From characters to words: the turning point of BPE merges

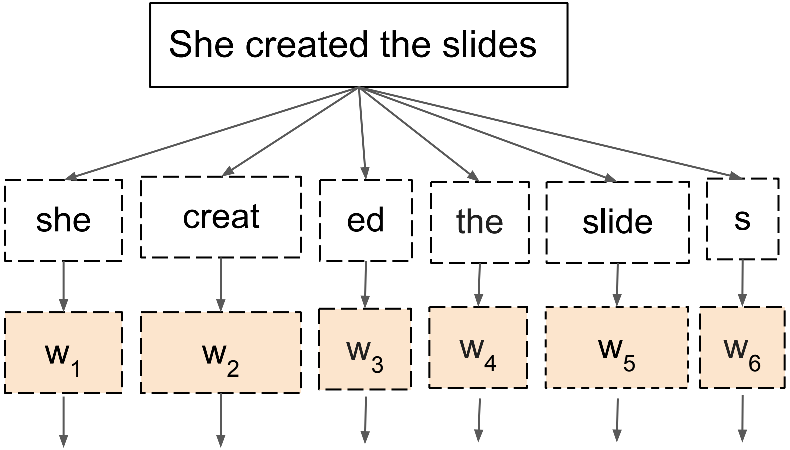
Gutierrez-Vasques Ximena*, Bentz Christian, Sozinova Olga and Samardzic Tanja

EACL, 2021

University of Zurich[UZH]

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Subword tokenization



Text

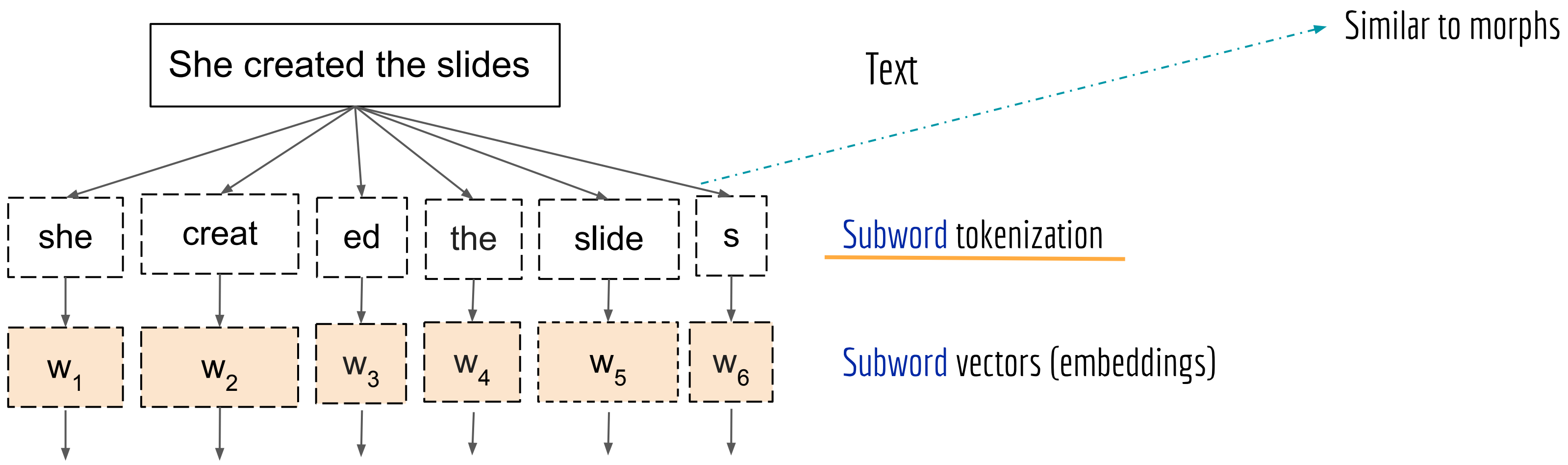
Subword tokenization

Subword vectors (embeddings)

NLP applications

- Statistical/Neural language modeling
- Machine translation
- ...
- Downstream tasks

Subword tokenization



NLP applications

- Statistical/Neural language modeling
- Machine translation
- ...
- Downstream tasks

Merge	BPE text tokenization
0	g-o-d c-r-e-a-t-e-d t-h-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d t-h-e l-i-g-h-t
1	g-o-d c-r-e-a-t-e-d **th**-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **th**-e l-i-g-h-t
2	g-o-d c-r-e-a-t-e-d **the** h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **the** l-i-g-h-t
3	g-**od** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d g-**od** d-i-v-i-d-e-d the l-i-g-h-t
4	**god** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d **god** d-i-v-i-d-e-d the l-i-g-h-t
5	god c-r-e-a-t-**ed** the h-e-a-v-e-n a-n-d god d-i-v-i-d-**ed** the l-i-g-h-t
6	god c-r-**ea**-t-ed the h-**ea**-v-e-n a-n-d god d-i-v-i-d-ed the l-i-g-h-t

.
.
.

