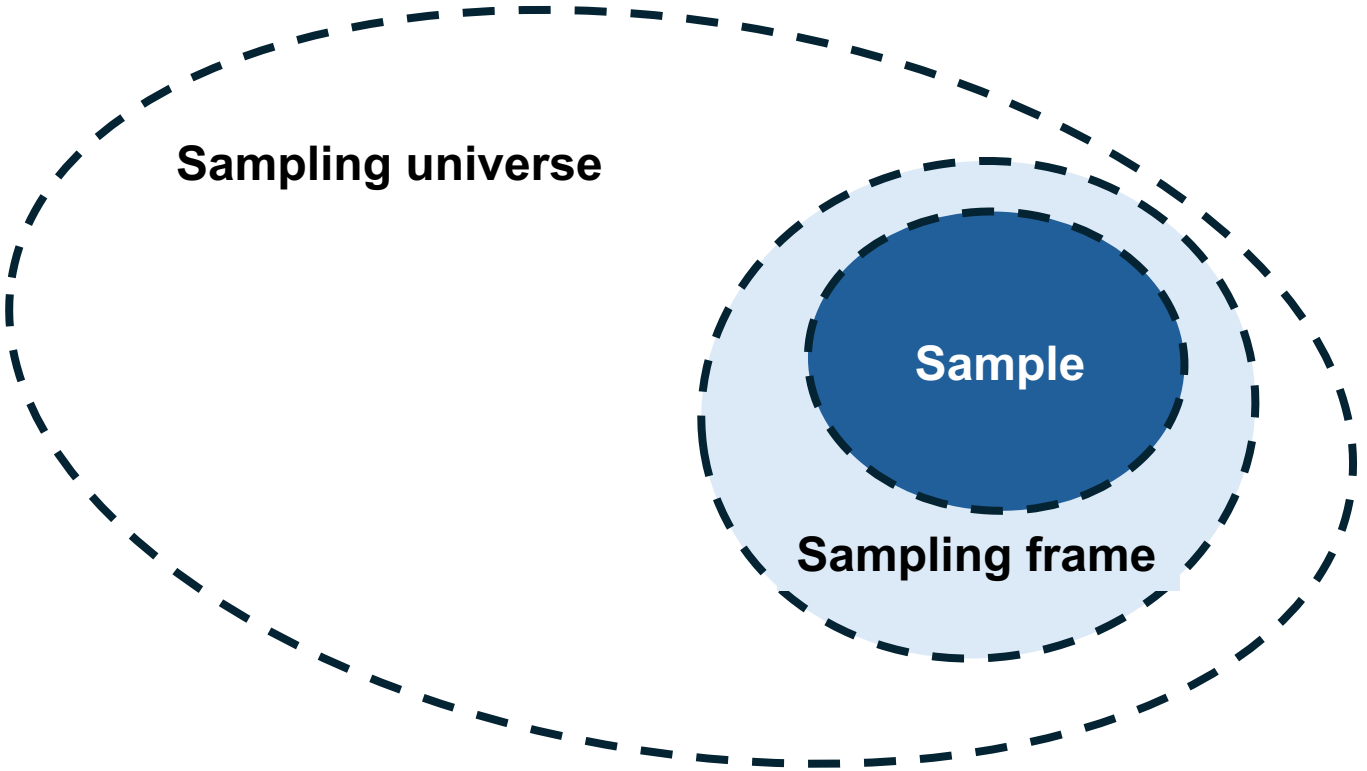
What about more realistic scenarios?



What about more realistic scenarios?



What about more realistic scenarios?



Dataset expansion

Possible Future Extensions

People have expressed interest in providing annotated data for the following languages but no valid data has been provided so far.

Akkadian	1	117K		Afro-Asiatic, Semitic
Amharic	2	–		Afro-Asiatic, Semitic
Archaic Irish	1	–		IE, Celtic
Assamese	1	–		IE, Indic
Balatipone	1	–		Bororoan
Bengali	3	–		IE, Indic
Bhojpuri	1	–		IE, Indic
Classical Nahuatl	1	–		Uto-Aztecan
Cuicatec	1	–		Oto-Manguean
Cusco Quechua	1	–		Quechuan
Czech	1	1,191K		IE, Slavic
Danish	1	–		IE, Germanic
Dargwa	1	–		Nakh-Daghestanian, Lak-Dargwa
English	1	–		IE, Germanic
Esperanto	1	–		Constructed
French	1	–		IE, Romance
Frisian	1	–		IE, Germanic
Gedeo	1	<1K		Afro-Asiatic, Cushitic

