From characters to words: the turning point of BPE merges

Ximena Gutierrez-Vasques[1], Christian Bentz[2], Olga Sozinova[1], Tanja Samardžić[1]

URPP Language and Space, University of Zürich[1] Department of General Linguistics, University of Tübingen[2]

EACL 2021

University of Zurich[UZH]

Eberhard Karls Universität Tübingen

1. Introduction

Subword tokenization

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 is merged

Closer to character level

Closer to word level

2. Research goals

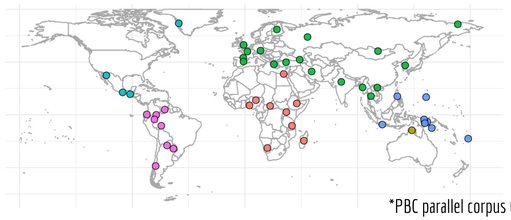
We want to know:

- 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

3. 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)

A text T with a vocabulary V of subword types $V = \{t_1, t_2, ..., t_V\}$ of size $|V|$

- Entropy:

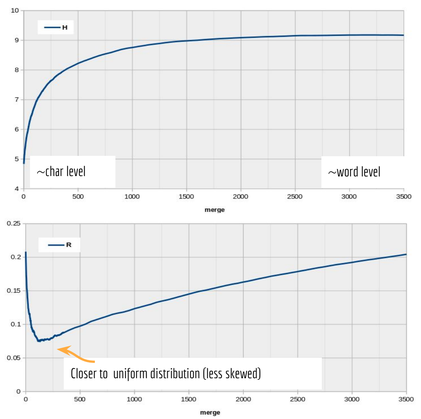
$$H(T) = -\sum_{i=1}^{V} p(t_i) \log_2 p(t_i)$$

- Redundancy:

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

4. Results

Entropy and Redundancy across BPE merges



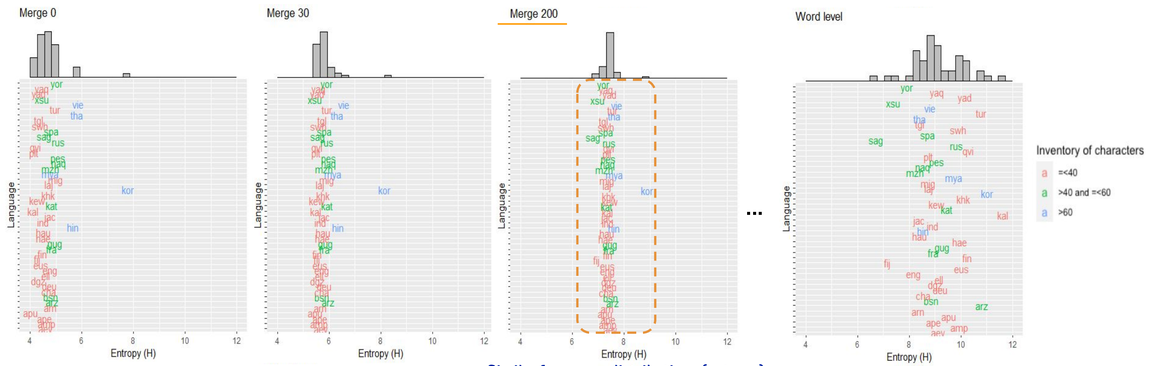
* Example for French (fra)

- First merges cause the most drastic changes:

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

5. Results

Cross-linguistic comparison of Entropy



Similar frequency distributions (entropy) across languages

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

6. The turning point (~merge 200)

- Entropy values are 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 (~before 200) , the entropy of texts is strongly correlated with a complexity measure based on modeling character trigrams sequences within a word.
- At the later merges (~after 200) the entropy gradually correlates with word unigram entropy,

7. 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

- The entropy over word-level types reflects one dimension of morphological diversity. However, at a more atomic subword level, a different dimension of morphological complexity is reflected.