Merge	BPE text tokenization
0	g-o-d c-r-e-a-t-e-d t-h-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d t-h-e l-i-g-h-t
1	g-o-d c-r-e-a-t-e-d **th**-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **th**-e l-i-g-h-t
2	g-o-d c-r-e-a-t-e-d **the** h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **the** l-i-g-h-t
3	g-**od** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d g-**od** d-i-v-i-d-e-d the l-i-g-h-t
4	**god** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d **god** d-i-v-i-d-e-d the l-i-g-h-t
5	god c-r-e-a-t-**ed** the h-e-a-v-e-n a-n-d god d-i-v-i-d-**ed** the l-i-g-h-t
6	god c-r-**ea**-t-ed the h-**ea**-v-e-n a-n-d god d-i-v-i-d-ed the l-i-g-h-t

.
.
.

Closer to **character** level

Closer to **word** level

Merge	BPE text tokenization
0	g-o-d c-r-e-a-t-e-d t-h-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d t-h-e l-i-g-h-t
1	g-o-d c-r-e-a-t-e-d **th**-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **th**-e l-i-g-h-t
2	g-o-d c-r-e-a-t-e-d **the** h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **the** l-i-g-h-t
3	g-**od** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d g-**od** d-i-v-i-d-e-d the l-i-g-h-t
4	**god** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d **god** d-i-v-i-d-e-d the l-i-g-h-t
5	god c-r-e-a-t-**ed** the h-e-a-v-e-n a-n-d god d-i-v-i-d-**ed** the l-i-g-h-t
6	god c-r-**ea**-t-ed the h-**ea**-v-e-n a-n-d god d-i-v-i-d-ed the l-i-g-h-t

.
.
.

Most **frequent pair**

Closer to **character** level

Closer to **word** level

Our Research Goals

→ How the distribution of these subwords **changes** across different merge operations (and across languages)?

Our Research Goals

→ How the distribution of these subwords **changes** across different merge operations (and across languages)?

→ Do languages get **‘closer’** in terms of their subword distributions under specific levels of tokenization?

Our Research Goals

→ How the distribution of these subwords **changes** across different merge operations (and across languages)?

→ Do languages get ‘**closer**’ in terms of their subword distributions under specific levels of tokenization?

→ Interpret these observations in light of previous findings regarding morphological **complexity**

Our approach

We quantify this cross-linguistic variation using information-theoretic measures

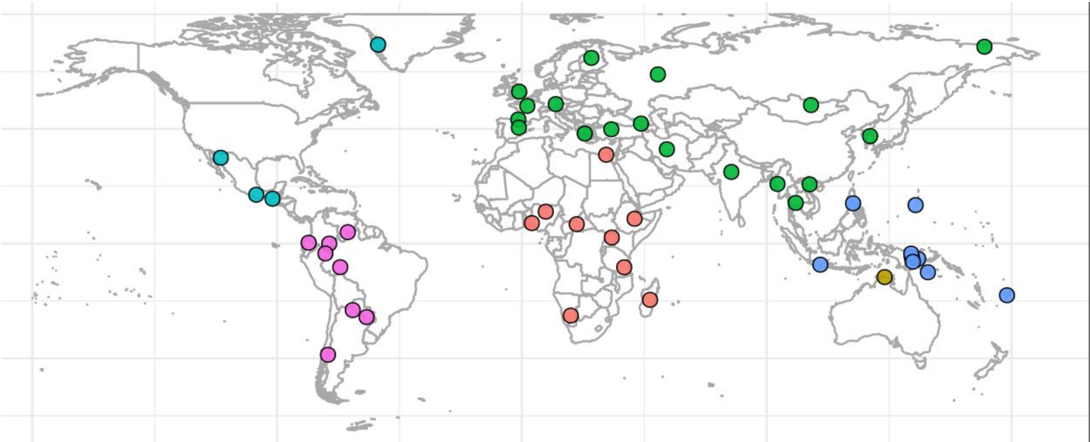
- ➔ We measure Shannon **entropy** and **redundancy** over varied subword tokenizations of texts obtained with BPE
- ➔ At each incremental merge, we compare the values across **47** typologically diverse languages

Our approach

We quantify this cross-linguistic variation using information-theoretic measures

➔ We measure Shannon **entropy** and **redundancy** over varied subword tokenizations of texts obtained with BPE

➔ At each incremental merge, we compare the values across **47** typologically diverse languages



*PBC parallel corpus (sample that aims to maximize both genealogical and areal diversity)

Entropy of a text T

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

A text T with a vocabulary of orthographic word types:

$$V = \{t_1, t_2, ..., t_V\} \text{ of size } |V|$$

Entropy of a text T

A text T with a vocabulary of orthographic word types:

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

$$V = \{t_1, t_2, \ldots, t_V\} \text{ of size } |V|$$

H (Kalallisut)= 11.63 bits
H (Finnish)= 10.27 bits
.
.
.
H (English)= 8.16 bits
H (Sango)= 6.71 bits

Usually correlated to **morphological complexity** measures

Languages with a greater diversity of word types will have **higher entropy** values --> word types are less predictable, **richer morphology**

Entropy of a text T

A text T with a vocabulary of subwords types:

$$V = \{t_1, t_2, ..., t_V\} \text{ of size } |V|$$

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

Entropy of a text T

A text T with a vocabulary of subwords types:

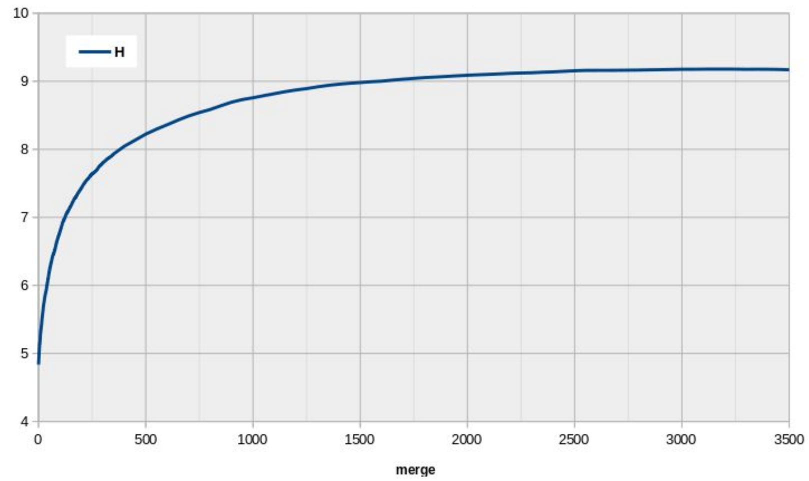
$$V = \{t_1, t_2, ..., t_V\} \text{ of size } |V|$$

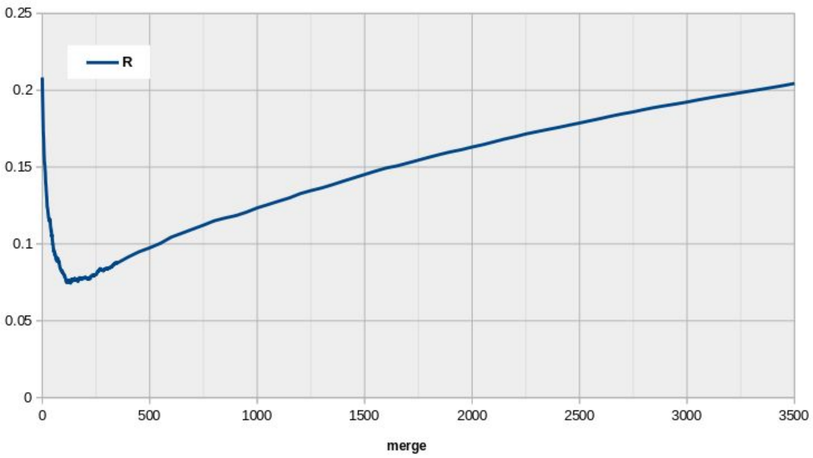
$$H(T) = -\sum_{i=1}^{V} p(t_i) \log_2 p(t_i)$$

Redundancy of a text T

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

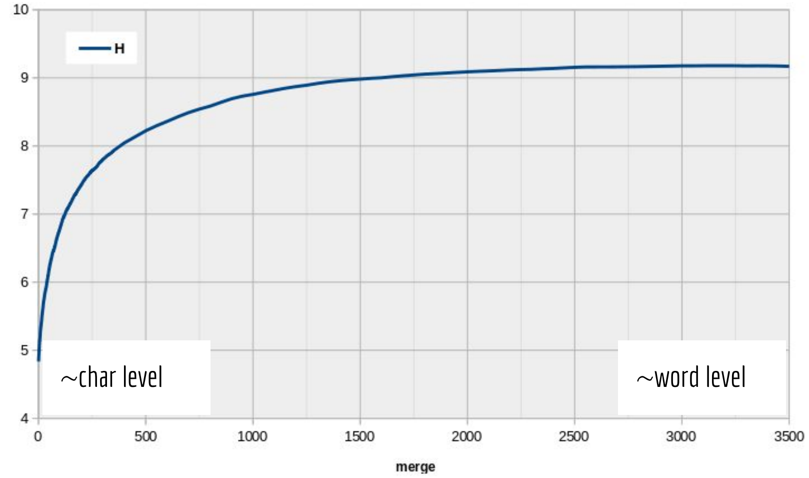
Entropy and Redundancy across BPE merges

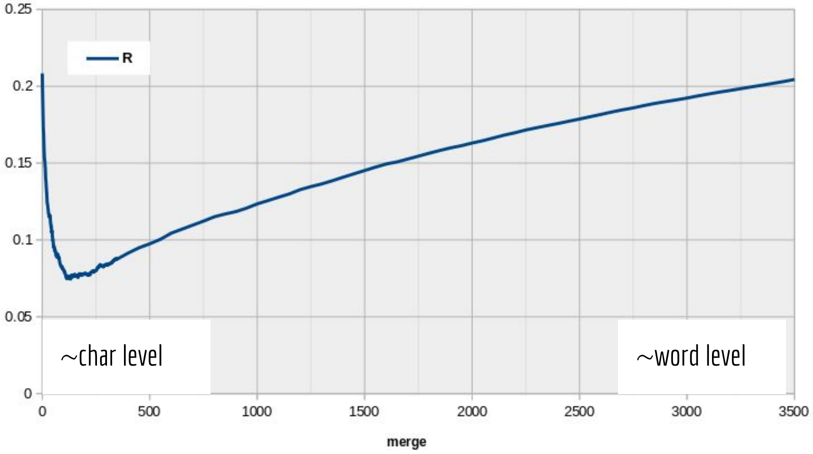




* French (fra)