https://universaldependencies.org

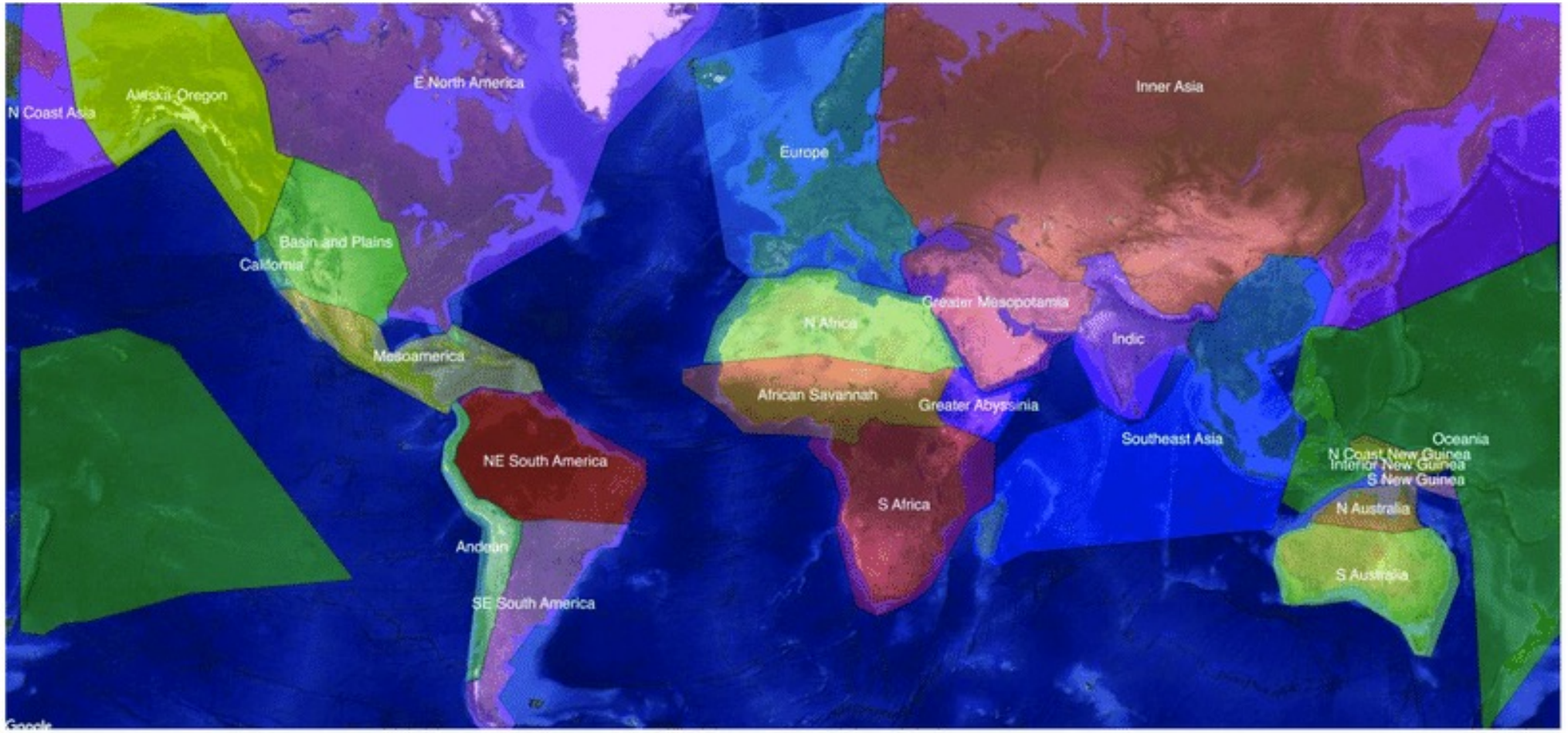
Dataset expansion

Table 3
Effects in diversity metrics from adding Seri to UD v2.14.

MPD ↑	**MPD′** ↑	**FVO** ↓	**FVO′** ↓	**FVI** ↑	**FVI′** ↑	$\mathcal{H}$ ↑	$\mathcal{H}'$ ↑
0.725	<u>0.728</u>	0.679	<u>0.677</u>	<u>0.985</u>	<u>0.985</u>	0.681	<u>0.685</u>

