A look to languages through the glass of BPE compression

Ximena Gutierrez-Vasques[1] **Christian Bentz**[2] **Tanja Samardžić**[1]
[1]URPP Language and Space, University of Zürich
[2]Department of General Linguistics, University of Tübingen
{ximena.gutierrezvasques, tanja.samardzic}@uzh.ch
chris@christianbentz.de

1 Introduction

One of the predominant methods for subword tokenization is Byte-pair encoding (BPE). Originally, this is a data compression technique based on replacing the most common pair of consecutive bytes with a new symbol (Gage, 1994). When applied to text, each iteration merges two adjacent symbols; this can be seen as a process of going from characters to subwords through iterations (Sennrich et al., 2016).

Regardless of the language, the first merge operations tend to have a stronger impact on the compression of texts, i.e., they capture very frequent patterns that lead to a reduction of redundancy and to an increment of the text entropy (Gutierrez-Vasques et al., 2021). However, the natural language properties that allow this compression are rarely analyzed , i.e., do all languages get compressed in the same way through BPE merge operations? We hypothesize that the type of recurrent patterns captured in each merge depends on the typology and other corpus-related phenomena. For instance, for some languages, this compression might be related to frequent affixes or regular inflectional morphs, while for some others, it might be related to more idiosyncratic, irregular patterns or even related to orthographic redundancies.

We propose a novel way to quantify this, inspired by the notion of morphological productivity.

2 Data and Methods

For each merge operation, we quantify whether the newly created subword has a tendency to be more productive (many different word types contain it) or more irregular/idiosyncratic (few word types contain it, but those types have high frequency).

We analyze 47 diverse languages[1], and we especially focus on the first 200 merges.

In morphology, we can think of a productive pattern as one that can be applied to many different lexemes (systematic). In contrast, a non-productive pattern won't appear in many different lexemes. Although, it can be very frequent, e.g., suppletion (Baayen, 1992; Bonami and Beniamine, 2016; Bybee, 2010).

Subword	|W|	Cum. freq.	Idiosyncrasy
ed</w>	271	917	3.38
had</w>	1	104	104

Table 1: Example of subwords, PBC corpus (English)

We propose the following rough operationalization of productivity to classify subwords. For each newly created subword we calculate what we have called idiosyncrasy index:

$$\text{idiosyncrasy(subword)} = \frac{\sum_{w \in W} freq(w)}{|W|} \quad (1)$$

Where $|W|$ is the number of word types that contain the current subword, $\sum_{w \epsilon W} freq(w)$ is the cumulative frequency of those word types.

Subwords that appear in many different word types will have lower values of the idiosyncrasy index. While subwords that appear in few word types, but the cumulative frequency of these types (or number of tokens) is very high will have higher values of idiosyncrasy. See example in Table 1.

3 Results

We represent each subword as a three-dimensional vector ($|W|$, $Cum.freq$, $idiosyncrasy$). The exact way subwords distribute across space and which ones gets merged first depends on the language; see example in Fig. 1.

In languages like Kalaallisut (typically seen as polysynthetic), many of its subwords seem to distribute around the area with a low level of idiosyncrasy index. Many of these subwords seem to be highly productive, i.e., they appear in a relatively high number of word types, and these types have a relatively high cumulative frequency. Subwords

[1]Parallel Bible Corpus (PBC) (Mayer and Cysouw, 2014)

