From characters to words: the turning point of BPE merges

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

Abstract

The distributions of orthographic word types are very different across languages due to typological characteristics, different writing traditions, and other factors. The wide range of cross-linguistic diversity is still a major challenge for NLP, and for the study of language more generally. We use BPE and information-theoretic measures to investigate if distributions become more similar under specific levels of subword tokenization. We perform a cross-linguistic comparison, following incremental BPE merges (we go from characters to words) for 47 diverse languages. We show that text entropy values (a feature of probability distributions) converge at specific subword levels: relatively few BPE merges (around 200 for our corpus) lead to the most similar distributions across languages. Additionally, we analyze the interaction between subword and word-level distributions and show that our findings can be interpreted in light of the ongoing discussion about different morphological complexity types.[1]

1 Introduction

In NLP, one of the predominant methods for obtaining subword units is Byte-Pair Encoding (BPE). These subwords have proven to be useful for improving several NLP tasks, most likely because they capture morphological patterns to some extent (and also phonological and orthographic ones).

BPE is based on a compression algorithm which finds frequently ocurring patterns in a text by means of incrementally merging adjacent symbols into longer strings (Gage, 1994; Sennrich et al., 2015). The granularity of the subword units is controlled by the number of merge operations applied to the text (few merges lead to a text tokenization closer to the character level, while more merges lead to a tokenization closer to the word level). Usually the number of BPE merges is chosen arbitrarily depending on the application.

It is rarely analyzed how the distribution of these subwords changes across different merge operations. Our goal is to investigate if languages get 'closer' in terms of their subword distributions under specific levels of tokenization. We quantify this cross-linguistic variation using information-theoretic measures.

Information theory provides a useful tool for exploring variation, and for quantifying the predictability/organization of patterns, e.g., in morphological systems (Ackerman and Malouf, 2013). We measure Shannon entropy and redundancy over varied subword tokenizatons of texts obtained with BPE. At each incremental merge, we compare the values across 47 typologically diverse languages.

Cross-linguistic corpora are widely used as a means of quantifying linguistic diversity. For instance, the range of entropy values measured over word-level types varies greatly across languages. This is a reflection of the diversity of morphological systems. However, we show that this cross-linguistic variation is not so pronounced at the subword level. Namely, a convergence of entropy values across languages is achieved at a relatively low number of merge operations. The entropy of subword distributions grows quickly before this turning point, while the growth is considerably slower after it.

Furthermore, in this turning point, the subword distributions start to correlate with the ones observed at the word-level. We interpret this change of trend in light of previous findings regarding the difference between subword and word-level complexity: a language that is complex at the word level (rich inflectional morphology), is not necessarily complex at a more atomic subword level (predictable subword patterns).

[1]Data and code available at https://github.com/ximenina/theturningpoint

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3454–3468
April 19 - 23, 2021.

2 Background

2.1 Byte-pair Encoding (BPE)

Originally, BPE is a data compression technique based on replacing the most common pair of consecutive bytes with a new symbol (Gage, 1994). This is currently one of the predominant approaches for subword tokenization (or morphological segmentation). It is widely used to improve tasks like machine translation or language modeling (Sennrich et al., 2015; Provilkov et al., 2020). Another popular method is provided by Morfessor (Smit et al., 2014).

When BPE is applied to text, each iteration merges two adjacent symbols. The main hyperparameter of BPE is the number of merge operations applied to the data, which controls the granularity of the subword units. In NLP, this hyperparameter is usually chosen empirically, e.g., based on the dataset size or on the task, regardless of the typological features of a specific language.

2.2 Text and information theory

Bentz et al., (2016) distinguish between corpus-based and paradigm-based approaches for quantifying morphological complexity. While the former approaches measure morphological productivity directly on raw text corpora, the latter make use of higher level language descriptions, i.e., grammars, and inflectional paradigms.

In corpus-based approaches, a text is usually regarded as a sequence of symbols. Each symbol is generated with a certain probability, and hence carries a certain information content (Juola, 1998; Ehret and Szmrecsanyi, 2016a; Ehret, 2016b; Koplenig et al., 2017; Bentz et al., 2017). The higher the probability of a symbol, the lower its information content. Against this backdrop, the average information content of a text can be estimated by Shannon entropy, and approximated with type-token-ratios (TTR). For instance, if we consider orthographic words as symbols, languages with a greater diversity of word types will have higher entropy (word types are less predictable, due to, e.g., richer morphology). In fact, such corpus-based measures have been shown to be correlated also with paradigm-based approaches that quantify morphosyntactic distinctions based on grammars (Bentz et al., 2016; Kirov et al., 2017).

The complexity of the morphological system of a language is not only related to the diversity of word types that can be produced, but also to the way in which subwords are organized within them.

On the corpus-based side, there are some studies which have focused on the predictability of internal word structure. For instance, there are cross-linguistic accounts illustrating the trade-off between the size of syllables and the size of words: languages with structurally simple and short syllables need more syllables for encoding the same content (Fenk-Oczlon and Fenk, 1999; Coupé et al., 2019). Another line of research proposes to quantify the amount of word-internal information by comparing the (character level) entropy of the original text with a version where the regularities within orthographic words have been masked (Juola, 1998; Ehret and Szmrecsanyi, 2016a; Ehret, 2016b; Koplenig et al., 2017).

Most recently, morphological complexity has been approached through the lense of neural language models, and their learning of subword structure (Vania and Lopez, 2017; Mielke et al., 2019). Gutierrez-Vasques and Mijangos (2020) propose a measure reflecting the predictability of the internal structure of words. It relies on the entropy rate of a neural language model that is trained to predict sequences of character n-grams within a word. We here compare the results of this latest neural network approach with the entropy of subword units based on BPE.

3 Data and methods

Our general proposal comprises calculating several measures over varied subword tokenizatons of texts obtained with BPE. In each consecutive tokenization, we hence regard a different set of strings of characters as symbols of our "alphabet". In merge 0, a text is a sequence of single UTF-8 characters; in the last merges, a text is a sequence closer to orthographic word types (i.e. original tokenization given by white spaces and punctuation). At each incremental step, we compare the values across parallel corpora in 47 typologically diverse languages.

3.1 Parallel corpus

Using parallel corpora facilitates meaningful comparisons across languages, as seen in cross-linguistic studies on morphological typology, lexical typology, and word order typology (Cysouw and Wälchli, 2007; Wälchli and Cysouw, 2012; Östling, 2015; Kelih, 2010; Mayer et al., 2014). In fact, the idea to compare language complexity through parallel corpora can be traced back to

Greenberg (1960).

