Languages through the Looking Glass of BPE Compression

Ximena Gutierrez-Vasques
URPP Language and Space, University of Zürich. ximena.gutierrezvasques@uzh.ch

Christian Bentz
Department of General Linguistics, University of Tübingen. chris@christianbentz.de

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

Byte-pair encoding (BPE) is widely used in NLP for performing subword tokenization. It uncovers redundant patterns for compressing the data, and hence alleviates the sparsity problem in downstream applications. Subwords discovered during the first merge operations tend to have the most substantial impact on the compression of texts. However, the structural underpinnings of this effect have not been analyzed cross-linguistically. We conduct in-depth analyses across 47 typologically diverse languages and three parallel corpora, and thereby show that the types of recurrent patterns that have the strongest impact on compression are an indicator of morphological typology. For languages with richer inflectional morphology there is a preference for highly productive subwords on the early merges, while for languages with less inflectional morphology, idiosyncratic subwords are more prominent. Both types of patterns contribute to efficient compression. Counter the common perception that BPE subwords are not linguistically relevant, we find patterns across languages that resemble those described in traditional typology. We thus propose a novel way to characterize languages according to their BPE subword properties, inspired by the notion of morphological productivity in linguistics. This allows us to have language vectors that encode typological knowledge induced from raw text. Our approach is easily applicable to a wider range of languages and texts, as it does not require annotated data or any external linguistic knowledge. We discuss its potential contributions to quantitative typology and multilingual NLP.

1. Introduction

One of the most striking differences between languages is the degree to which information is condensed in both speech and writing. While some languages concatenate short, repetitive chunks into long sequences, others form more condensed, shorter sequences. Consider the parallel sentences below.

(1) They wanted to catch a walrus English

Action editor: Carlos Gómez-Rodríguez. Submission received: 8 August 2022; revised version received: 26 April 2023; accepted for publication: 2 June 2023.

(2) aaffakkumasut Late Eastern Inuit[1]
Dahl (2017, p. 26) citing Fortescue (1992)

In the English sequence, we count 29 characters (including spaces), 7 morphemes, and 6 orthographic words. In the Late Eastern Inuit sequence, the same content is represented with 13 characters, 4 morphemes,[2] and 1 orthographic word. Despite such cross-linguistic differences, extensively studied in morphological typology, sequences in all natural languages contain redundant sub-sequences, and can be further compressed. This fact is exploited in modern Natural Language Processing (NLP) for improving text segmentation and encoding by means of subword tokenization (Gallé 2019; Mielke et al. 2021).

A popular method for uncovering subword units is Byte-Pair Encoding (BPE) (Gage 1994). This is a compression algorithm which proved helpful in machine translation and other downstream tasks (Sennrich, Haddow, and Birch 2016). Despite its usefulness in language processing, this method is commonly judged as not linguistically relevant, and pitched against other approaches which leverage explicit external linguistic knowledge, for instance, about the morphology of the respective languages.

We find this common view puzzling: to compress language data, BPE needs to merge sequences of co-occurring characters, i.e., subwords, that reduce redundancy. These patterns might not correspond to usual morphological analyses, but they are structural elements. What we aim to find out in this study is what kinds of structural elements are exploited by BPE for text compression across different languages and if this allows for data-driven induction of typological knowledge. We focus mostly on the first elements merged by BPE that have been identified as having the most substantial impact on the compression of texts (Gutierrez-Vasques et al. 2021) and put forward the following hypothesis:

> **Hypothesis**: The properties of subwords found in BPE compression depend on the morphological type of the language in question.

To test this hypothesis, we first need to quantify the subwords' properties. We achieve this with a novel method inspired by the notion of morphological productivity in linguistics. For each subword found in the incremental process of BPE compression, we quantify whether it has a tendency to be more *productive* (many different word types contain it) or more *idiosyncratic* (few word types contain it, but those types have high frequency). We obtain language vectors in a BPE subword productivity space based on this operationalization.

The second step in testing our hypothesis is to assess to what degree these language vectors, derived from the properties of BPE subwords, encode the known morphological types of languages. If indeed there is a connection between BPE compression and the structure of language, there should be a good agreement between language vectors in the BPE subword space and independent language vectors derived from external linguistic knowledge. We evaluate this agreement both in quantitative and qualitative terms.

If our hypothesis is confirmed, properties of subwords can distinguish automatically between different morphological types of languages using only raw text. By monitoring the outcome of compression steps, we can track the cross-linguistic differences in what kinds of redundancy are gradually removed in different languages. For example, are the most redundant patterns in English and Late Eastern Inuit of the same kind?

1 The language here called "Late Eastern Inuit" is likely associated with what is called "Eastern Canadian Inuktitut" (ISO 639-3: ike) in Glottolog (https://glottolog.org/).

2 These morphemes have become amalgamated over historical time such that morpheme boundaries are rather blurred.

For some languages, the compression can be achieved via productive subwords, those patterns resembling inflectional markers, affixes, and regular morphological phenomena. For other languages, the compression can be achieved by harnessing idiosyncratic subwords, those that correspond to frequent irregular patterns or whole orthographic words which are highly redundant due to their high frequency in the corpus.

We carry out experiments with several highly diverse multilingual data sets. We show that the BPE-based language vectors capture distinctions traditionally described in morphological typology. Furthermore, we find a good alignment between these vectors and those obtained from a typological database, showing that it is possible to induce typological knowledge from raw data and a compression algorithm. While compression is mainly applied in the context of data processing and storage, we here use it as a research method.

The possibility of comparing languages using typological information has proven beneficial in several current NLP domains, e.g., multilingual NLP and cross-lingual transfer of models. Our BPE-based vectors automatically induce typological knowledge from text. This opens up a way of easily extending the current typological language vectors used in NLP, which often rely on linguistic databases where features are incomplete. In terms of linguistic research, our study provides a quantitative tool for morphological typology, contributing to recent trends that go beyond predefined morphological categories towards continuous representations.

The article consists of five sections in addition to Introduction and Conclusion. In Section 2, we introduce all the theoretical and technical notions that constitute the background for our proposal together with the current state of the art. In Section 3, we describe in detail our approach to quantifying the properties of subwords and how we test the relationship between text compression and language typology. In Section 4, we present the results of the experiments showing the alignment between our proposed method and typological databases. In Section 5, we further corroborate the empirical results showing that they are in line with a wide body of literature in language typology. In Section 6, we discuss the limitations of our proposal and possible improvements in future work.

2. Background and Related Work

Compression is typically known as a standard tool in the context of data storage. In a broader sense, it also emerges as a principle underlying efficient communication in animals (Ferrer-i Cancho et al. 2013), and humans (Kirby et al. 2015; Ferrer-i Cancho 2018; Ferrer-i Cancho, Bentz, and Seguin 2022), as well as brain and cultural evolution more generally (Tamariz and Kirby 2015; Al Roumi et al. 2021; Johnston et al. 2022). There have been some proposals to harness text compression as a strategy for approximating linguistic complexity (Juola 1998; Ehret and Szmrecsanyi 2016; Ehret 2016). This line of research can help to answer some fundamental linguistic questions, for instance, which information encoding units, i.e., characters, morphemes, or orthographic words, are most useful to uncover regular patterns in written texts (Geertzen, Blevins, and Milin 2016). In all of these studies, standard compression tools like *gzip* are used. A more dynamic view of compression is currently employed in NLP systems for preprocessing text as input to neural models.

2.1 Subword tokenization as text compression

Subword tokenization has become a standard preprocessing step in NLP. It consists of splitting orthographic words into smaller units (subwords), leading to less varied types with higher frequencies. This helps to counter the sparseness problem for rare words, and, thus, optimizes the input for NLP systems, especially for morphologically rich languages.

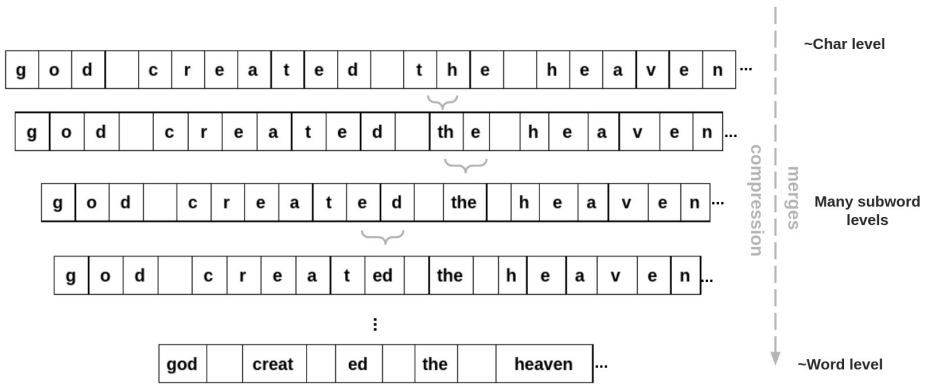


Figure 1: BPE compression example. Each (non-empty) box represents a single symbol.

There are several ways to decompose orthographic words into subwords given a text (Mielke et al. 2021). In particular, some unsupervised data-driven methods have approached this as a data compression task. Byte-Pair Encoding (BPE), for instance, is a lossless compression algorithm first applied to text processing by Sennrich, Haddow, and Birch (2016). Other methods, such as Morfessor (Creutz and Lagus 2002; Grönroos et al. 2014) or the SentencePiece Unigram model (Kudo 2018) are based on Minimum Description Length (MDL), a principle more widely used in statistical learning and information theory. According to MDL, the best model for learning about the data is the one that provides the shortest description, i.e., compresses the data the most (Rissanen 1978; Goldsmith 2001; Myung 2001; Grönroos et al. 2014)[3].

Data compression exploits redundant patterns or, in other words, *regularities* in the data. If we think of natural languages, these underlying regularities are recurrent patterns in the strings of characters, such as recurrent orthographic words, e.g. *the* or *and*, morphological markers, e.g. the *-ed* or *-ing* suffixes, or writing conventions, e.g. using *th* to represent the dental fricative /θ/ in English, or using *sch* to represent the postalveolar fricative /ʃ/ in German.

2.1.1 Byte-Pair Encoding (BPE). The term *Byte-Pair Encoding* refers to the initial idea of iteratively replacing the most common pair of consecutive bytes with a new symbol (Gage 1994). This technique belongs to the macro-based algorithms which achieve compression by replacing redundant strings or patterns with common pointers to a shorter reference (Storer and Szymanski 1978; Gallé 2019).

In its application to text, BPE creates subwords by means of iterative merges of two adjacent symbols with the highest frequency (Sennrich, Haddow, and Birch 2016). The algorithm starts by splitting words into a sequence of characters. We can think of this as characters separated by white spaces. The algorithm merges the most frequent pair of consecutive characters within the corpus in the first operation, e.g., (‘e’,‘d’) → (‘ed’). The merged characters become a new symbol. In each of the following operations, the algorithm calculates the co-occurrence frequency of pairs of all the current consecutive symbols and it merges the most frequent pair again (see Fig 1). When the algorithm merges a frequent pair of symbols, it

3 According to MDL, the model that provides the shortest description of the data should be chosen since this will be reflected in the generalization capability of the model: the more we can compress the data, the more we have learnt about it, and the better we can predict it.

shortens the text by replacing many instances of the pair of symbols (plus the white space between them) with a single symbol.

The merged characters are subwords. As more merges are applied, longer subwords are obtained – we are getting closer to the orthographic word level. The algorithm stops when a pre-specified number of merge operations has been reached, or when it cannot find a pair of consecutive symbols with frequency greater than one. Typically, BPE does not cross the orthographic word boundary. In other words, it relies on orthographic word boundaries as "upper bounds" also for subwords.

Due to the conceptual simplicity of BPE, the lack of encoding of explicit linguistic knowledge, as well as the lack of generalized stopping criteria to obtain the most appropriate subword tokenization, the NLP literature usually regards this method as not linguistically informed (Gallé 2019; Bostrom and Durrett 2020; Clark et al. 2022; Saleva and Lignos 2021; Mielke et al. 2021; Oncevay et al. 2022; Mager et al. 2022). An increasing amount of work has compared BPE versus approaches that use explicit linguistic knowledge, e.g., rule-based morphological analyzers, and semi-supervised Morfessor. Interestingly, using more linguistically informed methods does not necessarily lead to improvement in tasks like machine translation (Domingo et al. 2023; Macháček, Vidra, and Bojar 2018; Saleva and Lignos 2021). Beyond machine translation, BPE has shown to be a competitive strategy in language modeling (Mielke and Eisner 2019). State-of-the-art pre-trained language models like GPT use subword tokenization based on BPE (Radford et al. 2019).

It has been hypothesized that BPE's success in NLP is mainly due to its increased compression capability compared to similar algorithms (Gallé 2019). In this context, compression capability is understood as: Given two subword vocabularies of the same size (obtained with two different algorithms), which one is able to cover a text sequence with fewer symbols? BPE subwords outcompete other subword approaches in this sense.

2.1.2 BPE compression in the first merges. Gutierrez-Vasques et al. (2021) investigate the information-theoretic properties of varied BPE subword tokenizations. Particularly, they measure the entropy and redundancy of a text over several subword frequency distributions obtained through incremental BPE merges.

Figure 2 gives an example of the redundancy curves for three typologically different languages, using the Parallel Bible Corpus (PBC). Notice that the first BPE operations cause the most drastic changes, i.e., the redundancy of the texts drops sharply (and the entropy grows)[4]. The patterns captured on these merges are the most useful for compressing the text. After a relatively small number of operations (around 200 on average for this particular corpus) the changes become less pronounced, and redundancy reaches a minimum. After this minimum, redundancy starts to grow again slowly through merges.

Languages seem to share the universal property of having the greatest compression potential during the first couple hundred BPE merges. Since the most redundant patterns are always found at the beginning of the compression process, one might ask if this generalization is due to the frequency distribution of symbols in natural languages or a consequence of the algorithm itself. The fact that some elements have much higher probability than others makes the data compressible, e.g., the Lempel-Ziv (LZ) techniques (Ziv and Lempel 1977;

4 When highly recurrent patterns get merged, the redundancy of the texts is reduced. We can think of this in terms of skewed frequency distributions. At merge 0, the frequency distribution of subwords is more skewed since the subwords inventory is composed only of characters, some of them with high frequencies. This implies higher redundancy and less uncertainty. When BPE starts merging the most salient patterns, subwords' frequency distribution gets closer to a uniform distribution (more symbols, lower frequencies). The redundancy in this type of frequency distribution is low, while the entropy is high.

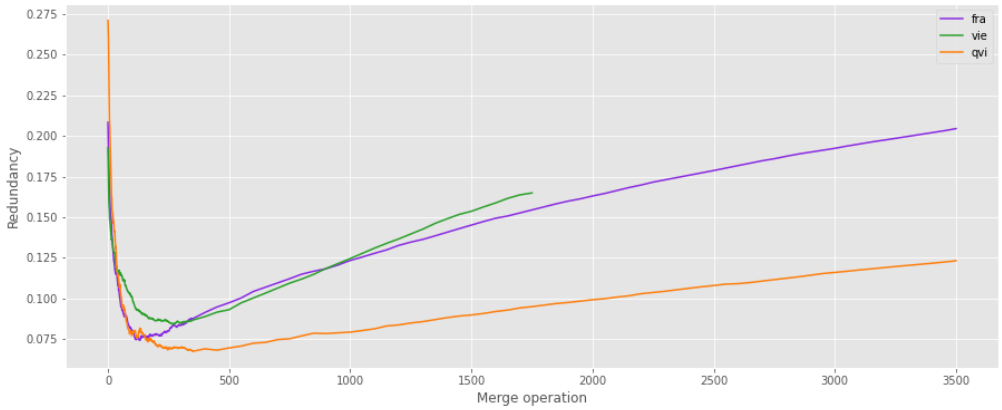


Figure 2: Redundancy (R) across BPE merges for French (fra), Vietnamese (vie) and Quechua (Imbabura) (qvi). Parallel bible data adopted from Gutierrez-Vasques et al. (2021).

Reynar et al. 1999) leverage the Zipf distribution of data to compress it. An extreme opposite would be a uniform distribution where everything has the same probability, and there are no redundancies to compress. BPE certainly exploits the fact that linguistic symbols in text follow a power law (Zipfian) distribution, which causes a rapid drop of redundancy (measured over the subword tokenizations obtained at each operation). It is worth remembering at this point that the Zipfian laws are formulated for stable units (characters, morphemes, words), while the redundancy curves concern subword vocabularies, which evolve with the number of merges.[5]

Inspecting the subwords merged up to the point of minimum redundancy more closely reveals that the types of patterns allowing compression are not the same across languages. Our work departs from this observation.

2.1.3 WordPiece tokenization as an alternative to BPE. Among alternative subword tokenization algorithms we here also consider WordPiece. It was originally designed for dealing with Korean and Japanese voice search (Schuster and Nakajima 2012). More recently, WordPiece subword tokenization can be found in popular pre-trained models like BERT (Devlin et al. 2019). WordPiece and BPE display some commonalities. For instance, both have initial vocabularies comprising characters, and both iteratively merge adjacent symbols, forming subwords. The difference is that, in the case of WordPiece, the merging criterion is not the most frequent symbol pair. Instead, WordPiece chooses the pair that maximizes the likelihood of the data upon merging (given an n-gram language model trained on the data). Once a specific number of merge operations have been applied, WordPiece takes the resulting subword vocabulary and follows a left-to-right longest-match-first strategy for tokenizing each word in a text.

5 To what extent Zipfian laws hold for these subword vocabularies is potentially a research question in itself, which remains outside of the scope of the current study.

2.2 Productive vs idiosyncratic patterns in linguistics

To analyze subwords created by algorithms like BPE and WordPiece in more depth, we turn to concepts from quantitative linguistics. The concept of *productivity*, for instance, is most often discussed in relation to word formation processes, although it is sometimes also mentioned in the study of phonology and syntax. Regular morphological patterns are productive by virtue of applying to many different lexemes. For example, an inflectional pattern like *-ed* in English, marking the past tense, applies to many verbs. We can also think of derivational suffixes like *-ly*, which combine with a wide variety of lexemes. The more productive a morphological pattern, the more likely we will apply it also to new lexemes, e.g., the speaker will tend to choose *-ed* to construct the past tense of an unseen verb or a borrowing (Bybee 2010), though particularly salient irregular forms might also be used (Cuskley et al. 2015). In contrast, there are also patterns of a more *idiosyncratic* nature, i.e., those combining with few lexemes. An example in Modern English is the *-en* pattern marking plural, which combines only with a very limited number of nouns, e.g. *ox-en*. In combination with its stem, it behaves like a fossilized unit: a speaker will hardly apply it to new stems. At the extreme end of idiosyncrasy we find irregular forms like *was*, *had*, etc. which occur as orthographic words by themselves.

To study morphological productivity more systematically, previous accounts have first laid out the difference between *type frequency* and *token frequency*. According to this terminology, the former refers to the number of word types containing a particular morphological pattern, while the latter is the cumulative frequency of occurrences of those words (Berg 2014). Taking this into account, several authors have argued that productivity should be measured as type frequency. If a pattern occurs in many words (high type frequency), then its productivity is said to be high (Bybee 2010, 2003).

Productivity in this sense can be captured using quantitative methods. Baayen (1992, 1993) proposes a measure based on counting the number of word types containing a particular affix. This is done in an incremental fashion, that is, calculating the counts in text chunks over a large corpus. Iterating through a corpus of 18 million word tokens of English, it is shown that productive affixes will keep appearing in the text samples, while less productive ones will reach an asymptote quickly, indicating the probability of observing new formations with the respective affix is close to zero.