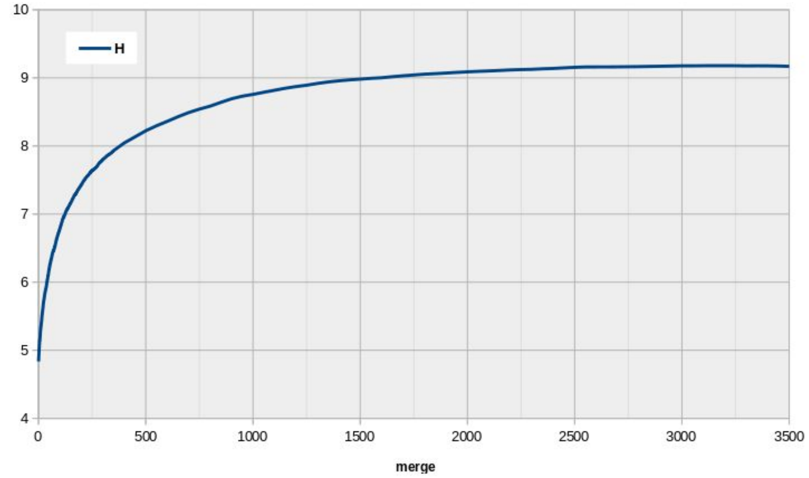
Entropy and Redundancy across BPE merges

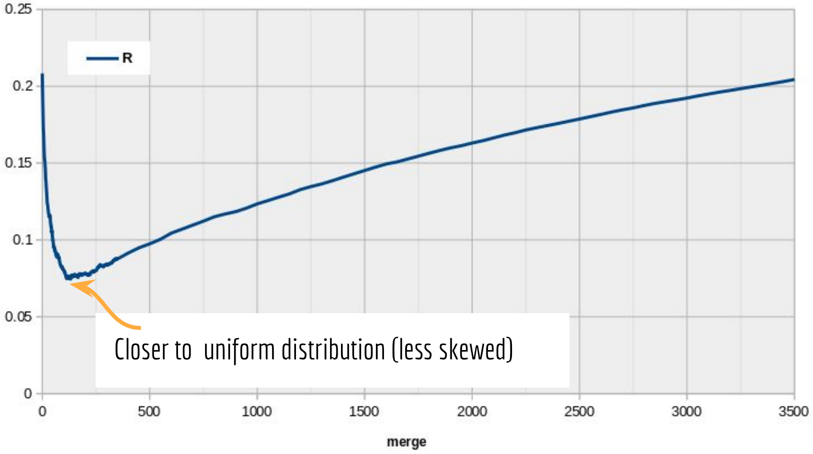




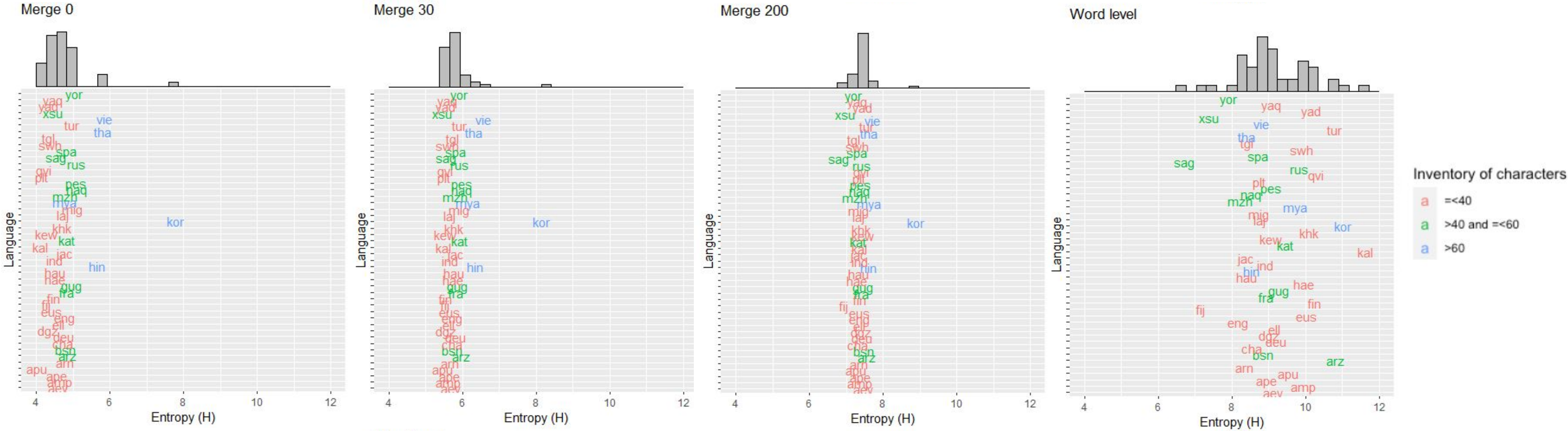
* French (fra)

