Text-based typology for modeling linguistic diversity in NLP

Ximena Gutierrez-Vasques
National Autonomous University of Mexico (UNAM)

March 2024

European Chapter | EACL

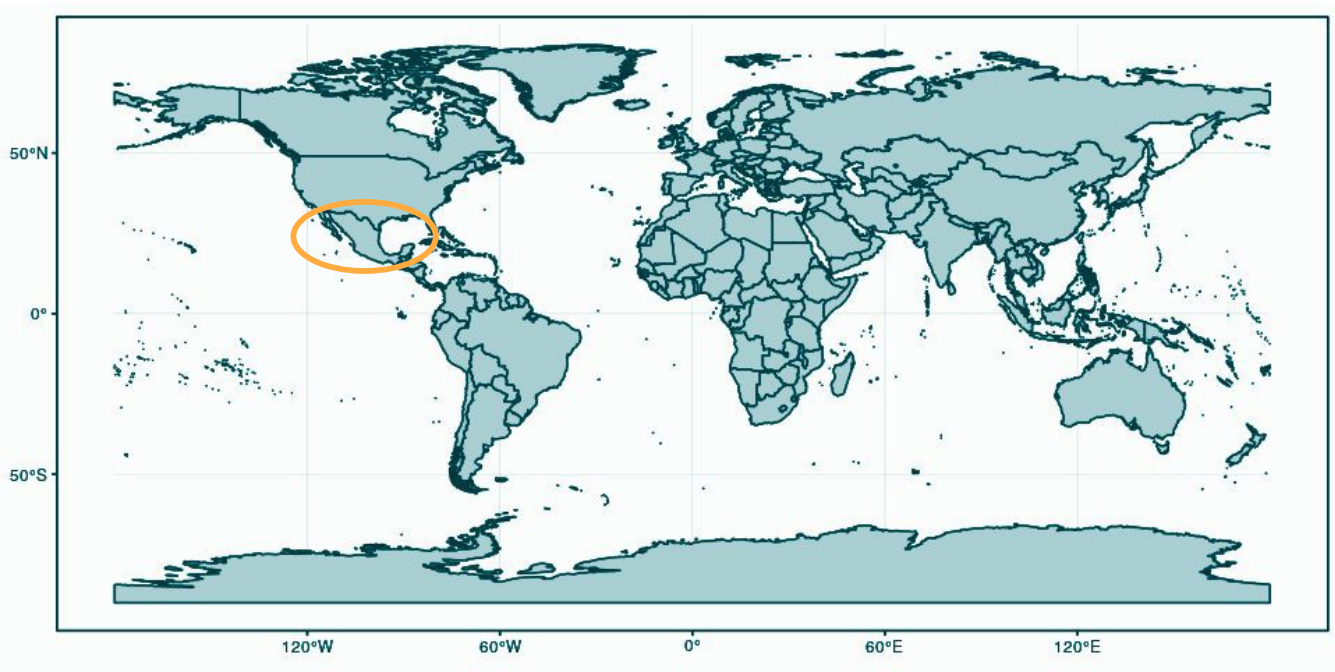
SIGTYP

1. <u>Motivation</u>

2. Approaching linguistic diversity in NLP

3. Text-based typology

How I got interested in the field…



- ~ 68 languages
- 11 language families
- More than 300 dialectal variations

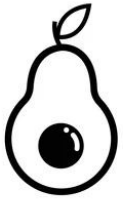
- Nahuatl

- **Nahuatl**

Xocolatl
Chocolate

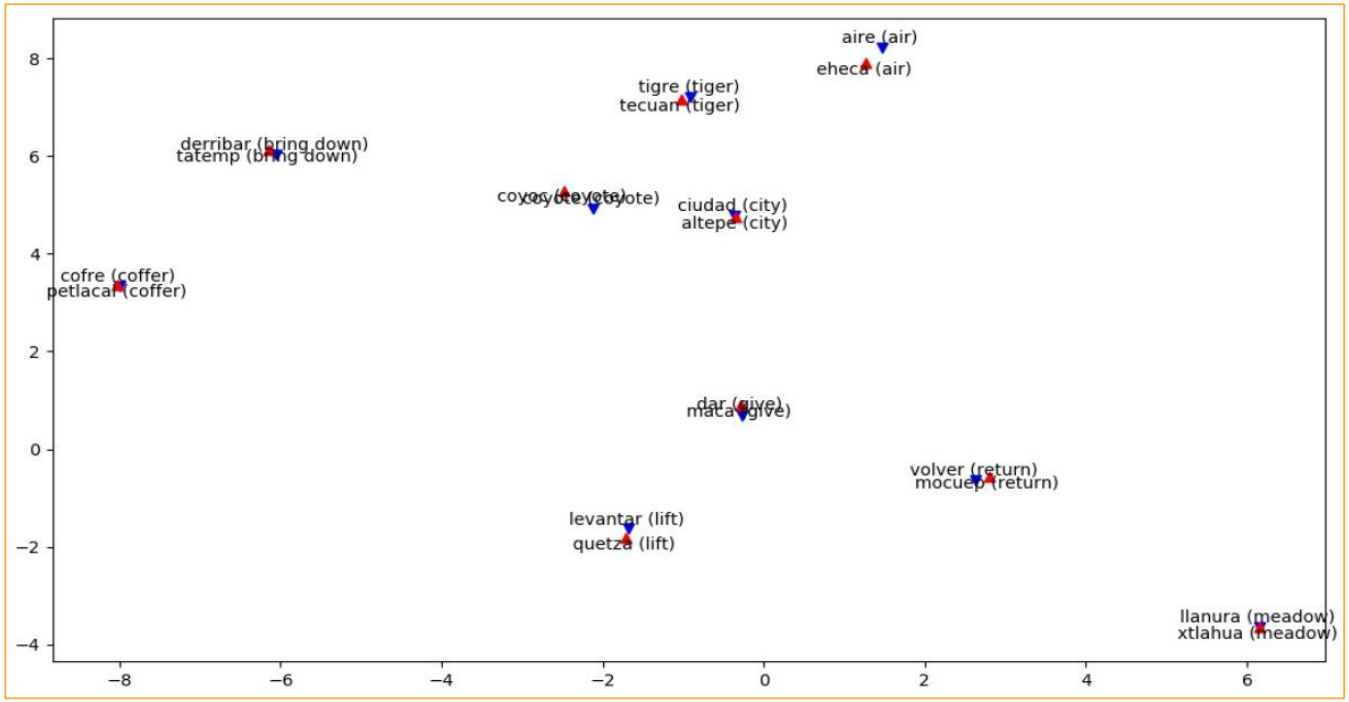
Tomatl
Tomato

Awacatl
Avocado



Awacamolli
Guacamole

1. Build a digital parallel corpus (Spanish-Nahuatl)
2. Learn multilingual vector representations and use them for bilingual lexicon extraction



Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. (LREC'16)
Gutierrez-Vasques, X., Medina-Urrea, A., & Sierra, G. (2019). Morphological segmentation for extracting Spanish-Nahuatl bilingual lexicon.

NLP struggles with typological diversity

Ti-nech-caki-s-neki

2.SG.S-1.S.O-‘listen’-FUT-‘want’

You want to hear me

NLP struggles with typological diversity

Ti-nech-caki-s-neki

2.SG.S-1.S.O-'listen'-FUT-'want'