Another common concept that type frequency and productivity are strongly linked with is *regularity*: "One regularizes to patterns used by many [different] lexemes, not to patterns used by frequent lexemes" (Bonami and Beniamine 2016). High token frequency, on the other hand, is related to *irregularity* (Greenberg 1966; Pinker 1991; Ullman 1999; Bybee 2010; Wu, Cotterell, and O'Donnell 2019). If a pattern occurs only in a few words (low type frequency), its token frequency can still be very high (if a frequent word contains that pattern). In fact, according to Bybee (2003), high token frequency encourages the autonomy of linguistic units. Blevins, Milin, and Ramscar (2017) argue that regular and irregular patterns coexist in languages since there is a trade-off between opposing communicative pressures: Irregular patterns enhance discriminability. For example, the relatively stark contrast between *go* and *went* facilitates the mapping to the grammatical functions of *present* and *past* tense respectively, while for *go* and *goes* this is harder, since the contrast between the forms is rather minor. Regular patterns, on the other hand, increase the predictability of unseen forms. If we have seen *go/goes*, *speak/speaks*, *see/sees* it will be easy to predict what the third person singular form of *play* is.

To sum up, productive patterns are associated with regular forms (which do not need to be frequent), and idiosyncratic patterns with irregular forms (which tend to be frequent). Productive processes are present in inflectional (affixes that encode grammatical or morphosyntactic distinctions) and derivational morphology (affixes that encode lexicosemantic distinctions that

can change the category of a word). However, there is some evidence that inflection tends to be more productive than derivation in natural languages (Stump 2017).

2.3 Typological knowledge in NLP

Structural differences across diverse languages, such as in regular and irregular morphological patterns, have long been studied in linguistic typology. In NLP, the relevance of typology has arisen from the need to compare languages in multilingual applications. One way to assess the similarity between languages is to leverage the features stored in typological databases. A representative example is the library lang2vec (Littell et al. 2017), which provides language vectors derived from the typological database URIEL. The integration of this type of typological information has proven to be useful in several NLP domains, e.g., selecting transfer languages for improving cross-lingual tasks (Lin et al. 2019; Lauscher et al. 2020), measuring the language diversity of NLP multilingual models (Ruder et al. 2021), adapting languages for Universal Dependency parsing (Üstün et al. 2020), and for investigating the language properties that are encoded in multilingual sentence encoders (Choenni and Shutova 2022).

One of the limitations of relying mainly on linguistic databases is that the information is often incomplete: some languages are fully described, while only a few feature values are known for others. Ponti et al. (2019) note in their comprehensive survey that the information extracted from typological databases has achieved consistent but modest improvements in NLP systems. They advocate for newer approaches that can go beyond the broad and discrete nature of current typological categories and adapt to the continuous nature of contemporary NLP algorithms.

There are several examples of these efforts to create unsupervised approaches to linguistic typology. This includes the prediction of missing typological features not available for many languages (Malaviya, Neubig, and Littell 2017; Bjerva and Augenstein 2018; Bjerva et al. 2020), or the usage of raw data to infer similarities between languages (Bjerva et al. 2019).

In terms of morphology, Bjerva and Augenstein (2018) retrieved continuous language vector embeddings from training for a NLP morphological task and showed that these embeddings are able to encode morphological features found in the World Atlas of Language Structures (WALS). Oncevay et al. (2022) quantify the degree of synthesis and fusion for English, German, Turkish, and Spanish based on different methods of morphological segmentation. Along similar lines, Rathi, Hahn, and Futrell (2021) propose an information-theoretic characterization of the degree of morphological fusion applied to twenty-one languages.

Our work builds upon this strand of new approaches, anticipated by Ponti et al. (2019), to data-driven induction of typological knowledge. We propose a computational light approximation based on BPE, which does not require large training data, manual annotation, or any external linguistic knowledge. Therefore it is easily applicable to a wide range of languages for which some textual material exists.

2.3.1 Main notions in language typology. Language typology aims to uncover patterns of variation, and to identify the different language types that exist in the world, independently of their genealogy (Haspelmath 2008). In fact, one of the first typological approaches to classify languages by Sapir (1921) is based on characterizing them through the lens of morphology, namely, by using two dimensions relating to word formation:

1. Degree of *fusion* of morphemes:

isolating	*agglutinative*	*fusional*
Mandarin Chinese	Turkish	Classical Latin
我來了 (**wǒ** lái **le**)	gel-**di**-**m**	ven-**i**
1P come **PF**	come-**PF**-**1P**	come-**1P.PF**
"I came."		

2. Degree of *synthesis* of words:

analytic	*synthetic*	*polysynthetic*
English	Turkish	Chinook (Wishram)
I came to give it to her	on-a vermey-e gel-di-m	i-n-i-a-l-u-d-am

The degree of *fusion* according to Sapir's typology[6] indicates to what extent morphemes with different grammatical functions are "fused" together. In the Mandarin Chinese example, three separate morphemes give the information about person, the type of action (i.e. the verb), and tense.[7] In Turkish, these are "glued" together in the sense that there is a stricter order in which certain suffixes occur. Also, markers sometimes depend on one another in terms of phonological processes such as vowel harmony. In Classical Latin, first person and perfect tense information is genuinely "fused" together in one suffix *-i*. Note that the difference between the glossings for Turkish and Classical Latin (i.e. PF-1P versus 1P.PF) reflects this difference in fusion: in the former case, two separate morphemes are identifiable, while in the latter there is just one morpheme mapped onto the different grammatical functions. As a consequence, the morpheme to grammatical function ratio in Classical Latin would here be $\frac{2}{3}$, while in Mandarin Chinese and Turkish it would be $\frac{3}{3}$.

Roughly speaking, *fusion* is relevant at the level of morphemes combining to words, while *synthesis* is relevant at the level of words combining to sentences. Of course, this presupposes the distinction between morphemes and words, which is a thorny issue in itself (Haspelmath 2017). Having said this, Sapir's original idea about analytic languages is that "the sentence is of prime importance, the word is of minor interest", while for synthetic languages "the concepts cluster more thickly, the words are more richly chambered", and in polysynthetic languages "the elaboration of the word is extreme [...] including the syntactic relations [...]" Sapir (1921, p. 110). Note that in the English example, each morpheme is considered an orthographic word by itself, such that the ratio of orthographic words to morphemes is $\frac{7}{7} = 1$. In the Chinook example,[8] roughly the same content is literally *compressed* into a single orthographic word, i.e. a string of characters, such that each character carries grammatical information by itself.

6 He actually defines a fourth category called *symbolic* which is disregarded here.

7 1P: first person; PF: perfect tense.

8 This is taken directly from Sapir (1921, p. 57-58). Sapir calls this version of the Chinook language "Wishram dialect". Glosses for individual morphemes can be derived from Sapir's description. The full glossing is:

(1) i-n-i-a-l-u-d-am
PAST-1P-3P.N-3P.F-IOBJ-ABL(?)-give-CAUS(?)

"I came to give it to her."

The ratio of orthographic words to morphemes is $\frac{1}{8} = 0.125$. The Turkish translation[9] ranges somewhere in between, with three orthographic words over seven morphemes (i.e. $\frac{3}{7} = 0.43$).

The three-way distinctions between *isolating* → *agglutinative* → *fusional*, as well as *analytic* → *synthetic* → *polysynthetic* are still in use today. However, Sapir (1921) himself already acknowledged that languages cannot be classified strictly into these fixed categories. Following up on this, Greenberg (1960, p. 182) put forward a gradual, quantitative account, rather assessing overall tendencies instead of assigning a language to a single category: "A language may well and indeed usually does contain some agglutinational as well as some non agglutinational construction." In his approach, languages are characterized by several indices reflecting morphological features (see Table 1), which can take a range of values. For instance, the spectrum for the feature *synthesis* (measured as the ratio of morphemes per word in a random text sample of the language) goes from 1.06 o 3.72 – given Greenberg's sample of languages. On this spectrum, Vietnamese (Annamite) is located at the lower end, while "Eskimo"[10] is located at the higher end.

Table 1: Quantitative typological features proposed by Greenberg (1960, p. 193)

Typological index	**Sanskrit**	**Anglo-Saxon**	**Persian**	**English**	**Yakut**	**Swahili**	**Annamite**	**Eskimo**
Synthesis	2.59	2.12	1.52	1.68	2.17	2.55	1.06	3.72
Agglutination	0.09	0.11	0.34	0.3	0.51	0.67	...	0.03
Compounding	1.13	1	1.03	1	1.02	1	1.07	1
Derivation	0.62	0.2	0.1	0.15	0.35	0.07	0	1.25
Gross inflection	0.84	0.9	0.39	0.53	0.82	0.8	0	1.75
Prefixing	0.16	0.06	0.01	0.04	0	1.16	0	0
Suffixing	1.18	1.03	0.49	0.64	1.15	0.41	0	2.72
Isolation	0.16	0.15	0.52	0.75	0.29	0.4	1	0.02
Pure inflection	0.46	0.47	0.29	0.14	0.59	0.19	0	0.46
Concord	0.38	0.38	0.19	0.11	0.12	0.41	0	0.38

While the work by Sapir (1921) and Greenberg (1960) is still relevant today, morphological typology has developed further into the 21st century. We want to briefly sketch a more recent proposal by Bickel and Nichols (2007). They elaborate on the classic "fusional" cline *isolating* → *agglutinative* → *fusional* by rather proposing a threefold distinction based on markers of inflectional information, which they term *formatives*. These are typically bound affixes (e.g. *-ed* in *walk-ed*), though depending on the language, these could also be realized as free particles or clitics (e.g. the English genitive clitic *'s* in the noun phrase *[the guy I saw yesterday]'s dog*). Given formatives as the basic elements of inflection, there are three clines laid out by Bickel and Nichols (2007):[11]

1. *Fusion*. Degree of phonological merging of formatives with their hosts.

9 The full glossing of this Turkish example using Göksel and Kerslake (2004) is:

(1) on-a vermey-e gel-di-m
he/she/it-DAT give-DAT come-PF-1P
"I came to give it to her."

10 Greenberg uses this rather vague (and nowadays sometimes interpreted as derogatory) language name to refer to one of the Inuit, Yupik, or Aleut languages.

11 3P: third person; ABL: ablative case; F: feminine; PF: perfect tense; PL: plural; PST: past tense; SG: singular.

isolating	*concatenative*	*nonlinear*
Mandarin Chinese	Turkish	Standard Arabic
她写了	yazdı	كتبت
tā xiě **le**	yaz-**dı**	k**ata**b-at
3P.F write **PF**	write-**PF**	write.**PST**-3P.F.SG
"She wrote."	(Göksel and Kerslake 2004, p. 285)	(Ryding 2005, p. 438)

2. *Flexivity*. Degree of allomorphy of formatives.

nonflexive	*flexive*
Quechua (Yauyos)	German
warmi-**kuna** (woman-PL)	Frau-**en**
wasi-**kuna** (house-PL)	Häus-**er**
karru-**kuna** (car-PL)	Auto-**s**
(Shimelman 2017, p. 70)	

3. *Exponence*. Degree to which different categories are expressed by the same formative.

cumulative	*separative*
Classical Latin	Turkish
(de) dom-**ibus**	ev-**ler-den**
house-**PL.ABL**	house-**PL-ABL**
"from the houses"	(Göksel and Kerslake 2004, p. 68)

The reason Bickel and Nichols (2007) distinguish between these three clines instead of the single classical cline *isolating* → *agglutinative* → *fusional* is that concepts such as *fusion*, *flexivity*, and *exponence* are in principle orthogonal to one another. For instance, while traditionally agglutinative (i.e. concatenative) patterns are associated with nonflexive markers, they can just as well be flexive. The Quechua plural marker *-kuna* is concatenative/nonflexive, i.e. it does not change according to the noun it modifies, while the German plural formatives are mostly concatenative/flexive, since they change according to the declension class of the noun.

Resonating with the early work by Sapir (1921), Bickel and Nichols (2007) too point out that all of the notions discussed above do not strictly apply to languages as a whole, but rather to inflectional domains (e.g. tense, number), or even just particular formatives. Languages are rarely (if ever) entirely isolating, concatenative, or nonlinear. Take the example of Standard Arabic above. Tense is here marked by vowel changes inside the consonant template (*k-t-b*), such that the past tense stem is *katab-*, while the present tense stem would be *-ktub-* (Ryding 2005, p. 439). Thus, tense is here marked in a nonlinear, insegmentable way. However, note that the *-at* pattern at the end of the word is a genuine suffix marking for person, gender, and number. Hence, in Standard Arabic, even within the same word forms we find both nonlinear and concatenative formatives of inflectional information. Likewise, nonlinear formatives are often seen as an unusual feature, typically associated with Semitic languages like Arabic or Hebrew, but we find a considerable number of irregular verbs in English following similar patterns. A case in point is the English verb *write/wrote* in the translation of the Standard Arabic example.

Coming back to the principle of BPE compression and the productivity of subwords, the typological categories above are not equally relevant. While *synthesis* and *flexivity* are certainly relevant, *fusion* is only partly relevant, and *exponence* seems rather irrelevant. Note that synthesis is *per definition* related to the productivity of subwords since it captures the degree to which inflectional formatives are recurring in different word types. For example, in Turkish, the subword *-dim*, marking first person and perfective aspect/past tense, will repeat across many different verb types (*gel-dim* 'I came', *ye-dim* 'I ate', *ver-dim* 'I gave', etc.), while the respective

English pronoun 'I' stands by itself as an orthographic word, and is not very productive as a subword (though frequent). For similar reasons, flexivity is highly relevant for productivity too. Namely, nonflexive formatives like the Quechua plural marker *-kuna* are very productive, while flexive formatives like German *-er*, *-en*, and *-s* are less productive by virtue of their restriction to particular declension classes. Similar arguments also apply to the isolating versus concatenative distinction on the fusion cline, but for the third category, nonlinear formatives, the picture is more complicated. If a language displays consonant patterns like the *ktb* template in the case of Standard Arabic, then these are potentially very productive across different word types. However, this also depends on whether the respective vowels are explicitly coded in writing. If they are, then the nonlinear insertion of vowel formatives, e.g. in *katab-* and *-ktub-*, will "break" the productivity of the consonant template. Finally, exponence is not obviously linked to productivity, since the question of whether formatives are cumulative or separative is related to paradigmatic considerations, but subword patterns are "blind" to those. Note that the Classical Latin cumulative formative *-ibus* (ABL.PL) might be just as productive as the concatenation of separative formatives *-ler-den* (ABL-PL) in Turkish.

In summary, languages high on the synthesis scale, with concatenative and nonflexive formatives, are expected to display high morphological productivity. On the other hand, languages low on the synthesis scale, with isolating and flexive formatives, are expected to display low productivity. We set out to empirically test these expectations in our analyses.

Importantly, all the typological approaches sketched above depend on access to linguistically annotated data, e.g., grammars, morphological paradigms, dictionaries, etc. However, such data is, firstly, not always readily available for many languages, and, secondly, relying on various conventions which are not easily implementable and reproducible. Given this state of affairs, unsupervised approaches like BPE can provide a proxy to quantify and test typological hypotheses cross-linguistically with reproducible methods.

3. Data and Methods

We propose to analyze the subwords found by BPE with regards to how *productive*, or inversely, how *idiosyncratic* they are. In this work, subwords are the patterns that result from the BPE merging criteria through incremental operations. We initially perform 200 BPE merge operations on text samples from 47 languages and estimate the degree of productivity, idiosyncrasy, and cumulative frequency for each obtained subword. We then aggregate the values per language to obtain a vector representation. This allows us to cluster languages along these dimensions, and compare the resulting clustering with traditional typological classifications that rely on grammars and general knowledge about languages.

We focus on a few hundred BPE subwords since we conjecture these are enough for discriminating languages in terms of their structural properties. However, our methodology includes experimenting with different BPE merge operations and an alternative subword tokenization technique. In the remainder of this section, we describe each step in more detail.

3.1 Corpora

We use parallel corpora to facilitate meaningful comparisons between languages. Parallel texts are typically used in cross-linguistic studies on morphological typology, lexical typology, and word order typology (Greenberg 1960; Cysouw and Wälchli 2007; Wälchli and Cysouw 2012; Östling 2015; Kelih 2010; Mayer et al. 2014).

Our main data set is a selection of 47 diverse languages from the Parallel Bible Corpus (PBC) (Mayer and Cysouw 2014). This selection is a subset of the WALS 100 language

sample,[12] specifically designed to represent languages from diverse families and areas.[13] Since our focus is linguistic typology, our selection of languages is not based on simply selecting the ones for which text data is readily available online. We put more weight on representing a wide range of language families, areas, and structural features.

In particular, the corpus we use includes 1150 verses that overlap over the 47 languages. The complete list of languages and their respective ISO-639-3 codes are included in Appendix A. We want to mention that even though this corpus is verse aligned, and hence fully parallel, there are of course differences in the exact wordings that translators have chosen across different languages. As an example, consider the following verse in its Korean and English translation.[14]

Korean (*kor*)

(3) 저희가 예수의 말씀을 기억하고
jeo-hui=ga yei-su=ui mal-sseum=eul gi-eog=ha-go
1P-PL=SUBJ Jesus=POSS speak.HON=OBJ remember=COM
Literal translation: “And we remember Jesus’ speech.”
English verse: “And they remembered his words.”

While the Korean translation uses the first person plural pronoun (*jeo-hui* ‘we’), the English verse uses the third person plural pronoun (*they*). Also, the lack of tense specification in the Korean verse contrasts with explicit past tense marking (*remember-ed*) in English. Finally, the Korean verse gives the proper noun with possessive marking (*ye-su=ui*), while in English we encounter the anaphora *his*.

However, this is a rather extreme example of divergence, as other translations (e.g. Georgian[15]) are closer to the English one (see Example 4), despite the fact that Georgian is a language typologically very different from English. More generally, parallel verses as in the PBC are certainly much closer in content than arbitrary text chunks of different registers and styles. Hence, they are more stable testing ground for quantitative language comparison.

Georgian (*kat*)

(4) და მოეკსენნეს სიტყუანი მისნი
da mo-e-qsenn-es sit’q’va-n-i misni
and PREV-3P.PL-mention-3P.PL.AOR word-PL-NOM 3P.POSS.NOM.SG
Literal translation: “And they mentioned his words.”
English verse: “And they remembered his words.”

To ensure that our observations are not heavily dependent on the peculiarities of this specific Bible corpus, we repeat the measurements on two additional parallel corpora that differ in register and style, and vary in size. This includes the JW300 corpus, i.e. a compilation

12 https://wals.info/languoid/samples/100
13 http://www.christianbentz.de/MLC2019_data.html
14 The verse ID is 42024008. See also Table 3. The transliteration and glossing are here based on the Korean grammar by Yeon and Brown (2011). Note that here hyphens indicate syllable boundaries (corresponding to the syllable blocks in Hangul writing) , while the equal sign indicates morpheme boundaries. 1P: first person; COM: comitative particle (here translated as “and”); HON: honorific marker; OBJ: syntactic object marker; PL: plural; POSS: possessive marker; SUBJ: marker of syntactic subject. Misinterpretations and errors remain our own.
15 This example is transliterated and glossed according to the Georgian grammar by Hewitt (1995). Misinterpretations and errors remain our own. Note that some forms in this Georgian text, for instance, the plural marker *-ნი* ‘-ni’, suggest that it is written in an archaic style inspired by Old Georgian (Hewitt 1995, p. 38). PREV: preverb; 3P: third person; PL: plural; AOR: aorist; NOM: nominative; POSS: possessive; SG: singular.

of magazine articles (from the Jehovah's Witnesses website) for around 300 languages (Agić and Vulić 2019). In this case, we extracted a parallel corpus for 25 languages, namely, the ones sharing at least 68 parallel magazine articles, and which overlap with the PBC sample. Furthermore, we include the Universal Declaration of Human Rights (UDHR),[16] a parallel corpus with very short texts (only a couple thousand word tokens per language), but a wide variety of languages. In this case, we found 31 languages overlapping with the PBC sample.