Beyond Typology: Geographical Sampling

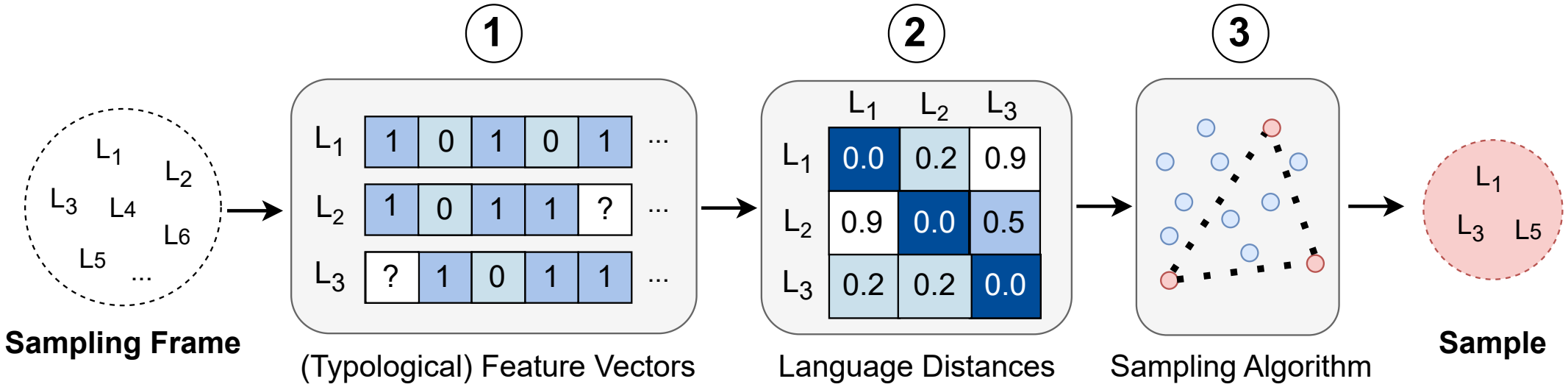
Beyond Typology: Geographical Sampling



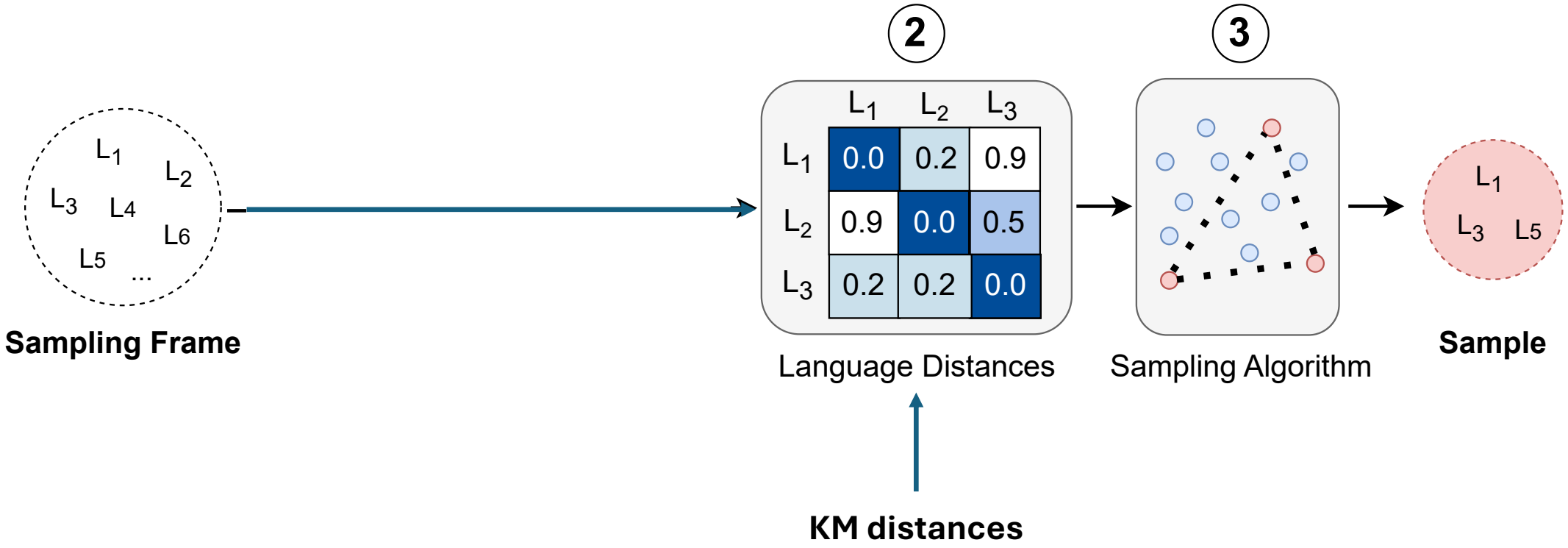
AUTOTYP areas

Bickel et al., 2017

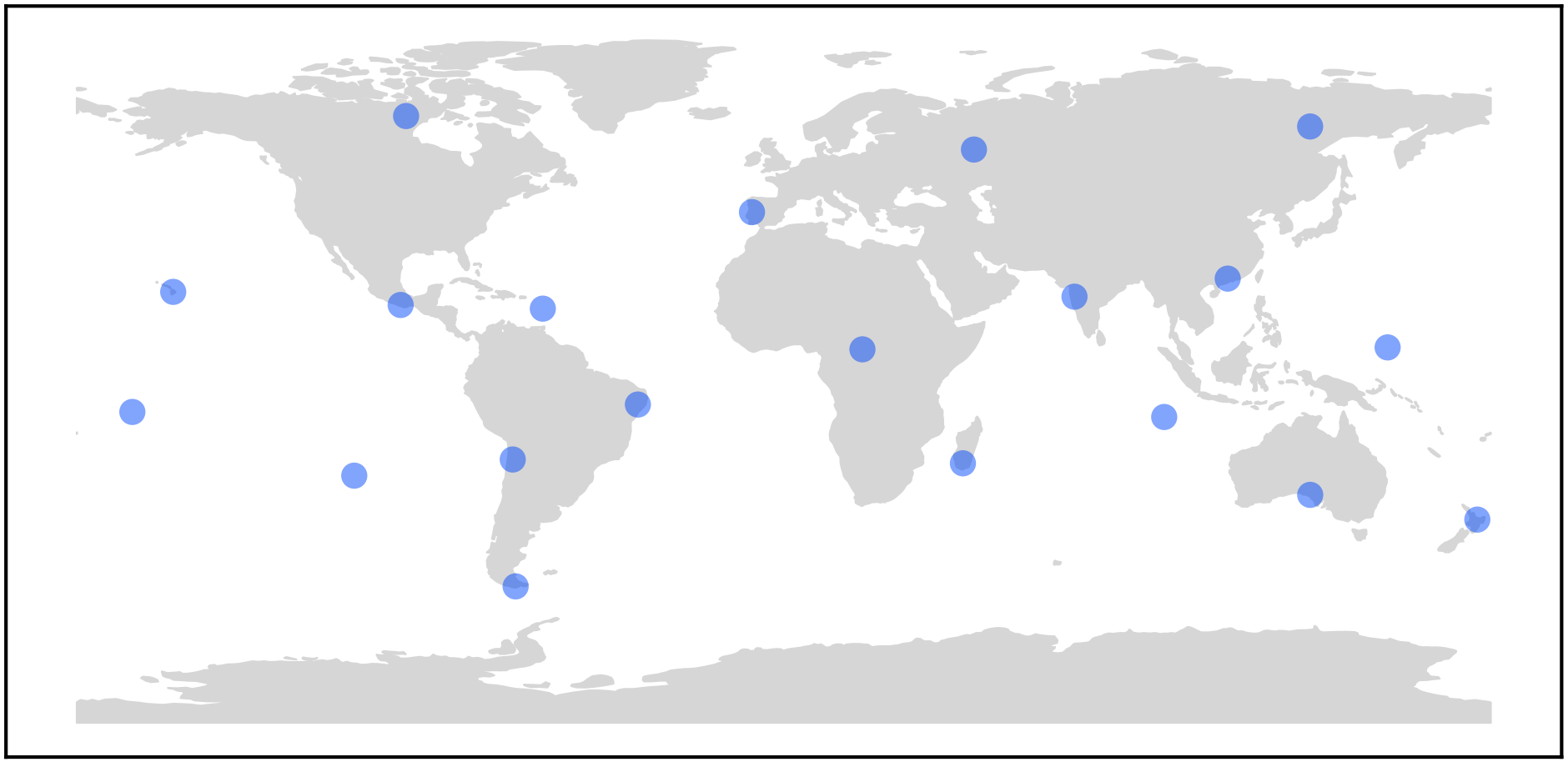
Beyond Typology: Geographical Sampling



Beyond Typology: Geographical Sampling



Beyond Typology: Geographical Sampling



Limitations

Limitations

- Data coverage (e.g. Spanish and German are currently not in Grambank)
- Should all features have an equal weight?
- Languages are more than morphosyntax

Limitations

- Data coverage (e.g. Spanish and German are currently not in Grambank)
- Should all features have an equal weight?
