Subword Geometry: Picturing Word Shapes

Olga Sozinova
URPP Language and Space
University of Zurich
olga.sozinova@uzh.ch

Tanja Samardžić
URPP Language and Space
University of Zurich
tanja.samardzic@uzh.ch

Abstract

In this work in progress, we are investigating the structural properties of subwords in 20 languages by extracting *word shapes*, i.e. sequences of subword lengths.

Words in natural languages consist of subword units. In traditional linguistics, such units are described as morphemes, roots and affixes, and are mostly studied from the functional point of view. Their structural and combinatorial properties are far less studied. Such properties include the length and the order of subwords, which are especially important in the information theoretic view of language as an efficient code. In this view, subwords are expected to be ordered following a specific efficient strategy.

The goal of our paper is to find out how subword segments are organized in terms of length. More specifically, is there a preference for uneven or even segment lengths? If yes, is a specific preference language-dependent or universal?

Contribution Our study highlights a new facet of word structures. We formulate a novel framework for quantitative comparison of languages and identify cross-linguistic patterns which can help improve unsupervised multilingual subword tokenization.

Approach The starting point of our investigation is the work by Menzerath (1954) and Altmann (1980), which resulted in a formulation of Menzerath-Altmann's law: "The larger the whole, the smaller the parts" (Menzerath, 1954, p. 101). Menzerath-Altmann's law shows a zoomed out view of word structures, namely, that longer words tend to have shorter segments. But are all segments equally shorter? Is there any difference in segments behavior depending on their order? These questions have not been tackled yet.

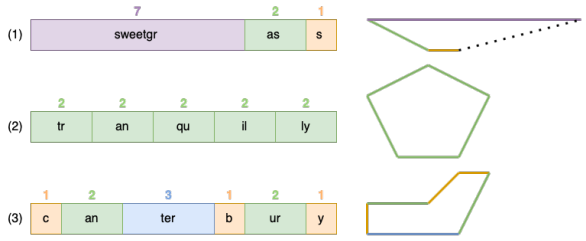

Figure 1: Word shapes examples for the BPE segmented words of length 10: *sweetgrass*, *tranquilly*, *canterbury*.

In addition to this, we want to know whether our observations converge on a typologically diverse language sample, including low resource languages. To our knowledge, there has been no study conducted in this direction on multilingual data.

In our approach, we distinguish between **even** and **uneven** subword sequences and compare their distributions across 7 high resource (HR) and 13 low resource (LR) languages.

For example, given a word consisting of 10 characters, what are the most likely segmentations that we get: 1) one segment of length 7 and two segments of length 2 and 1, 2) 5 segments each of length 2 or 3) 10 segments of lengths within a narrow range [1, 2, 3, 1, 2, 1]? (Figure 1)

The first option represents uneven lengths, the second one is strictly even and the last one is relatively even. If we imagine the resulting segments marked on a wire that we could bend into geometric figures, we could get an uneven open figure in the first case, an equilateral pentagon in the second case and a closed polygon in the last case, leaning towards even shapes. Following this metaphor, we call our object of study, i.e. the sequences of segments' lengths, the *word shapes*, which we characterize by their *evenness*.

Our hypothesis is that the word shapes of the third type are most common across languages.