Table 2 shows the parallel corpora size for the selection of languages included in the current analyses.

Table 2: Parallel corpora information.

Corpus	Languages	Total Tokens	Avg. tokens per language
PBC	47	1.1 M	25.1 K
JW300	25	4.7 M	188.9 K
UDHR	31	56.1 K	1.8 K

3.2 Scripts and writing systems

In the PBC sample of 47 languages, overall 9 different scripts are represented (see Table 3). While the majority of texts (42/47 or 89%) is written with alphabetic scripts, we also encounter so-called abugidas and abjats (see, for instance, Daniels and Bright 1996, p. 8 for a discussion of these terms). Given these different types of scripts, there are two levels of segmentation which are relevant to BPE compression: the level of *UTF-8 characters*, and the level of *orthographic words*.

Table 3: Different scripts (with respective ISO identification code), writing systems (Writ. Sys.), and number of languages using these (No.) in the PBC sample of 47 languages.

ISO 15924	Script	Writ. Sys.	No.	Example*
Latn	Latin	Alphabet	38	And they remembered his words ,
Arab	Arabic	Abjat	2	فَتَذَكَّرْنَ كَلَامَهُ ،
Grek	Greek	Alphabet	1	Και ενεθυμηθησαν τους λογους αυτου .
Deva	Devanagari	Abugida	1	तब उस की बातें उन को स्मरण आईं ।
Geor	Georgian	Alphabet	1	და მოეკსენნეს სიტყუანი მისნი .
Hang	Hangul	Alphabet	1	저희가 예수의 말씀을 기억하고
Mymr	Burmese	Abugida	1	မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ ၊ -
Cyrl	Cyrillic	Alphabet	1	И они вспомнили эти слова Его .
Thai	Thai	Abugida	1	พวก นาง จึง ระลึก ถึง พระ ดำรัส ของ พระองค์

*Verse number 42024008 of the New Testament.

BPE merges operate at the level of UTF-8 characters, and the particularities of script encodings matter here. For instance, the word *remembered* in English consists of 10 UTF-8 character tokens and 5 types ('r','e','m','b','d'). English texts typically contain 26 of such UTF-8 character types (bare punctuation). In comparison, Korean Hangul UTF-8 characters work in a different way. The Korean word corresponding to *remembered* in Example (3) is 기억하고 *gi-*

16 www.unicode.org/udhr

eog-ha-go which could be translated as 'and remember'.[17] The Korean orthographic word hence consists of four syllable blocks. It is these syllable blocks – rather than individual consonant and vowel characters of the Korean alphabet (called *jamo*)[18] – which are represented as UTF-8 characters in our texts,[19] and hence merged by BPE. For example, in the Korean PBC text, the most frequently co-occurring – and hence first merged – syllable blocks are 예수 *yei-su*, while for English this is *th*.

Due to the syllabified nature of Korean Hangul, the respective texts can display hundreds and thousands of UTF-8 characters. A similar proliferation is found in texts written with further abugidas (e.g. Hindi, Thai, Burmese). We here expect somewhat lower productivity of subwords, since a wider range of UTF-8 characters (in the case of abugidas representing syllables rather than individual phonemes) means more combinatorial possibilities which re-occur with lower probability. In latinized scripts with many special characters and diacritics (e.g. Vietnamese) a similar effect is expected.

Another structural property relevant at the level of UTF-8 characters is the presence or absence of diacritics to indicate vowels. This is the case in abjats (and also to some extent in abugidas). For example, in Modern Standard Arabic writing, short vowels are left unwritten in some registers, and indicated by diacritics in others (Ryding 2005, p. 25). Compare the word for *peace* in the Egyptian Arabic (arz) PBC and the UDHR of Example (5).

Egyptian Arabic (*arz*)

(5) PBC UDHR
سَلَام سلام
salam slm
'peace'

In the PBC version, the short vowels are explicitly coded as so-called *fatha* diacritics above the consonant, such that سَ represents the syllable *sa*, while in the UDHR version, only the consonantal template *s-l-m* is given. From the perspective of BPE this difference is crucial, since the diacritic can be coded as a separate UTF-8 character, and hence merged with the consonant when frequently co-occurring, e.g. *sa*, whereas in the texts without diacritics, the merging would take place only between consonants, e.g. *sl*.

The second level of segmentation relevant to BPE is the orthographic word. Word boundaries are adhered to by the algorithm in the sense that merges are not allowed across them. For example, for the English character string *the year*, word boundaries (white spaces in writing) would be explicitly coded as in: <w>the</w><w>year</w>. Despite the fact that *the+y* is a very frequent co-occurrence pattern, it would not be allowed to be merged across the word boundaries in this case.

The reliance on orthographic word boundaries is a problem for scripts where there are (almost) no boundaries (i.e. white spaces) at all, e.g. Burmese (mya) and Thai (tha). We applied the Python library Polyglot (Al-Rfou 2015) for the texts that were not originally tokenized at the orthographic word level. An example with the original text, and the tokenized version is given in (6).[20]

17 The dictionary form is 기억하다 *gi-eog-ha-da*, but the -(하) 고 ending is here a particle indicating the commitative, which is typically translated as 'and' in English (Yeon and Brown 2011, p. 118).
18 See a list of these *jamo* here: https://unicode-table.com/en/blocks/hangul-jamo/.
19 See a complete list of these overall 11184 syllable blocks in UTF-8 here: https://unicode-table.com/en/blocks/hangul-syllables/.
20 This example is transliterated, glossed and translated with the help of the Burmese grammars by Jenny and Hnin Tun (2016) and Lonsdale (1899). However, misinterpretations and errors remain our own. HON is a

Burmese (*mya*)

(6) [...] မိန့်တော်မူခဲ့သောစကားများကိုပြန်သတိရ ၍ ၊ –
[...] မိန့် တော် မူ ခဲ့ သော စ ကား များ ကို ပြန် သ တိ ရ ၍

[...] mín-tɔ-mu-gé θɔ sá.kà-myà-ko pyan θá.ti.rá ywé
[...] speak-HON-perform-DISPL REL word-PL-OBJ return remember and

"[...] and (they) remember the words that (he) spoke."

Note that the original Burmese text is "undersegmented", namely, there is only one white space between the conjunction ၍ *ywé* and the rest of the sentence. However, the tokenized version is somewhat "oversegmented". For instance, the verb သတိရ *θá.ti.rá* 'remember' is now split into three separate syllable symbols. In the undersegmented version of the text, we would expect high productivity of subwords, while in the tokenized version we use here, we rather expect lower productivity of subwords.

The problems and pitfalls of UTF-8 encodings (Moran and Cysouw 2018) and orthographic word boundaries should be kept in mind when using BPE, and compression algorithms more generally. We will mention particular problems of our application in more detail in the discussion section.

3.3 Measuring productivity and idiosyncrasy

Given that we have a handle on the different scripts in a particular BPE implementation, we can start to measure the productivity of subwords generated by it. As we have discussed in Section 2.2, we think of a *productive* subword as one that is found in many different orthographic word types. In contrast, some subwords appear in rather few different word types. These subwords can still be very frequent, e.g., when they occur frequently by themselves. We will call these *idiosyncratic*.[21] We propose a straightforward operationalization of the degree of productivity and idiosyncrasy of a BPE subword. For each merge operation, we calculate:

$$\text{productivity}(s) = |W_s|\,, \tag{1}$$

$$\text{c.freq}(s) = \sum_{w \epsilon W_s} \text{freq}(w), \tag{2}$$

$$\text{idiosyncrasy}(s) = \frac{\text{c.freq}(s)}{\text{productivity}(s)}. \tag{3}$$

honorific affix, which always coincides with the verb *mu* "do/perform" (Lonsdale 1899, p. 194). DISPL is a displacement marker (Jenny and Hnin Tun 2016, p. 219), which might be translated as past tense here. REL is a relative marker (Jenny and Hnin Tun 2016, p. 258) which marks the preceding verbal clause as an attribute (of a noun), i.e. "words that (he) spoke".

21 Roughly speaking, the linguistic notions of productivity and idiosyncrasy could be thought of in terms of the TF-IDF measure: productivity is in some way similar to inverse document frequency, idiosyncrasy to term frequency.

Table 4: Example of subwords obtained through BPE merge operations based on the PBC corpus (English), as well as their productivity ($|W|$), cumulative frequency (c.freq), and idiosyncrasy.

Subword	|W|	c.freq.	idiosyncrasy
ed</w>	271	917	3.38
had</w>	1	104	104
and</w>	11	2197	199.72

Where s is a given subword, $|W_s|$ is the number of orthographic word types that contain the subword[22], freq(w) is the function which assigns the raw frequency to a word, and $\sum_{w \epsilon W_s}$ freq(w) is the cumulative frequency over all the word types a given subword is part of. Thus, the productivity measure in Equation (1) is simply the number of word types[23] that contain the sequence BPE chose to merge at the current operation.

To capture idiosyncrasy, we need to incorporate the cumulative frequency of the word types in which a subword appears. Therefore, the idiosyncrasy measure in Equation (3) takes the cumulative frequency of these word types and divides it by the respective productivity $|W_s|$. Thus, subwords that appear in few word types, but have a high cumulative frequency (number of tokens summed over the types), will have high values of idiosyncrasy, whereas subwords that appear in many different word types will tend to have lower idiosyncrasy values.

Regarding the range of values of these measures, the minimum value we can get from productivity(s) is 1 (when the current subword s is contained in only one word type), while the maximum is the size of the word vocabulary (when the current subword s is contained in all the word types). Similarly, for c.freq(s) the minimum value is 1 (when only one word type contains the current subword s and its frequency is 1), and the maximum value is the total number of word tokens in the corpus (the subword s is contained in all word types, thus their cumulative frequency is the total number of word tokens). For idiosyncrasy(s), the lowest value is 1, this happens when a subword s is distributed in a given number of word types, and each of these word types occurs only once in the corpus (same value of productivity(s) and c.freq(s)), while the highest values are reached when the subword is contained only in one word type (the productivity(s) is 1) but its frequency is very high (c.freq(s)).

We can see an example in Table 4 and Figure 3. A subword like *ed</w>*[24] in English is contained in many word types (high productivity $|W|$), with relatively high cumulative frequency (c.freq) but low idiosyncrasy. In contrast, the subword *had</w>* is contained in fewer word types (actually just one, the word itself), but the frequency of this word type is high. Another example of a highly idiosyncratic subword is *and</w>*. We can see that even though it is distributed in several word types, most of its occurrences are concentrated in just one single type that is highly frequent (see Figure 3).

We represent each subword as a three-dimensional vector characterizing its productivity, cumulative frequency, and idiosyncrasy, and we visualize these vectors in a 3D-space (see Figure 4). We decided to keep c.freq as one of these three dimensions, since it provides

22 The set of orthographic words that contain a given subword depends on the merge operation. For instance, in English, the subword *-re-* is apparently contained in word types like: *ordered*, *answered*. However, these types are not included in the counts since in BPE *-ed* gets merged first, i.e., the algorithm does not consider that *-re-* occurs in those word types, since a new "symbol", *-ed*, has been introduced earlier: *a-n-s-w-e-r-ed, o-r-d-e-r-ed.*

23 In the rare cases where a subword appears more than once in the same word type, we do not increase the counts.

24 *</w>* indicates that the subword is located at the end of a word.

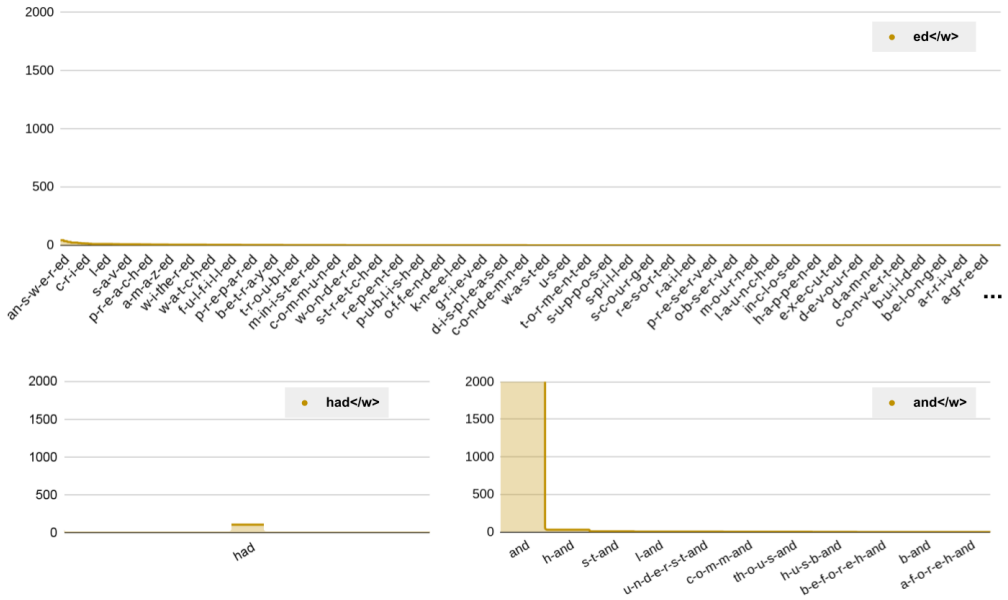


Figure 3: Examples of subword frequency distributions. The x-axes give word types which contain the substring merged by BPE (PBC corpus).

important information for characterizing each subword, e.g., two subwords might have similar idiosyncrasy values but different c.freq.

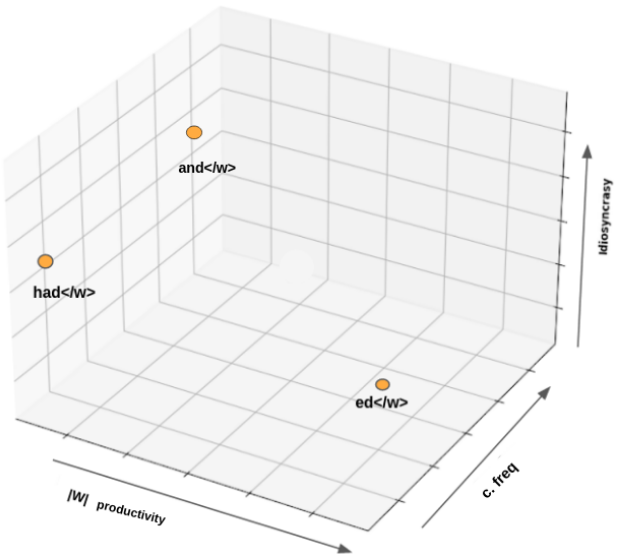


Figure 4: Example of subwords represented as points in a three-dimensional space.

3.4 Obtaining subwords

3.4.1 Subwords per merge. We apply an existing BPE implementation to generate subwords.[25] In the main analyses, we limit the number of merge operations to 200 (for all languages and corpora). This decision is based on previous research (Gutierrez-Vasques et al. 2021) which shows that early merges capture the subwords that achieve the most significant compression (reduction of redundancy, see also Section 2.1).

For each incremental BPE merge, we then visualize the resulting subword using the operationalizations of productivity, idiosyncrasy and cumulative frequency described in the previous section. We now have a vector representation that captures these properties for each pattern merged by BPE. In this way, we can appreciate the basic quantitative properties of subwords that aid BPE compression depending on the respective language. This will allow us to characterize languages later.

We hypothesize that just a few hundred of BPE merges are sufficient for capturing the most salient patterns that differentiate languages in terms of their morphological typology. However, as a sanity check, we explore different numbers of merge operations, as well as WordPiece as an alternative subword tokenization algorithm. We give further details in Section 3.6.

3.4.2 Averaging across subwords per language. For cross-linguistic comparisons, we provide a single three-dimensional vector representation for each language by averaging the values (productivity, idiosyncrasy, cumulative frequency) obtained during the first 200 merge operations. Each dimension captures the central tendency of the respective measure through merges. We center these vectors around zero, and scale them with respect to the standard deviation[26]. Each of the features is standardized independently by removing the mean (μ) and scaling to unit variance (σ):

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \tag{4}$$

These vector representations facilitate the comparison of languages in terms of their subword properties, i.e., how much more productive or idiosyncratic the BPE subwords of one language are compared to others.

3.5 Comparison to WALS features

The World Atlas of Language Structures (WALS) (Dryer and Haspelmath 2013) serves as an external criterion to assess how relevant the BPE induced space is to the morphological typology of languages. In a sense, WALS represents human expert judgments. The overall 144 Chapters (written by 55 authors) condense information about structural features (phonology, lexicon, morphology, syntax) for languages across the world. We harness this information to characterize the morphological profile of languages in our sample. Our starting point is the set of 28 WALS features which are relevant to describing morphological complexity (Bentz et al. 2016). Unfortunately, the coverage of WALS is incomplete, such that certain features are not equally available for all languages. We therefore include only a subset of 15 WALS features: those with feature codings for at least 46 languages of our sample (see Table 5).

25 https://github.com/rsennrich/subword-nmt

26 This does not modify the distribution; it just centers and scales the data points.

Table 5: Subset of WALS features that we use for characterizing the morphological typology of languages. The column "Languages" gives the number of languages in the PBC sample for which a given feature is available.

Feature	Name	Categories	Languages
20A	Fusion of Selected Inflectional Formatives	7 (non-ordinal)	47
22A	Inflectional Synthesis	7 (ordinal)	47
26A	Prefixing vs. Suffixing in Inflectional Morphology	6 (non-ordinal)	47
28A	Case Syncretism	4 (ordinal)	47
29A	Syncretism in Verbal Person/Number marking	3 (ordinal)	47
49A	Number of Cases	9 (ordinal)	46
59A	Possessive Classification	4 (ordinal)	47
65A	Perfective/Imperfective Aspect	binary	47
66A	The Past Tense	4 (ordinal)	47
67A	The Future Tense	binary	47
69A	Position of Tense/Aspect Affixes	5 (non-ordinal)	46
70A	The Morphological Imperative	5 (partially ordinal)	46
78A	Coding of Evidentiality	6 (non-ordinal)	47
102A	Verbal Person Marking	5 (partially ordinal)	47
112A	Negative Morphemes	6 (non-ordinal)	46

Given these features, we represent each language as a vector of 15 dimensions[27] by using the numeric values provided (see column "Categories" in Table 5). We apply centering and scaling to these. Our text-induced BPE space is not directly comparable with the WALS space since the dimensions are different. However, we assess whether languages close in WALS space are also close in our BPE space. To this end, we firstly perform a k-means clustering analysis[28] in the WALS feature space, and then check if the languages clustering together are also neighbors in our BPE space. The measure *mean silhouette coefficient* is used for the second step. The silhouette coefficient is a common intrinsic measure of cohesion and separation of clusters (Rousseeuw 1987). It returns a coefficient $s(i)$ in the range $[-1, 1]$ for each data point i in the data set. This is calculated using the mean *intra-cluster* Euclidean distance $a(i)$, i.e., the mean distance between the data point i and all other data points in the same cluster; and the *inter-cluster* distance $b(i)$, measured as the mean distance between the data point and the data points of the nearest cluster (excluding the cluster to which this data point belongs). We then have

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{5}$$

with

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \epsilon C_I, i \neq j} d(i, j), \tag{6}$$

27 In case a WALS feature coding is not available for a given language, we assign a zero in the corresponding vector dimension.

28 We use *kmeans++* for smarter centroid initialization and improved clustering quality. We use $k = 4$.

and

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \epsilon C_J} d(i, j). \quad (7)$$

C_I is the WALS-based cluster a data point i belongs to. C_J is any other cluster. Values of $s(i)$ closer to -1 indicate that the data point was assigned to the wrong cluster, i.e., it is relatively far away – in BPE space – from the other data points of the WALS-based cluster. A value close to 1, on the other hand, indicates that the data point is close to points in the assigned cluster.