Entropy and Redundancy across BPE merges

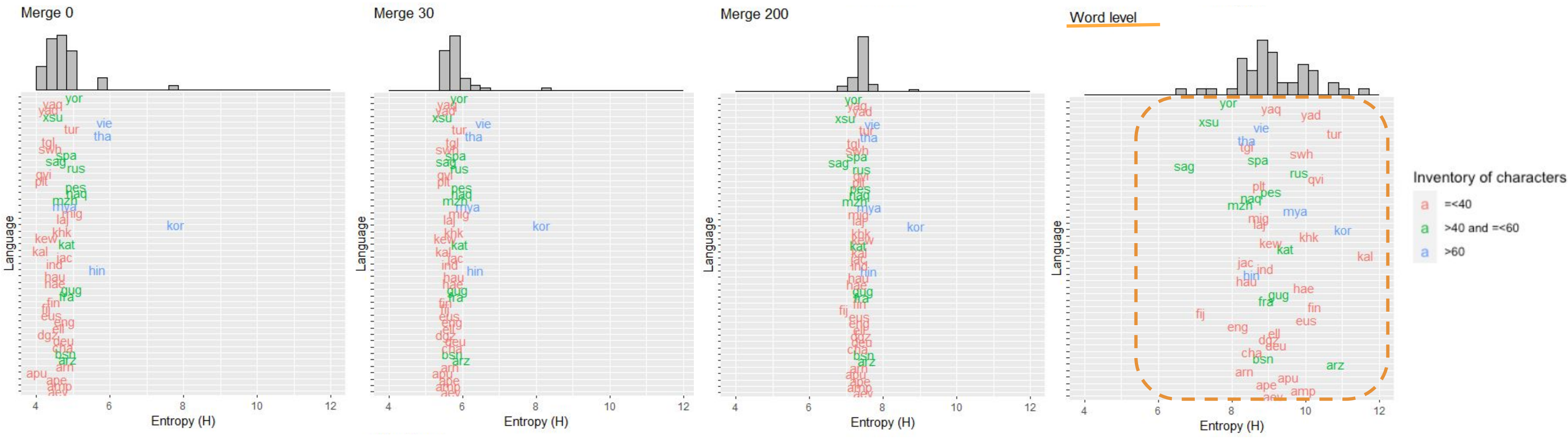




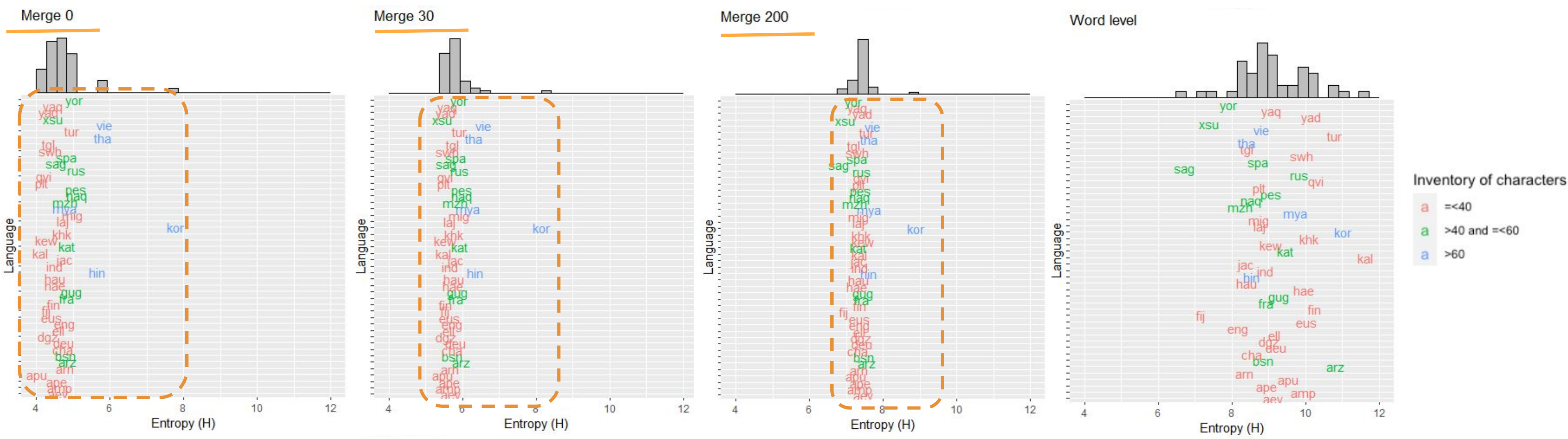
* French (fra)