In this work, we use a publicly available parallel corpus for 47 languages that was extracted from the Parallel Bible Corpus (PBC) (Mayer and Cysouw, 2014). This specific dataset[2] contains 115 preprocessed parallel verses per language consistently coded in UTF-8. The set of 47 languages is a subset of the WALS 100-language sample, which aims to maximize both genealogical and areal diversity. See the list of languages, their ISO639-3 code and linguistic families in Appendix A.

3.2 Scripts and Writing Systems

The respective texts are written in different scripts (Arabic, Cyrillic, Devanagari, Georgian (Mkhedruli), Korean Hangul, Latin, Modern Greek, Myanmar (Burmese), Thai), and reflect different writing systems (abugida, abjad, alphabet, syllabary). Since BPE starts to operate at the level of UTF-8 characters, the ideosyncrasies of encodings are relevant for our analyses. For instance, the word *beginning* in English consists of 9 UTF-8 character tokens and 5 types ('b','e','g','i','n'), while the corresponding word written in Korean Hangul 시작 (transliterated as 'sijak') consists of two syllable blocks, namely 시 ('si') and 작 ('jak'). It is these syllable blocks – rather than individual letters of the Korean alphabet – which are represented as UTF-8 characters. Thus, while English texts typically contain 26 UTF-8 character types (bare punctuation), Korean texts might display hundreds and thousands. A similar proliferation of UTF-8 characters is found in texts written with Abugidas (e.g. Hindi, Thai, Burmese), or in latinized scripts with many special characters and diacritics (e.g. Vietnamese).

3.3 BPE merge operations

The BPE algorithm starts by splitting words into a sequence of characters. We can think of this as characters separated by white spaces. In the first operation, the algorithm merges the most frequent pair of consecutive characters within the corpus, e.g., ('e','d') $\rightarrow$ ('ed'), thus creating a new symbol that is added to the vocabulary. In each of the following operations, the algorithm calculates the co-occurrence frequency of pairs of all the current consecutive symbols and it merges again the most frequent pair.

When the algorithm merges a frequent pair of symbols, it automatically removes many of the white spaces in the text (this is one aspect of how BPE achieves text compression). As more merges are applied, longer symbols (in terms of number of characters) are obtained – we are getting closer to the 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 1.

A worked out toy example can be found in Table 1. Note that symbols occurring at the end of a word are considered different from the ones that occur at any other position. The symbols that are merged in BPE are hence character sequences of variable sizes. These can be interpreted as subword units. This is why BPE is usually seen as a morphological segmentation technique in NLP.

We applied an existing BPE implementation[3] to the texts of the parallel corpus. For each language, we obtain many different segmented versions of the text depending on the number of merges applied. In specific, we go from merge 0 to merge 10K.

We traverse the range of merge operations by using different step sizes. We simply do this to ease the computational load. Moreover, the different trends that we observe are already stable by merge 350.

- Fine-grained merges:
  0 to 350 (step size: 1)
- Coarse-grained merges:
  350 to 5K merges (step size: 50)
  5k to 10K merges (step size: 1K)

3.4 Information-theoretic measures

Once the texts are segmented, we apply two different information-theoretic measures. Both take as input a text T with a vocabulary of types $V = \{t_1, t_2, ..., t_V\}$ of size $|V|$. At the word level, these types correspond to words (strings separated by spaces). Analogously, at the subword level, the types are the subword units obtained at a specific number of merge operations. We distinguish between the subword units that are at the end of a word and the rest of them, e.g. *-ed* and *-ed-* are considered different types in the vocabulary.[4]

We use **entropy** as a measure of the average information content of types in a text. We can

[2] Dataset from the Interactive Workshop on Measuring Language Complexity (IWMLC 2019)

[3] https://github.com/rsennrich/subword-nmt

[4] We follow this distinction since it is made by the BPE implementation that we use.

Merge	Text Version	Alphabet (Vocabulary of Symbols)
0	g-o-d c-r-e-a-t-e-d t-h-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d t-h-e l-i-g-h-t	a-, c-, d, d-, e, e-, g-, h-, i-, l-, n, n-, o-, r-, t, t-, v-
1	g-o-d c-r-e-a-t-e-d **th**-e h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **th**-e l-i-g-h-t	a-, c-, d, d-, e, e-, g-, h-, i-, l-, n, n-, o-, r-, t, t-, **th-**, v-
2	g-o-d c-r-e-a-t-e-d **the** h-e-a-v-e-n a-n-d g-o-d d-i-v-i-d-e-d **the** l-i-g-h-t	a-, c-, d, d-, e-, g-, h-, i-, l-, n, n-, o-, r-, t, t-, **the**, v-
3	g-**od** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d g-**od** d-i-v-i-d-e-d the l-i-g-h-t	a-, c-, d, d-, e-, g-, h-, i-, l-, n, n-, **od**, r-, t, t-, the, v-
4	**god** c-r-e-a-t-e-d the h-e-a-v-e-n a-n-d **god** d-i-v-i-d-e-d the l-i-g-h-t	a-, c-, d, d-, e-, g-, **god**, h-, i-, l-, n, n-, r-, t, t-, the, v-
5	god c-r-e-a-t-**ed** the h-e-a-v-e-n a-n-d god d-i-v-i-d-**ed** the l-i-g-h-t	a-, c-, d, d-, e-, **ed**, g-, god, h-, i-, l-, n, n-, r-, t, t-, the, v-
6	god c-r-**ea**-t-ed the h-**ea**-v-e-n a-n-d god d-i-v-i-d-ed the l-i-g-h-t	a-, c-, d, d-, e-, **ea-**, ed, g-, god, h-, i-, l-, n, n-, r-, t, t-, the, v-

Table 1: Example of BPE merge operations. Original text: *God created the heaven [...] and God divided the light [...]*

calculate the entropy as follows (Shannon, 1948):

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

Where the probability of a type $p(t)$ is estimated using the so-called maximum likelihood method (i.e. its relative frequency in the text). Higher values of entropy indicate higher complexity (less predictability). We take this as our main measure of text-based morphological complexity through merges.