- Languages are more than morphosyntax
- What about within-language variation?

Limitations

- Data coverage (e.g. Spanish and German are currently not in Grambank)
- Should all features have an equal weight?
- Languages are more than morphosyntax

- What about within-language variation?

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

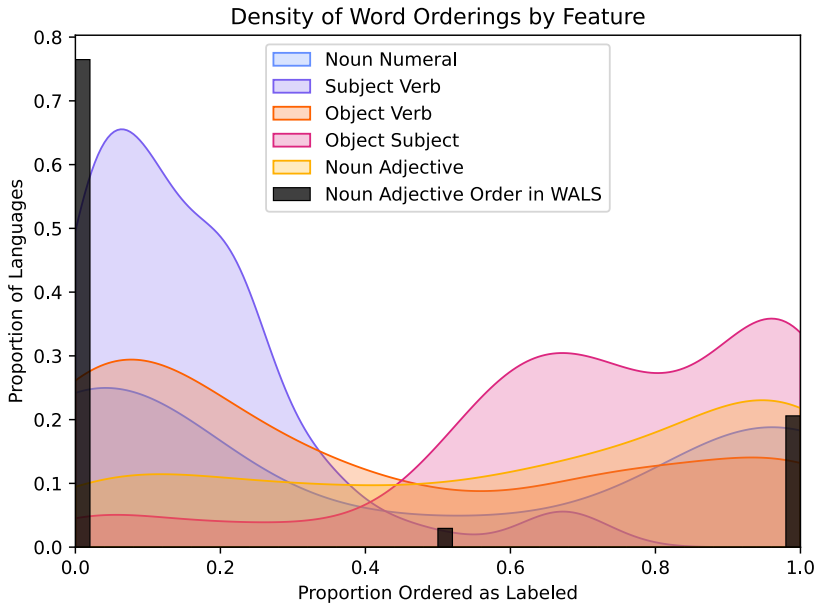
Limitations

for all $d \in$ UD Datasets **do**
  $na \leftarrow 0$ ▷ na is the Noun-Adj count
  $an \leftarrow 0$ ▷ 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

Multilingual Gradient Word-Order Typology from Universal Dependencies (Baylor et al., EACL 2024)