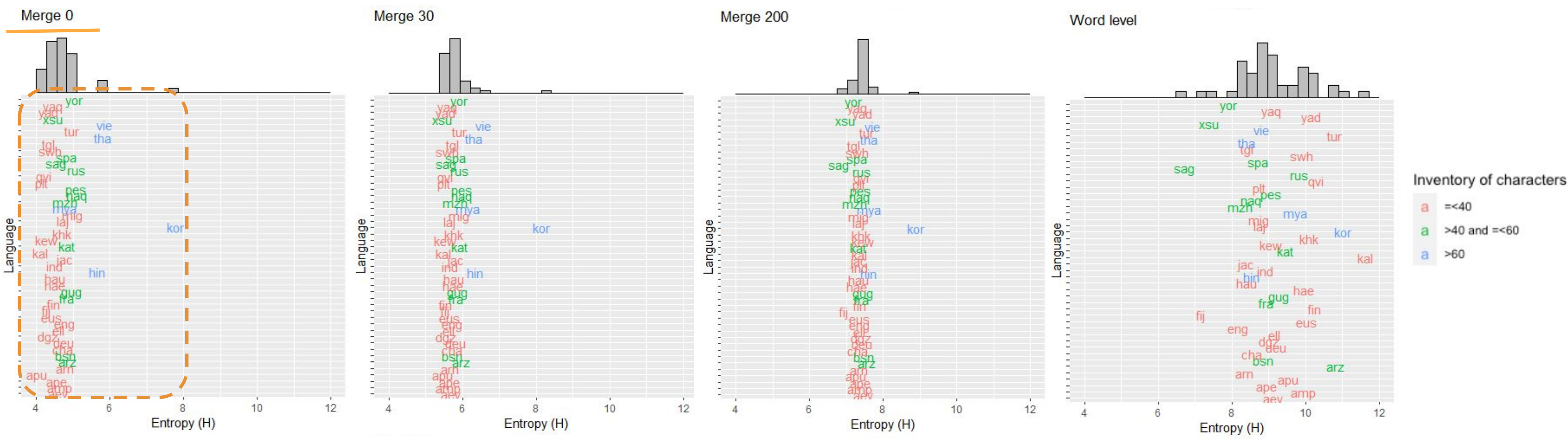
The turning point



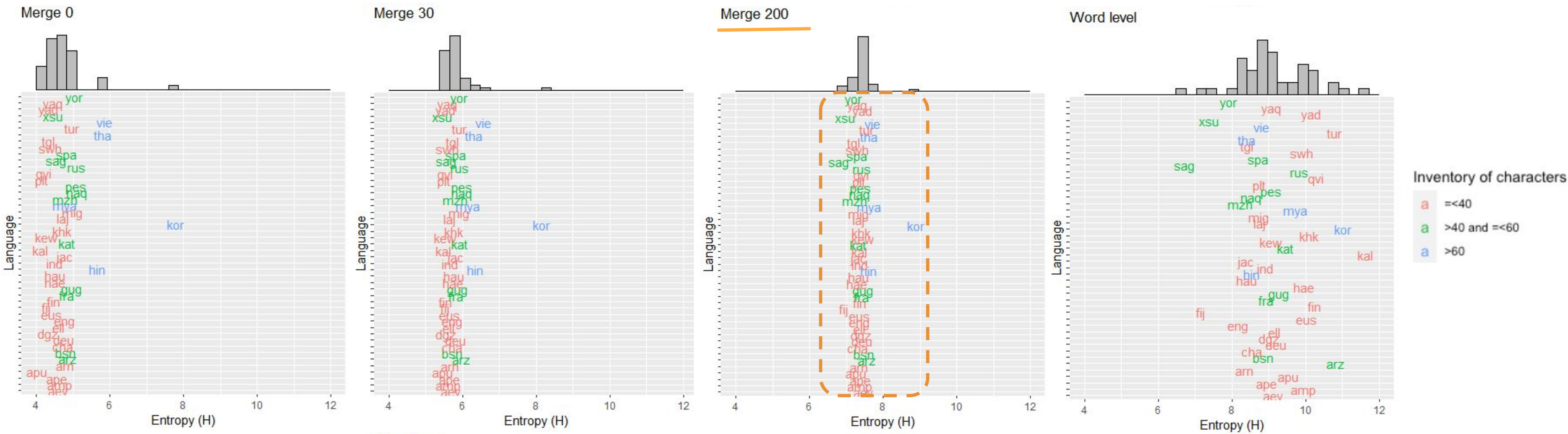
The turning point



The turning point

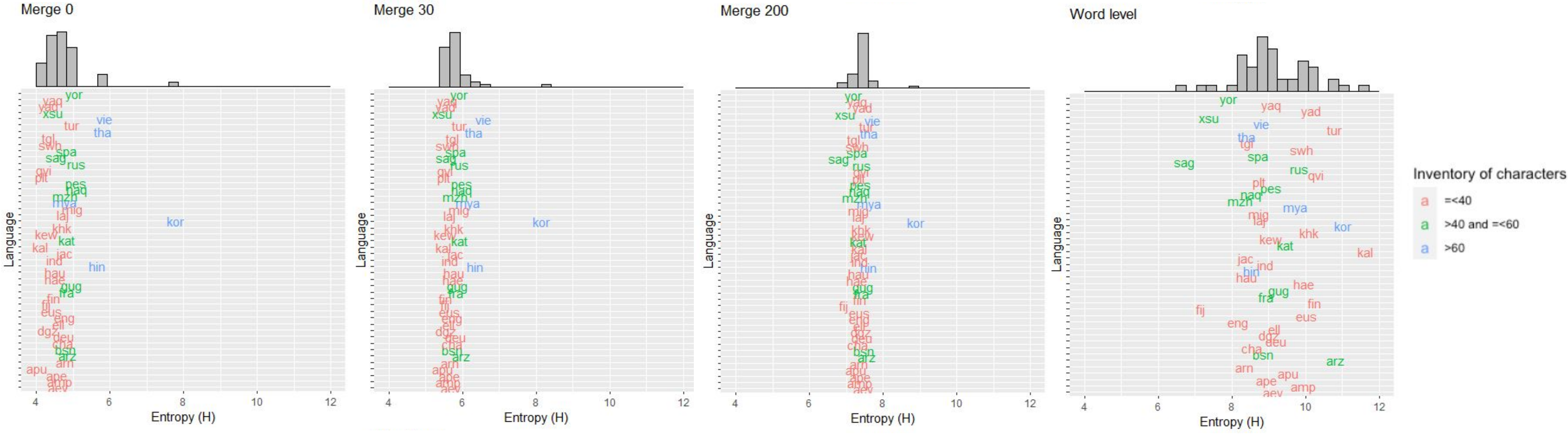


The turning point



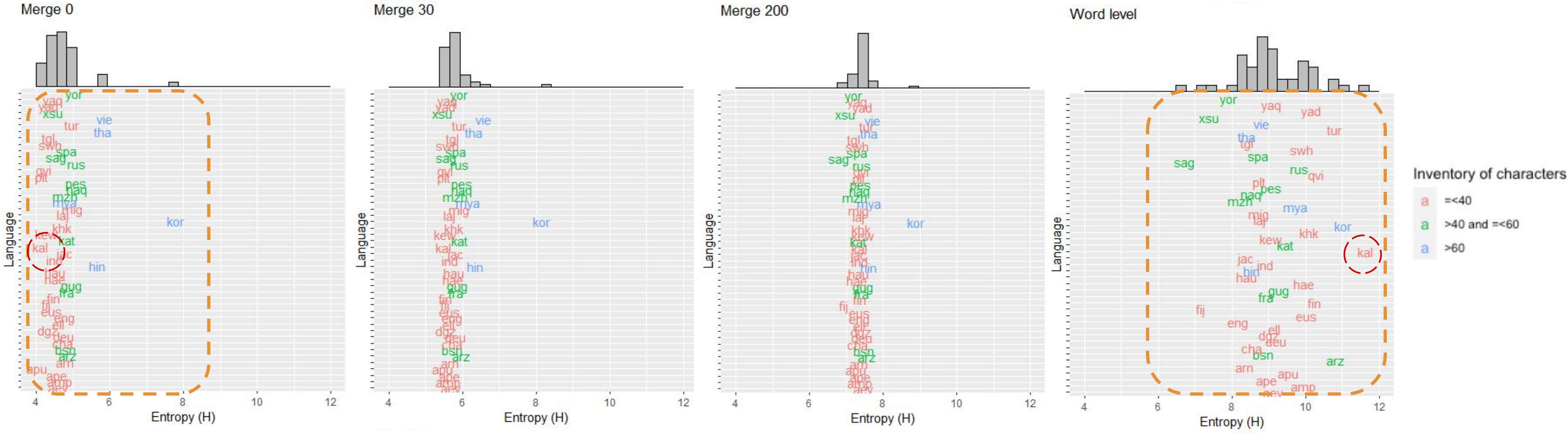