Finally, we take the mean $\tilde{s}$ over all data points, that is, the *mean silhouette coefficient* of the entire data set – given a specific number of clusters. Although this is an intrinsic evaluation measure, we do not use it in a strictly intrinsic way, i.e., we measure $\tilde{s}$ of the clusters obtained with WALS but using the three-dimensional vector representations of the BPE space.

3.6 Hyperparameter setting

Our methodology as described so far is based on a fixed BPE hyperparameter, i.e., we use the first 200 subwords that emerge from the BPE compression (the number of subwords is the same as the number of merge operations applied). A natural question is if the language vector representations that we obtain are affected by the number of subwords taken into account. To assess this, we vary the BPE hyperparameter, i.e., the number of merge operations. Namely, we go from 1 to 1000 merge operations for all languages. For example, when the number of merge operations is 50, this implies measuring productivity, idiosyncrasy and cumulative frequency for the first 50 subwords. Then we average across these subwords to obtain a single vector representation per language. We then analyze how the distribution of the data points changes with different numbers of merges, and how this impacts our comparison to the WALS-based arrangement of languages. In particular, we calculate the mean silhouette coefficient for each configuration obtained at each merge.

3.7 Subword tokenization by WordPiece

In our analyses, the main focus is on BPE subwords. However, there are alternative subword tokenization algorithms. WordPiece, for instance, also relies on iterative merging. It is therefore also compatible with our approach, since we can explicitly extract the merged subwords at each step and inspect them along the dimensions of our measures. Importantly, the merging criterion in WordPiece is different from BPE. Instead of merging the two adjacent subwords with the highest co-occurrence frequency, the following criterion[29] is applied:

$$\text{score}(s_1, s_2) = \frac{\text{freq}(s_1, s_2)}{\text{freq}(s_1) \times \text{freq}(s_2)}, \quad (8)$$

where s_1, s_2 is a pair of adjacent subwords in the corpus. WordPiece merges the pair of subwords that maximize Equation (8). Note that this score includes the co-occurrence

29 Based on https://huggingface.co/course/chapter6/6?fw=pt. In this implementation, WordPiece distinguishes between the subwords that are the beginning of a word, and the rest of them. While in BPE, there is a distinction between the ones that are at the end of a word, and the rest of them.

frequency of two consecutive subwords, as in BPE (i.e. $\text{freq}(s_1, s_2)$), but it normalizes this frequency by the product of individual frequencies. In other words, BPE uses an absolute measure of co-occurrence frequencies as criterion, while WordPiece penalizes co-occurences of symbols which are highly frequent by themselves. For example, the first merged characters for English according to BPE are *th*, while for WordPiece this is *ex*. Another difference is the boundary markers that are used in each algorithm. The BPE implementation distinguishes between the subwords at the end of a word and the rest of them, while WordPiece distinguishes between the subwords at the beginning of a word and the rest of them[30]. See Appendix C for an extended example of the English subwords across BPE and WordPiece.

We compare the distributions obtained with these two subword tokenization methods, and assess which approach leads to vector representations more representative of the morphological typology of languages, according to the WALS-based clustering.

4. Results

4.1 Productivity and idiosyncrasy of BPE subwords

Figure 5 shows the positions of subwords in a three-dimensional space for four languages (English, Sango, Turkish, and Kalaallisut). They belong to different linguistic families, are typologically different, and they also represent opposing trends in terms of the quantitative properties of their subwords. Visualizations for all languages in the sample can be found on github.[31]

All languages share the universal property of achieving the most significant compression during the first BPE merge operations (green color in Figure 5). In other words, the character combinations which get merged during the first operations reflect highly recurrent subwords that cause the most drastic decrease in the redundancy of the texts (remember also Figure 2). At later merges (red and finally blue color) the subwords start accumulating near the origin. These subwords appear in fewer word types, and those word types are not very frequent anymore. Thus, they have low values in all three dimensions and are not among the top candidates for compression, so their "compression potential" goes to zero.

In general, the redundant patterns which get merged first, and contribute most to compression, stand out in one of two dimensions:

- High productivity (many word types contain them and they accumulate relatively high frequency),
- High idiosyncrasy (few word types contain them, but those are highly frequent).

Besides these generalisations that hold for all the languages in our sample, the central insight here is that languages are systematically different with regards to how their subwords distribute in the productivity/idiosyncrasy space. This distribution is characteristic for each language. Namely, there are languages with more prominent presence of subwords around the area with high values of productivity (and a low degree of idiosyncrasy), e.g. Turkish (tur) and Kalaallisut (kal). In this case, highly productive subwords are precisely the ones that get merged early. In contrast, in other languages, e.g. Sango (sag) and English (eng), the subwords are less productive but concentrate more in areas where the idiosyncrasy is high. Still, these few types

30 These distinctions may introduce a certain preference towards capturing more suffixal or prefixal patterns.
31 The code an other resources can be found at https://github.com/ximenina/bpe-morphology

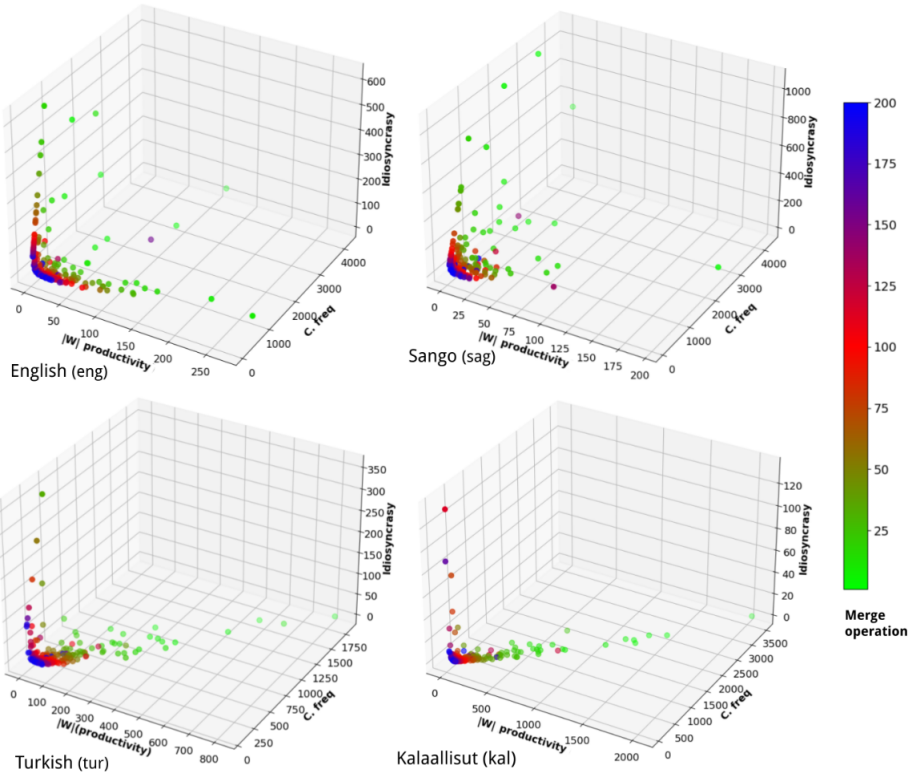


Figure 5: BPE subwords for four languages (PBC corpus). Each data point represents a subword. The colors are a visual cue for the merge operation (i.e. from 1 to 200) at which the respective subword was created.

have a relatively high cumulative frequency, and are good candidates for compression, i.e. they tend to get merged first.

Notice that the axes are not scaled here: for Sango, the idiosyncrasy index runs up to more than 1000. In contrast, for Kalaallisut the highest values are approximately ten times smaller, i.e. around 120, and none of these idiosyncratic subwords are merged early. English is closer to Sango with values running up to 600, while Turkish is closer to Kalaallisut, with values of maximally 350. In terms of productivity, however, we have the inverse pattern: Kalaallisut subwords can reach values of around 2000, while in Sango only a single subword reaches productivity of 200, with all others below 100. Again, English is closer to Sango, with values running up to 250, while Turkish – with values up to 800 – gets (somewhat) closer to Kalaallisut.

Note that the fact that a subword has low productivity does not necessarily imply that it will be high in idiosyncrasy. In fact, there can be subwords that are low in productivity but also low in idiosyncrasy, e.g., data points that concentrate near the origin.

4.2 Languages in the BPE subword productivity space

Figure 6 shows points for individual languages plotted in 3D space.[32] Remember that these points represent mean values of productivity, idiosyncrasy, and cumulative frequency for the subwords created in 200 merge operations. Following up on the example shown in Figure 5, we notice that languages like English (eng) and Sango (sag) concentrate around the region where productivity values are low, but with Sango having considerably higher idiosyncrasy. Turkish (tur) and Kalaallisut (kal), on the other hand, are found in the region with low levels of idiosyncrasy and high levels of productivity.

More generally, languages like Fijian (fij), Yoruba (yor), Vietnamese (vie), and Thai (tha) are amongst the ones with lowest productivity of subwords. Korean is also among these, however, this is an artefact of the writing system (discussed above). Namely, when the PBC text is transliterated into a Latin script, the subword productivity of Korean is actually amongst the highest (see Appendix F). At the high end of productivity we further find Imbabura Quechua (Imbabura) (qvi), Yagua (yad), and Barasano (bsn). Interestingly, these are spoken in a relatively confined geographic space (on a global scale) of north-western South America. Figure 7 illustrates the geographical distribution of all the languages. Egyptian Arabic (arz) is also among the languages with highly productive subwords, even surpassing some Eurasian languages such as Russian, Finnish, and Turkish in this dimension. This is certainly related to the productive consonant templates in Arabic writing, though note that, in the PBC texts, vowels are also indicated (see also Section 3.2 for discussion).

Some of the languages with highest idiosyncrasy indices include Burmese (mya), Sanumá (xsu), Thai (tha), and Sango (sag). In the case of Burmese, high idiosyncrasy is to some extent a reflection of the tokenization, which "oversplits" orthographic words (see also Section 3.2). In Sango and Sanumá, on the other hand, the high idiosyncrasy of subwords is more clearly driven by their morphological structure.

Finally, we repeat the same analyses for the JW300 and UDHR corpora to assess the dependence of the results on a specific corpus. Appendix D gives the respective plots, and a comparative analysis using the distribution of Euclidean distances across the datapoints. As a general trend, languages maintain similar positions in the BPE space despite the different corpora sizes and registers. We observe that in the UDHR corpus, featuring small text sizes, there is greater variation in the Euclidean distances between the data points that belong to the same language. However, even here we find that a similar arrangement of languages is maintained with respect to PBC.

Similarity patterns between languages that emerge from our analysis of the first 200 BPE merges correspond rather closely to the known properties of the languages and are rather robust across different text samples.

4.3 Comparison between BPE subword productivity space and WALS-based clustering

To evaluate the degree to which the arrangement of languages in the BPE subword productivity space agrees with their known typological features, we turn to the WALS database. Figure 8 shows the BPE space again, with the data points for languages in the same positions as before. This time, however, the points are colored according to four clusters obtained by using morphological features from WALS. If the BPE characterization of languages, on one hand, and the WALS features, on the other, had little in common, then the colors of points should

32 Appendix A contains information about all the languages and their vector representations. Additionally, Appendix B shows the 2D planes.

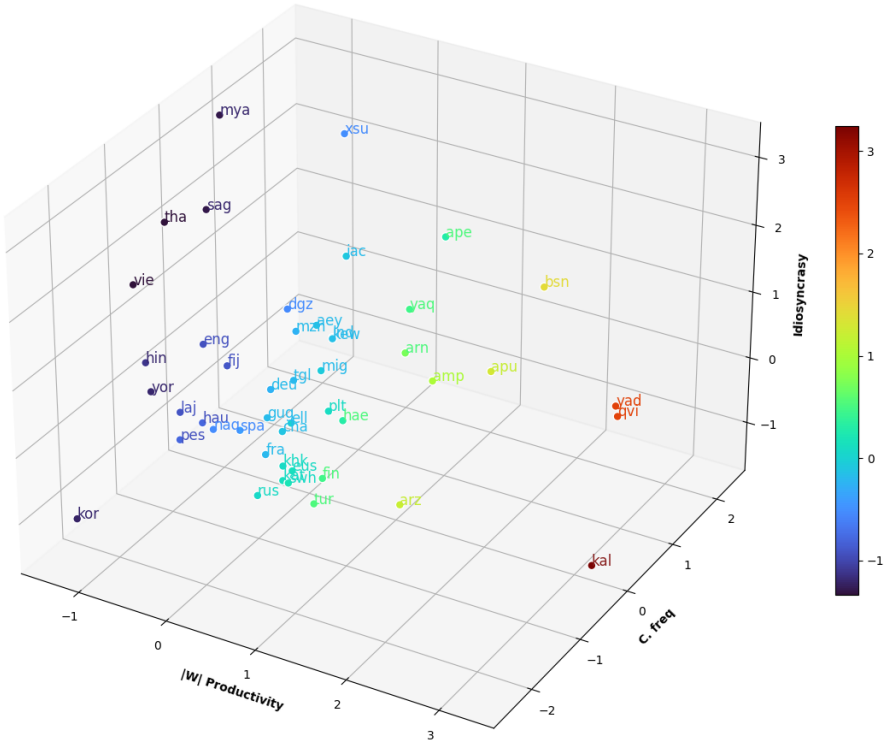


Figure 6: Languages represented by a single vector that averages the values obtained during the first 200 merges (PBC). The color reflects the variation of the x-axis (Productivity $|w|$).

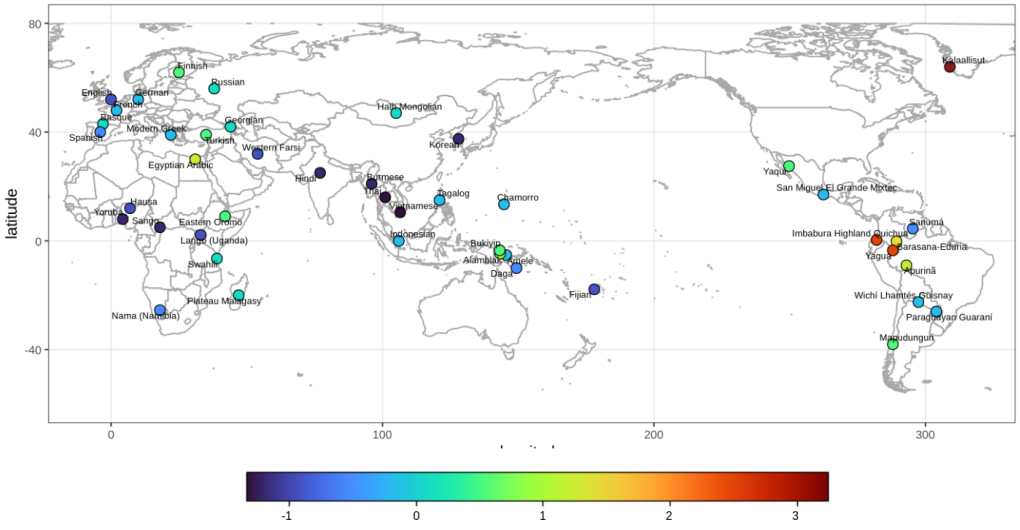


Figure 7: Geographical distribution (PBC). Color indicates the variation of Productivity $|w|$.

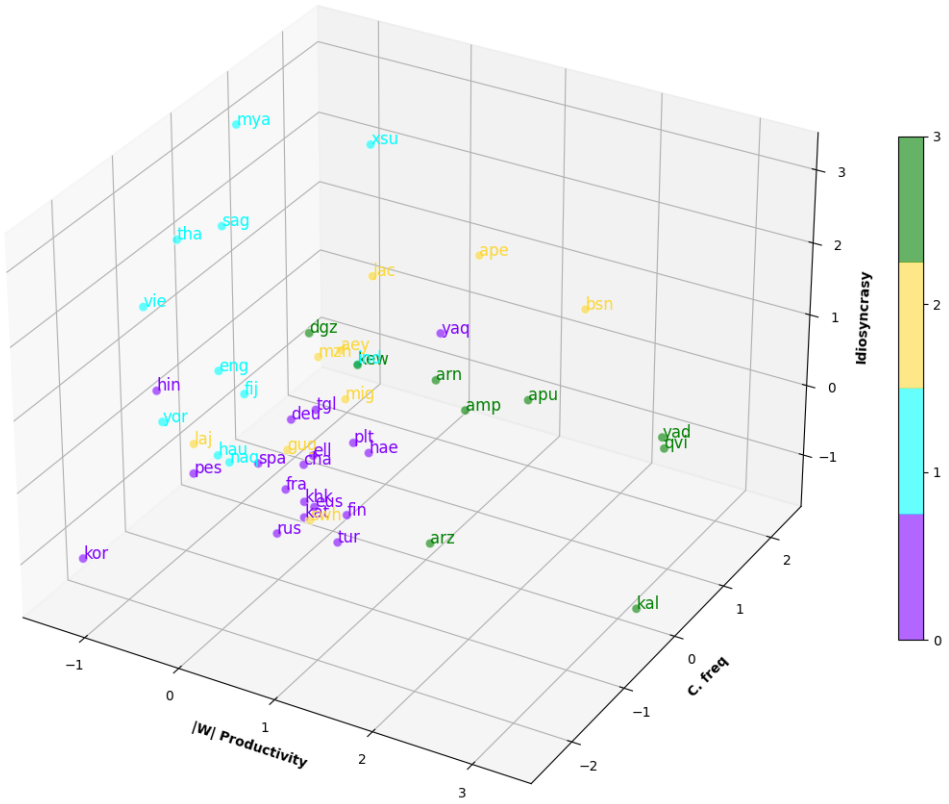


Figure 8: BPE language space (PBC), the colors represent the clusters obtained with WALS-based vector representations using k-means $(k = 4)$.

have a random distribution in Figure 8. In the opposite scenario, i.e. if the BPE and WALS characterizations agree, we should see clusters distinguished by colors. The visual impression in Figure 8 is that, indeed, there is considerable clustering of colors.

We back up this finding by calculating the mean silhouette coefficient (see Section 3.5). Table 6 gives $\tilde{s}$ calculated for three different scenarios. First, we define the upper bound to be the score calculated using the WALS vector representations only ($\text{WALS}_{\text{original}}$). This score shows how compact the WALS clusters are in their own space. Second, we define the baseline score given randomly drawn values of the BPE vector representations ($\text{WALS}_{\text{BPErandom}}$). In this setting, each language is represented in the BPE three-dimensional space, with the value of each dimension generated by randomly drawing samples from a uniform distribution of the original intervals of values. The BPE space is overlaid with the WALS clusters for which the silhouette coefficient is calculated. Third, the score that quantifies the agreement between the features extracted from WALS and the BPE productivity vectors ($\text{WALS}_{\text{BPE200}}$) is calculated in the same way as the baseline, but using the actual (observed) values of BPE vectors instead of the random values.

The upper bound value of $\tilde{s}$ (over the WALS feature space itself) is 0.23. This score is in the upper part of the range ([-1,1]), but still far from 1, meaning that the WALS clusters are moderately compact and distinct in the most favorable scenario. In the case of the BPE space overlaid with WALS feature clusters, $\tilde{s}$ decreases to 0.12. This is due to some languages

Table 6: Mean silhouette coefficient $\tilde{s}$ for WALS-based k-means clustering ($k = 4$) but measured over different vector representations, or characterizations, of languages. The second column includes additional BPE characterizations using different merge operations. Similarly, the third column contains characterizations obtained from WordPiece subwords.