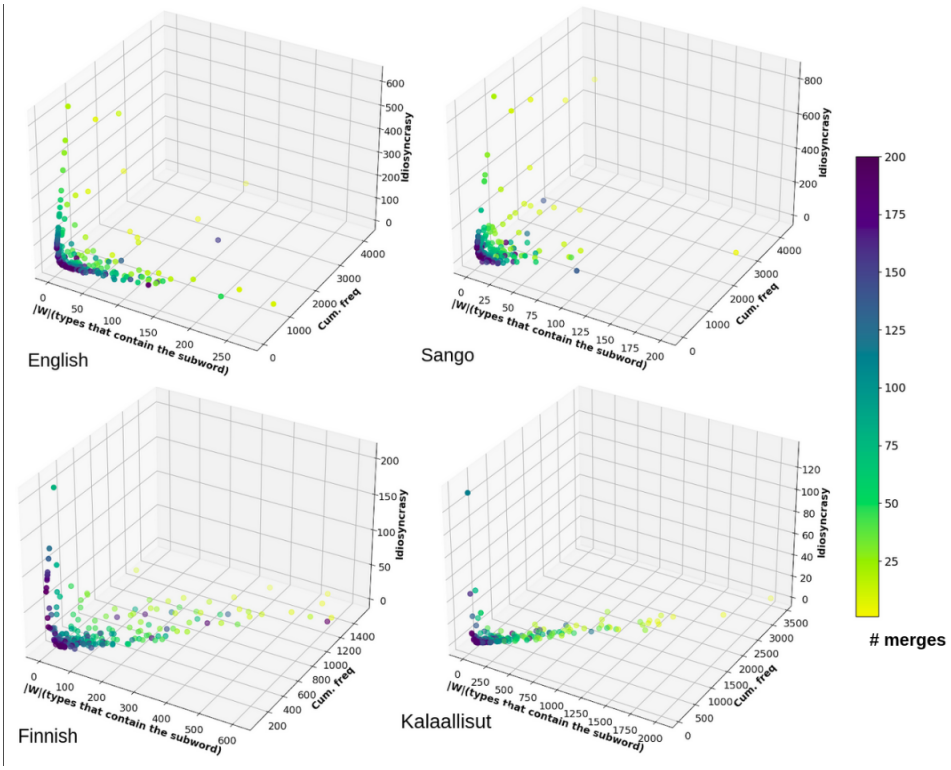

Figure 1: BPE subwords. Color indicates the number of merge operation in which that subword was created.

that correspond to regular morphological patterns might be around this area. There are also some subwords with a high level of idiosyncrasy; however, not only are they fewer, but if we compare them against languages like English, the idiosyncrasy's range is relatively low (120 vs. 600).

Moreover, in languages like Kalaallisut or Finnish, the subwords with higher idiosyncrasy index are not captured during the first merges, probably they are not the best candidates for compression, while for languages like English or Sango the subwords with the highest levels of idiosyncrasy seem to be merged during the first 50 operations.

We can also see that in languages like Sango (typically seen as isolating) or English, the subwords are less productive, with a more prominent concentration in the area where the idiosyncrasy index is high, i.e., subwords that are part of very few types. Still, these few types have a relatively high cumulative frequency. In general, subwords with a high idiosyncrasy level can correspond to cases where: a) whole words are merged in relatively quickly since their frequency on the corpus is high; b) morphological phenomena that may not be productive but are quite frequent, like irregular or suppletive patterns.

We also notice that, for some languages, orthography seems to influence the symbols that get merged during some of the first operations.

The more merge operations, the more the languages tend to behave similarly: the subwords start to accumulate closer to the origin, since they appear in fewer word types and those word types are not very frequent. This explains why the subwords formed at later merges have smaller effect on the change on the compression.

4 Conclusions and future work

Our preliminary findings suggest that the type of patterns that emerge through the first merges, and that allow compression, are mainly an interaction between a) Subwords with high value of idiosyncrasy index; b) Subwords highly productive. For morphologically rich languages with regular inflectional morphology, BPE shows a tendency to merge subwords that are productive. For languages with poorer morphology or less regular patterns, BPE will tend to merge subwords that are less productive but with a high idiosyncrasy index.

As a future step, we can cluster the different types of subwords per each language, e.g., regular affixes, irregular stems, etc. Moreover, the different subword distributions obtained for each language, can be used to cluster languages according to their morphological properties.

References

Harald Baayen. 1992. Quantitative aspects of morphological productivity. In *Yearbook of morphology 1991*, pages 109–149. Springer.

Olivier Bonami and Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of bpe merges. In *European Chapter of the Association for Computational Linguistics, Long Papers. 2021 (to appear)*. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania*, 135(273):40.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.