- **Similar frequency distributions (entropy)** across languages

The turning point



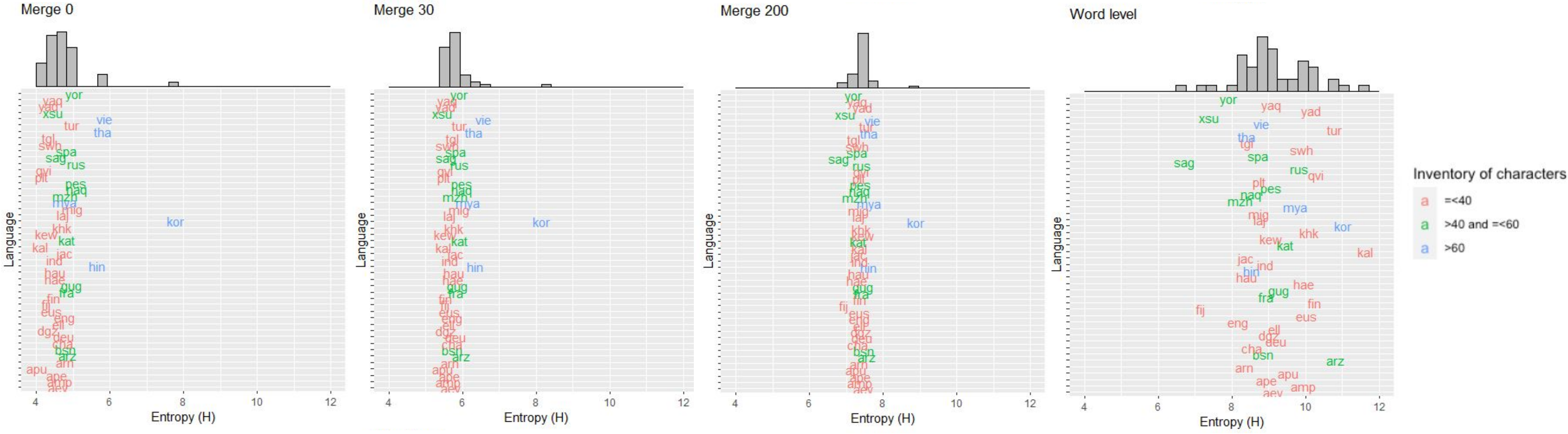
- The **ranking** of languages (by their entropy) is different before and after 200 merges (approx.)