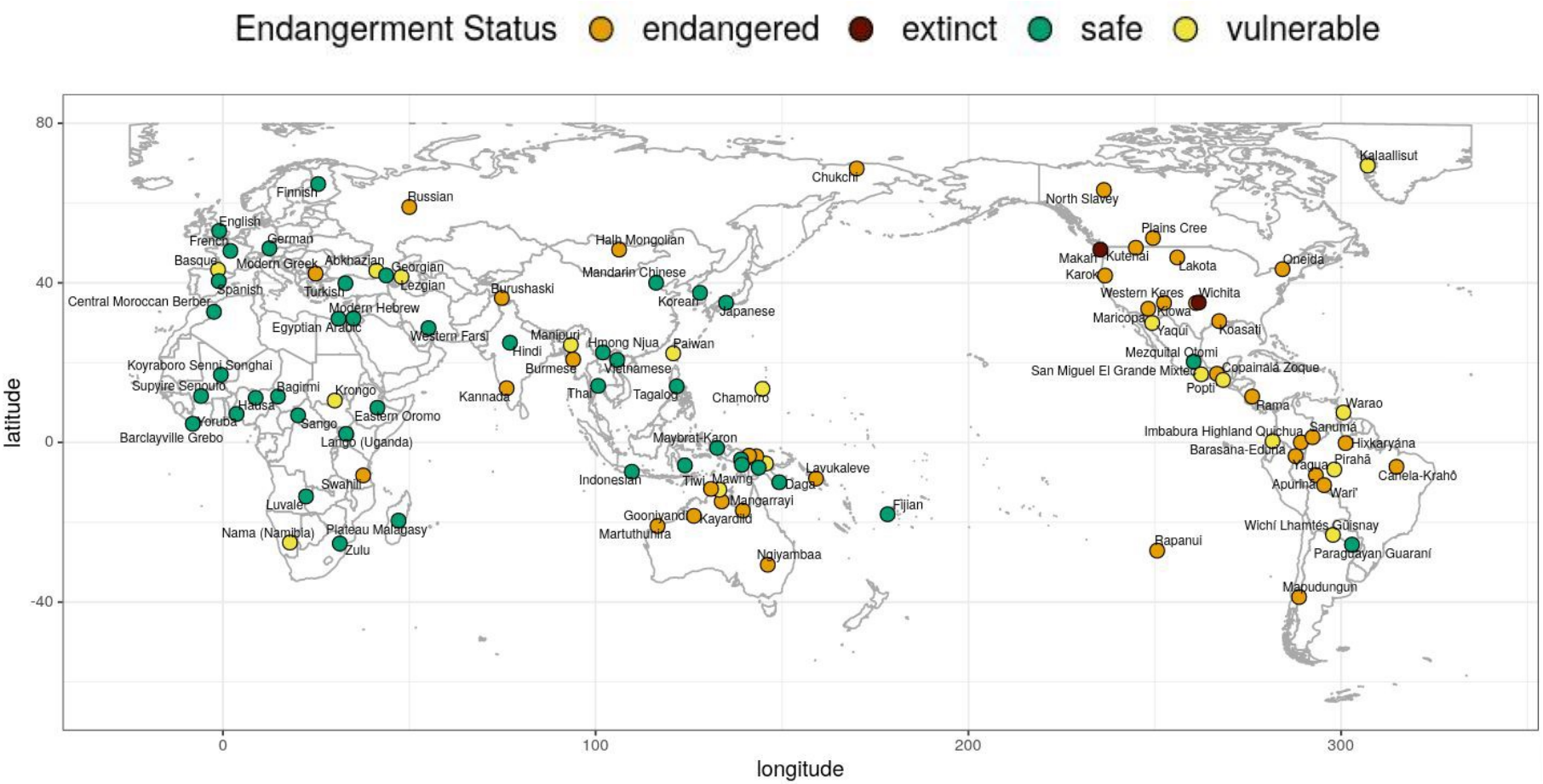
You want to hear me

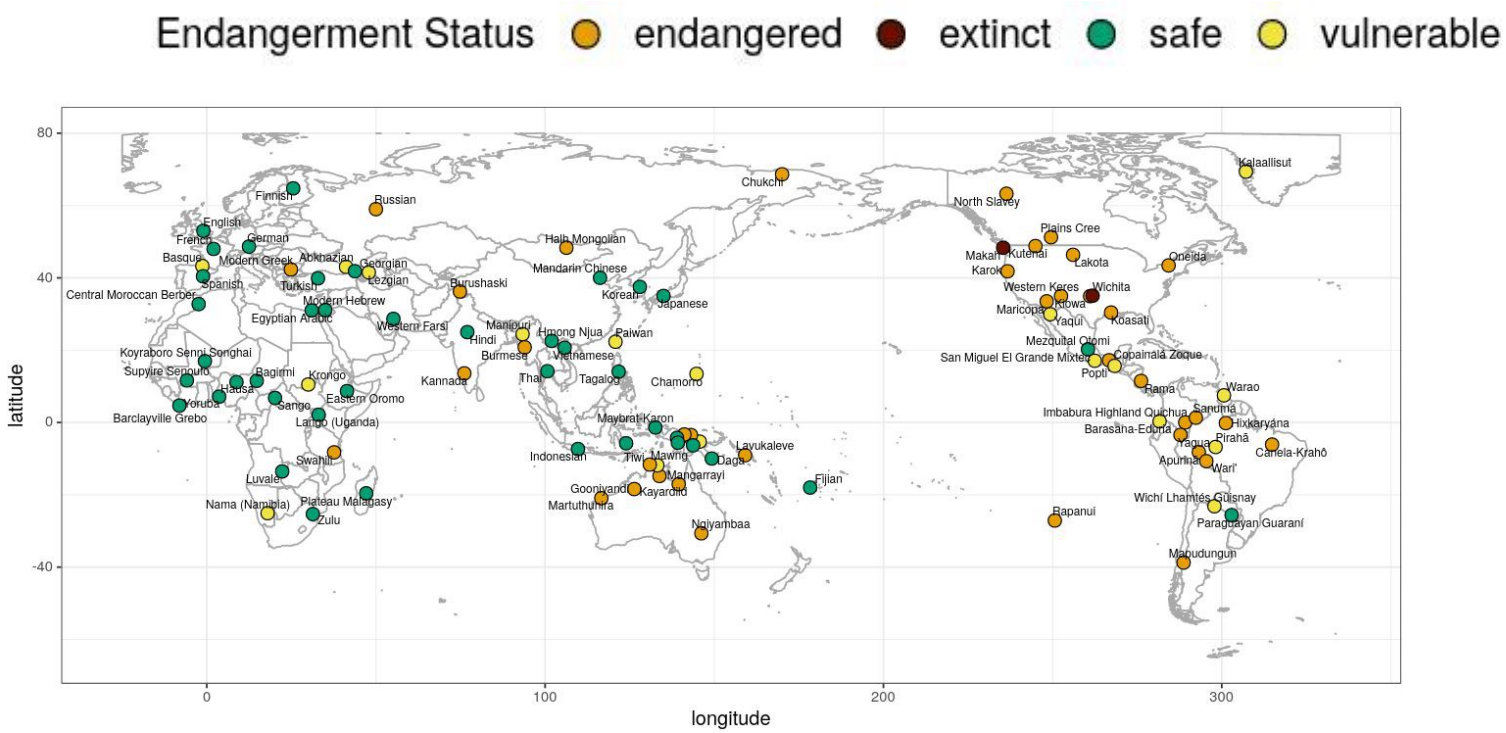
- Otomi language

High tone /dá-tsot'e/ *(1.CPL-arrive)* 'I arrived'
Low tone /da-tsot'e/ *(3.IRR-arrive)* 'He would arrive'

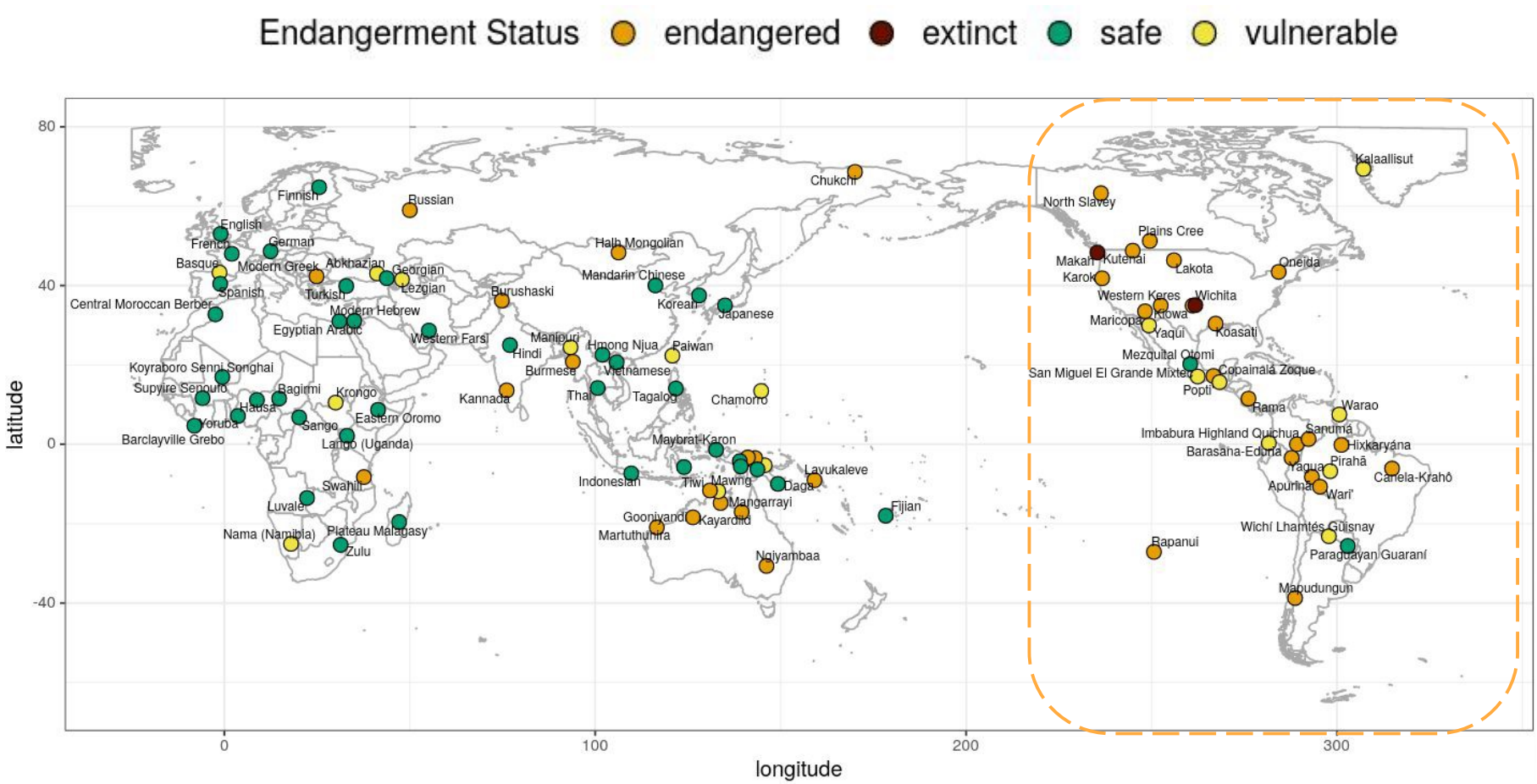
- Mixtec language

$nu^{3}mi^{3}$ *(3.IRR-hug)* 'He would hug'
$nu^{14}mi^{3}$ *(3.NEG.IRR-hug)* 'He would not hug'
$nu^{13}mi^{3}$ *(3.CPL-hug)* 'He hugged'


Endangerment Status
endangered
extinct
safe
vulnerable
latitude
longitude
80
40
0
-40
0
100
200
300
Finnish
Russian
Chukchi
Kalaallisut
North Slavey
English
French
German
Basque
Modern Greek
Abkhazian
Georgian
Lezgian
Turkish
Spanish
Central Moroccan Berber
Modern Hebrew
Egyptian Arabic
Western Farsi
Burushaski
Halh Mongolian
Mandarin Chinese
Korean
Japanese
Plains Cree
Makah
Kutenai
Lakota
Oneida
Karok
Western Keres
Wichita
Kiowa
Maricopa
Yaqui
Koasati
Mezquital Otomi
San Miguel El Grande Mixtec
Copainalá Zoque
Popti
Rama
Warao
Sanumá
Hixkaryána
Pirahã
Canela-Krahô
Wari'
Apurinã
Yagua
Barasana-Eduria
Imbabura Highland Quichua
Wichí Lhamtés Güisnay
Paraguayan Guaraní
Mapudungun
Rapanui
Manipuri
Hindi
Hmong Njua
Paiwan
Burmese
Vietnamese
Thai
Tagalog
Kannada
Chamorro
Koyraboro Senni Songhai
Supyire Senoufo
Bagirmi
Krongo
Hausa
Yoruba
Barclayville Grebo
Sango
Eastern Oromo
Lango (Uganda)
Swahili
Luvale
Nama (Namibia)
Plateau Malagasy
Zulu
Maybrat-Karon
Indonesian
Tiwi
Mawng
Mangarrayi
Kayardild
Gooniyandi
Martuthunira
Daga
Lavukaleve
Fijian
Ngiyambaa



Move beyond a limited set of languages to achieve deeper and more complete understanding of language


Endangerment Status
endangered
extinct
safe
vulnerable
latitude
longitude
80
40
0
-40
0
100
200
300
Finnish
Russian
Chukchi
Kalaallisut
English
French
German
Basque
Modern Greek
Abkhazian
Georgian
Lezgian
Turkish
Spanish
Central Moroccan Berber
Modern Hebrew
Egyptian Arabic
Western Farsi
Burushaski
Halh Mongolian
Mandarin Chinese
Korean
Japanese
Manipuri
Hindi
Hmong Njua
Paiwan
Burmese
Vietnamese
Kannada
Thai
Tagalog
Chamorro
Koyraboro Senni Songhai
Supyire Senoufo
Bagirmi
Krongo
Hausa
Yoruba
Sango
Eastern Oromo
Barclayville Grebo
Lango (Uganda)
Maybrat-Karon
Lavukaleve
Indonesian
Tiwi
Mawng
Daga
Mangarrayi
Swahili
Luvale
Gooniyandi
Kayardild
Fijian
Nama (Namibla)
Plateau Malagasy
Zulu
Martuthunira
Ngiyambaa
North Slavey
Plains Cree
Makah
Kutenai
Lakota
Oneida
Karok
Western Keres
Wichita
Maricopa
Kiowa
Yaqui
Koasati
Mezquital Otomi
San Miguel El Grande Mixtec
Copainalá Zoque
Popti
Rama
Warao
Imbabura Highland Quichua
Sanumá
Barasana-Eduria
Hixkaryána
Yaqua
Pirahã
Canela-Krahô
Apurinã
Wari'
Wichí Lhamtés Güisnay
Bapanui
Paraguayan Guaraní
Mapudungun

We still have some way to go…

Gemini

Here are some examples of how you might use these translations in a sentence:

- **Nican nica ninoca, inin yehuatl Miguel.** (Here I am, this is Miguel.)
- **Nican nochihua, ¿tlazocamati inon altepetl?** (I am here, do you know this town?)
- **Nican nochi, ¿tlein ic tlahtlauhtia?** (I am here, what are you doing?)
- **Nde'e ni'i, ¿tlein ic tlahtlauhtia?** (I am here, what are you doing?)