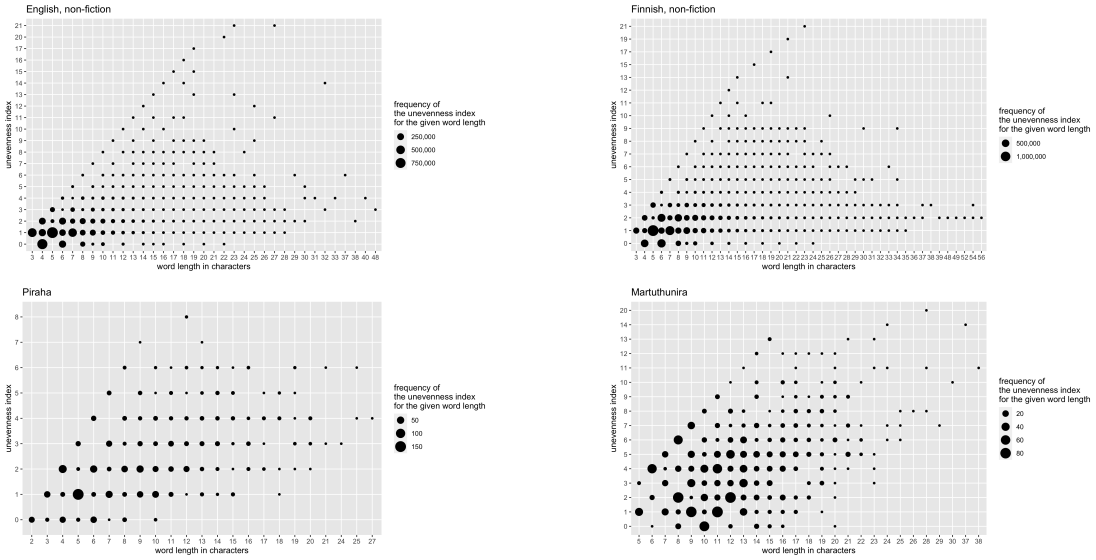

Figure 2: Unevenness index (UI) distribution in: (a) English, (b) Finnish, (c) Pirahã and (d) Martuthunira texts.

Data and Methods We use the corpus for which we collect the data ourselves. Our corpus is based on the 100 language sample of typologically diverse languages [1] and contains texts of different genres.

For this study, we select 20 languages, 7 high resource (HR) and 13 low resource languages (LR). HR languages include German, English, Finnish, French, Modern Greek, Russian and Spanish. LR languages in our study are Bagirmi, Burushaski, Dani (Lower Grand Valley), Imonda, Kayardild, Lavukaleve, Makah, Martuthunira, Maybrat, Ngiyambaa, Pirahã, Rama, Tiwi.

HR data includes texts from the Parallel Bible Corpus (Mayer and Cysouw, 2014) and from the Open Subtitles corpus (Lison and Tiedemann, 2016). LR data consists of manually prepared texts, extracted from grammars and fieldwork materials. LR texts contain manual segmentations, which we use directly. For HR languages, we obtain the segmentations by applying the BPE algorithm (Gage, 1994; Sennrich et al., 2016). The preprocessing step is finalized by calculating the word shapes.

Our main analysis consists of establishing a relationship between evenness of the word shapes and the word length. In order to formalize evenness, we introduce the following *unevenness index (UI)*: $UI = max(segments) - min(segments)$, where *segments* is a set of segments' lengths for a given word.

The lower UI, the more even is the word shape. When UI equals zero, all segments are of equal length, the word shape is strictly even. The word shapes on the Figure 1 are quantified by UI as: (1) $7 - 1 = 6$, (2) $2 - 2 = 0$ and (3) $3 - 1 = 2$.

Results Figure 2 shows the distribution of the unevenness index with regards to the word length in English, Finnish (both BPE segmented) and Pirahã, Martuthunira (both manually segmented). The frequency of the UI values for the given word length is depicted by the size of the points.

We can see that in HR languages, shorter words (left part of the triangle) can be both even and uneven, however even shapes are preferred. Longer words (right part of the triangle), on the contrary, don't have much variance and are restricted to be even. LR languages are the same regarding shorter words, but the long words behave differently, preferring uneven shapes.

Our initial hypothesis is proved. Both in HR and LR languages, the word shapes tend to be even, especially in shorter words.

Conclusion In this ongoing study, we have found a universal trend: word shapes tend to be even. This tendency is strong in shorter words, while longer words have more unpredicted behavior cross-linguistically. In HR languages, they are restricted to even word shapes only, while in LR languages they tend to have uneven word shapes. The patterns we observe are strikingly uniform across languages, making our findings valuable for the understanding of the subwords organization processes.

[1] https://wals.info/languoid/samples/100

References

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2(2):1–10.

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

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

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

Paul Menzerath. 1954. *Die Architektonik des deutschen Wortschatzes*, volume 3. F. Dümmler.

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.