We also use **redundancy,** a measure that is related to entropy. The entropy of a source of data is maximum when the symbols comprising a message can be chosen freely and they are equiprobable (maximum uncertainty). The redundancy, as defined here, quantifies how close the empirically estimated entropy $H(T)$ is to the maximum value it can take, assuming that we utilize the same alphabet (or types in our case). It can be defined as follows (Partridge, 1981; Karmeshu, 2003):

$$R(T) = 1 - \frac{H(T)}{max\{H(T)\}} = 1 - \frac{H(T)}{\log_2 |V|} \quad (2)$$

Where $H(T)$ is the entropy of a text, calculated as in (1). The maximum entropy can be calculated as $max\{H(T)\} = log_2|V|$, i.e, the entropy when the probability distribution of types is uniform $p(t_i) = \frac{1}{|V|}$. The values of R range from 0 to 1. Values closer to 1 indicate higher redundancy.

3.5 Spearman's rank correlation

We use correlations for exploring the connection between values yielded by the complexity measures described above. We rank languages according to these measures. In particular, we use Spearman's rank correlation, which tests for a correlation between the rankings of two variables (monotonic relationships, not necessarily linear).

We apply Spearman's rank correlation to the following variables:

1. Final merge (the number of merge operations needed to reach the final step of the BPE algorithm);
2. Average word length (at the word level);
3. Size of the vocabulary of characters (of the original texts);
4. Entropy and redundancy measured over the texts at different merges;
5. Two external morphological complexity measures (based on unigrams and trigrams of characters).

Since we have many variables, and hence pairwise correlations, we apply the Bonferroni correction on p-values. We select the correlations that are still significant after correcting for multiple testing, see Appendix C.

4 Results

4.1 From characters to words: the turning point

Figure 1 shows the entropy of languages at different merge operation stages: merge 0, merge 30, merge 200, and the word level. We choose these specific merges for the sake of illustration. In particular, merge 200 is representative of the turning point of several trends.

At the very first merges, texts are closer to character level tokenizations, i.e., small subword units. The initial point is merge 0 (roughly corresponding to the character level). Here, the texts' entropies range between $4.01 - 7.77$ bits. We notice that languages with a larger inventories of UTF-8 characters start with higher entropy values. For instance,

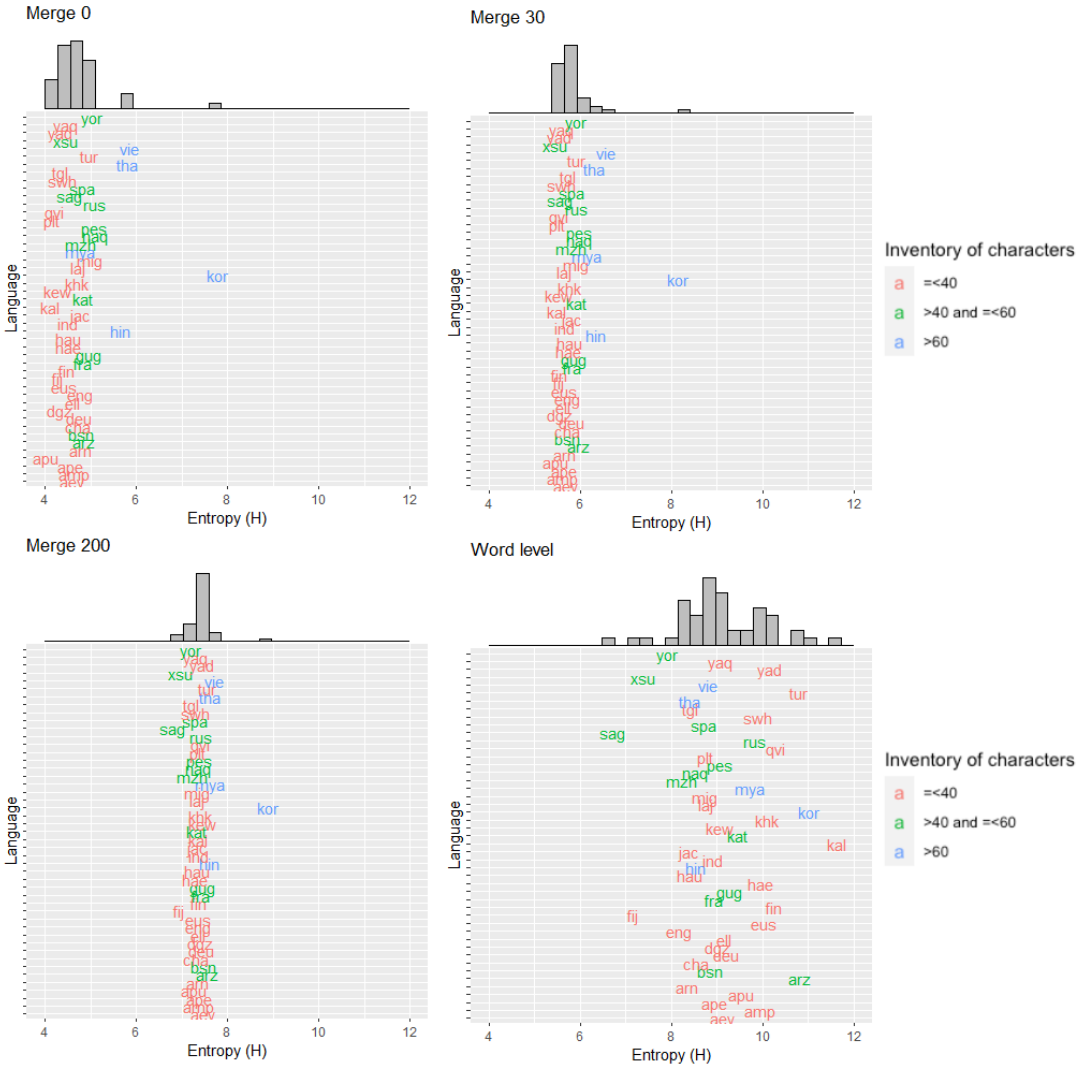

Figure 1: Entropies of languages at different merge operations. Histograms of the distributions are shown above the panels. The sizes of the original UTF-8 character sets are indicated by colors.

Korean *(kor)* is an outlier due to its alpha-syllabary writing system (Section 3.2).

In the subsequent merges, all languages start increasing their entropy. However, the values also start to become less dispersed across languages. In fact, we can see that, in merge 200, the majority of languages are centered around 7.3 bits. This means that, at this merge, the frequency distributions of subwords are similar across languages. Moreover, the size of the initial inventory of characters seems not to affect the texts' entropy at these later merges.