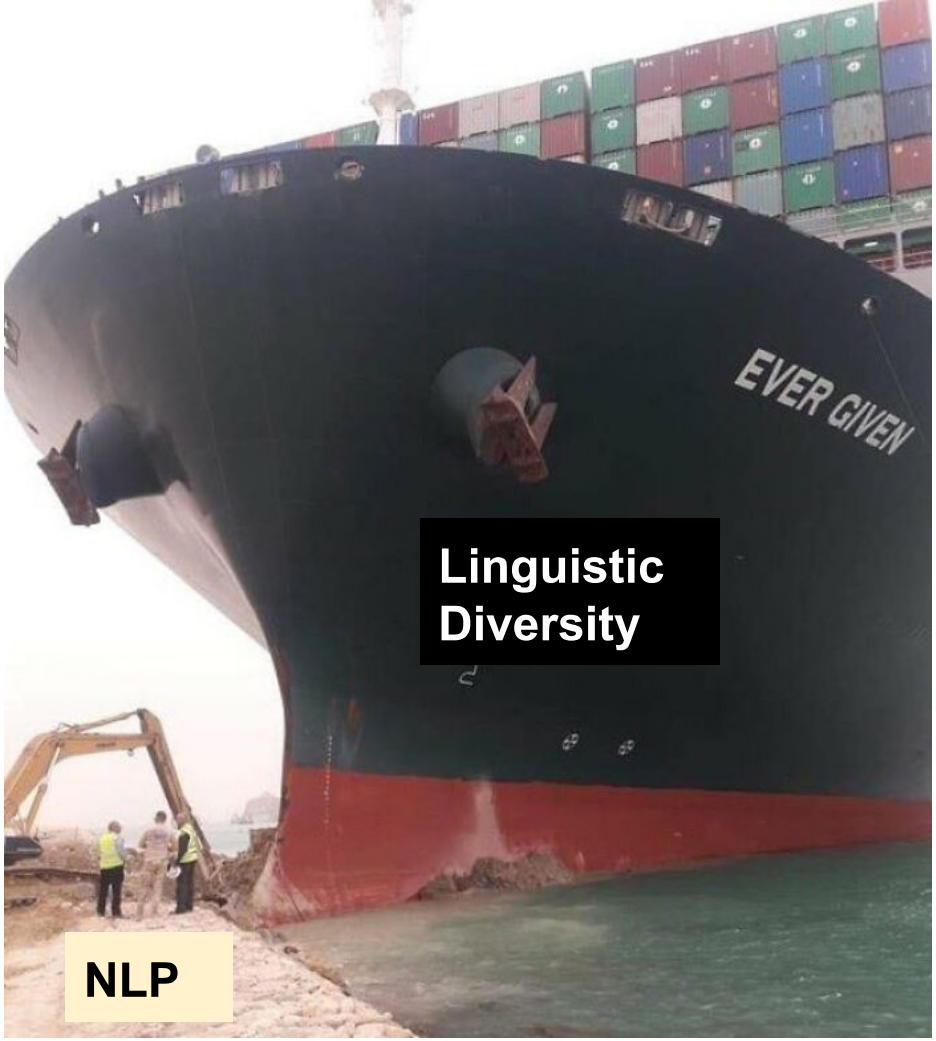
I hope this helps!

"Thank you that town?"

(wrong translations)

Mixing up languages (Nahuatl and Otomi)
Making up verbs
…

** Consulted February 2024*


EVER GIVEN
Linguistic Diversity
NLP

1. Motivation

2. Approaching linguistic diversity in NLP

3. Text-based typology

Linguistic diversity in NLP

→ Multilingual datasets and models are becoming increasingly important in NLP

How to select languages?

Linguistic diversity in NLP

How to select languages?

NLP perspective

- Favors languages for which text data is readily available online.

Linguistic diversity in NLP

How to select languages?

NLP perspective

- Favors languages for which text data is readily available online.

Linguistics perspective

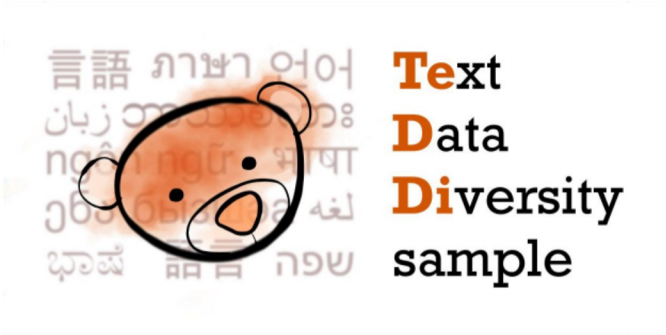
- More weight on representing a wide range of language families, areas, and structural features

WALS 100-language sample

- Maximizes genealogical (language family) and areal (geographic) diversity

- Selection is independent of text data availability

TeDDi sample

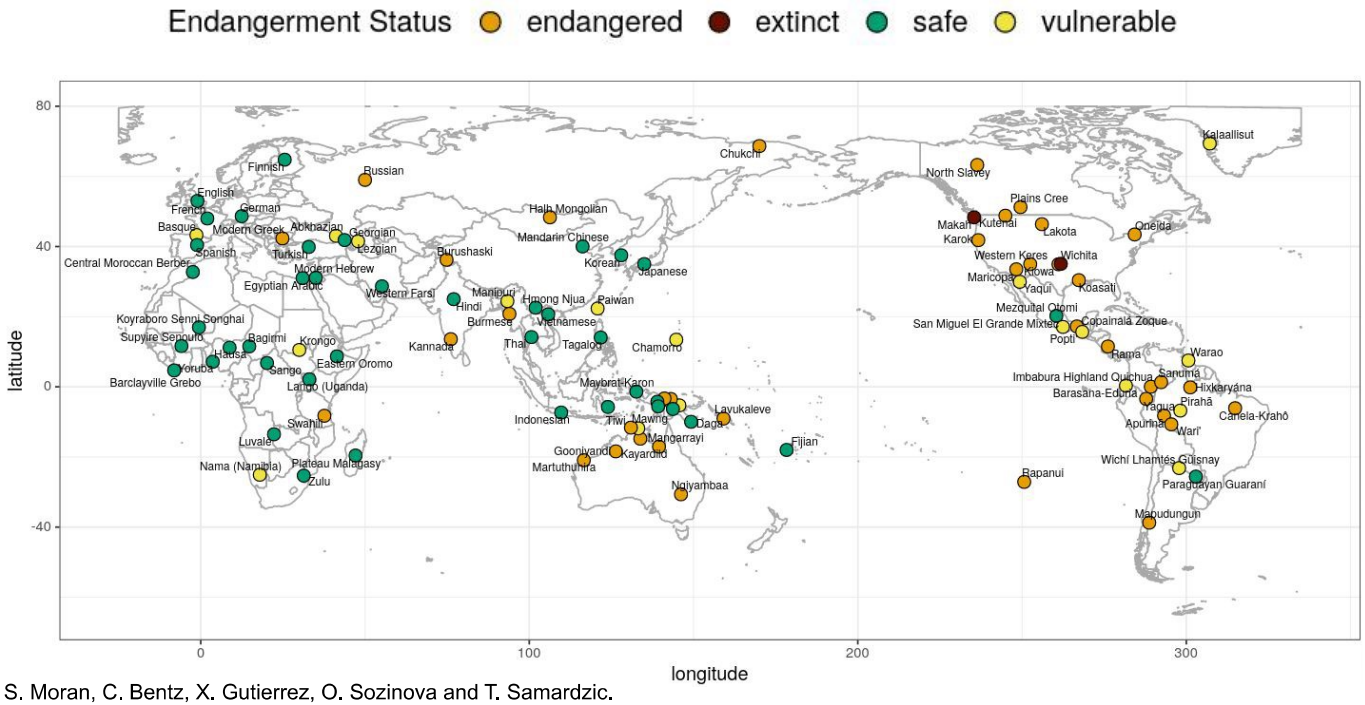


- Based on WALS 100 language sample
- Approximate the diversity of languages across the world by means of text samples

Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, Tanja Samardzic (2022). TeDDi Sample: Text Data Diversity Sample for Language Comparison and Multilingual NLP. *LREC*

TeDDi sample

Data collection and curation



. S. Moran, C. Bentz, X. Gutierrez, O. Sozinova and T. Samardzic.