Version	$\tilde{s}$	Version	$\tilde{s}$	Version	$\tilde{s}$
$WALS_{original}$	0.23	$WALS_{BPE10}$	0.03	$WALS_{WP10}$	-0.19
$WALS_{BPErandom}$	-0.17	$WALS_{BPE50}$	0.09	$WALS_{WP50}$	-0.02
$WALS_{BPE200}$	0.12	$WALS_{BPE100}$	0.10	$WALS_{WP100}$	-0.01
$WALS_{WP200}$	0.05	$WALS_{BPE1000}$	0.13	$WALS_{WP1000}$	0.11

being in the same WALS cluster, but distant in our BPE space. For example, Swahili (swh) is close to Turkish (tur), Finnish (fin), and Russian (rus) in the BPE space, but clusters with other languages such as Lango (laj), Guaraní (gug), and Mixtec (mig) in terms of WALS features. In the baseline case of random arrangement of language vectors in the BPE space, $\tilde{s}$ further drops to -0.17. To put the outcome of our comparison into perspective, the mean silhouette coefficient for the BPE space overlaid with WALS clusters only drops by ca. 27%[33] on the scale from the original WALS space to the random baseline. In other words, the fit of WALS clusters to the points in the BPE space is much closer to the upper bound than to the baseline. We thus interpret the $\tilde{s}$ value of 0.12 as a solid quantitative evidence of the agreement between language representations extracted from WALS and those that result from analyzing BPE subwords.

4.3.1 Additional BPE subword productivity spaces. The distribution of languages in the productivity space changes depending on the number of BPE merge operations applied to characterize them. This is because language vectors are based on the mean values of productivity, idiosyncrasy, and cumulative frequency of the subwords created up to a specific merge operation, e.g., 200.

To understand this effect better, we obtain varied BPE productivity spaces derived from incrementally applying 1 to 1000 merge operations. The most visible changes in language distributions happen during the first merges. By merge 200 the arrangement of data points in the productivity space is stable. Language vectors obtained using 200 merges are almost identical to the ones obtained using more subwords, e.g., 1000. Detailed graphs can be found in Appendix H. This behavior is expected. We saw before that subwords of later merges tend to accumulate near the origin, i.e., they are not very discriminative anymore (Figure 5). Therefore, even though we keep merging, the mean values are mainly influenced by the salient data points captured on the first merges.

We also assess the impact of changing the BPE hyperparameter in the agreement between the BPE subword productivity space and the clustering based on WALS typological features. The second column of Table 6 contains the mean silhouette coefficient ($\tilde{s}$) obtained with a selected number of BPE merge operations. The results for the complete range of merges can be appreciated in Figure 9. As a general trend, the more subwords we use to characterize languages, the stronger the agreement with WALS – until we reach a certain point where there is no more improvement. Again, the most noticeable changes occur in the first merge operations. By choosing around 200 merge operations, we will obtain a productivity space similar to the

33 Given the values in Table 6 we get $\frac{0.23-0.12}{0.23-(-0.17)} = \frac{0.9}{0.38} = 0.27$.

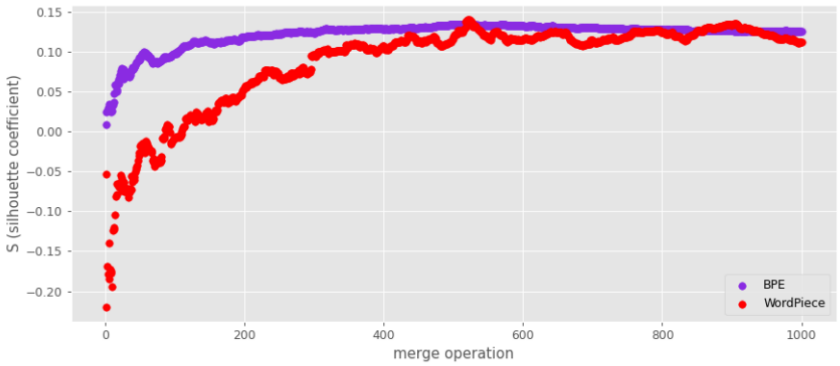


Figure 9: Mean silhouette coefficient $\tilde{s}$ for WALS-based k-means clustering ($k = 4$) measured over the vector representation obtained at different merge operations for BPE and WordPiece.

ones of subsequent merges, but using less computational resources. Just a few hundred BPE subwords are needed to characterize the morphological typology of languages.

4.4 WordPiece results

Although BPE and WordPiece both incrementally merge subwords, their quantitative properties are different. This is reflected in the productivity spaces. The arrangements of languages obtained using WordPiece subwords systematically diverge from the ones obtained with BPE when few merges are applied. Moreover, the positioning of the language vectors in the WordPiece space shows less agreement with the WALS space, reflected in lower values of the mean silhouette coefficient $\tilde{s}$ (Table 6). In other words, the quantitative properties of the first subwords found by WordPiece are less representative of the morphological typology of languages. In fact, using few merge operations we obtain negative values of $\tilde{s}$ ($\text{WALS}_{\text{WP10}}$ is even below the random baseline). But if we keep merging, the arrangement of languages in the WordPiece productivity space slowly becomes more similar to the WALS space (Figure 9). This means that WordPiece can also be used to extract typological features, but not in the same way as BPE.

WordPiece merges adjacent subwords that are frequently encountered together, but with their marginal frequencies being low. On the first merges, this causes a preference towards patterns that contain rare characters in the respective language. Also, compared to BPE, there is a greater predominance of longer subwords – similar to stems – from the very first operations. It makes sense that stem-like patterns are less indicative of the morphological typology of a language, i.e., the subwords do not resemble regular morphological phenomena like affixes.

A note is needed here to underline that we are not comparing the quality of subword tokenizations. In fact, for tokenizing a text, WordPiece utilizes the merged patterns in a different way compared to BPE.[34] Our comparison concerns solely the subwords that emerge iteratively, and what these patterns reveal about language structure.

34 It keeps only the vocabulary of the last iteration, and applies a longest-match-first strategy.

5. Discussion

5.1 BPE subwords as typological features

What emerges from our analyses is that the same subwords that are the most useful for compressing texts, are also useful for differentiating languages. For example, English subwords formed in the first two hundred BPE merges reach productivity scores of ca. 250, while in Turkish these range up to ca. 800 (remember Figure 5). Also, these subword properties match the classifications of WALS chapters on morphological typology.

But how exactly do these observations relate to the morphology of languages? At first sight, it seems like the patterns harnessed for compression are not related to the classic idea of morpheme structure at all. For instance, in English, the first five merges create the following set of subwords: {*th*, *an*, *and*<\w>, *the*, *the*<\w>}. A pattern like *th* is an orthographic convention of English writing, representing a phoneme, rather than a morpheme. However, notice that in subsequent merges, the two-character subwords become three-character subwords which indeed represent morphemes. The definite article *the* as well as the conjunction *and* in English are morphemes – as well as orthographic words.

To illustrate this point further, Table 7 gives the 10 most *productive* subwords in English and Turkish of *more than two characters*. Notice that BPE compression uncovers inflectional suffixes (e.g. *-ing*, *-eth*, *-est*),[35] stems (e.g. *com-*), and prefixes (e.g. *for-*), while other subwords (e.g. *-oun-*, *-ent*, *-ght*) are less straightforwardly analyzable as morphemes – at least from the perspective of Modern Standard English.[36] Interestingly, in Turkish, the picture is even clearer. The most productive subwords are actually inflectional morphemes of standard grammar. For instance, *-lar* and *-ler* are mostly used as allomorphs for plural marking (e.g. *adam-lar* 'man-PL') (Göksel and Kerslake 2004, p. 65); *-den/-dan* are ablative case suffixes (e.g. *sen-den* 'you-ABL', i.e. 'from you') (Göksel and Kerslake 2004, p. 67); and *-yor* is an imperfective tense suffix (Göksel and Kerslake 2004, p. 69).

Table 7: Most *productive* subwords (of more than two characters) for English and Turkish up to 200 merges. We color the merge numbers roughly as in Figure 3, i.e. green (merge 0-50), red (merge 50-100), violet (merge 100-150), blue (merge 150-200).

English (eng)				Turkish (tur)			
prod.	subword	merge	examples	prod.	subword	merge	examples
110	ing<\w>	29	beginn**ing**	355	lar	8	on**lar**a
67	eth<\w>	103	nazar**eth**, eat**eth**	309	ler	13	gün**ler**de
38	est<\w>	166	**lest**, car**est**	150	ler<\w>	32	gitti**ler**
26	led<\w>	122	fil**led**, cal**led**	145	lar<\w>	38	adam**lar**
26	com	137	**com**ing, **com**e	131	den<\w>	40	sen**den**
23	oun	129	r**oun**d, f**oun**d	129	yor	69	öğreti**yor**d
21	for	86	**for**sook	116	dan<\w>	60	tarafın**dan**
21	ent<\w>	91	w**ent**, garm**ent**	110	ini<\w>	64	indiğ**ini**
21	ght<\w>	102	tau**ght**, mi**ght**	107	ların	96	ağ**ların**ı
19	ing	90	th**ing**s, br**ing**ing	80	ine<\w>	50	üzer**ine**

35 Arguably, *-led* is also an inflectional suffix. The *-l* of the stem is merged to *-ed* which has already been merged before.

36 In fact, in many cases, these patterns are probably related to inflectional and derivational marking, which has changed and fossilized over time.

Table 8: Most *idiosyncratic* subwords (of more than two characters) for English and Turkish up to 200 merges. We color the merge numbers roughly as in Figure 3, i.e. green (merge 0-50), red (merge 50-100), violet (merge 100-150), blue (merge 150-200).

English (eng)			Turkish (tur)			
idiosyncrasy	subword	merge	idiosyncrasy	subword	merge	translation†
617	him</w>	18	352	isa</w>	34	“Jesus”
489	that</w>	23	258	dedi</w>	56	“said”
486	the</w>	5	152.5	bir</w>	43	“a, one”
442	they</w>	27	121	onlara</w>	127	“them”
376	them</w>	38	99	sonra</w>	160	“after”
298	said</w>	47	86	size</w>	184	“to/for you”
262	shall</w>	57	82	şöyle</w>	189	“such (a)”
237	his </w>	63	73.5	için</w>	106	“for”
233	for </w>	66	61.5	ama</w>	126	“but”
224	unto</w>	26	40	diye</w>	128	“that”

† According to the Turkish grammar by (Göksel and Kerslake 2004).

The most *idiosyncratic* subwords uncovered by BPE compression up to 200 merges (see Table 8) are full blown orthographic words. In fact, the words occurring in this list are often semantically related across the two languages, e.g. Turkish *dedi* corresponds to English *said*, *onlara* to *them*, and *için* to the preposition *for*.

To appreciate the structural differences in the two languages, note the general trend in Turkish toward productive inflectional markers in the early merged subwords – which contribute most to compression – while in English, the trend is toward idiosyncractic words. This is visible when inspecting the merge numbers (color-coded for visual reference) in Tables 7 and 8. In other words, the concatenative and (largely) nonflexive nature of Turkish morphology, on one hand, and the less productive nature of English inflectional morphemes, on the other, are captured by the quantitative properties of their BPE subwords.

Our conjecture is that the distributional differences in subwords are linked to the morphological typology of the respective languages. Languages with rich inflectional morphology exhibit a more prominent concentration of subwords in the area with high productivity and low idiosyncrasy values. In contrast, languages with poorer inflectional morphology show a more prominent subword concentration in areas where idiosyncrasy is high.

In the example of English and Turkish, the most productive subwords in each case largely correspond to patterns discussed as inflectional affixes in standard grammars. While this might not be immediately apparent in early merges of pairs of characters, it becomes clear when merges of more than two characters are formed. Further merges (beyond 200) might blur this picture again, leading to the general impression of BPE subwords not being linguistically relevant.

5.2 Text-based morphological typology

As already pointed out by Sapir (1921) and later implemented by Greenberg (1960) (see also Section 2.3), language diversity should be measured on a continuum, rather than broken down into discrete categories. However, traditional typology often works with categorical features. With this in mind, our features can provide a quantitative yardstick for the categorical features of traditional typology. As a first step in this direction, we here interpret our compression-based

Table 9: Average productivity of BPE subwords alongside categorical information on *synthesis* and *fusion* of languages from the typological literature. Productivity values are colored roughly according to the scale in Figure 6. This is a subsample of the 47 languages of the PBC sample for which we were able to find a reference giving a categorical synthesis value (references given in a separate column). The reference for fusion is WALS Feature 20A (Bickel and Nichols 2013a).

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

productivity measure in terms of the notions of *synthesis* and *fusion* as they are established in the typological literature (Table 9).

Low productivity. We observe in Table 9 that languages with low subword productivity (i.e. around one standard deviation lower than the mean or zero after standardization) are generally categorized as *analytic* by typologists. An analytic language tends to have a low ratio of morphemes per word, i.e., each word can consist of a single, independent morpheme. This type of language has a preference to encode relations and meanings through syntactic structures instead of using word-internal structures. Take the examples for Sango (7) and Sanumá (8). Grammatical categories like person and tense are here not encoded by means of inflectional affixes (as it happens in more synthetic languages), but by free morphemes. Interestingly, in the case of Sanumá, the form *töpö* that appears glossed as 3PL, is actually the subword that obtains the highest idiosyncrasy using our BPE operationalization.

Sango (sag)

(7) löndö mo gä
rise 2S come
"get up and come (here)" (Karan 2006, p. 247)

Sanumá (xsu)

(8) Sama **töpö** se kite
1PL.EXCL **3PL** hit FUT
"We will hit them." (Borgman 1990)

Further languages categorized as analytic in our sample are English (eng), Fijian (fij), Vietnamese (vie), Thai (tha), Hindi (hin), and Yoruba (yor). These are the languages with cooler colors (low productivity) also in Figure 6. Two borderline cases are French (fra), and Persian/Western Farsi (pes). The subword productivity of the latter is somewhat closer to analytic languages, but it is categorized as synthetic.[37]

The terms "analytic" and "isolating" are often used interchangeably. But as pointed out above with reference to Bickel and Nichols (2007), in a strict sense, these are independent dimensions of morphological typology. Vietnamese (vie), Thai (tha) and Yoruba (yor) are examples of languages with very little (or no) inflectional morphology. According to WALS feature 20A on the fusion of inflectional formatives, Vietnamese is exclusively isolating, while Thai is isolating/concatenative, since it features some derivational formatives in the form of reduplication, affixing, compounding. Yoruba, on the other hand, is classified as tonal/isolating, since tones are here also relevant as grammatical information (not purely for lexical distinctions). Hence, most of the analytic languages are said to be at least partly isolating. In contrast, English and Sango are assigned the label "exclusively concatenative". In the case of English, there are some formatives of grammatical information, e.g. for tense and person/number, which are analyzed as concatenative.[38] Quoting Haspelmath and Sims (2010, p. 4) "Quite generally, we can say that English makes more use of morphology than Yoruba. But there are many languages that make more use of morphology than English."[39] This statement is nicely matched by the BPE productivity measure, yielding -1.21 for Yoruba and -0.94 for English.

Medium productivity. In the medium range of subword productivity, we find languages like Chamorro (cha), French (fra), German (deu), Spanish (spa), Russian (rus), and – with somewhat higher values – Swahili (swh), Turkish (tur), and Yaqui (yaq). These languages are fairly consistently categorized as synthetic with concatenative formatives by typologists (see Table 9). For Chamorro, we were not able to find an explicit mention of its synthesis. It is described as a largely concatenative language with some fusional phenomena in a recent morphological overview article (Stolz 2015). There is some disagreement on its exact morphological status, however. According to WALS feature 26A, it has little affixation (Dryer 2013), and WALS feature 20A classifies its morphology as concatenative/isolating (Bickel and Nichols 2013b). Its productivity value of -0.11 (see Appendix A) reflects this in-between-status, as it falls between languages typically called "analytic" or "synthetic".

French, German, Spanish, and Russian are Indo-European languages with productive inflectional morphology. Furthermore, they are often named as examples of cumulative exponence, i.e. one single morpheme encodes several grammatical functions, rather than each morpheme having a separate function (separative exponence). Traditionally, this was referred to as fusional language, contrasting with agglutinative language (see also Section 2.3). Cumulative and separative languages represent two sides of synthetic morphology. In both cases, words can contain several morphemes, however, in separative (agglutinative) languages, the morphemes remain almost unchanged after concatenation, while in cumulative (fusional) languages, they are "fused" into a single morpheme, including phonological alternations that obscure the boundaries. The difference between cumulative and separative exponence seems to play into the BPE space as well. Languages of the former type (English, Spanish, German,

37 We should be aware that Persian (Farsi) is here written with an abjat script without indication of vowels. This can have an impact on the productivity of subwords. See also discussion in Section 3.2.

38 The auxiliary *will* in periphrastic future tense constructions could be seen as an isolating formative in English though.

39 We have added the term *bound* here.

French) are located in the lower productivity, and higher idiosyncrasy regions than typical languages of the latter type (Swahili, Turkish, Finnish, Basque). Unfortunately, there is no WALS chapter which would code the exponence of languages across different word categories,[40] and which could hence be used to investigate the link of exponence to higher/lower subword productivity more systematically.

High productivity. At the high end of productivity (i.e. more than one standard deviation above the mean) we find languages such as Alamblak (amp), Apurinã (apu), Barasano (bsn), Quechua (Imbabura)(qvi), and Kalaallisut (West Greenlandic, kal). These have warmer colors in Figure 6. For better appreciation, Appendix E contains the zoomed-in regions of the space.

Languages in this area are generally considered morphologically rich, concatenative, and polysynthetic by typologists (see Table 9). Kalaallisut stands out with a subword productivity more than three standard deviations higher than the mean. Furthermore, its idiosyncrasy is very low. It encodes many morphosyntactic distinctions at the word level by concatenating morphemes, see Example (9), which is given as a paradigm example of polysynthesis (Haspelmath and Sims 2010, p. 5). Note that the entire main clause of the English translation (*I didn't understand at all (that)* [...]) is here "synthesized" into a single orthographic word (*Paasinngilluinnarpara*).

Kalaallisut/West Greenlandic (*kal*)

(9) Paasi-nngil-luinnar-para ilaa-juma-sutit.
understand-not-completely-1SG.SBJ.3SG.OBJ.IND come-want-2SG.PTCP
"I didn't understand at all that you wanted to come along." (Fortescue 1984, p. 36)

Some of the other highly productive languages, for instance, Yagua (yad) and Quechua (Imbabura) (qvi), are located in a somewhat different area of the space, since they have higher degrees of cumulative frequency. This might reflect further particularities of their morphology. For instance, a phenomenon frequently found in Quechua varieties is root reduplication applying to both nouns and verbs, as in the example *kaša kaša* meaning "place full of thorny plants", where *kaša* by itself means "thorny plant" (Adelaar 2004, p. 1455). Note that this morphological strategy increases the cumulative frequency of subwords.

Within the spectrum of high productivity, we additionally find Apurinã (apu), Alamblak (*amp*), and Barasana-Eduria (*bsn*). These are also named among polysynthetic languages (see Table 9). The former two are described as predominantly suffixal and with some fusional elements (Facundes 2014; Palmer 2017; Bruce et al. 1984), while the morphology of Barasana-Eduria (bsn) is peculiar for having tonal phenomena and also allomorphy (Gomez-Imbert 2004).

Mapudungun (arn) is an interesting case. It is categorized as a polysynthetic language with remarkably rich verbal morphology – including noun-incorporation (i.e. the subject of a sentence might be represented morphologically inside the verb). This contrasts with its almost non-existent inflectional morphology of nouns (Zúñiga 2017), as illustrated in the following example:

Mapudungun (*arn*)

(10) anü-m-ka-i pinu yengu.
sit.down-CAUS-CONT-IND cane 3DU
"Both of them planted cane." (Bickel and Zúñiga 2017, p. 12)

40 WALS feature 21A codes exponence only for "selected inflectional formatives", i.e. case markers.

While the verb consists of four morphemes, the noun consists of one. In a sense, Mapudungun is a polysynthetic language in its verbal domain, but analytic in the nominal domain. This might explain why it has an overall subword productivity (0.73) somewhat lower than other languages categorized as "polysynthetic", and rather approaching the score of synthetic languages like Turkish (0.54).