As BPE approaches tokenizations which are closer to orthographic words, entropy values start to disperse again. If we measure entropies over the original texts (without any subword tokenization), we can see that the cross-linguistic variation is considerably wider than the one obtained when the texts are represented by subwords (fewer merges).

This trend is also observable in Fig. 2, where the standard deviation (σ) of entropy across languages is shown as a function of the number of merges. A minimum is reached at merge 200; in fact, between merges 190-240 there is practically no variation of σ. This means that, around this number of merges, the entropies of subword distributions' across languages are closest to one another. After 240 merges, the values start to slowly disperse again and keep dispersing up to the final merges.

Moreover, if we rank languages by their entropy, the rankings obtained before merge 200 are not correlated with the ones observed in later merges (see Section 4.3). From this point onward, the rankings start to gradually correlate with the one observed at the word-level. For instance, at merge 350 (a relatively low number of merges) the vocabularies still contain many short subword units, and the variance

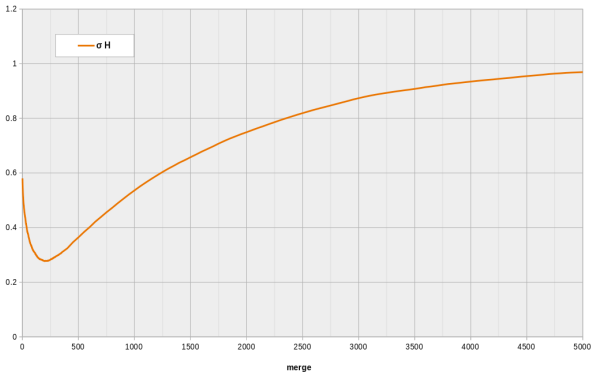

Figure 2: Standard deviations (σ) of entropies across languages per each BPE merge operation.

of entropies across languages is still close to the minimum. Despite of this, the language rankings obtained at this merge are already similar to the rankings observed at the word level.

For some languages, the entropy at the subword level (fewer merges) is systematically different from that at the word level (higher number of merges). For instance, Kalaallisut (*kal*), typically seen as polysynthetic, starts with low H in the first merges – it ranks almost last. However, at merge 200, it is rather in the middle range, and it further increases in entropy with subsequent merges, to the point where it ranks highest of all languages, namely at the word level (Fig. 1).

As we discussed earlier, the entropy (H) has its minimum at merge 0. After each merge operation, H increases. The first merges cause the most drastic changes in the entropy. After the first hundreds of merges, the entropy increases more slowly, i.e., each merge does not cause a big increment of the text entropy anymore. In contrast, redundancy starts decreasing since the first merges. R reaches a minimum after a certain number of merges for all languages (297 merges on average), and then it starts increasing again as the word level is approached. Figure 3 shows an example of the entropy (H) and redundancy (R) across merges for the French text. Appendix B contains the entropy and redundancy curves for all languages.

The first operations merge very frequently adjacent symbols, which impacts the subword distributions of the texts. This is the reason why redundancy and entropy are changing quickly for the early merges. These first merges find the most frequent and productive patterns, e.g., inflectional markers ('-ed' and '-ing'), and orthographic practices for representing sounds (e.g. 'th') in English.

When highly recurrent patterns get merged, the redundancy of the texts is reduced. We can think of this in terms of skewed distributions. At merge 0, the vocabulary's initial distribution of elements is skewed (few UTF-8 characters, high frequencies). When BPE starts merging the most salient patterns, the distribution of subwords gets closer to a uniform distribution (more symbols, lower frequencies), and a minimum of redundancy is reached. After this, the merge operations lead again to skewed distributions (redundancy grows again). However, these latter distributions across languages are correlated to the ones observed at the word level and not to the first merges' skewed distributions.

Interestingly, the number of merges at which entropies start to grow slower, and redundancies reach a minimum, are in the same range of merges in which languages start to change their trends between subword and word level, ca. 200-300 merges. This is also the turning point in which the cross-linguistic standard deviation of entropies reaches a minimum.

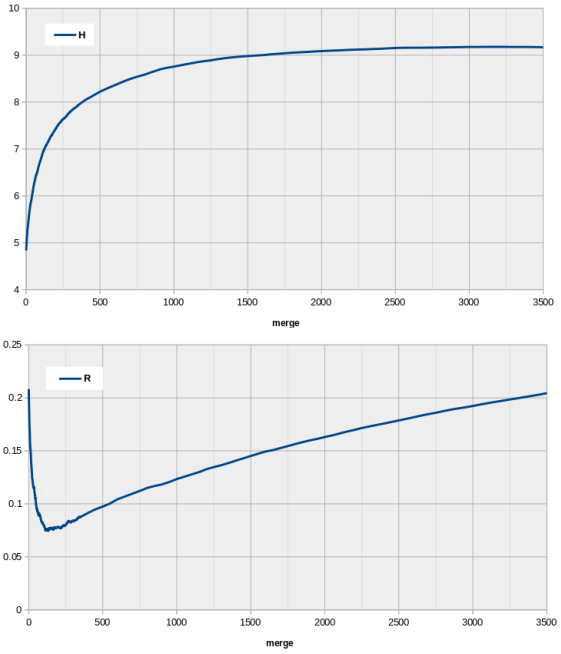

Figure 3: H and R across BPE merges for French (fra).

4.2 Max. number of merges per language

Not all languages require the same number of operations to reach the point where no pair of subwords can be merged anymore. In our corpus, languages needed between 1.1K and 7K operations for reaching this final merge (Figure 4).

We can see that languages with lower values

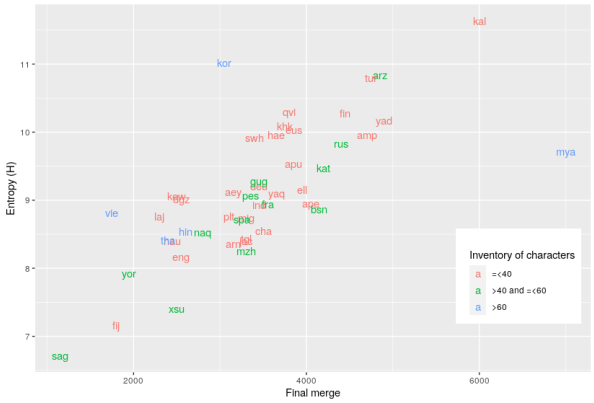

Figure 4: Maximum number of BPE merges for each language (x-axis) against word level entropy (y-axis).

of word entropy require fewer merges. This is the case for languages with isolating tendencies: Sango (*sag*), Vietnamese (*vie*), Fijian (*fij*), Yoruba (*yor*). The Indo-European language that requires the least merges is English (*eng*). These are also the languages that reach the entropy plateau faster, i.e., the first merges capture very productive recurrences, but after a relatively small number of operations, the text entropy does not change that much anymore.