TeDDi sample

It features:

- Text corpora for languages on the typological diversity sample in WALS
- Currently 89 languages (58 language families). More than 20k texts
- 6 broad genres, 2 modalities (spoken, written)
- Open-source corpus processing tools

https://github.com/MorphDiv/TeDDi_sample/

More languages do not necessarily mean more diversity…

- We assess the diversity of several NLP datasets in reference to the WALS diversity sample

More languages do not necessarily mean more diversity…

→ We assess the diversity of several NLP datasets in reference to the WALS diversity sample

Comparison based on:

- Typological features
- Text-based features calculated from text samples (TeDDi)

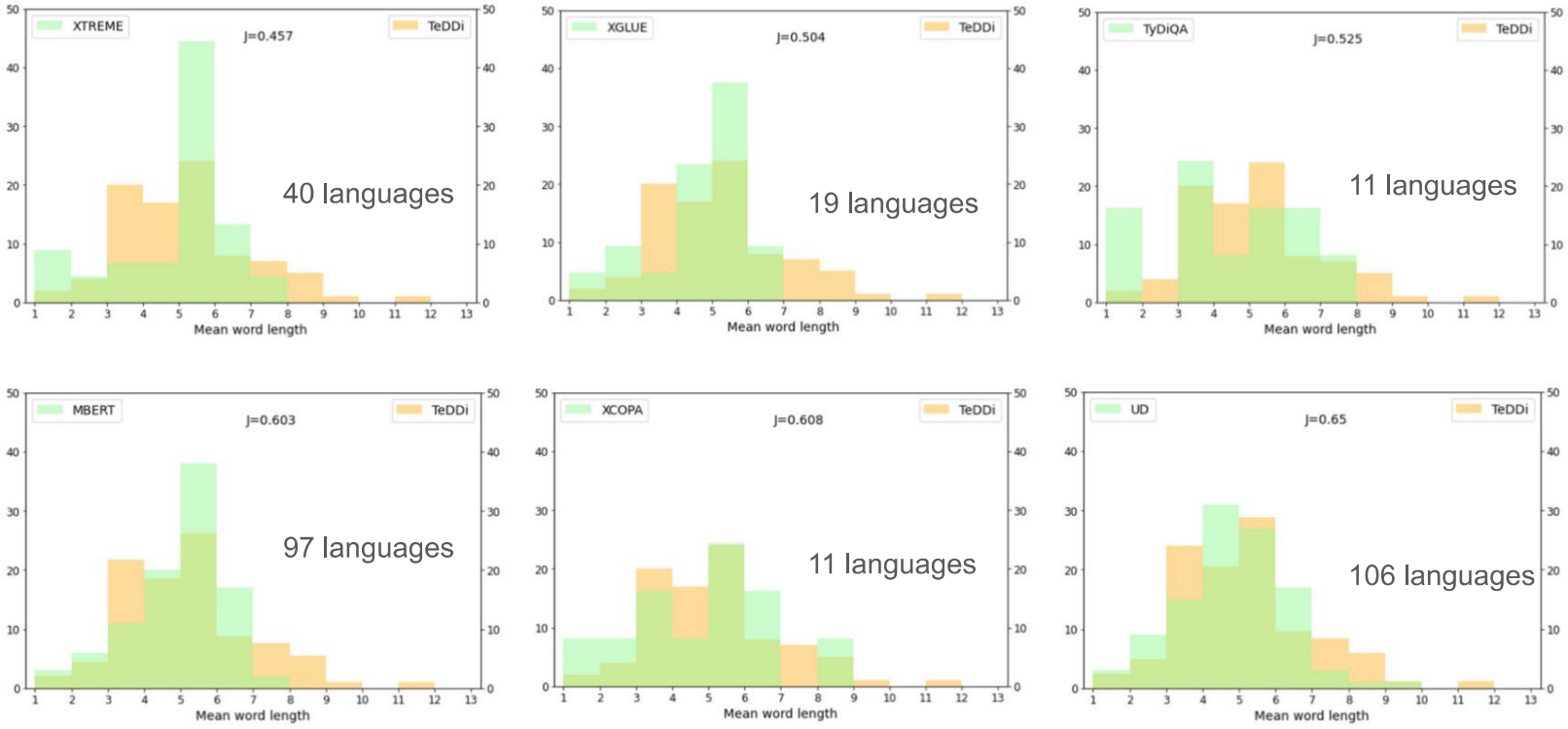
More languages do not necessarily mean more diversity…

→ We assess the diversity of several NLP datasets in reference to the WALS diversity sample

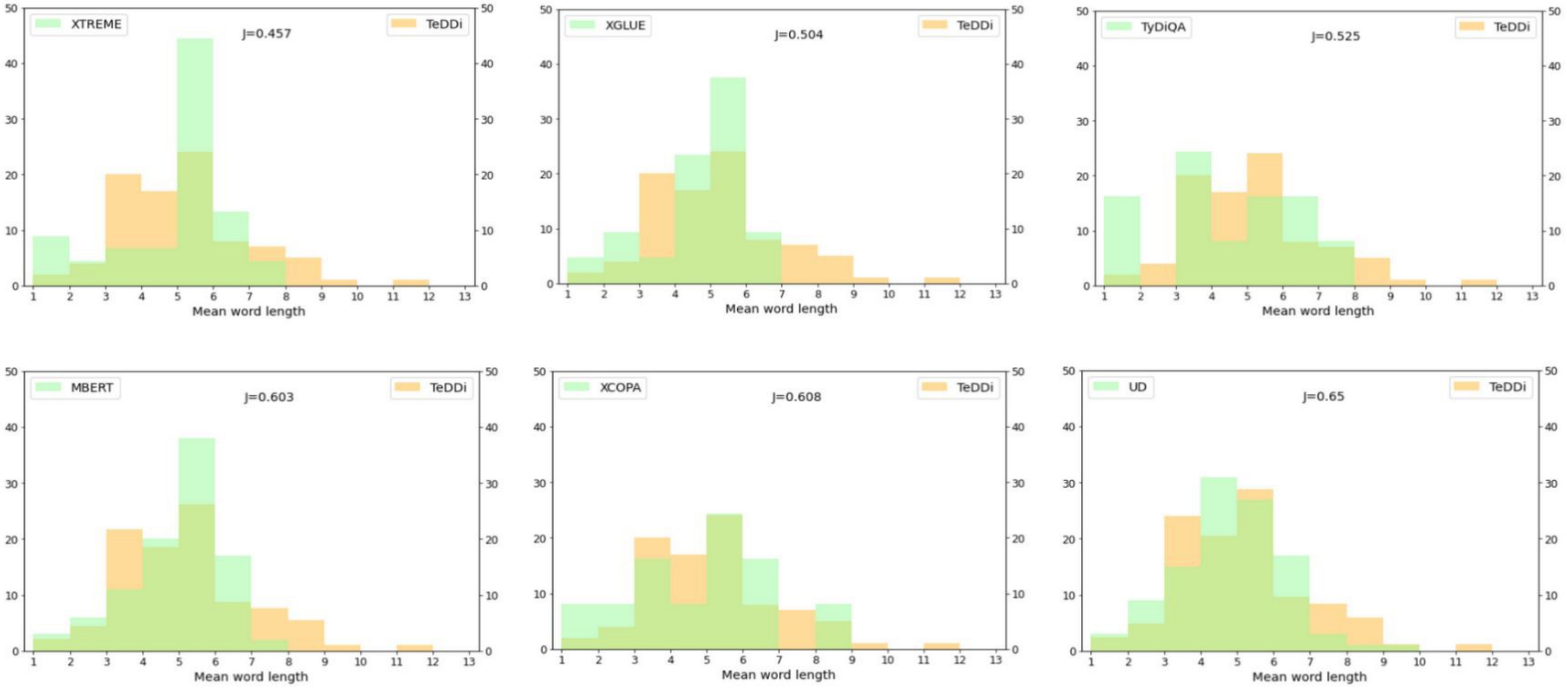
Comparison based on:

- Typological features
- Text-based features calculated from text samples (TeDDi)

Word length (captures morphology, reliable in very small samples)



Union and intersection between the distributions of the mean word length in TeDDi and NLP data sets



Samardzic, T., Gutierrez, X., Bentz, C., Moran, S., & Pelloni, O. A Measure for Transparent Comparison of Linguistic Diversity in Multilingual NLP Data Sets. *NAACL 2024 (Findings)*

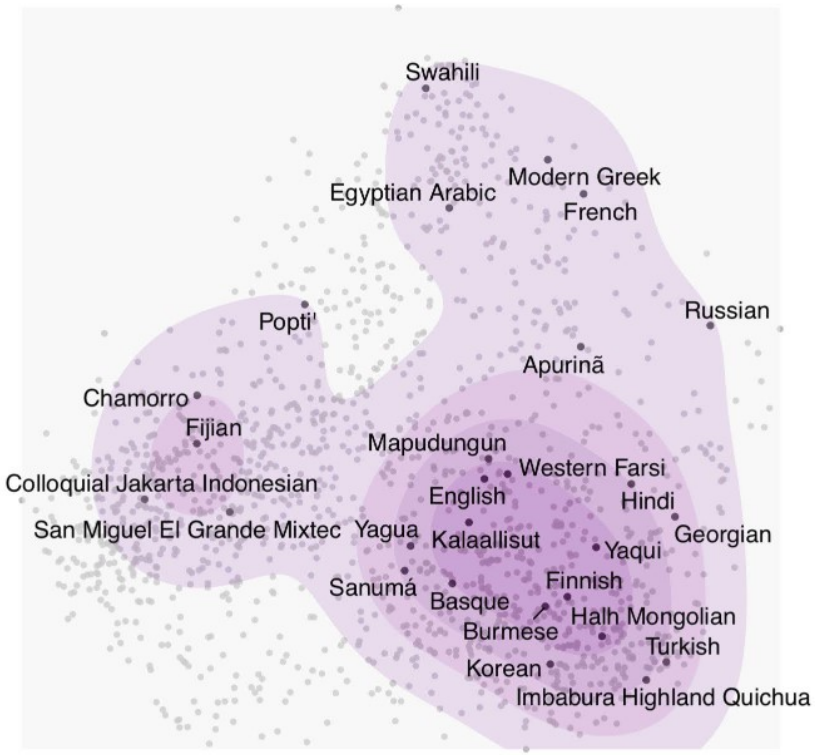


Figure 6: PCA plot for the paper with the highest absolute feature coverage (Gutierrez-Vasques et al., 2021), where each point represents a language in Grambank.

What is 'Typological Diversity' in NLP?

Ploeger, E., Poelman, W., de Lhoneux, M., & Bjerva, J. (2024)

1. Motivation

2. Approaching linguistic diversity in NLP

3. Text-based typology

Typology

Typological databases: Important resource for assessing linguistic diversity and comparing languages

Typology

Typological databases: Important resource for assessing linguistic diversity and comparing languages

It requires deep linguistic knowledge or access to annotated data (not easily available for many languages)

↓

Text-based approaches?