The turning point



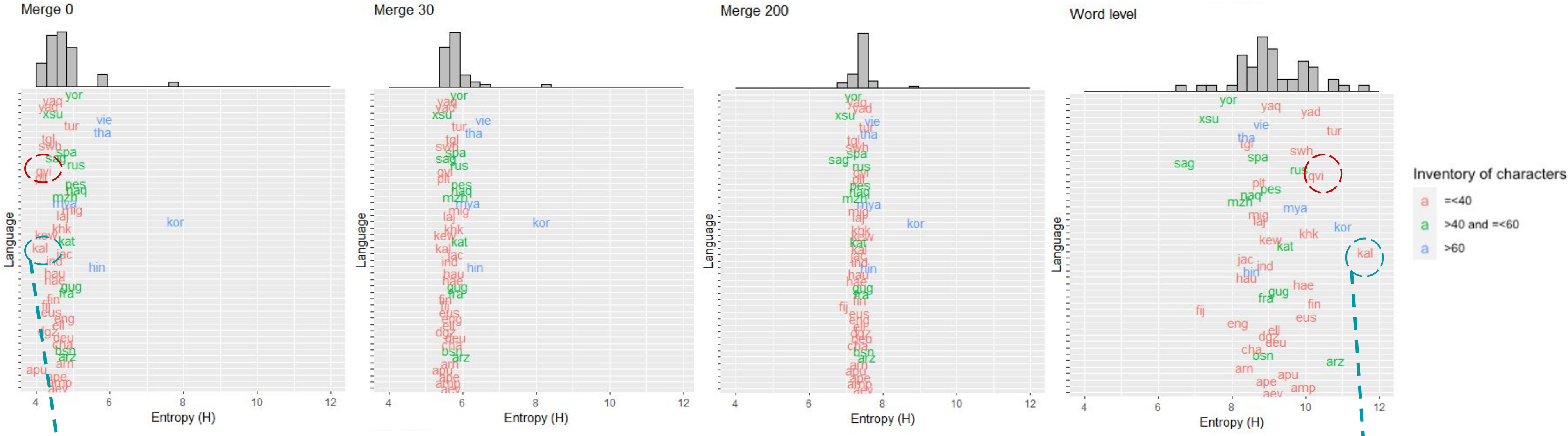
- The **ranking** of languages (by their entropy) is different before and after 200 merges (approx.)

The turning point



- The rankings obtained on the first merges (before 200) are **correlated** with the predictability of **sequences of trigrams within a word**

The turning point



predictable subword patterns

rich inflectional morphology

- Research on **linguistic complexity**

The turning point

Summary

Turning point (~merge 200)

➔ Entropy values **least dispersed** across languages

Summary

Turning point (~merge 200)

- ➔ Entropy values **least dispersed** across languages
- ➔ **Subword** token distributions gradually start to look like **word-level** distributions

Summary

<u>Turning point (~merge 200)</u>

- ➔ Entropy values **least dispersed** across languages
- ➔ **Subword** token distributions gradually start to look like **word-level** distributions
- ➔ Text **redundancy** start to grow after an initial drop
- ➔ Text **Entropy** slows down after initial fast growth

Summary

<u>**Turning point (~merge 200)**</u>

→ Entropy values **least dispersed** across languages

→ **Subword** token distributions gradually start to look like **word-level** distributions

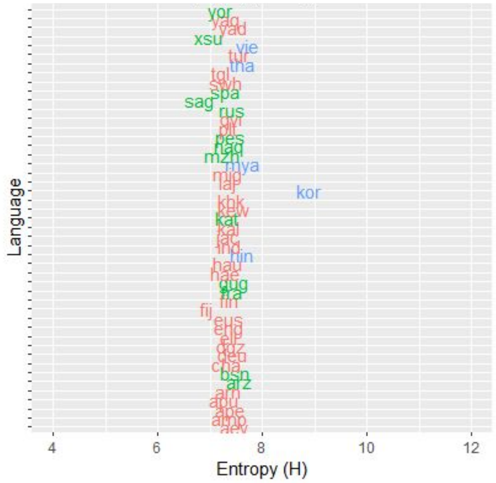
→ Text **redundancy** start to grow after an initial drop

→ Text **Entropy** slows down after initial fast growth

→ At the **early merges,** the entropy of texts is strongly correlated with a complexity measure based on modeling character **trigrams sequences**

Conclusions

- ➔ Some subword tokenizations led to **surprisingly similar** entropy across languages
    - ➔ This could be beneficial for NLP multilingual tasks, e.g., choose the **number of BPE merge** operations



Thank you for your attention!

Data and code:
https://github.com/ximenina/theturningpoint