On the other hand, languages with richer morphological processes (polysynthetic, agglutinative, or template morphology) require more merges to reach this final point. Namely, Kalaallisut (*kal*), Burmese (*mya*), Yagua (*yad*), Egyptian Arabic (*arz*), Turkish (*tur*), Alamblak (*amp*), Finnish (*fin*), require the most merges. Note that Burmese (*mya*) is not generally considered a morphologically complex language; it has very long orthographic words on average. This is related to the Burmese script, which uses white spaces differently from other scripts (therefore, it has very long character string sequences).

Notice that despite this outlier, the figure illustrates that characters' inventory size is not strongly influencing the final number of merge operations (Fig. 4). A language that has a small character inventory to start with can still require many merge operations to reach the final merge, as exemplified by Kalaallisut (kal), Turkish (tur), Yagua (yad), and Finnish (fin).

4.3 Correlations between measures

In order to investigate relationships between measures, a correlation matrix is shown in Figure 5 with the variables explained in Section 3.5. We only include correlations for the entropy and redundancy taken at merges 0, 30, 200, 350, and the word level in this matrix. See Appendix C for the complete correlation matrix, which includes a wider range of merges.

The final merge is strongly correlated with the entropy at the word level and the average word length of a language. This is expected since languages with higher entropy/TTR (at the word level) tend to have longer words because they encode more morphosyntactic distinctions within a word. Therefore, under this conceptualization of complexity, complex languages will require more merges during BPE encoding. We did not find any strong correlations between word length and any of the early merges' measures.

The entropy of the texts on the first operations (from merge 0 to merge 100) showed no strong correlation with the entropy at the word level, either positive or negative ($\rho < \pm 0.14$). Therefore, there is not a general trade-off between subword and word level complexities. Some languages display such a trade-off, e.g., being very complex at the word level, while having low complexity on the first merges. However, others are more stable across merges.

Interestingly, it seems that after the 'turning point', the correlation between subword tokenizations and the word level starts to be more prominent. At merge 200 the correlation is $\rho = 0.47$, this gradually increases, e.g., by merge 350, there is already a strong correlation with the word level ($\rho = 0.72$). As we saw in Section 4.1, the rankings obtained at the first merges (somewhere below 200) differ from the trend observed at the subsequent merges, .i.e., after merge 200 the complexity rankings of languages start to be more similar to the one observed at the word level.

Regarding redundancy, there is a strong trade-off between entropy and redundancy at the word level ($\rho = 0.72$). This is understandable since a language with high entropy at the word level will have a wide diversity of word forms, few repetitions, hence, less possibility of compression (low redundancy). However, entropy and redundancy are not always correlated. As shown in Fig. 3 (and Appendix B), the entropy tends to grow through merges, while redundancy first decreases and then grows again. In fact, cross-linguistically, we did not find a strong correlation between H and R during the early merges. By merge 200, H and R

already start to be negatively correlated. This trade-off is maintained in further operations and at the word level.

Even though R and H are not correlating on the first merges, they show similar behavior, i.e., both of them seem to follow a trend on the first merges, not correlated with their respective values at the word level, but then this trend changes (around the turning point).

We can also see that the size of a language's character inventory is correlated with the entropy on the first merges (around 0.6). While for redundancy, this correlation is not significant.

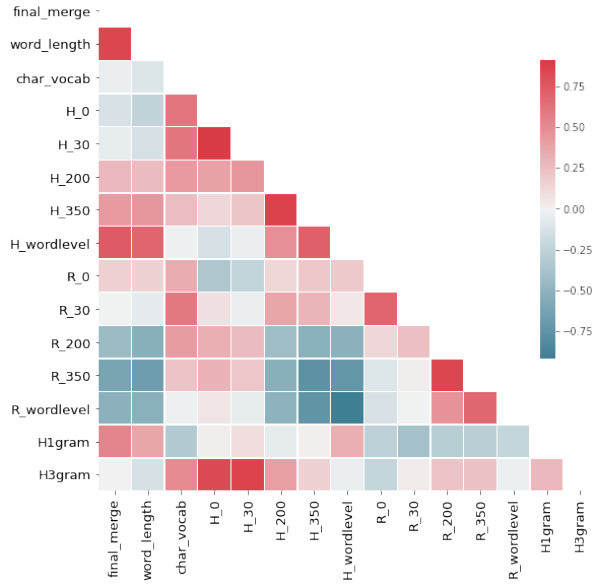

Figure 5: Spearman's rank correlation (subset of variables)

4.3.1 Predictability of sequences

Entropy is reflecting the degree of organization or predictability of subword types across several merges. This provides a glimpse of the morphological complexity of languages. However, one might argue that to really approach morphological complexity, our measures would have to consider the restrictions given by the allowed sequences of subword units within a word.

To address this concern, we compare our results with a corpus-based morphological complexity measure that aims to quantify the predictability of subword sequences within a word. This external approach uses a neural language model for estimating the predictability of sequences of n-grams within a word (Gutierrez-Vasques and Mijangos, 2020). H_{1gram} is the entropy rate obtained at the character level, H_{3gram} is the entropy rate using sequences of character trigrams.

The entropies of texts at merge 0 are already strongly correlated with H_{3gram} ($\rho = 0.81$). A higher correlation is obtained at merge 30 ($\rho = 0.86$). This means that at this point, the subword distribution is reflecting a complexity that is related to the predictability of its morphs – or at least of its character trigrams. This correlation starts to vanish in later merges, especially after the turning point (since we are probably starting to capture predictability more related to the word level).

Interestingly, H_{1gram} does not correlate with BPE entropies, not even at merge 0. This could be related to the fact that, even at merge 0, we distinguish between the characters at the end of an orthographic word versus any of the remaining positions. This already captures some degree of sequentiality that seems more related to H_{3gram}.

The measure H_{3gram} is restricted to predicting fixed-size overlapping sequences of characters within a word, while the tokenizations that we obtain with BPE contain subwords of variable lengths across merges. Despite this, it is interesting that these two measures strongly correlate at the first merge operations, suggesting that they reflect a similar phenomenon at the subword level.