Text-based typology

Morphological level: we would like to group languages according to their morphological structures and study some relationships

Discover these structural properties from text corpus in a resource-cheap way

Text-based typology

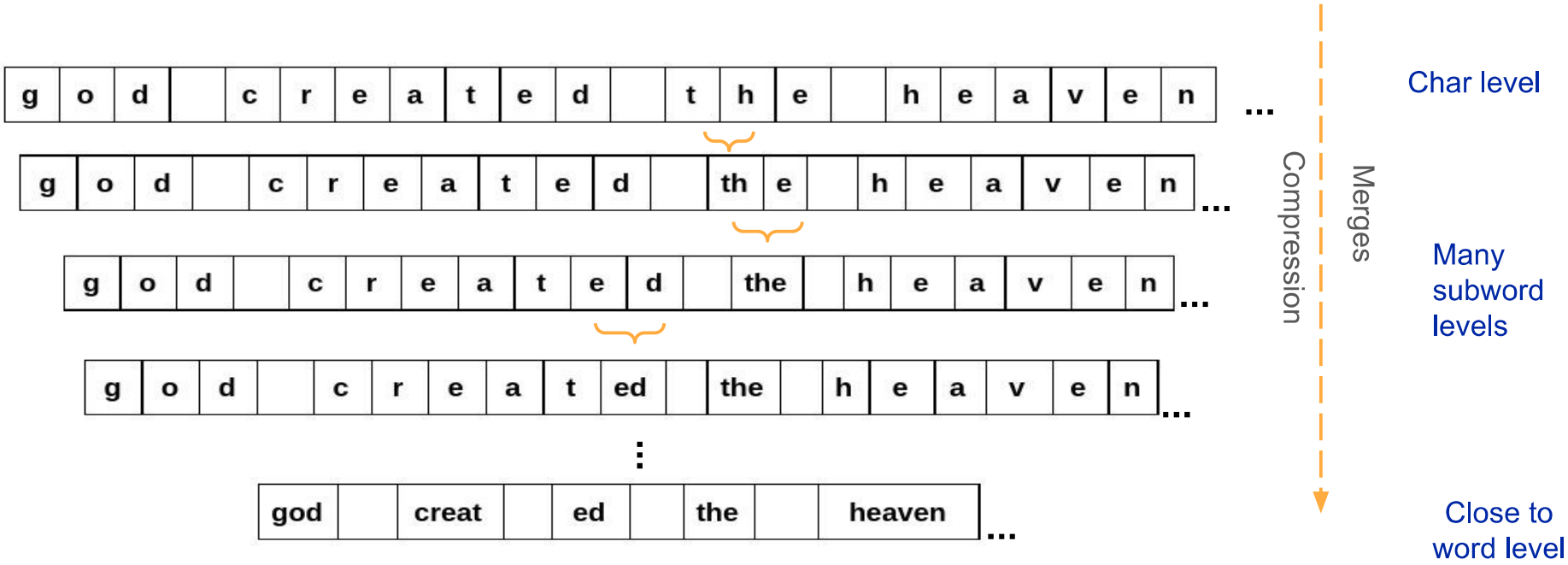
Morphological level: we would like to group languages according to their morphological structures and study some relationships

Discover these structural properties from text corpus in a resource-cheap way

Subword tokenization → uncovers **regularities** in the data.

Subword tokenization

BPE merges sequences of co-occurring characters (subwords)



Subword tokenization and Information Theory

- We measure Shannon entropy and redundancy of the unigram distributions over varied subword tokenizations of texts obtained with BPE
- At each incremental merge, we compare the values across 47 typologically diverse languages.

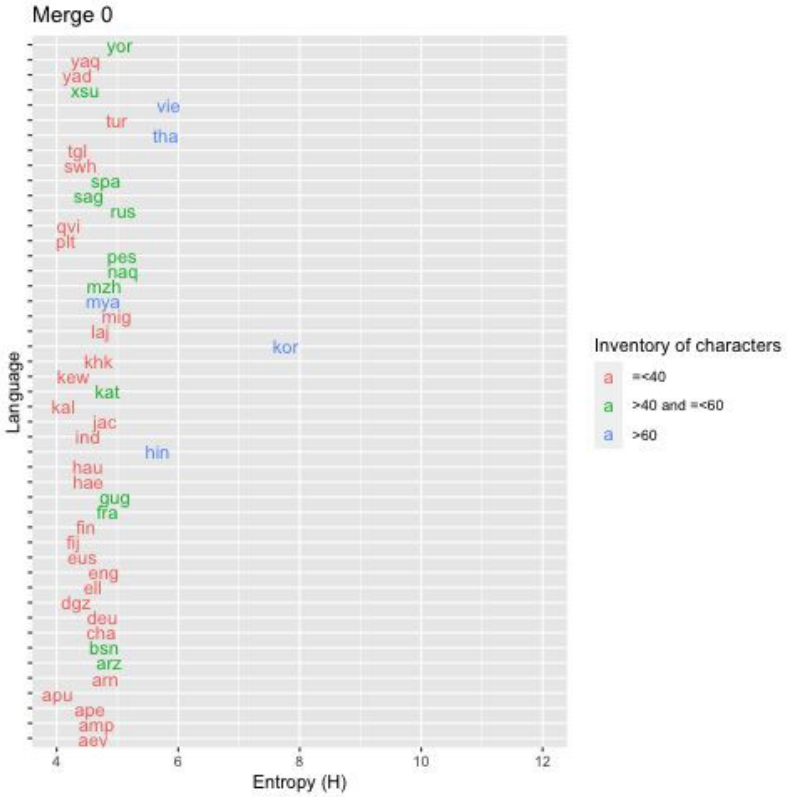
$$H(T) = -\sum_{i=1}^{V} p(t_i) \log_2 p(t_i)$$

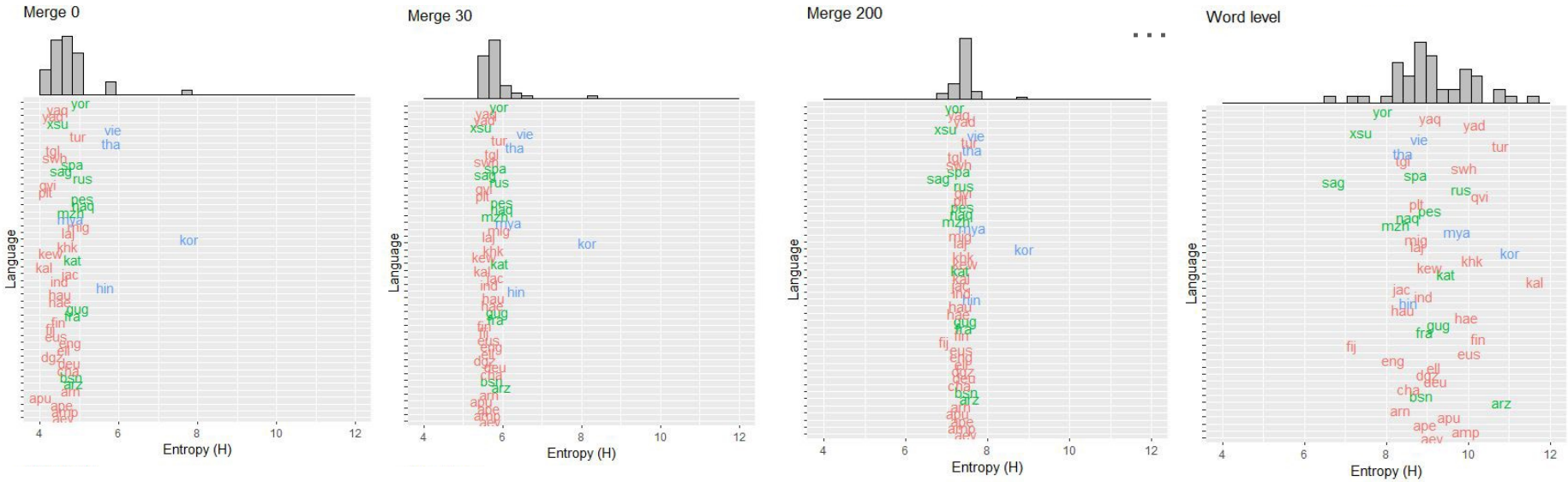
Entropy

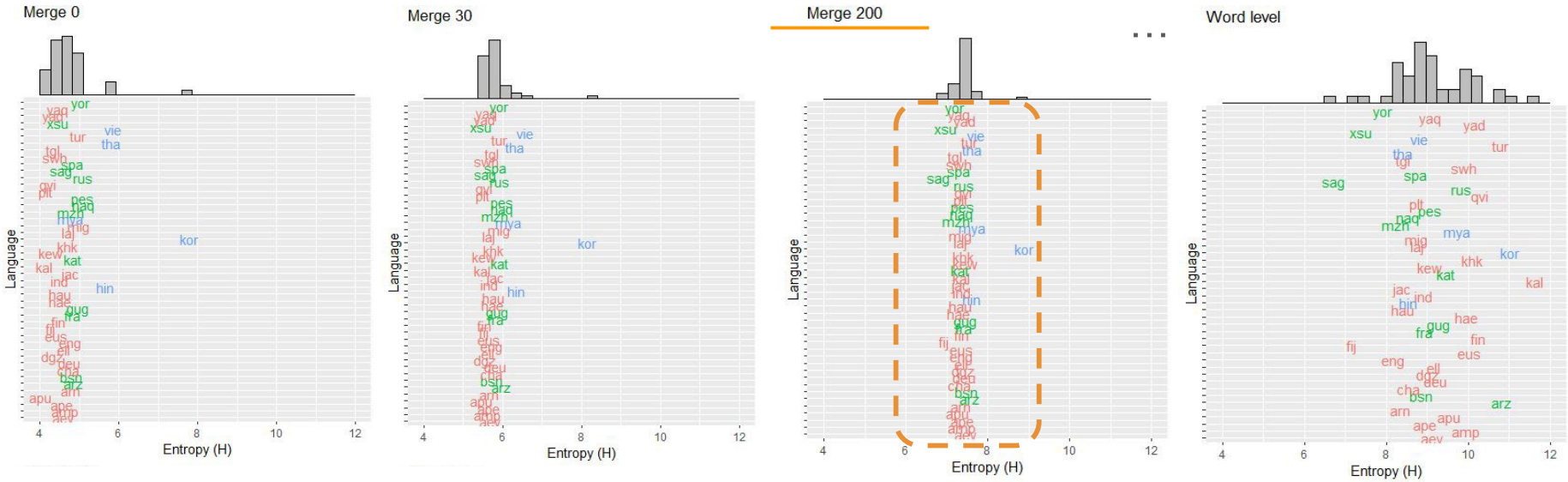
$$R(T) = 1 - \frac{H(T)}{max\{H(T)\}} = 1 - \frac{H(T)}{\log_2 |V|}$$