Limitations

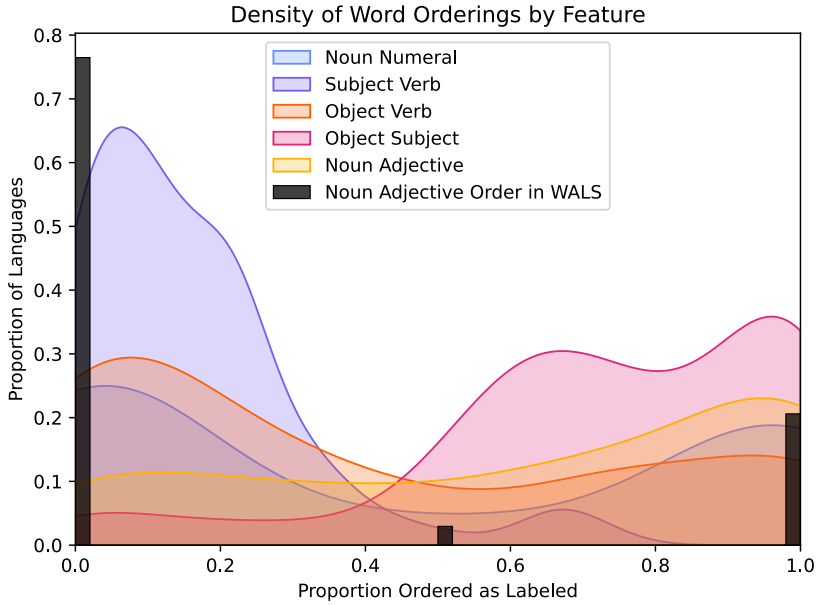
for all $d \in$ UD Datasets **do**
  $na \leftarrow 0$ ▷ na is the Noun-Adj count
  $an \leftarrow 0$ ▷ an is the Adj-Noun count
  for all sentence $s \in d$ **do**
    $na \leftarrow na+$ **count** Noun-Adj in s
    $an \leftarrow an+$ **count** Adj-Noun in s
  end for
  $na\_proportion \leftarrow \frac{na}{na+an}$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



Multilingual Gradient Word-Order Typology from Universal Dependencies (Baylor et al., EACL 2024)

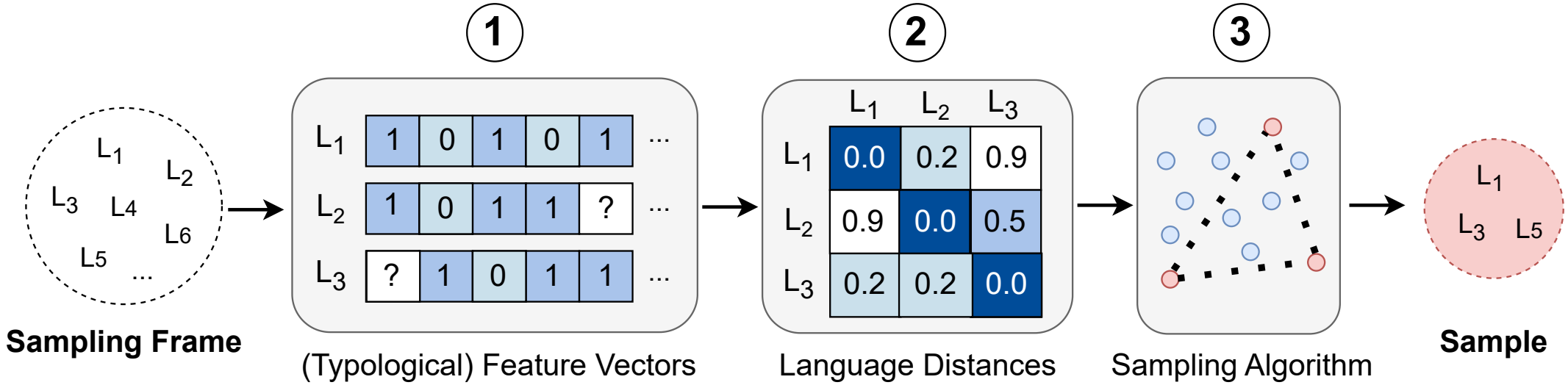
Limitations

for all $d \in$ UD Datasets **do**
  $na \leftarrow 0$ ▷ na is the Noun-Adj count
  $an \leftarrow 0$ ▷ 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