5 Discussion

The "turning point" we discussed here reflects several phenomena. Firstly, languages become more similar in terms of their subword entropies around this point. Secondly, the trends of entropy and redundancy start to change (redundancy starts to grow, while entropy growth slows down). Thirdly, there is a shift in several cross-linguistic correlations around the same region of merges. It seems that two different subword distributions emerge, one before the turning point, and one afterward.

The entropy measured over word-level types reflect one dimension of diversity, i.e., some languages have rich inflectional morphology while others do not mark grammatical information word-internally. However, in another dimension, at the subword level, this variation is reduced.

If we rank languages by their entropy, the first merges' rankings are not correlated to the ones closer to the word level. In some of the languages, we see a clear trade-off across merges. For instance, some languages with high word entropy (low predictability, long words with the potential of com-

bining many different morphological distinctions), have highly predictable subword distributions.

Languages where this trade-off is observed include: Basque (*eus*), Imbabura Highland Quichua (*qvi)*, Finnish (*fin*), Yagua (*yad*), Kalaallisut (*kal*). There are also examples of languages that start with high complexity on the first merges but they are not the most complex ones at the word level: Vietnamese (*vie*), Thai (*tha*), Hindi (*hin*), Yoruba (*yor*), Nama (*naq*). English (*eng*) is also an example of a language that has comparatively low entropy at the word level, probably due to a relative lack of productive inflections, but is not one of the least complex at the subword level (first merges).

However, this trade-off was not general, as we did not find a significant negative correlation between the entropies at first merges with the entropies at last merges.

Taking a linguistic perspective, several accounts have focused on the structure of the morphological paradigm and the predictability between the inflected forms (Blevins, 2006, 2016; Ackerman and Malouf, 2013; Cotterell et al., 2018). For instance, Ackerman and Malouf (2013) distinguish between two types of complexity: a) enumerative complexity (E-complexity), reflecting the diversity of morphological distinctions, word forms, paradigm size; and b) integrative complexity (I-complexity), related to systematic paradigmatic organization underlying the morphological surface patterns.

The entropy values of I-complexity tend to be lower and less disperse than the ones exhibited by E-complexity (Ackerman and Malouf, 2013). According to these observations, a morphological paradigm could grow (many different word forms, many morphosyntactic distinctions) as long as it maintains its predictive structure. This is argued to be the reason why languages vary more widely in the dimension of E-complexity, while being more constrained in the I-dimension. Even though our work is not based on paradigms, and it did not require the use of linguistically annotated data, our findings seem to point in a similar direction, at least in the sense that the internal predictability of words is more similar across languages than than their word-level predictability.

However, note that there is some evidence from language learning experiments – with neural networks and human participants – which suggests that both are more sensitive to E-complexity than I-complexity (Johnson et al., 2020). It is an open question how our word-internal predictability measures relate to language learning.

On a practical note, the fact that certain numbers of BPE merges lead to more similar entropies across languages could be beneficial for NLP multilingual tasks. To our knowledge, the entropy and redundancy of tokenized texts have not been used as a criterion for choosing an appropriate number of BPE merge operations. There is recent work that investigates how the number of merges can lead to more balanced distributions of subwords, improving tasks like NMT (Gowda and May, 2020).

Another important question is to what extent our findings can be generalized to other corpora. The corpus size, type of register, etc., are likely to influence the turning point. As a general trend we expect this convergence to arise in a relatively low number of BPE operations.

6 Conclusions

In this paper, we went from single characters to orthographic words through incremental merges of BPE. We observed that text entropy values across 47 typologically diverse languages are less dispersed at the subword level than at the word-level. Our findings revealed a curious turning point, around the merge 200, where the values are least dispersed. Around this point, subword token distributions gradually start to look like word-level distributions (subword- and word-level entropy rankings are correlated only after this point). Additionally, this is approximately the point where text redundancy starts to grow after an initial drop and also where entropy growth slows down considerably after initial fast growth.

At the early merges, the entropy of texts is strongly correlated with an independent measure based on modeling character trigrams sequences. This provides new evidence that contributes to the ongoing discussion regarding different types of linguistic complexity.

Finally, our analysis could provide a useful insight for NLP processing. Choosing the number of merges that result in more similar distributions across languages could lead to more suitable subword representations for multilingual settings.

Acknowledgments

We thank the EACL reviewers. This work has been partially supported by the SNSF grant no. 176305 and CONACYT.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, pages 429–464.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

Christian Bentz, 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.

James P Blevins. 2006. Word-based morphology. *Journal of Linguistics*, pages 531–573.

James P Blevins. 2016. *Word and paradigm morphology*. Oxford University Press.

Ryan Cotterell, Sebastian J Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? *arXiv preprint arXiv:1806.03743*.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

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

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

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

Gertraud Fenk-Oczlon and August Fenk. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology*, 3(2):151–177.

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

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

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

Ximena Gutierrez-Vasques and Victor Mijangos. 2020. Productivity and predictability for measuring morphological complexity. *Entropy*, 22(1):48.

Tamar Johnson, Kexin Gao, Kenny Smith, Hugh Rabagliati, and Jennifer Culbertson. 2020. Predictive structure or paradigm size? Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems.

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

Jawaharlal Karmeshu. 2003. *Entropy measures, maximum entropy principle and emerging applications*, volume 119. Springer Science & Business Media.

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

Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell, and Matt Post. 2017. A rich morphological tagger for English: Exploring the cross-linguistic tradeoff between morphology and syntax. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 112–117, Valencia, Spain. Association for Computational Linguistics.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Mueller-Spitzer. 2017. The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, 12(3):e0173614.

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

Thomas Mayer, 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.

Sabrina J Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? *arXiv preprint arXiv:1906.04726*.

Robert Östling. 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.

Derek Partridge. 1981. Information theory and redundancy. *Philosophy of Science*, 48(2):308–316.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? *arXiv preprint arXiv:1704.08352*.