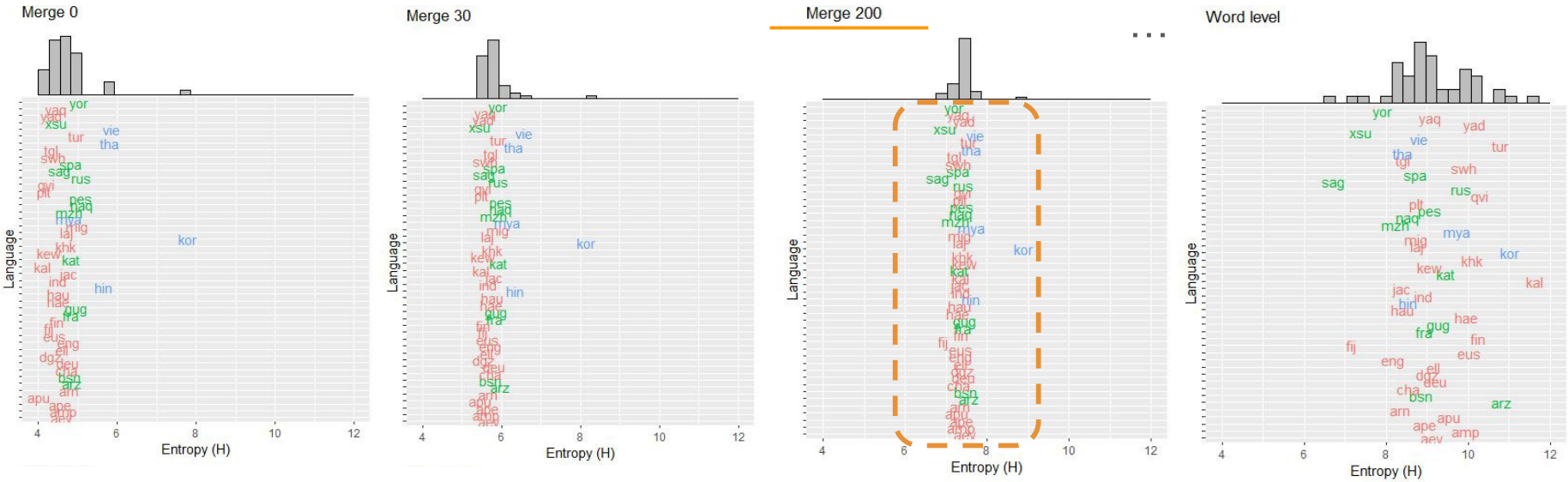
Redundancy

Variation of entropy through BPE merges

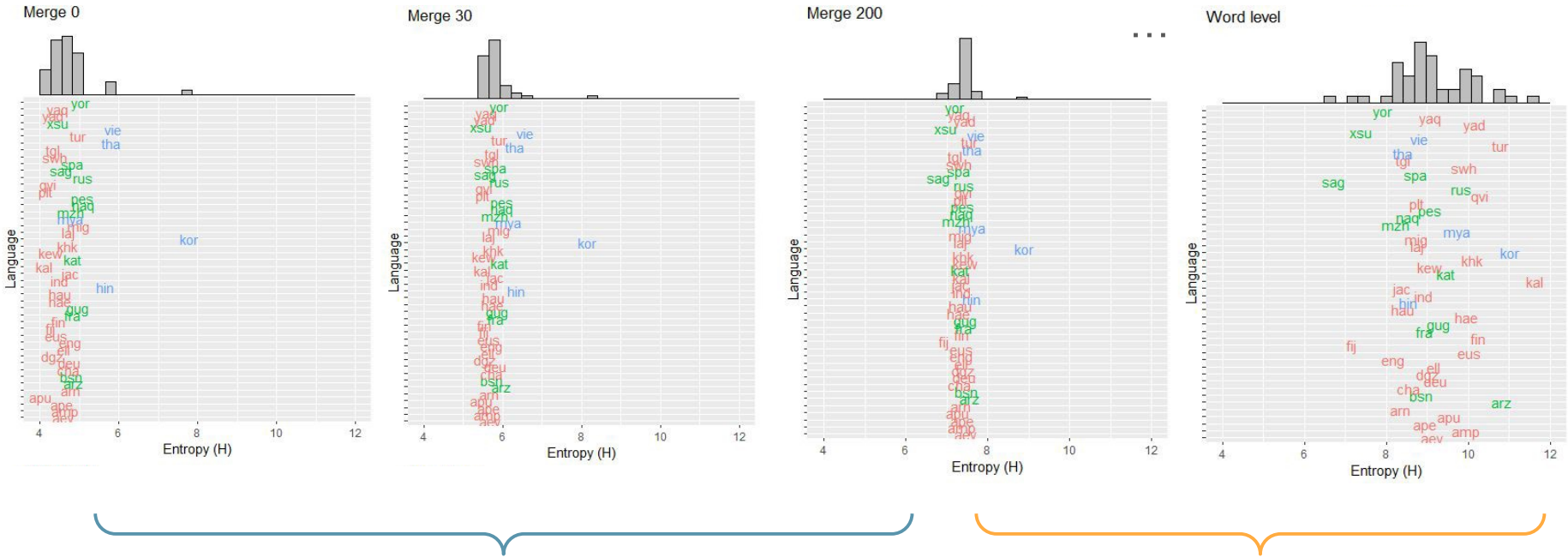



Merge 0
Merge 30
Merge 200
Word level
Language
Entropy (H)
4
6
8
10
12
yor
xsu
vie
tha
tur
spa
rus
pes
mya
kor
khk
kat
hin
kal
jac
ind
hae
gug
fra
fin
eus
eng
ell
deu
bsn
arz
arn
apu
ape
amp
yaq
yad
swh
sag
qvi
pli
mzh
mig
kew
hau
fij
cha
aev


Merge 0
Merge 30
Merge 200
Word level
Language
Entropy (H)
4
6
8
10
12

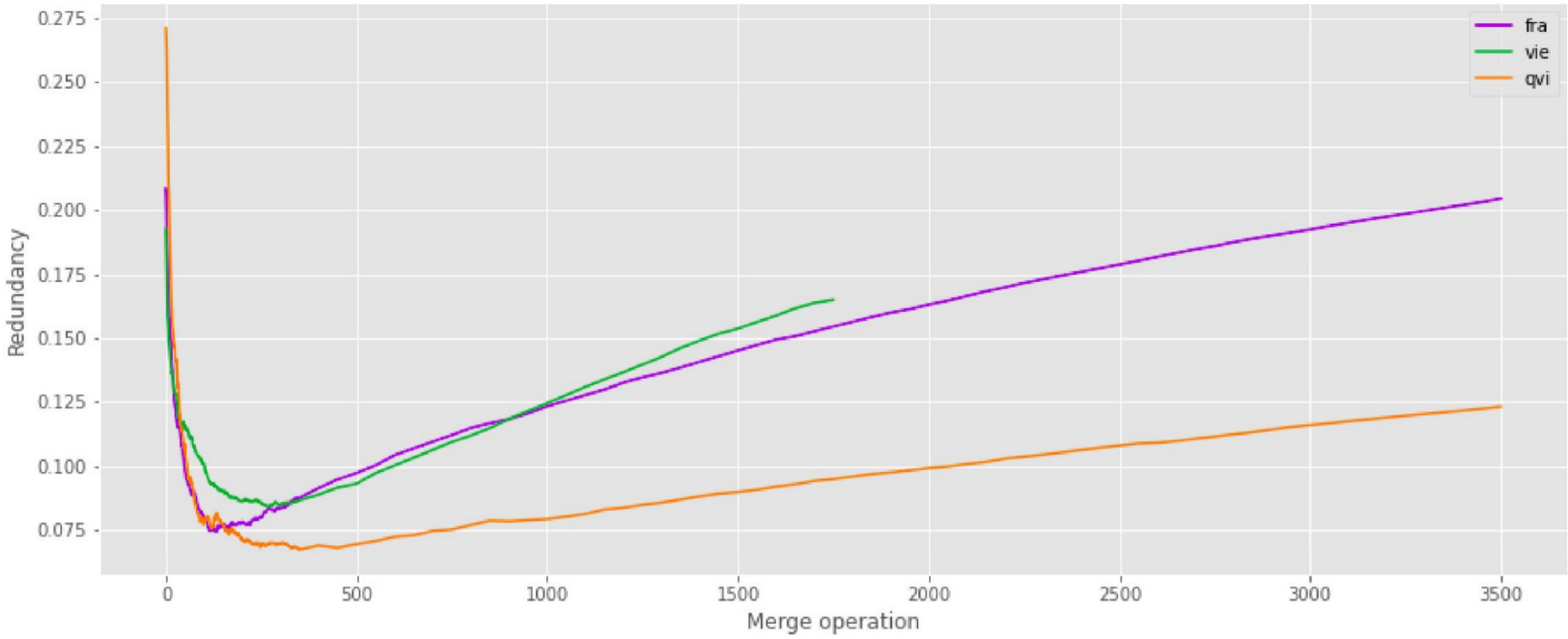


Gutierrez-Vasques, X., Bentz, C., Sozinova, O., & Samardzic, T. (2021). From characters to words: the turning point of BPE merges. *EACL*

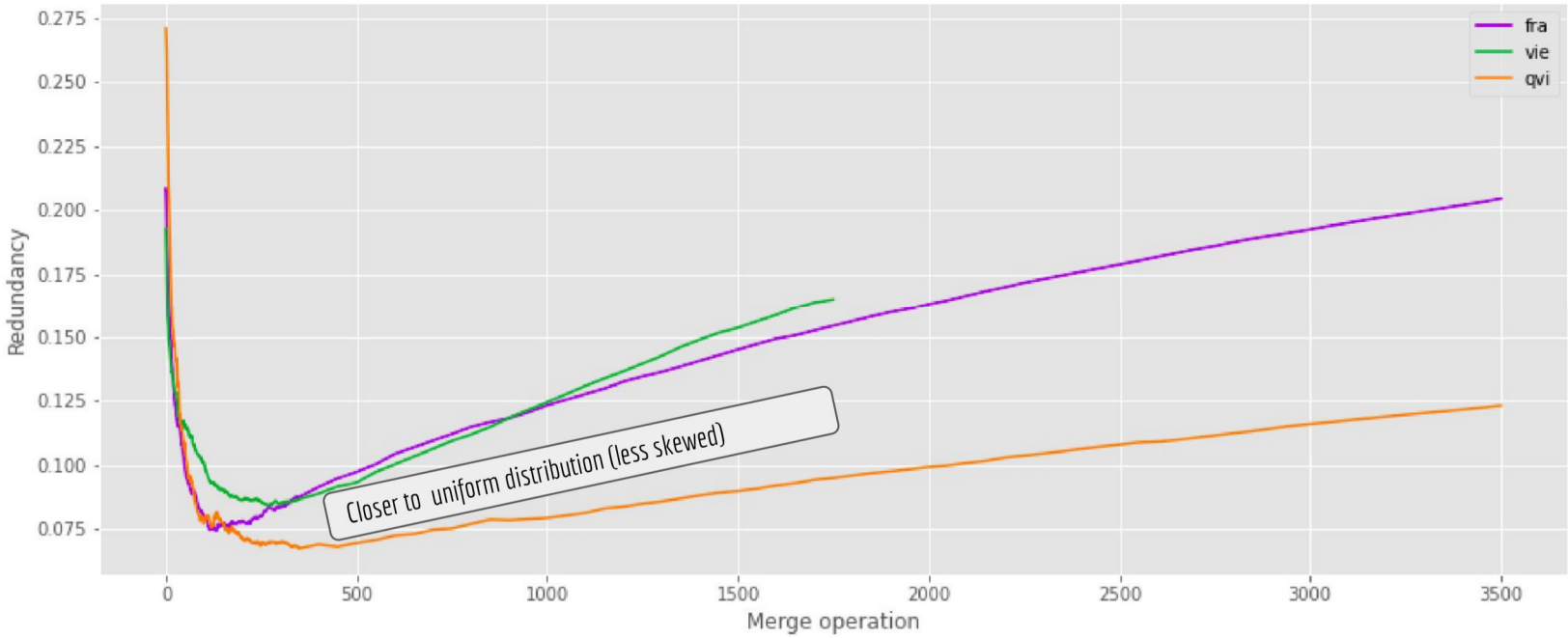


The **ranking** of languages is **different** before and after 200 merges (approx.)