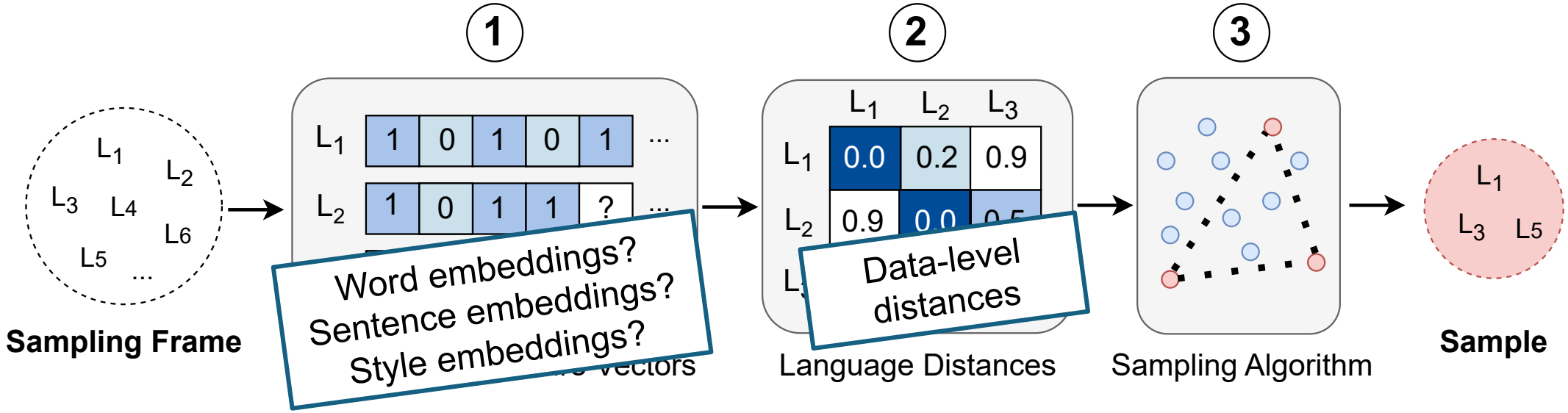


treebank_file	language	noun_adj_percent	adj_noun_percent	noun_num_percent	num_noun_percent	subj_verb_percent	verb_subj_percent
hr_set-ud-test.conllu	hr	0.0520547945205479	0.947945205479452	0.8928571428571429	0.1071428571428571	0.6995798319327731	0.30042016806722
cs_cltt-ud-train.conllu	cs	0.477859778597786	0.522140221402214	0.1	0.9	0.8208955223880597	0.179104477611940
da_ddt-ud-train.conllu	da	0.1899827288428324	0.8100172711571675	0.6839080459770115	0.3160919540229885	0.7042518837459634	0.295748116254036
nl_alpino-ud-test.conllu	nl	0.0588235294117647	0.9411764705882352	0.9411764705882352	0.0588235294117647	0.7037861915367484	0.29621380846325

Future Work


1
2
3
L1
L2
L3
L4
L6
L5
...
Sampling Frame
L1 1 0 1 0 1 ...
L2 1 0 1 1 ? ...
L3 ? 1 0 1 1 ...
(Typological) Feature Vectors
L1 L2 L3
L1 0.0 0.2 0.9
L2 0.9 0.0 0.5
L3 0.2 0.2 0.0
Language Distances
Sampling Algorithm
L1
L3 L5
Sample

Future Work


1
2
3
L1
L2
L3
L4
L6
L5
...
Sampling Frame
L1 1 0 1 0 1 ...
L2 1 0 1 1 ? ...
Word embeddings?
Sentence embeddings?
Style embeddings?
L1 L2 L3
L1 0.0 0.2 0.9
L2 0.9 0.0
Data-level distances
Language Distances
Sampling Algorithm
L1
L3 L5
Sample

Conclusion

All samples are diverse, but some samples are more diverse than others?

Can we approach typological diversity more systematically?

All samples are diverse, but some samples are more diverse than others?

Probably.

Can we approach typological diversity more systematically?

All samples are diverse, but some samples are more diverse than others?

Probably.

Can we approach typological diversity more systematically?

Yes.

Thank You!

And thanks to my co-authors:
Emi Baylor, Johannes Bjerva, Miryam de Lhoneux, Andreas Holck Høeg-Petersen, Wessel Poelman, Anders Schlichtkrull

Beyond Post-Hoc Typological Diversity in NLP

Esther Ploeger
Dept. of Computer Science
Aalborg University
AAU-NLP
espl@cs.aau.dk

9 January 2025

References

Baylor, Emi, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from Universal Dependencies. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 42–49.

Bell, Alan. 1978. Language samples. Universals of human language, 1:123–156.

Clark, Jonathan H., Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics, 8:454–470.

Dahl, Östen. 2008. An exercise in a posteriori language sampling. Language Typology and Universals, 61(3):208–220.

Dryer, Matthew S. 2013. Order of Adjective and Noun. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.4)

Greenberg, Joseph Harold. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Universals of Language. MIT press, Cambridge, MA, pages 40–70.

Haynie, Hannah J., Damián Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank's typological advances support computational research on diverse languages. In Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 147–149.

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

Majewska, Olga, Ivan Vulic, Diana McCarthy, and Anna Korhonen. 2020. Manual clustering and spatial arrangement of verbs for multilingual evaluation and typology analysis. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4810–4824.

Miestamo, Matti, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. Linguistic Typology, 20(2):233–296.

Nichols, Johanna and Balthasar Bickel. 2009. The autotyp genealogy and geography database: 2009 release. URL: https://github.com/autotyp/autotyp-data

Ploeger, Esther, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2024. A Principled Framework for Evaluating on Typologically Diverse Languages. *arXiv preprint arXiv:2407.05022*

References

Ploeger, Esther, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is ”typological diversity” in NLP? In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5681–5700.