Paraguayan Guaraní (gug) is an extreme example of discrepancy between the typological categorization and our quantitative measure. In Aikhenvald (2017), it is discussed under the rubric of polysynthetic language.[41] Counter expectation, it has a subword productivity even lower than the global mean, i.e. -0.19 (the same as French in our estimation). This is likely related to a "clash" between the definition of polysynthesis based on a feature like noun-incorporation, on one hand, and a quantitatively oriented criterion like the subword productivity, on the other. For instance, in Paraguayan Guaraní, the word *mba'e* 'thing' can be incorporated into a complex verb form:

Paraguayan Guaraní (*gug*)

(11) A-**mba'e**-jogua-ta ko-ka'aru
1ACTIVE-**thing**-buy-FUTURE this-afternoon
"I'll go shopping this afternoon." (Aikhenvald 2017, p. 297)

Thus, the language fulfills the noun-incorporation criterion of polysynthesis. However, while *mba'e* as a subword has a relatively high cumulative frequency in our PBC text – occurring 387 times[42] – it only occurs in 38 different word types. This is rather low productivity compared to an inflectional suffix like *-lar* in Turkish, which occurs overall in 500 different word types (remember Table 7). Thus, at least in this particular case, noun-incorporation yields subwords which have a different quantitative profile compared to inflectional morphemes.

Finally, we also want to mention Egyptian Arabic (arz) as another particular case. It is located in the area with very high productivity (1.21, see Appendix A), and relatively low idiosyncrasy. However, it is normally not named among polysynthetic languages. Its morphology is also not considered exclusively concatenative, but rather "non-linear" (Bickel and Nichols 2007), or "Ablaut/concatenative" (Bickel and Nichols 2013a). In this language, BPE captures many patterns on the first merges with high productivity. These also involve consonant/vowel combinations, not just consonant templates, since in the Egyptian Arabic texts used in our analyses, vowels are indicated by diacritics.

In the map (Figure 7), we overlaid our productivity $|W|$ measure, since we have discussed that it relates easily to the concepts of synthesis and fusion. Without making any general claims for the moment, we note that languages with the highest degree of productivity seem to be found in the Americas (polysynthetic and concatenative tendency). In contrast, languages with the lowest degree are found in some regions of Asia and Africa (analytic and isolating tendency). This is expected from the typological literature on morphology. We leave a more systematic study of geographic patterns based on subword productivity for future work.

In summary, there is generally a good fit between categorical distinctions along the clines of *analytic* → *synthetic* → *polysynthetic*, as well as *isolating* → *concatenative*, on one hand, and quantitative text-based measures like subword productivity, on the other. Languages with low subword productivity are fairly consistently categorized as "analytic" and "isolating" by typologists, while medium subword productivity rather maps onto the categories of "synthetic"

41 She also uses the term "highly synthetic" in some instances.
42 This includes both *mba'e* as part of a larger word, as well as *mba'e</w>* at the end of a word (and as an orthographic word itself).

and "concatenative", and high productivity onto "polysynthetic" and "concatenative". There are some counter-examples to these mappings, which reveal interesting discrepancies between the traditional typological ideas about morphological categories, and our quantitative measures. Representing the notion of cumulative frequency as a separate dimension in the space allows us to identify some subtypes of languages along the productivity/idiosyncrasy scale in accordance with the facts cited in the typological literature.

5.3 Comparison to WALS feature clustering

In the previous subsections, we have zoomed into quantitative subword properties of individual languages (English and Turkish), as well as compared average subword productivities to certain typological categories (synthesis, fusion). To get a more general view on the link between morphological properties as defined by typologists, and our compression-based account, we now turn to the results of the WALS-based feature clustering projected into BPE space (Figure 8).

As a general trend, languages that get clustered according to WALS criteria, are also found in contiguous regions in the BPE space. For example, cluster 0 (violet data points) is located roughly around average productivity and idiosyncrasy values. It is the most populated cluster, containing languages like Finnish (fin), Basque (eus), Georgian (kat), Halh Mongolian (khk), Russian (rus), Turkish (tur), Modern Greek (ell), Chamorro (cha), German (deu), French (fra), Korean (kor), Western Farsi (pes), and Spanish (spa). In fact, this corresponds mostly to languages falling under the umbrella terms "synthetic" and "concatenative".

English (eng), Fijian (fij), Hausa (hau), Indonesian (ind), Burmese (mya), Nama (Namibia) (naq), Sango (sag), Thai (tha), Vietnamese (vie), Sanumá (xsu), and Yoruba (yor) get assigned to cluster 1 (cyan data points). In our BPE space, this corresponds mostly to the region with low values of productivity paired with high values of idiosyncrasy – in some cases. The respective languages are squarely associated with the term "analytic" in the discussion above.

Cluster 2 (yellow data points) contains languages that tend to have relatively high values of idiosyncrasy and also cumulative frequency in the BPE space. This cluster includes Amele (aey), Bukiyip (ape), Barasana-Eduria (bsn), Popti' (jac), and Wichí Lhamtés Güisnay (mzh), but also several cases of languages whose position in the BPE space does not correspond to their WALS clusters, mainly Swahili (swh), Paraguayan Guaraní (gug), and Lango (Uganda) (laj).

Finally, cluster 3 (green data points) contains languages that are mostly distributed in the area of the BPE space characterized by high cumulative frequency, irrespective of their average productivity. Quechua (Imbabura) (qvi), Yagua (yad), Alamblak (amp), and Apurinã (apu) have both high productivity and cumulative frequency, while Daga (dgz) and West Kewa (kew) are considerably less productive than the other cluster members. The highly productive language Kalaallisut (kal) belongs to this group as well, though having somewhat lower cumulative frequency.

Appendix G shows the silhouette coefficient of each data point $s(i)$ in the BPE space ($\text{WALS}_{\text{BPE200}}$). Languages with low values (between -0.7 and -0.6) can be considered to be misplaced in the BPE space, i.e., they are far from other languages that belong to the same WALS cluster and closer to languages that belong to other clusters. Here we see languages like Swahili (swh), Nama (Namibia) (naq), Paraguayan Guaraní (gug), Hausa (hau), Daga (dgz), Lango (Uganda) (laj). In fact, Swahili gets the lowest silhouette coefficient. We notice in Figure 8 that Swahili is similar to the Eurasian languages in terms of quantitative subword properties, but it contrasts with them, for instance, when it comes to the number of nominal case markers (WALS chapter 49). In the nominal domain, it rather clusters with the Subsaharan,

Mesoamerican and South American languages in our sample, which lack nominal case marking all together.

On the other hand, there are languages with high $s(i)$ (greater than 0.7), meaning these BPE productivity data points are well matched to their respective WALS clusters, i.e., languages that end up close to the other languages in their WALS cluster and far from languages that belong to other WALS clusters. They remain in a relatively well delimited and compact area in the BPE space. Here we see cases like Basque (eus), French (fra), Russian (rus), Halh Mongolian (khk), Georgian (kat).

5.4 Subwords beyond 200 merge operations

Our primary analysis is based on the first 200 redundant patterns uncovered by BPE compression. This hyperparamenter can be tuned, i.e. the analysis can be conducted using a different number of merge operations. However, we find that taking very few subwords (e.g. 10) is not enough to characterize the languages typologically, while taking more than 200 does not lead to considerable changes in the language vector representations anymore.

5.5 BPE vs. WordPiece

Interestingly, the first 200 subwords obtained from the WordPiece merging criteria are less indicative of morphological structure. Our analyses suggest that BPE subwords tend to have a more grammatical or functional nature, allowing identification of the type of morphology. As we discussed throughout this work, these can be productive patterns that resemble inflectional markers, affixes, very frequent irregular morphological patterns, or function words.

6. Limitations and future work

The method of subword tokenization we have chosen here is relatively straightforward, text based, reproducible, and it works without strong theoretical assumptions about language structure. Still, our approach is susceptible to the intrinsic shortcomings of the BPE method (and WordPiece) applied to diverse textual material. For instance, an assumption which is hard-coded into this implementation is that orthographic word boundaries work as strict delimiters not to be transitioned in subword generation. In future research, it would be interesting to remove even this restriction, and generate subwords over strings of characters without word delimiters. However, this would also require the reconceptualization of our productivity measure, which currently hinges upon counts of orthographic word types. We also showed that some types of scripts, like in the case of Korean, due to its syllabified nature, would require additional adjustment of our approach.

Non-standard languages represent another known challenge for subword tokenization techniques in general, including BPE. Also, previous articles have highlighted that BPE encoding may not be suitable for non-linear morphology, arguing that linguistically supervised strategies may achieve better subword tokenizations in this case (Shapiro and Duh 2018; Amrhein and Sennrich 2021; Nzeyimana and Niyongabo Rubungo 2022). We do see some impact in the case of Arabic in our findings, but this impact depends on whether vowels are indicated in the respective script, and whether the respective UTF-8 characters are handled correctly by the respective BPE implementation. Currently, our analysis covers several different writing systems in a rather robust way, but some specific cases (e.g. Korean Hangul and Burmese) require special attention.

It has recently been discussed that multilingual corpora in NLP tend to have "language contamination" (Blevins and Zettlemoyer 2022). In our case, this type of noise is more

controlled due to the nature of the corpora (relatively small, relatively well curated sources). However, even in the presence of mixed material, this will likely not have much impact on the BPE patterns captured in the first merges. The patterns of other languages are unlikely as frequent as an inflection marker or a highly frequent word captured on the first operations.

With regards to our comparison between WALS-based k-means clustering and the BPE space, there are two main limitations: firstly, we choose $k = 4$ since it offers a compromise between decent mean silhouette coefficients, on one hand, and keeping feasible the interpretability of the resulting clusters (not too many), on the other. A more systematic assessment of this trade-off for different values of k would be useful also for other studies using this method. In the future, more sophisticated methods for estimating the number of clusters can be incorporated, e.g., x-means (Pelleg, Moore et al. 2000), and for measuring the agreement between different vector spaces (Kriegeskorte, Mur, and Bandettini 2008).

Second, one challenging aspect of the quantitative evaluation is the lack of a traditional gold standard. WALS represents a valuable source of linguistic knowledge for cross-linguistic comparison, but it has limitations. It is just one option of using a reduced subset of (available) features to characterize languages in terms of their morphology. Using other typological databases such as AUTOTYP (Bickel et al. 2022) can provide another yardstick. In any case, the feature values tend to be discrete, e.g., binary features, which conceal the gradient nature of language.

More generally, clustering or categorization techniques induce rigid boundaries over a continuum. In our analyses, these boundaries can be affected when new languages are added. This emphasizes the importance of working with a diverse sample of languages to ensure that we obtain a representative snapshot of how natural languages tend to distribute according to their BPE compression properties, i.e., making it less likely to find a new language that will be very far away from the populated regions we have worked with.

6.1 NLP applications

On a practical note, aside from providing a quantitative tool for morphological typology, our approach has the potential to benefit NLP tasks. Several NLP applications use typological language vectors, e.g., lang2vec, for various purposes. Our approach also results in language vectors. This opens the possibility of extending the current typological vectors used in NLP into the morphological domain (in a resource-cheap way).

This seems especially promising in the context of transfer learning in highly multilingual settings, where the success of the transfer depends on language similarity, the amount of annotated resources, and other factors (Pires, Schlinger, and Garrette 2019; Lauscher et al. 2020). In fact, the criteria for determining the languages more appropriate to transfer from constitutes a current research problem (Lin et al. 2019; Malkin, Limisiewicz, and Stanovsky 2022). Our operationalization could provide an informed criterion to facilitate cross-linguistic transfer, especially in challenging scenarios like zero-shot cross-lingual model transfer between under-resourced and distant languages.

It is noteworthy that for both subword tokenization algorithms (BPE and WordPiece), we used a number of merge operations considered small for practical use in NLP downstream applications (Mielke et al. 2021). However, our aim was not to improve subword tokenization or the performance of a downstream task; here we use the patterns that emerge from text compression as a tool for comparing languages. An interesting research direction is to leverage these subword properties to achieve more efficient subword tokenization (Pelloni et al. 2022).

Along a similar line, Rust et al. (2021) propose measures like fertility (the average number of subwords produced for every tokenized word) for inspecting the properties of subword vocabularies. This is performed in the context of pretrained multilingual language models

and the impact of tokenization on downstream tasks. Future research is needed to explore how our characterization of subwords relates to these types of measures, e.g., languages with a tendency to more idiosyncratic subwords on the first merges might have low fertility. In contrast, productive languages could score higher in fertility.

7. Conclusions

In this work, we started from the concept of data compression as a way to reduce redundancy. When BPE encoding is applied to natural languages, a universal property emerges: the most significant compression is achieved with the patterns captured on the first merge operations (this is likely related to the Zipfian nature of languages). We inspected these subwords more closely and discovered that the types of patterns allowing effective compression are not the same across languages. As a matter of fact, they are representative of language structure and an indicator of the morphological typology of languages. Our findings show that, in some languages, text compression is achieved via productive subwords, those resembling inflectional markers, affixes, and other regular morphological phenomena. In other languages, the best candidates for compression tend to be idiosyncratic subwords, i.e., frequent irregular patterns or whole orthographic words.

We propose a novel way to characterize the BPE subwords inspired by the notion of morphological productivity in linguistics. We show that as few as 200 merge operations are already suitable for capturing the most relevant subwords patterns that allow us to characterize languages. Interestingly, the same amount of incremental patterns obtained from the WordPiece merging criteria are less indicative of morphological structure.

The language vector representations that result from this method are a reflection of phenomena discussed in traditional typology, even though our approach does not use annotated data or any external linguistic knowledge. It relies merely on a common text compression technique (subword tokenization) applied to the written representation of languages, i.e., raw text. No further preconceptions or assumptions about the structure of languages are necessary. Despite the simplicity of this approach, the implications are far-reaching: through the looking glass of compression we see more clearly the commonalities and differences in languages down to the atoms of information encoding.

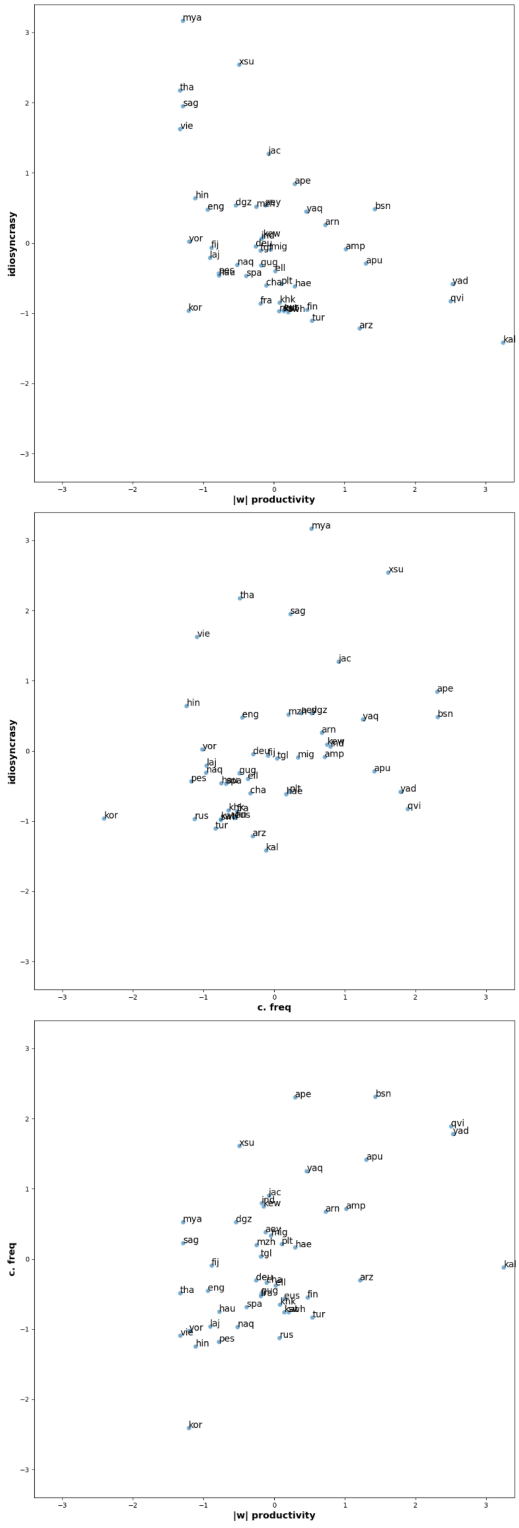
Our research lies at the nexus of computational linguistics and linguistic typology, and enables improvements in both directions. It advances text-based morphological typology, complementing traditional analyses, which are not always straightforwardly reproducible and scalable to diverse languages. In turn, the possibility of comparing languages with automatically induced typological knowledge is especially interesting for various downstream applications developed in the face of linguistic diversity.

8. Appendices

A. Complete list of languages (PBC)

iso639_3	Language	Family	$\|W\|$	$c.freq$	$idiosyncrasy$
cha	Chamorro	Austronesian	-0.106	-0.337	-0.602
gug	Paraguayan Guaraní	Afro-Asiatic	-0.186	-0.494	-0.318
fra	French	Indo-European	-0.190	-0.529	-0.861
deu	German	Indo-European	-0.258	-0.299	-0.045
spa	Spanish	Indo-European	-0.390	-0.680	-0.466
naq	Nama (Namibia)	Khoe-Kwadi	-0.523	-0.968	-0.312
hau	Hausa	Afro-Asiatic	-0.781	-0.752	-0.460
pes	Western Farsi	Indo-European	-0.781	-1.179	-0.430
fij	Fijian	Austronesian	-0.887	-0.091	-0.065
laj	Lango (Uganda)	Eastern Sudanic	-0.908	-0.961	-0.210
kor	Korean	Korean	-1.213	-2.410	-0.961
eng	English	Indo-European	-0.937	-0.454	0.479
hin	Hindi	Indo-European	-1.115	-1.243	0.644
yor	Yoruba	Niger-Congo	-1.206	-1.021	0.025
vie	Vietnamese	Austro-Asiatic	-1.329	-1.093	1.629
tha	Thai	Tai-Kadai	-1.333	-0.485	2.176
mig	San Miguel El Grande Mixtec	Oto-Manguean	-0.046	0.334	-0.090
tgl	Tagalog	Austronesian	-0.192	0.039	-0.103
jac	Popti'	Mayan	-0.078	0.909	1.273
aey	Amele	Trans-New Guinea	-0.121	0.383	0.542
kew	West Kewa	Trans-New Guinea	-0.152	0.750	0.089
ind	Indonesian	Austronesian	-0.179	0.798	0.063
mzh	Wichí Lhamtés Güisnay	Matacoan	-0.252	0.198	0.519
xsu	Sanumá	Yanomam	-0.495	1.613	2.542
dgz	Daga	Dagan	-0.542	0.524	0.538
sag	Sango	Niger-Congo	-1.288	0.225	1.951
mya	Burmese	Sino-Tibetan	-1.293	0.525	3.171
bsn	Barasana-Eduria	Tucanoan	1.429	2.311	0.489
arn	Mapudungun	Araucanian	0.728	0.674	0.258
yaq	Yaqui	Uto-Aztecan	0.457	1.252	0.450
ape	Bukiyip	Torricelli	0.293	2.305	0.845
yad	Yagua	Peba-Yaguan	2.528	1.784	-0.581
qvi	Quechua (Imbabura)	Quechuan	2.499	1.891	-0.825
apu	Apurinã	Arawakan	1.304	1.416	-0.286
amp	Alamblak	Sepik	1.017	0.715	-0.083
hae	Eastern Oromo	Pama-Nyungan	0.298	0.167	-0.612
plt	Plateau Malagasy	Austronesian	0.104	0.217	-0.578
kal	Kalaallisut	Eskimo-Aleut	3.246	-0.119	-1.415
arz	Egyptian Arabic	Afro-Asiatic	1.214	-0.305	-1.211
tur	Turkish	Altaic	0.540	-0.835	-1.106
fin	Finnish	Uralic	0.472	-0.546	-0.945
swh	Swahili	Niger-Congo	0.204	-0.755	-0.982
kat	Georgian	Kartvelian	0.141	-0.757	-0.968
eus	Basque	Basque	0.137	-0.564	-0.955
khk	Halh Mongolian	Altaic	0.081	-0.649	-0.845
rus	Russian	Indo-European	0.071	-1.127	-0.969
ell	Modern Greek	Indo-European	0.018	-0.374	-0.399

B. Two-dimensional XZ,YZ and XY planes (PBC)



C. BPE vs WordPiece, English example

Table 10: The first 30 merge operations in BPE and Wordpiece (English). The </w> symbol indicates the end of a word (BPE). The ## symbol indicates any position that is not the beginning of a word (WordPiece).