Redundancy



→ **First merges** cause the most **drastic** changes:



→ **First merges** cause the most **drastic** changes:

inflectional markers, very frequent stems, orthographic practices, etc.

Compression and morphological structure

- Regardless of the language: the first elements merged by BPE have the most substantial impact on the compression of texts
- Are these highly recurrent patterns similar across languages?

Compression and morphological structure

- Regardless of the language: the first elements merged by BPE have the most substantial impact on the compression of texts

- Are these highly recurrent patterns similar across languages?

Our hypothesis:

The types of subwords allowing effective compression depend on the morphological type of the language in question.

Subwords' properties

For each subword **s** found by BPE (at each merge operation):

$$\text{productivity}(s) = |W_s|$$

$$\text{c.freq}(s) = \sum_{w \in W_s} \text{freq}(w)$$

$$\text{idiosyncrasy}(s) = \frac{\text{c.freq}(s)}{\text{productivity}(s)}$$

Example:

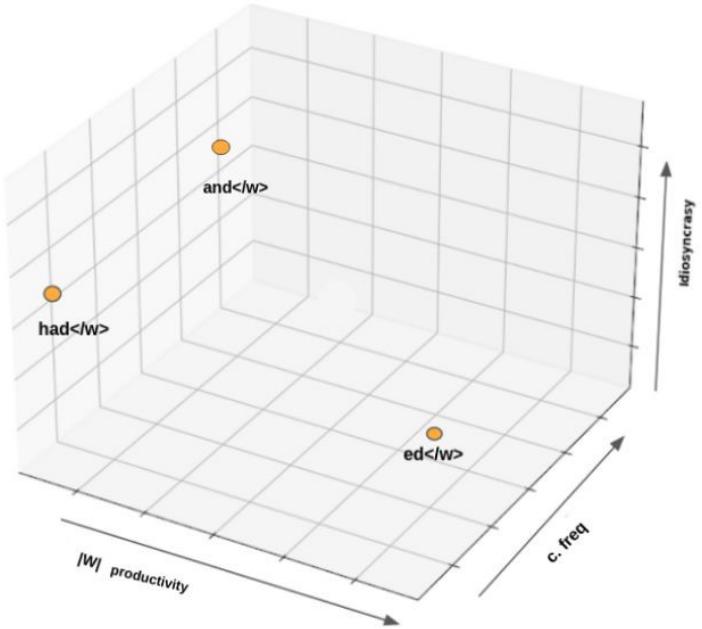
Subword	|W|	c.freq.	idiosyncrasy
ed</w>	271	917	3.38
had</w>	1	104	104
and</w>	11	2,197	199.72

PBC Corpus (English)

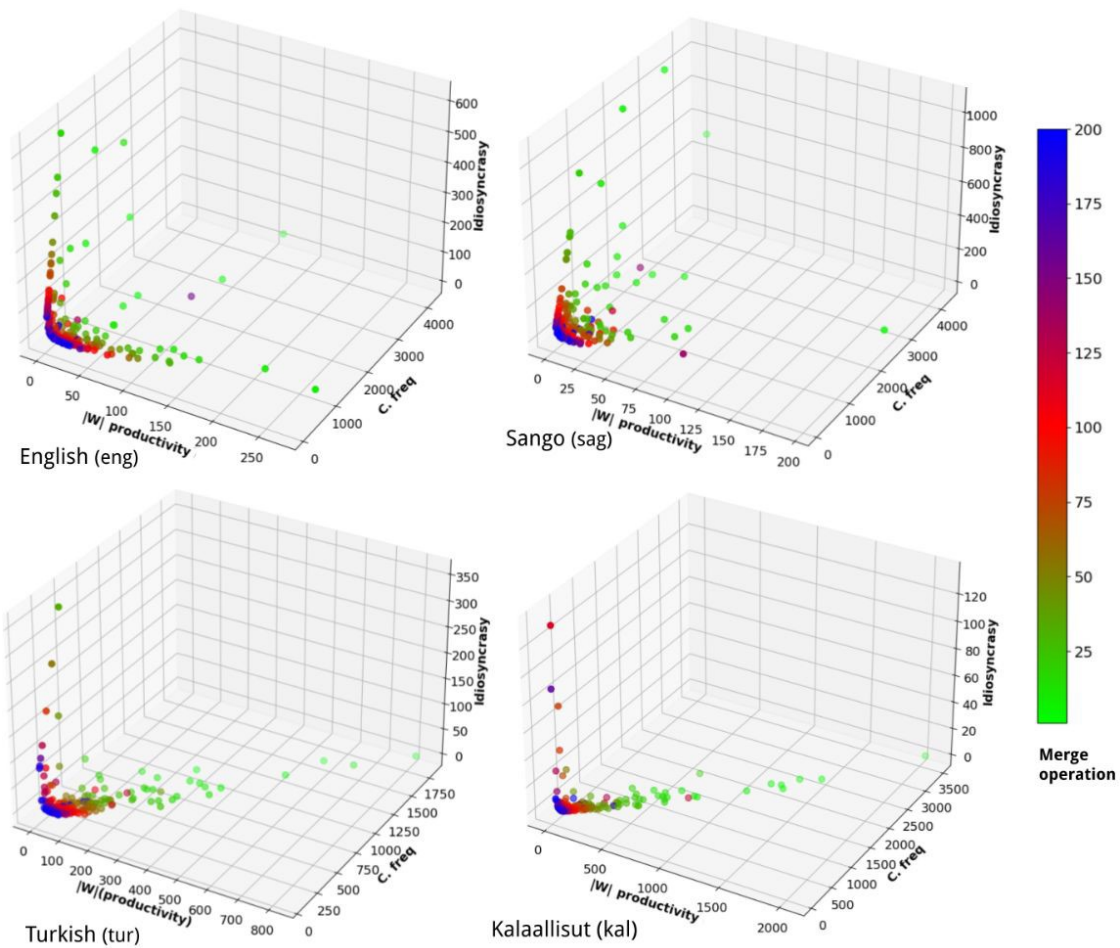
Subwords' properties

Subword	|W|	c.freq.	idiosyncrasy
ed</w>	271	917	3.38
had</w>	1	104	104
and</w>	11	2,197	199.72

PBC Corpus (English)



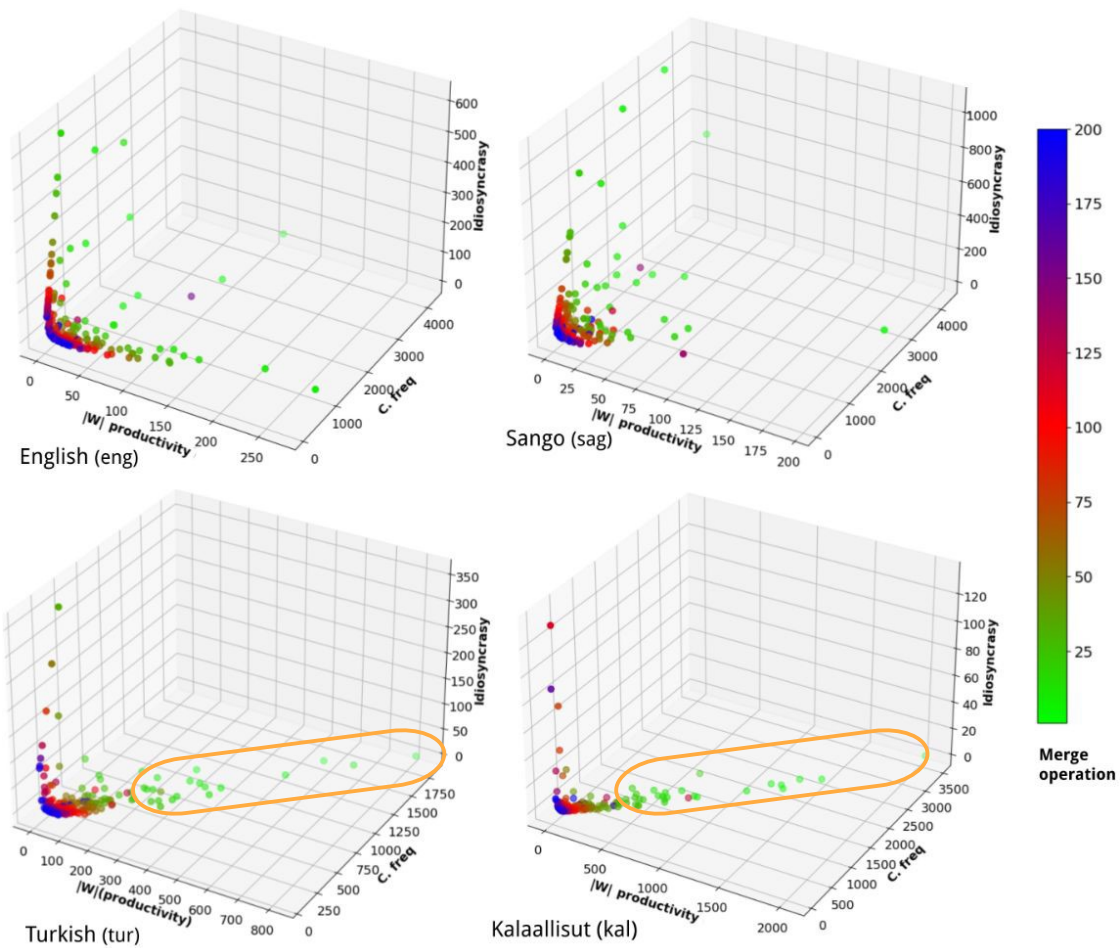
Subwords' properties



BPE subwords for four languages (PBC corpus).