Ponti, Edoardo Maria, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2362–2376.

Rijkhoff, Jan and Dik Bakker. 1998. Language sampling. Linguistic Typology, 2(3):263–314.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld, and Peter Kahrel. 1993. A method of language sampling. Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”, 17(1):169–203.

Singh, Shivalika, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. arXiv preprint arXiv:2402.06619.

Skirgård, Hedvig, Hannah J Haynie, Damián E Blasi, Harald Hammarström, Jeremy Collins, Jay J Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. Science Advances, 9.

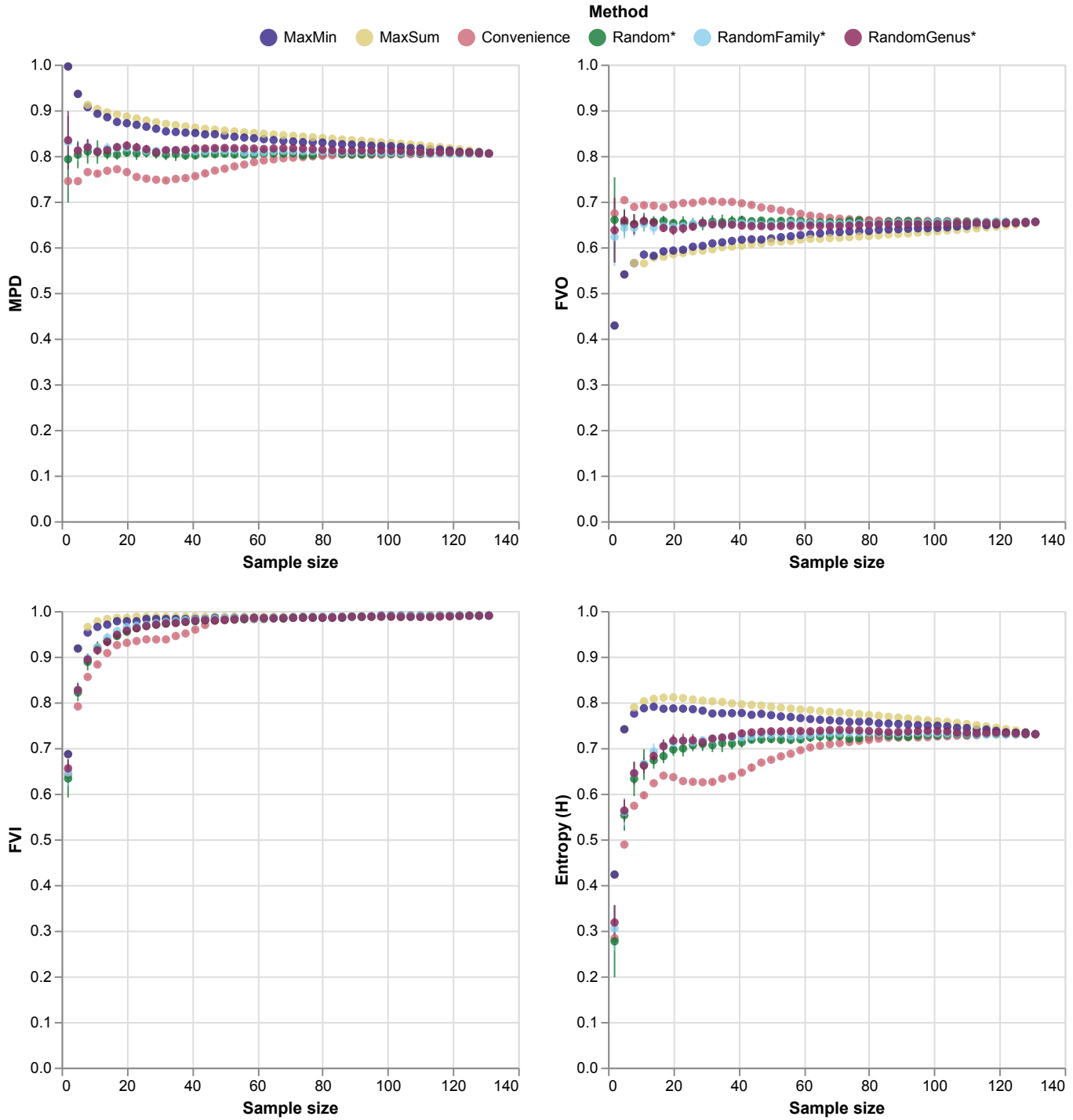
Skirgård, Hedvig, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, et al, 2023. Grambank v1.0. Dataset.