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

A Languages

iso639_3	language	family
aey	Amele	Trans-New Guinea
amp	Alamblak	Sepik
ape	Bukiyip	Torricelli
apu	Apurinã	Arawakan
arn	Mapudungun	Araucanian
arz	Egyptian Arabic	Afro-Asiatic
bsn	Barasana-Eduria	Tucanoan
cha	Chamorro	Austronesian
deu	German	Indo-European
dgz	Daga	Dagan
ell	Modern Greek	Indo-European
eng	English	Indo-European
eus	Basque	Basque
fij	Fijian	Austronesian
fin	Finnish	Uralic
fra	French	Indo-European
hae	Eastern Oromo	Pama-Nyungan
gug	Paraguayan Guaraní	Afro-Asiatic
hau	Hausa	Afro-Asiatic
hin	Hindi	Indo-European
ind	Indonesian	Austronesian
jac	Popti'	Mayan
kal	Kalaallisut	Eskimo-Aleut
kat	Georgian	Kartvelian
kew	West Kewa	Trans-New Guinea
khk	Halh Mongolian	Altaic
kor	Korean	Korean
laj	Lango (Uganda)	Eastern Sudanic
mig	San Miguel El Grande Mixtec	Oto-Manguean
mya	Burmese	Sino-Tibetan
mzh	Wichí Lhamtés Güisnay	Matacoan
naq	Nama (Namibia)	Khoe-Kwadi
pes	Western Farsi	Indo-European
plt	Plateau Malagasy	Austronesian
qvi	Imbabura Highland Quichua	Quechuan
rus	Russian	Indo-European
sag	Sango	Niger-Congo
spa	Spanish	Indo-European
swh	Swahili	Niger-Congo
tgl	Tagalog	Austronesian
tha	Thai	Tai-Kadai
tur	Turkish	Altaic
vie	Vietnamese	Austro-Asiatic
xsu	Sanumá	Yanomam
yad	Yagua	Peba-Yaguan
yaq	Yaqui	Uto-Aztecan
yor	Yoruba	Niger-Congo

B Entropy and redundancy graphs

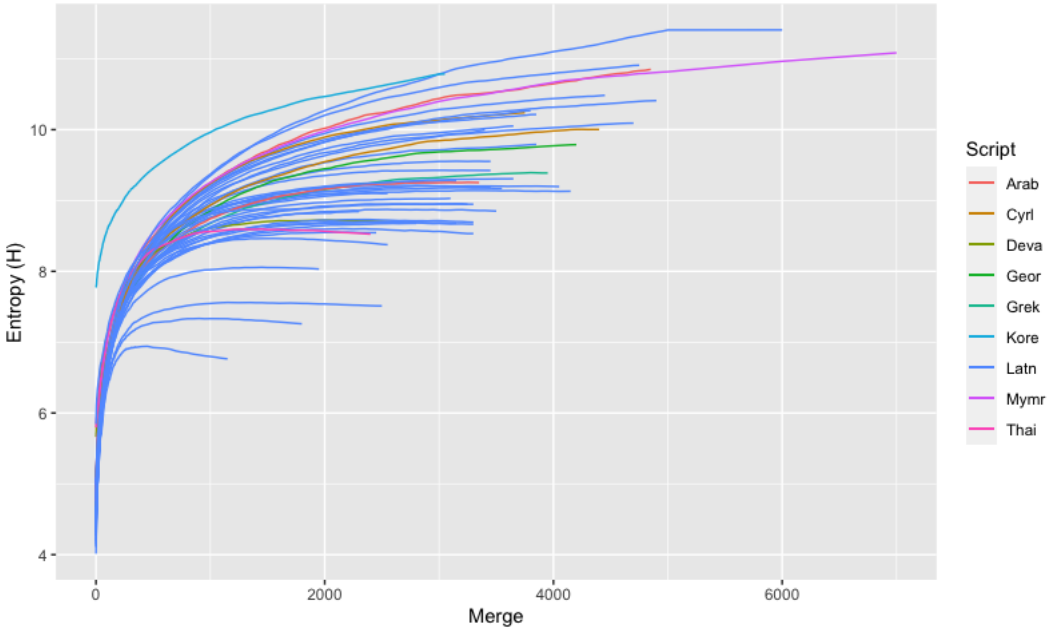

Figure 6: Entropy of all languages.

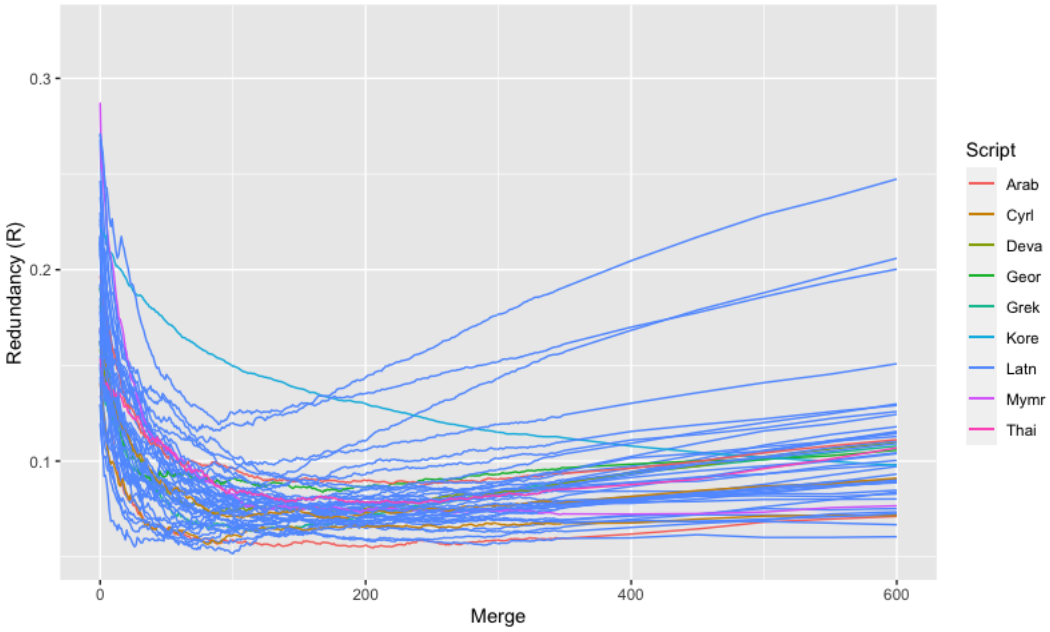

Figure 7: Redundancy of all languages (only the first 600 merges are shown in order to illustrate the merges in which the minimums are reached for most languages)