BPE				**WP**			
Subword	$\|W\|$	c. freq	idiosyncrasy	Subword	$\|W\|$	c. freq	idiosyncrasy
th	116	4188	36.10	ex	14	23	1.64
an	102	2630	25.78	of	12	736	61.33
and</w>	11	2197	199.73	exc	5	10	2.00
the	46	1536	33.39	qu	7	13	1.86
the</w>	3	1459	486.33	##qu	4	7	1.75
hi	42	1206	28.71	ev	9	66	7.33
to</w>	5	1088	217.60	##bb	3	27	9.00
in	212	996	4.70	up	5	125	25.00
ed</w>	270	912	3.38	exp	2	2	1.00
ha	49	781	15.94	##bj	1	1	1.00
sa	48	777	16.19	##ubj	1	1	1.00
he</w>	5	2211	442.20	subj	1	1	1.00
ou	119	743	6.24	##'s	22	40	1.82
of</w>	3	712	237.33	##s'	3	4	1.33
er	149	709	4.76	##ws'	1	2	2.00
ea	133	679	5.11	##gs'	1	1	1.00
th</w>	107	651	6.08	##us'	1	1	1.00
him</w>	1	617	617.00	##sus'	1	1	1.00
ll</w>	23	592	25.74	##abb	2	26	13.00
at</w>	4	550	137.50	##ubb	1	1	1.00
un	38	536	14.11	rubb	1	1	1.00
en</w>	50	508	10.16	sabb	1	25	25.00
that</w>	1	489	489.00	##rabb	1	1	1.00
or	91	467	5.13	rubbi	1	1	1.00
es	125	458	3.66	sabba	1	25	25.00
unto</w>	2	448	224.00	sabbat	1	25	25.00
they</w>	1	442	442.00	##arabb	1	1	1.00
in</w>	21	442	21.05	barabb	1	1	1.00
ing</w>	110	438	3.98	barabba	1	1	1.00
er</w>	79	437	5.53	barabbas	1	1	1.00

D. Other corpora

Results for the JW300 and UDHR corpora. We only focus on the subsets of languages that intersect with the PBC languages. We overlay the data points corresponding to the PBC corpus as a reference. Languages maintain similar positions in the space despite the different corpora sizes and registers.

D.1 JW300

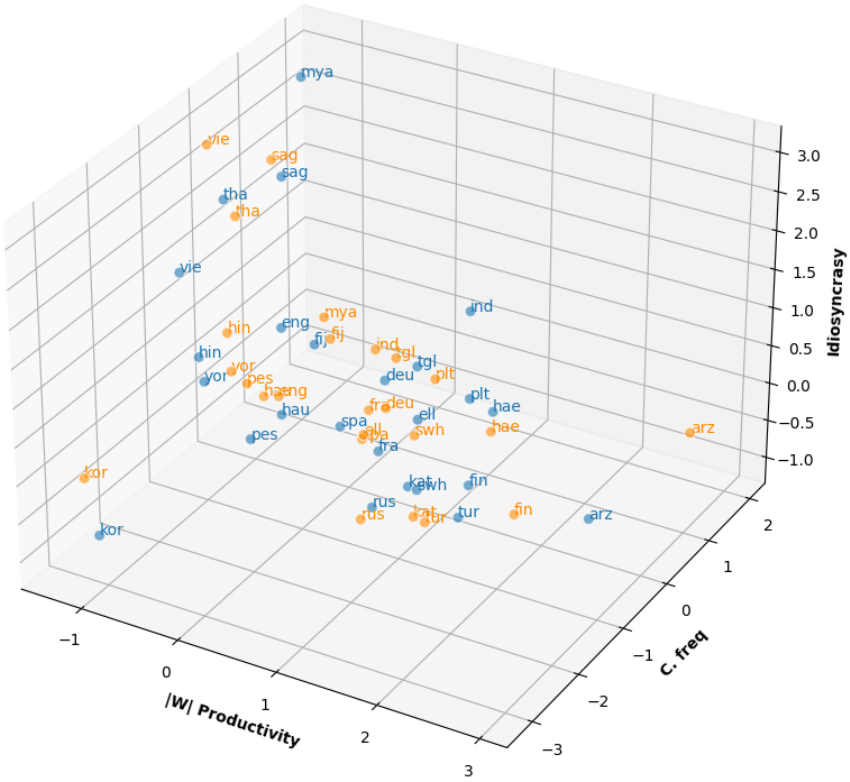


Figure 10: Languages corresponding to the intersection between JW300 (orange dots) and PBC (blue dots).

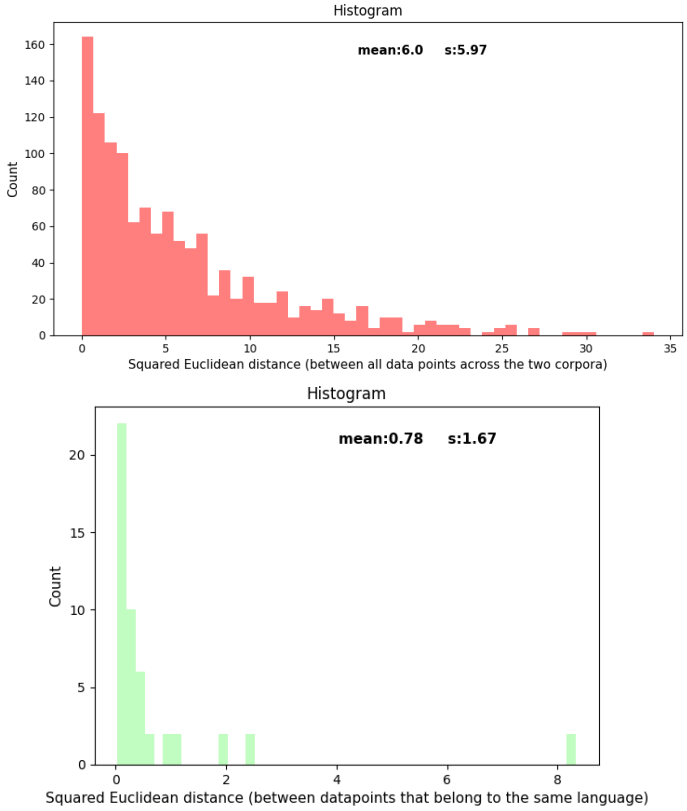


Figure 11: The upper figure shows the distribution of the Euclidean distances across all data points between JW300 and PBC. The lower figure, in contrast, shows the distribution taking into account only the distances between the data points belonging to the same language. If the BPE vector representations for the same languages are similar in the two corpora, then it is expected that the latter distance distribution has generally lower values, i.e. a lower mean compared to the overall distance distribution. This is what we find indeed. Note that the scales on the x-axes are not the same.

D.2 UDHR

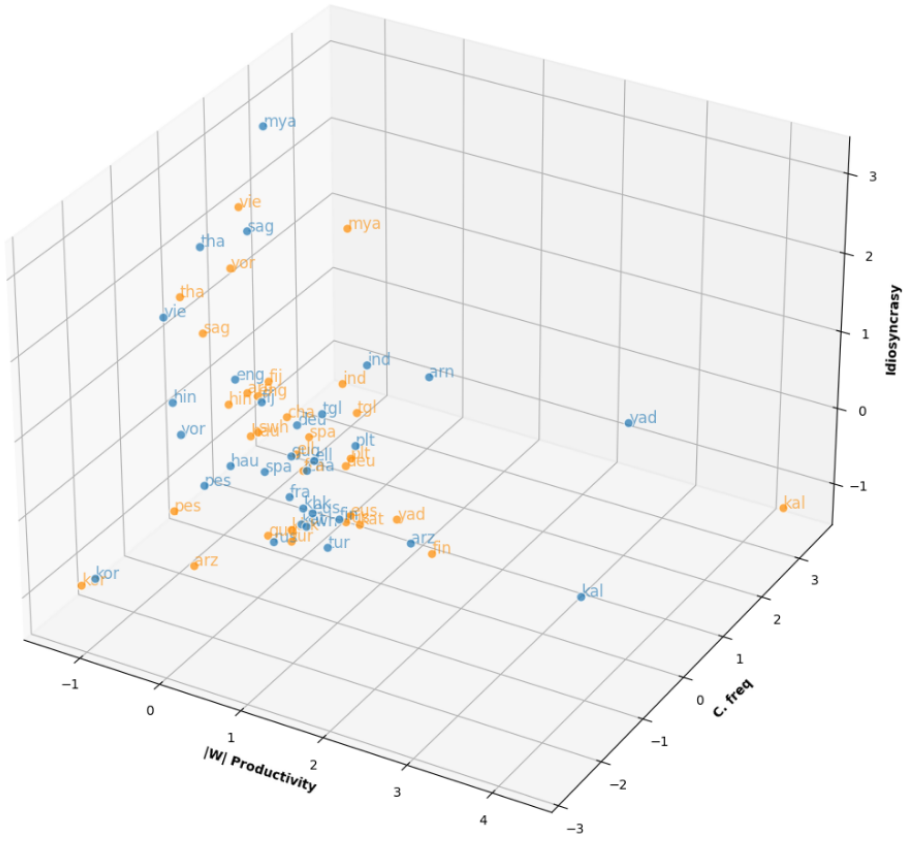


Figure 12: Languages corresponding to the intersection between UDHR (orange dots) and PBC (blue dots).

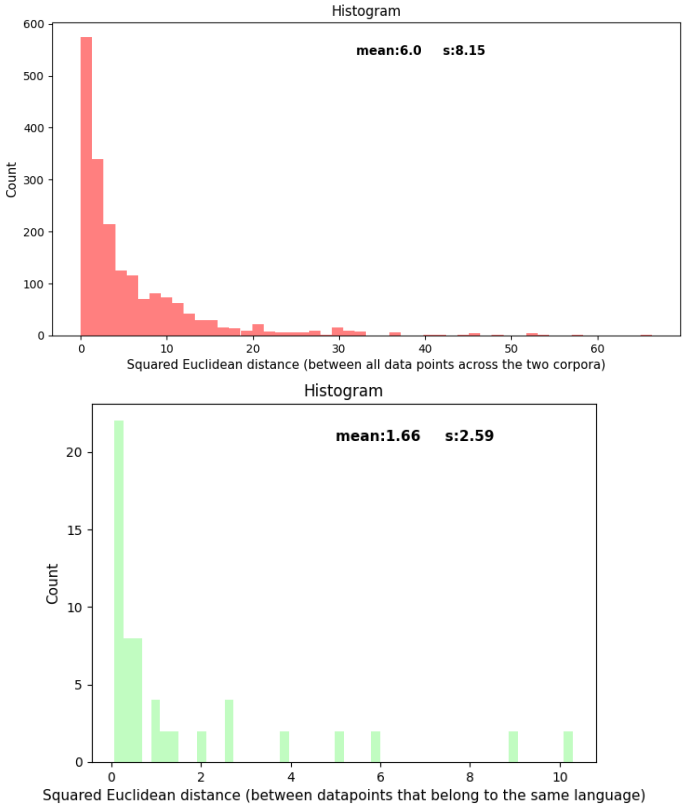


Figure 13: The upper figure shows the distribution of the Euclidean distance across all data points between UDHR and PBC. The lower figure shows the distribution taking into account the distances only between the data points belonging to the same language in the different corpora. If the BPE vector representations for the same languages are similar in the two corpora, then it is expected that this distance distribution has lower values, i.e. a lower mean compared to the overall distance distribution. This is what we find indeed. Note that the scales on the x-axes are not the same.

E. Zoomed-in regions of the space

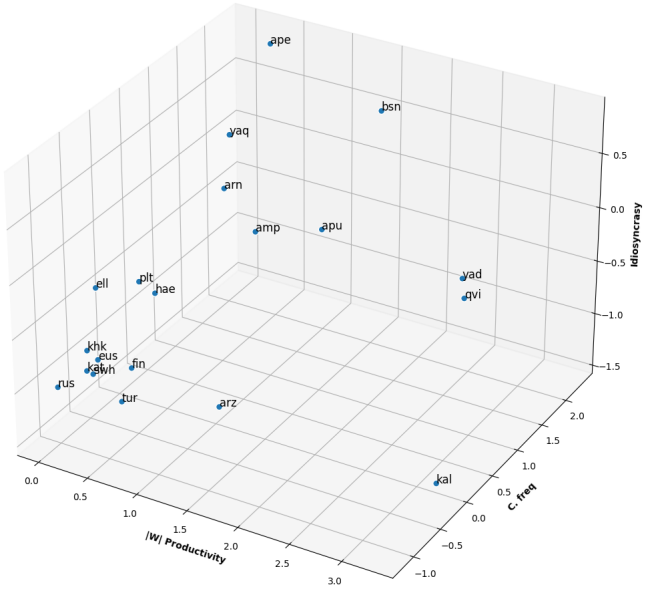


Figure 14: Zoom into the languages with productivity higher than the mean (PBC).

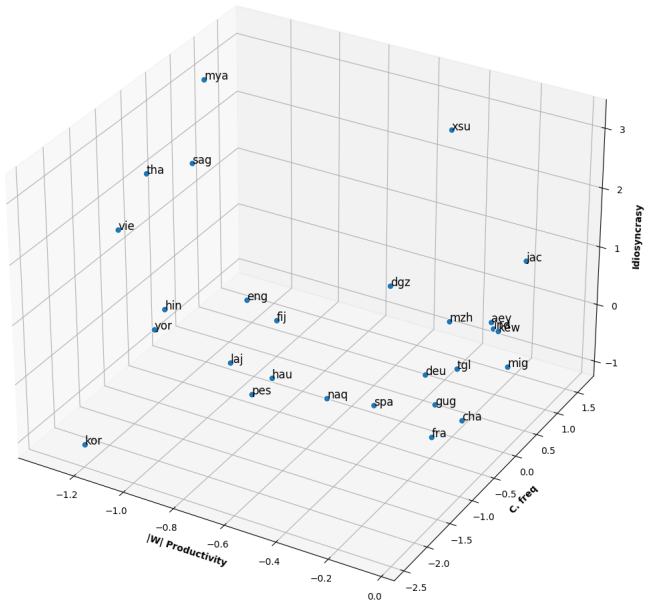


Figure 15: Zoom into the languages with productivity lower than the mean (PBC).

F. BPE space with transliterated Korean

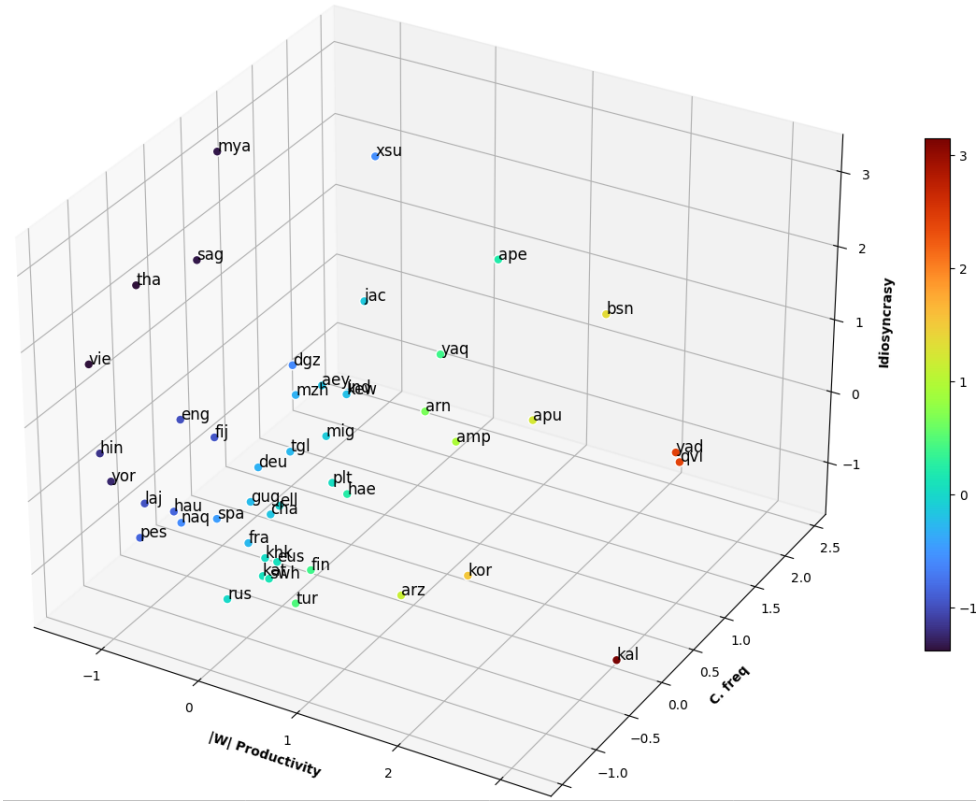


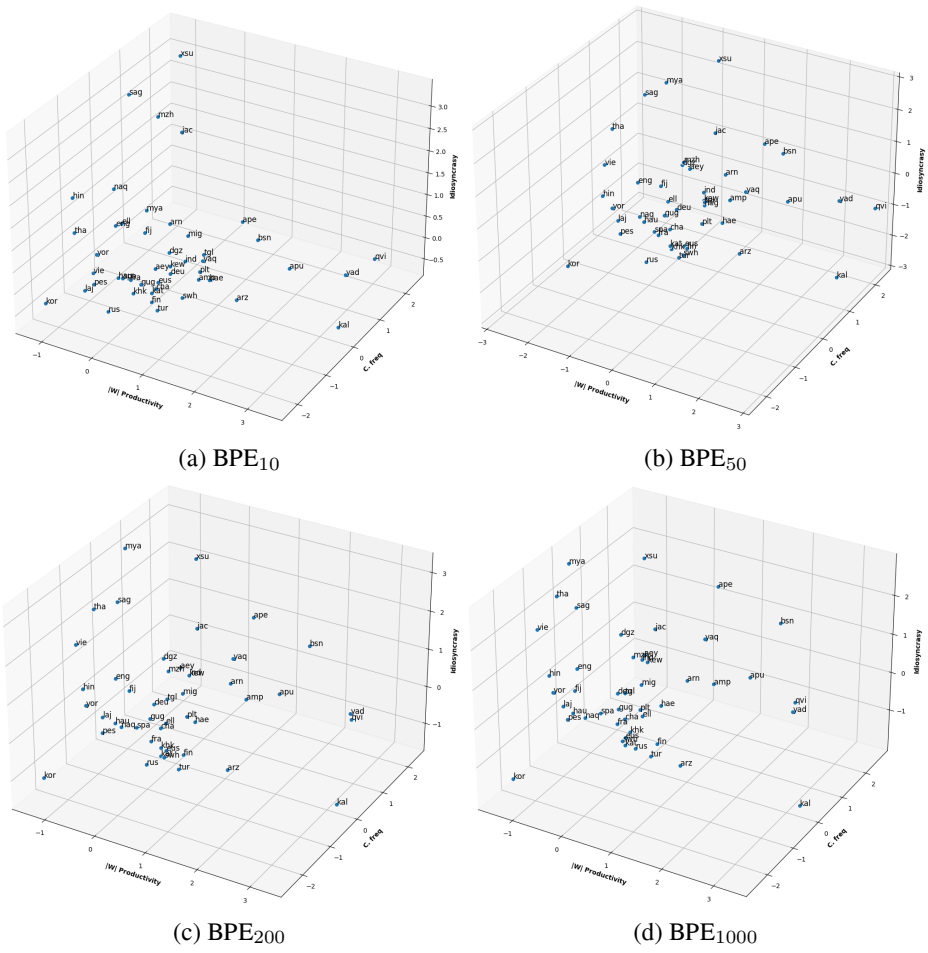
Figure 16: BPE space using a transliterated version of Korean texts (PBC). The transliteration was performed using the Google Cloud translation API.