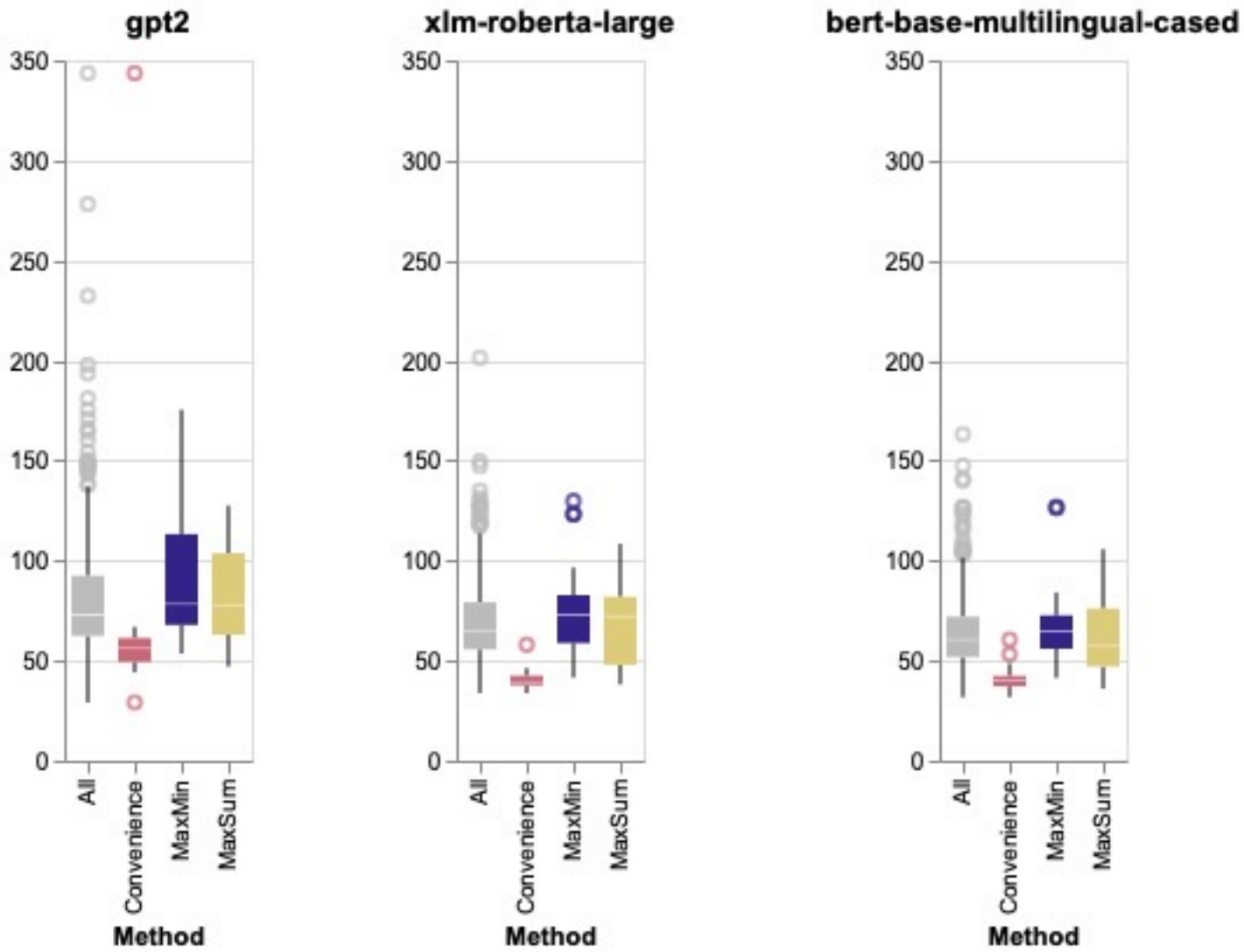
Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Viveka Velupillai. 2012. An Introduction to Linguistic Typology, reprinted 2017 with corrections edition. John Benjamins Publishing Company, Amsterdam Philadelphia, PA

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agic, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, et al. 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.

Extra Slides


Method
MaxMin
MaxSum
Convenience
Random*
RandomFamily*
RandomGenus*
MPD
FVO
FVI
Entropy (H)
Sample size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
20
40
60
80
100
120
140


gpt2
xlm-roberta-large
bert-base-multilingual-cased
350
300
250
200
150
100
50
0
All
Convenience
MaxMin
MaxSum
Method

Input: k: number of languages to sample, $\mathcal{L}$: sampling frame, $dist$: function giving the pairwise distance between languages in $\mathcal{L}$ (e.g., Equation 1)

Algorithm 1 MaxSum Sampling

$l \leftarrow \arg\max_{l \in \mathcal{L}} \sum_{l' \in \mathcal{L}} dist(l, l')$
$L \leftarrow \{l\}$
while $|L| < k$ **do**
  $L \leftarrow L \cup \{\arg\max_{l \in \mathcal{L} \setminus L} \sum_{l' \in L} dist(l, l')\}$
end while
return L

Algorithm 2 MaxMin Sampling

$l \leftarrow \arg\max_{l \in \mathcal{L}} \sum_{l' \in \mathcal{L}} dist(l, l')$
$l' \leftarrow \arg\max_{l' \in \mathcal{L}} dist(l, l')$
$L \leftarrow \{l, l'\}$
while $|L| < k$ **do**
  $L \leftarrow L \cup \{\arg\max_{l \in \mathcal{L} \setminus L} (\min_{l' \in L} dist(l, l'))\}$
end while
return L