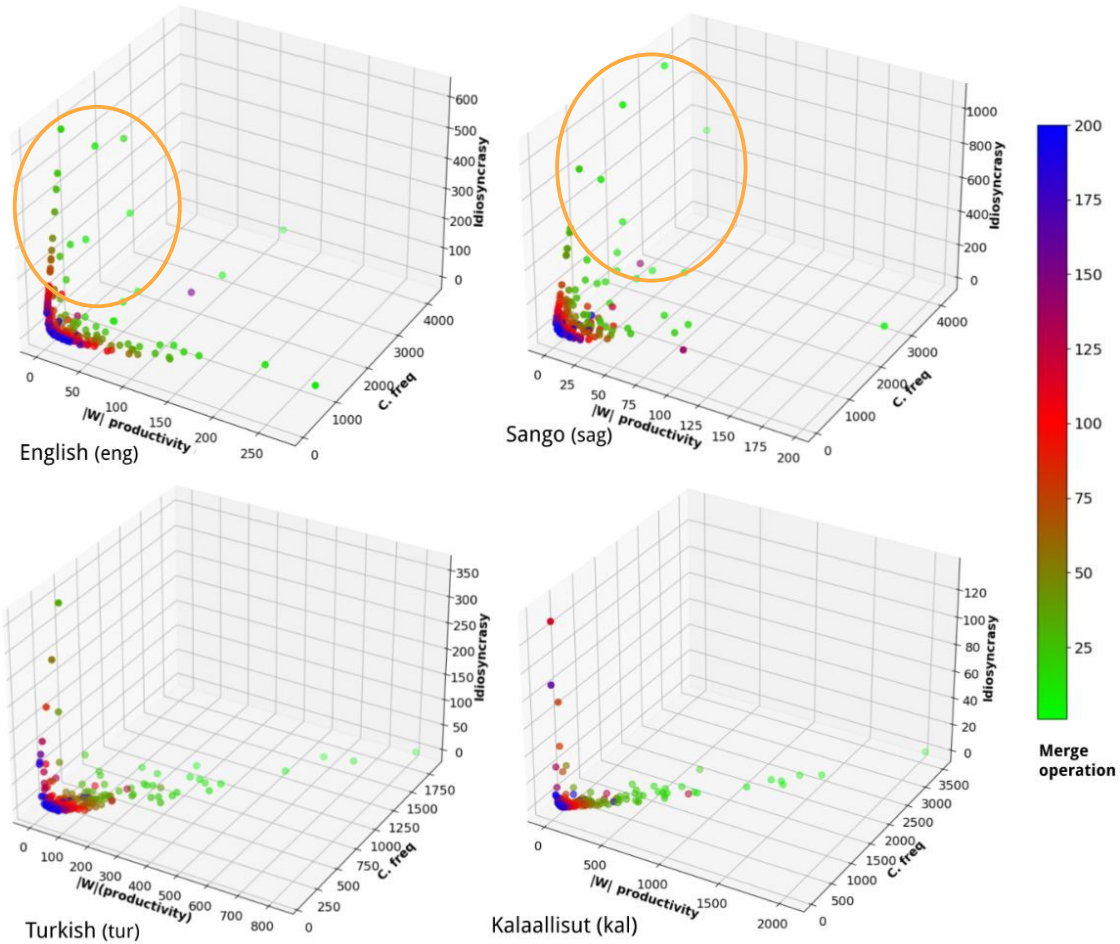
Each data point represents a subword.

The colors are a visual cue for the merge operation (i.e., from 1 to 200) at which the respective subword was created



Depending on the language, compression can be achieved via:

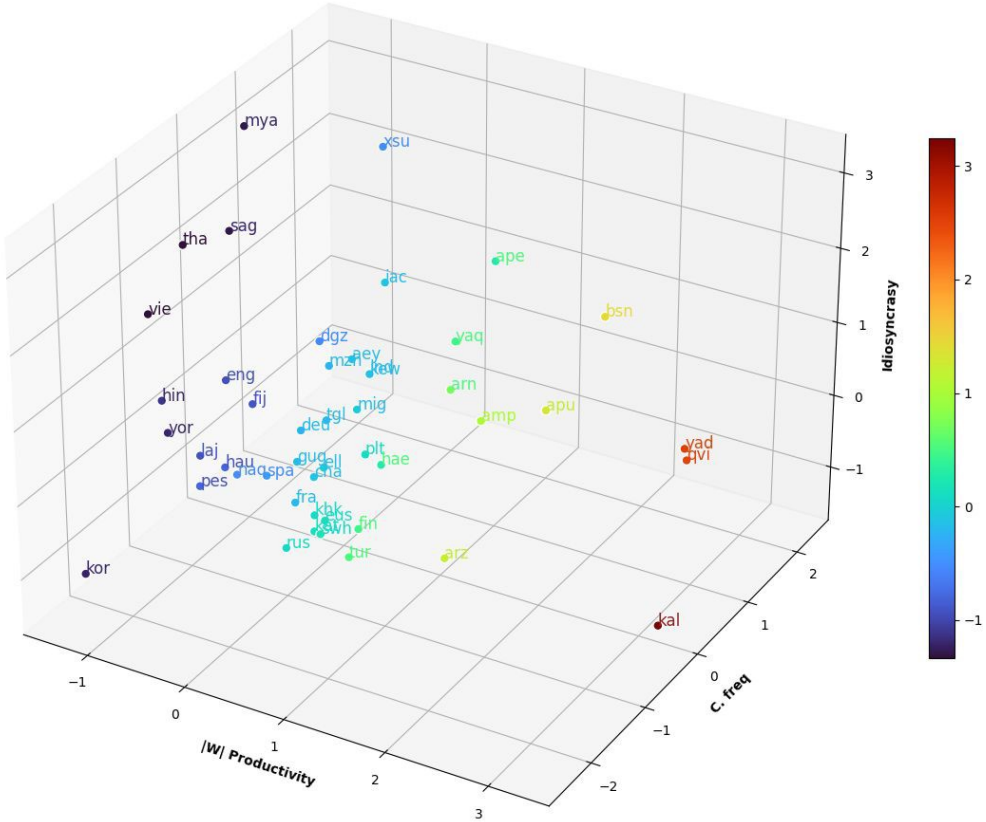
- **Productive subwords** (markers, affixes, and regular morphological phenomena)



Depending on the language, compression can be achieved via:

- Productive subwords (markers, affixes, and regular morphological phenomena)
- **Idiosyncratic subwords** (frequent irregular patterns or whole orthographic words which are highly redundant)

Compression and morphological structure



- Language vectors in the BPE subword space
- Data-driven induction of typological knowledge

Agreement between BPE and typological knowledge

Evaluation

Quantitative

Agreement between language vectors we created and independent language vectors derived from external linguistic knowledge (WALS)

Qualitative

Exhaustive analysis on how our BPE observations relate to the morphology of languages

Synthesis and fusion

ISO	Name	Prod.	Synthesis	Reference	Fusion
vie	Vietnamese	-1.33	analytic	Haspelmath and Sims (2010)	isolating
tha	Thai	-1.33	analytic	Moravcsik (2012)	isolating/concat.
sag	Sango	-1.29	analytic	Karan (2006)	concatenative
yor	Yoruba	-1.21	analytic	Haspelmath and Sims (2010)	tonal/isolating
eng	English	-0.94	analytic	Haspelmath and Sims (2010)	concatenative
fij	Fijian	-0.89	analytic	Dixon (1988)	isolating
pes	Persian/Farsi	-0.78	synthetic	Greenberg (1960)	concatenative
fra	French	-0.19	synthetic	Dixon and Aikhenvald (2003)	concatenative
ell	Greek (Modern)	0.02	synthetic	Dixon and Aikhenvald (2003)	concatenative
rus	Russian	0.07	synthetic	Aikhenvald (2007)	concatenative
swh	Swahili	0.2	synthetic	Haspelmath and Sims (2010)	concatenative
yaq	Yaqui	0.46	synthetic	Guerrero (2019)	concatenative
tur	Turkish	0.54	synthetic	Bickel and Nichols (2007)	concatenative
gug	Paraguayan Guaraní	-0.19	polysynthetic	Aikhenvald (2017)	concatenative
arn	Mapudungun	0.73	polysynthetic	Bickel and Zúñiga (2017)	concatenative
amp	Alamblak	1.02	polysynthetic	Bruce et al. (1984)	concatenative
apu	Apurinã	1.3	polysynthetic	Facundes (2014)	concatenative
bsn	Barasano	1.43	polysynthetic	Gomez-Imbert (2004)	concatenative
kal	Kalaallisut	3.25	polysynthetic	Haspelmath and Sims (2010)	concatenative

Agreement between BPE and typological knowledge

Evaluation

Quantitative

Agreement between language vectors we created and independent language vectors derived from external linguistic knowledge (WALS)

Qualitative

Exhaustive analysis oh how our BPE observations relate to the morphology of languages

Gutierrez-Vasques, X., Bentz, C., & Samardžić, T. (2023). Languages through the looking glass of BPE compression. *Computational Linguistics*

Final remarks

- Interdisciplinary view on linguistic diversity is helpful

Final remarks

- Interdisciplinary view on linguistic diversity is helpful

- Data-driven methods open up a way of complementing traditional typology
  - For example, we obtained typological knowledge from raw text and a compression algorithm

Final remarks

- Interdisciplinary view on linguistic diversity is helpful

- Data-driven methods open up a way of complementing traditional typology
  - For example, we obtained typological knowledge from raw text and a compression algorithm

- The possibility of comparing languages using typological information can be beneficial in several downstream applications in the face of linguistic diversity
  - For example, multilingual NLP and cross-lingual transfer of models.

Final remarks

- Interdisciplinary view on linguistic diversity is helpful

- Data-driven methods open up a way o al typology
  - For example, we c ext and a compressi

Limitations:
- Assumption of orthographic word boundaries
- Some types of writing systems require adjustments (e.g., Korean)
- Non-standardized writing

- T guages using typological information can be be nstream applications in the face of linguistic diversity
  - xample, multilingual NLP and cross-lingual transfer of models.

Final remarks

- Interdisciplinary view on linguistic diversity is helpful

- methods open up a way of complementing traditional typology
  knowledge from raw text and a

Limitations:
- No conclusive evidence of the benefits of including typological criteria in the NLP neural multilingual models

- The possibility of comparing languages using typo be beneficial in several downstream applications in the face of linguistic rsity
  - For example, multilingual NLP and cross-lingual transfer of models.

Acknowledgments

Tanja Samardžić

University of Zürich

Olga Pelloni

University of Zürich
Telepathy labs

Christian Bentz

Eberhard-Karl-Universität Tübingen

Steven Moran

Université de Neuchâtel