G. Silhouette coefficients for $\text{WALS}_{\text{BPE200}}$

iso639_3	**Language**	$s(i)$
khk	Halh Mongolian	0.72670988
rus	Russian	0.72581551
kat	Georgian	0.72460184
fra	French	0.72283858
eus	Basque	0.70944521
spa	Spanish	0.70728078
cha	Chamorro	0.67089151
tur	Turkish	0.66468223
fin	Finnish	0.65591473
pes	Western Farsi	0.64035977
ell	Modern Greek	0.63801056
yad	Yagua	0.49543719
tha	Thai	0.49362442
qvi	Quechua (Imbabura)	0.49228889
deu	German	0.49199166
kor	Korean	0.45584722
vie	Vietnamese	0.44809369
sag	Sango	0.43579276
hae	Eastern Oromo	0.38523573
plt	Plateau Malagasy	0.37097495
apu	Apurinã	0.36937643
mya	Burmese	0.33072413
kal	Kalaallisut	0.3208212
tgl	Tagalog	0.31217467
aey	Amele	0.19794441
jac	Popti’	0.12265152
amp	Alamblak	0.10459679
ape	Bukiyip	0.07443282
mzh	Wichí Lhamtés Güisnay	0.0520436
xsu	Sanumá	0.02065723
eng	English	-0.08609748
hin	Hindi	-0.10442254
mig	San Miguel El Grande Mixtec	-0.18355666
arn	Mapudungun	-0.26703156
arz	Egyptian Arabic	-0.31923804
bsn	Barasana-Eduria	-0.3608085
yor	Yoruba	-0.39711486
fij	Fijian	-0.47197221
yaq	Yaqui	-0.50979933
ind	Indonesian	-0.57322525
kew	West Kewa	-0.58901974
laj	Lango (Uganda)	-0.66789
dgz	Daga	-0.67224107
gug	Paraguayan Guaraní	-0.71348287
hau	Hausa	-0.72020938
naq	Nama (Namibia)	-0.74049713
swh	Swahili	-0.76276699

H. Varying the BPE hyperparameter and merging criteria

H.1 BPE space at different merges

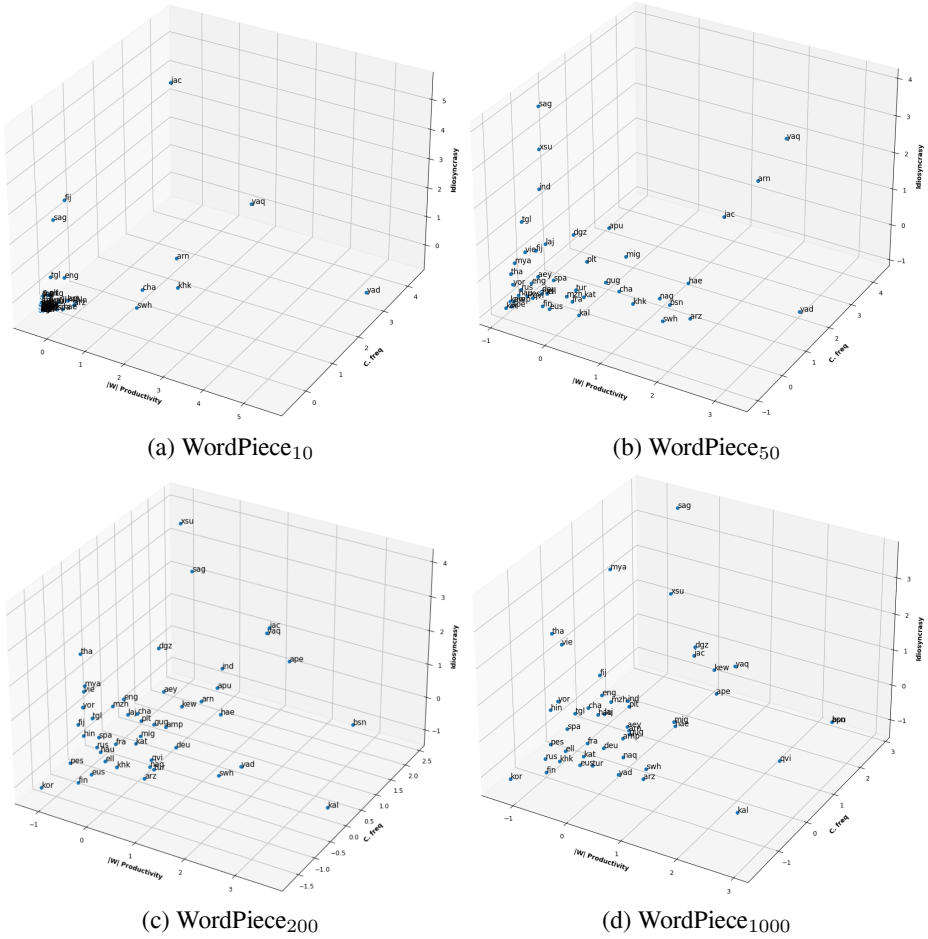


(a) BPE_{10} (b) BPE_{50}

(c) BPE_{200} (d) BPE_{1000}

Figure 17: BPE spaces using different numbers of merge operations. While the first merge operations have a stronger impact on the arrangement of lamguages, later merges (around 200) do not cause drastic changes anymore.

H.2 WordPiece space at different merges



(a) WordPiece$_{10}$ (b) WordPiece$_{50}$

(c) WordPiece$_{200}$ (d) WordPiece$_{1000}$

Figure 18: WordPiece spaces using different numbers of merge operations. Unlike BPE, WordPiece exhibits more substantial changes through merge operations.

9. Acknowledgments

We thank the reviewers for their useful feedback. This research was supported by the Swiss National Science Foundation (SNSF) grant 176305.

References

Adelaar, Willem. 2004. Quechua. In Geert Booij, Christian Lehmann, Joachim Mugdan, and Stavros Skopeteas, editors, *Morphology: An international handbook of inflection and word-formation*, volume 2. Walter de Gruyter, Berlin, pages 1454–1463.

Agić, Željko and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Association for Computational Linguistics, Florence, Italy.

Aikhenvald, Alexandra Y. 2007. Typological distinctions in word-formation. In Timothy Shopen, editor, *Language typology and syntactic description. Volume 3: Grammatical categories and the lexicon*. pages 1–26.

Aikhenvald, Alexandra Y. 2017. Polysynthetic structures of Lowland Amazonia. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*. Oxford University Press, Oxford, pages 284–311.

Al-Rfou, Rami. 2015. *Polyglot: a massive multilingual natural language processing pipeline*. Ph.D. thesis, State University of New York at Stony Brook.

Al Roumi, Fosca, Sébastien Marti, Liping Wang, Marie Amalric, and Stanislas Dehaene. 2021. Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron*, 109(16):2627–2639.

Amrhein, Chantal and Rico Sennrich. 2021. How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Association for Computational Linguistics, Punta Cana, Dominican Republic.

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

Baayen, Harald. 1993. On frequency, transparency and productivity. In *Yearbook of morphology 1992*. Springer, pages 181–208.

Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pages 142–153.

Berg, Thomas. 2014. On the relationship between type and token frequency. *Journal of quantitative linguistics*, 21(3):199–222.

Bickel, Balthasar and Johanna Nichols. 2007. Inflectional morphology. In Timothy Shopen, editor, *Language typology and syntactic description. Volume 3: Grammatical categories and the lexicon*. pages 169–240.

Bickel, Balthasar and Johanna Nichols. 2013a. Fusion of selected inflectional formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bickel, Balthasar and Johanna Nichols. 2013b. Fusion of selected inflectional formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B Lowe. 2022. The autotyp database (v1.0.1).

Bickel, Balthasar and Fernando Zúñiga. 2017. The word in polysynthetic languages: phonological and syntactic challenges. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*. Oxford University Press, Oxford, pages 158–185.

Bjerva, Johannes and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, Association for Computational Linguistics, New Orleans, Louisiana.

Bjerva, Johannes, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1529–1540, Association for Computational Linguistics, Minneapolis, Minnesota.

Bjerva, Johannes, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Association for Computational Linguistics, Online.

Blevins, James P, Petar Milin, and Michael Ramscar. 2017. The zipfian paradigm cell filling problem. In *Perspectives on Morphological Organization*. Brill, pages 139–158.

Blevins, Terra and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

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

Borgman, Donald M. 1990. Sanuma. In Desmond C. Derbyshire and Geoffrey K. Pullum, editors, *Handbook of Amazonian Languages 2*. Mouton de Gruyter, Berlin, pages 15–248.

Bostrom, Kaj and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Association for Computational Linguistics, Online.

Bruce, Les et al. 1984. *The Alamblak language of Papua New Guinea (East Sepik)*. Dept. of Linguistics, Research School of Pacific Studies, The Australian ⋯.

Bybee, Joan. 2003. *Phonology and language use*, volume 94. Cambridge University Press.

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

Ferrer-i Cancho, Ramon. 2018. Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3):207–237.

Ferrer-i Cancho, Ramon, Christian Bentz, and Caio Seguin. 2022. Optimal coding and the origins of zipfian laws. *Journal of Quantitative Linguistics*, 29(2):165–194.

Ferrer-i Cancho, Ramon, Antoni Hernández-Fernández, David Lusseau, Govindasamy Agoramoorthy, Minna J Hsu, and Stuart Semple. 2013. Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8):1565–1578.

Choenni, Rochelle and Ekaterina Shutova. 2022. Investigating Language Relationships in Multilingual Sentence Encoders Through the Lens of Linguistic Typology. *Computational Linguistics*, 48(3):635–672.

Clark, Jonathan H., Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Creutz, Mathias and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30, Association for Computational Linguistics.

Cuskley, Christine, Francesca Colaiori, Claudio Castellano, Vittorio Loreto, Martina Pugliese, and Francesca Tria. 2015. The adoption of linguistic rules in native and non-native speakers: Evidence from a wug task. *Journal of Memory and Language*, 84:205–223.

Cysouw, Michael and Bernhard Wälchli. 2007. Parallel texts: using translational equivalents in linguistic typology. *STUF-Sprachtypologie und Universalienforschung*, 60(2):95–99.

Dahl, Östen. 2017. Polysynthesis and complexity. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*. Oxford University Press, Oxford, pages 19–29.

Daniels, Peter T and William Bright. 1996. *The world's writing systems*. Oxford University Press, Oxford.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota.

Dixon, Robert M. W. 1988. *A Grammar of Boumaa Fijian*. University of Chicago Press, Chicago.

Dixon, Robert MW and Alexandra Y Aikhenvald. 2003. *Word: A cross-linguistic typology*. Cambridge University Press, Cambridge.

Domingo, Miguel, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2023. How much does tokenization affect neural machine translation? In *Computational Linguistics and Intelligent Text Processing*, pages 545–554, Springer Nature Switzerland, Cham.

Dryer, Matthew S. 2013. Prefixing vs. suffixing in inflectional morphology. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dryer, Matthew S and Martin Haspelmath. 2013. The world atlas of language structures online.

Ehret, Katharina. 2016. *An information-theoretic approach to language complexity: variation in naturalistic corpora*. Ph.D. thesis, Albert-Ludwigs-Universität Freiburg.

Ehret, Katharina and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler and Guido Seiler, editors, *Complexity, isolation and variation*. de Gruyter, Berlin.

Facundes, Sidi. 2014. Negation in apurinã (arawak). In *Negation in Arawak languages*. Brill, pages 121–146.

Fortescue, Michael. 1992. The development of morphophonemic complexity in eskimo languages. *Acta Linguistica Hafniensia*, 25(1):5–27.

Fortescue, Michael D. 1984. *West Greenlandic*. Croom Helm London.

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

Gallé, Matthias. 2019. Investigating the effectiveness of bpe: The power of shorter sequences. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1375–1381.

Geertzen, Jeroen, James Blevins, and Petar Milin. 2016. The informativeness of linguistic unit boundaries.

Göksel, Aslı and Celia Kerslake. 2004. *Turkish: A comprehensive grammar*. Routledge, London/New York.

Goldsmith, John. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

Gomez-Imbert, Elsa. 2004. Fonología de dos idiomas tukano del piraparaná: barasana y tatuyo. *Amerindia*, 29(30):43–80.

Greenberg, J. 1966. Language universals: With special reference to feature hierarchies.

Greenberg, Joseph H. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3):178–194.

Grönroos, Stig-Arne, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185.

Guerrero, Lilián. 2019. Supleción en yaqui y wixárika. *Lingüística Mexicana*, 1(2):119–140.

Gutierrez-Vasques, Ximena, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Association for Computational Linguistics, Online.

Haspelmath, Martin, editor. 2008. *Language Typology and Language Universals / Sprachtypologie und sprachliche Universalien / La typologie des langues et les universaux linguistiques: Eu - L*. De Gruyter Mouton.

Haspelmath, Martin. 2017. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica*, 51(s1000):31–80.

Haspelmath, Martin and Andrea Sims. 2010. *Understanding morphology*. Hodder Education.

Hewitt, Brian George. 1995. *Georgian: A structural reference grammar*, volume 2. John Benjamins Publishing.

Jenny, Mathias and San San Hnin Tun. 2016. *Burmese: A comprehensive grammar*. Routledge, New York/London.

Johnston, Iain G, Kamaludin Dingle, Sam F Greenbury, Chico Q Camargo, Jonathan PK Doye, Sebastian E Ahnert, and Ard A Louis. 2022. Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution. *Proceedings of the National Academy of Sciences*, 119(11):e2113883119.

Juola, Patrick. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Karan, Elke. 2006. Writing system development and reform: A process.

Kelih, Emmerich. 2010. The type-token relationship in slavic parallel texts. *Glottometrics*, 20:1–11.

Kirby, Simon, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Kudo, Taku. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Association for Computational Linguistics, Melbourne, Australia.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Association for Computational Linguistics, Online.

Lin, Yu-Hsiang, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Association for Computational Linguistics, Florence, Italy.

Littell, Patrick, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Association for Computational Linguistics, Valencia, Spain.

Lonsdale, Arthur Walter. 1899. *Burmese Grammar and Grammatical Analysis*. British Burma Press, Rangoon.

Macháček, Dominik, Jonáš Vidra, and Ondřej Bojar. 2018. Morphological and language-agnostic word segmentation for nmt. In *International Conference on Text, Speech, and Dialogue*, pages 277–284, Springer.

Mager, Manuel, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Association for Computational Linguistics, Dublin, Ireland.

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Association for Computational Linguistics, Copenhagen, Denmark.

Malkin, Dan, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Mayer, Thomas, Bernhard Wälchli, Christian Rohrdantz, and Michael Hund. 2014. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. *Language Processing and Grammars. The role of functionally oriented computational models*, pages 13–38.

Mielke, Sabrina J, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Mielke, Sebastian J and Jason Eisner. 2019. Spell once, summon anywhere: A two-level open-vocabulary language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6843–6850.

Moran, Steven and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press.

Moravcsik, Edith A. 2012. *Introducing language typology*. Cambridge University Press, Cambridge.

Myung, IJ. 2001. Computational approaches to model evaluation.

Nzeyimana, Antoine and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 5347–5363, Association for Computational Linguistics, Dublin, Ireland.

Oncevay, Arturo, Duygu Ataman, Niels Van Berkel, Barry Haddow, Alexandra Birch, and Johannes Bjerva. 2022. Quantifying synthesis and fusion and their impact on machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1308–1321, Association for Computational Linguistics, Seattle, United States.

Östling, Robert. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211.

Palmer, Bill, editor. 2017. *The Languages and Linguistics of the New Guinea Area: A Comprehensive Guide*. De Gruyter Mouton.

Pelleg, Dan, Andrew W Moore, et al. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734.

Pelloni, Olga, Anastassia Shaitarova, Olga Sozinova, and Tanja Samardzic. 2022. Subword evenness (sue) as a predictor of cross-lingual transfer to low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Main Volume*, pages 7428––7445, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Pinker, Steven. 1991. Rules of language. *Science*, 253(5019):530–535.

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Association for Computational Linguistics, Florence, Italy.

Ponti, Edoardo Maria, Helen O' Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rathi, Neil, Michael Hahn, and Richard Futrell. 2021. An information-theoretic characterization of morphological fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10115–10120, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Reynar, Jeffrey C, Fred Herz, Jason Eisner, Lyle Ungar, et al. 1999. Lempel-ziv data compression technique utilizing a dictionary pre-filled with frequent letter combinations, words and/or phrases. US Patent 5,951,623.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rousseeuw, Peter J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Ruder, Sebastian, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Rust, Phillip, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Association for Computational Linguistics, Online.

Ryding, Karin C. 2005. *A reference grammar of Modern Standard Arabic*. Cambridge University Press, Cambridge.

Saleva, Jonne and Constantine Lignos. 2021. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Association for Computational Linguistics, Online.

Sapir, Edward. 1921. *Language: An introduction to the study of speech*. Harcourt, Brace.

Schuster, Mike and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics,*

speech and signal processing (ICASSP), pages 5149–5152, IEEE.

Sennrich, Rico, 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, Association for Computational Linguistics, Berlin, Germany.

Shapiro, Pamela and Kevin Duh. 2018. Bpe and charcnns for translation of morphology: A cross-lingual comparison and analysis. *arXiv preprint arXiv:1809.01301*.

Shimelman, Aviva. 2017. *A grammar of Yauyos Quechua*. Language Science Press, Berlin.

Stolz, Thomas. 2015. Chamorro inflection. In *The Oxford Handbook of Inflection*.

Storer, James A and Thomas G Szymanski. 1978. The macro model for data compression. In *Proceedings of the tenth annual ACM symposium on Theory of computing*, pages 30–39.

Stump, Gregory T. 2017. *Inflection*, chapter 1. John Wiley Sons, Ltd.

Tamariz, Mónica and Simon Kirby. 2015. Culture: copying, compression, and conventionality. *Cognitive science*, 39(1):171–183.

Ullman, Michael T. 1999. Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive processes*, 14(1):47–67.

Üstün, Ahmet, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Association for Computational Linguistics, Online.

Wälchli, Bernhard and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs.

Wu, Shijie, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Association for Computational Linguistics, Florence, Italy.

Yeon, Jaehoon and Lucien Brown. 2011. *Korean: A comprehensive grammar*. Routledge, New York.

Ziv, Jacob and Abraham Lempel. 1977. A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343.

Zúñiga, Fernando. 2017. Mapudungun. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*. Oxford University Press, Oxford, pages 696–713.