C Correlations

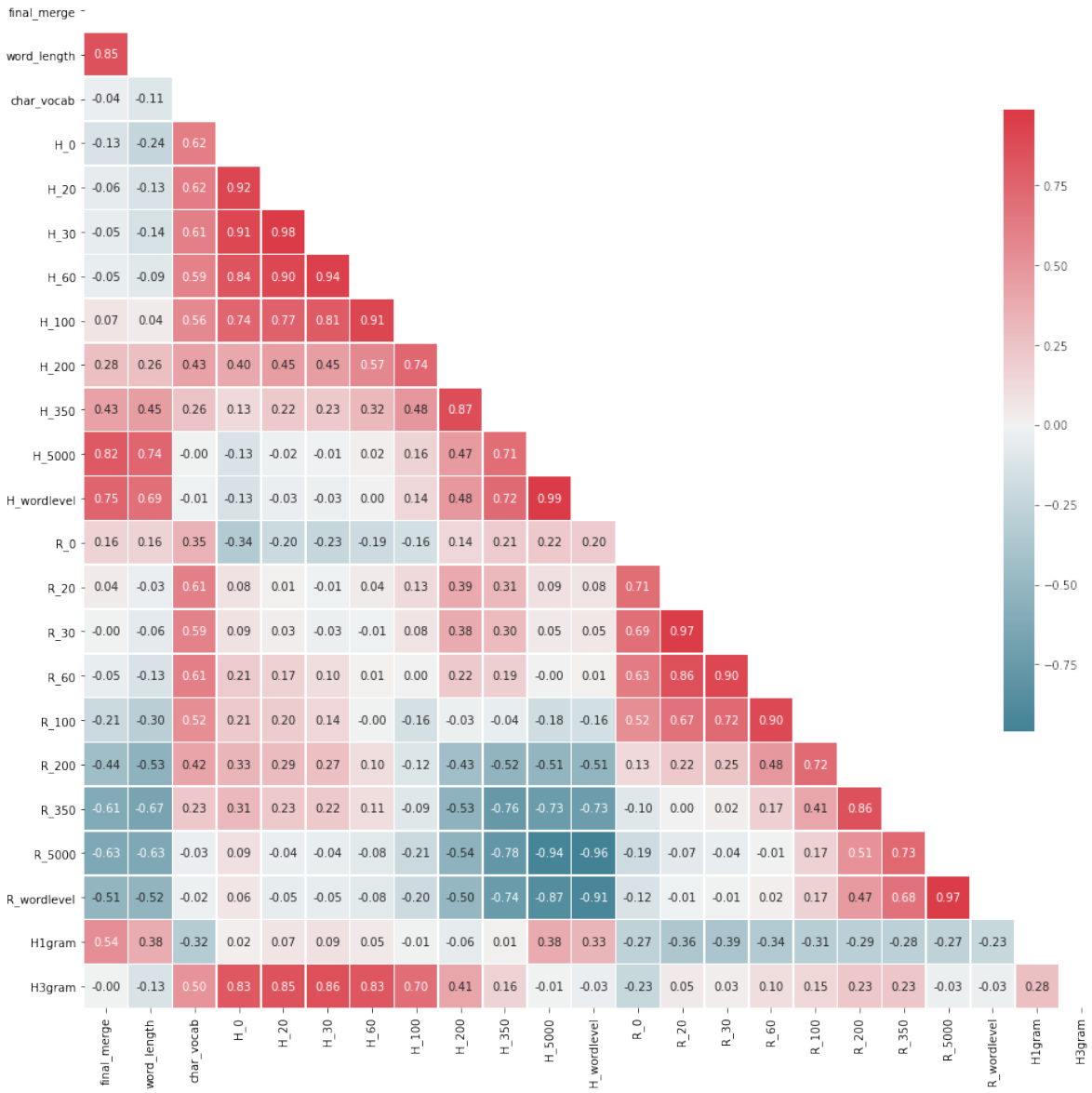

Figure 8: Correlation matrix

C.1 Significant correlations

Correlations still significant after Bonferroni correction (ordered from the highest to lowest coefficient).

var 1	var 2	pvalue	corr	num	pvalue.correct
H_5000	H_wordlevel	0	0.99	47	0
H_20	H_30	0	0.98	47	0
R_5000	R_wordlevel	0	0.97	47	0
R_20	R_30	0	0.97	47	0
H_30	H_60	0	0.94	47	0
H_0	H_20	0	0.92	47	0
H_60	H_100	0	0.91	47	0
H_0	H_30	0	0.91	47	0
R_30	R_60	0	0.9	47	0
H_20	H_60	0	0.9	47	0
R_60	R_100	0	0.9	47	0
H_200	H_350	1.55E-15	0.87	47	7.86E-13
H_30	H3gram	6.66E-15	0.86	47	3.37E-12
R_20	R_60	1.02E-14	0.86	47	5.17E-12
R_200	R_350	1.91E-14	0.86	47	9.66E-12
final_merge	word_length	3.46E-14	0.85	47	1.75E-11
H_20	H3gram	5.20E-14	0.85	47	2.63E-11
H_0	H_60	2.29E-13	0.84	47	1.16E-10
H_60	H3gram	6.39E-13	0.83	47	3.23E-10
H_0	H3gram	9.40E-13	0.83	47	4.75E-10
final_merge	H_5000	2.13E-12	0.82	47	1.08E-09
H_30	H_100	5.99E-12	0.81	47	3.03E-09
H_20	H_100	2.99E-10	0.77	47	1.51E-07
final_merge	H_wordlevel	1.10E-09	0.75	47	5.57E-07
word_length	H_5000	2.03E-09	0.74	47	1.03E-06
H_0	H_100	2.20E-09	0.74	47	1.11E-06
H_100	H_200	2.75E-09	0.74	47	1.39E-06
R_350	R_5000	6.46E-09	0.73	47	3.27E-06
R_30	R_100	1.03E-08	0.72	47	5.23E-06
H_350	H_wordlevel	1.07E-08	0.72	47	5.40E-06
R_100	R_200	1.63E-08	0.72	47	8.26E-06

var 1	var 2	pvalue	corr	num	pvalue.correct
H_350	H_5000	2.66E-08	0.71	47	1.35E-05
R_0	R_20	2.85E-08	0.71	47	1.44E-05
H_100	H3gram	4.84E-08	0.7	47	2.45E-05
word_length	H_wordlevel	7.18E-08	0.69	47	3.63E-05
R_0	R_30	7.28E-08	0.69	47	3.68E-05
R_350	R_wordlevel	1.52E-07	0.68	47	7.69E-05
R_20	R_100	1.97E-07	0.67	47	9.98E-05
R_0	R_60	2.05E-06	0.63	47	0.00104
char_vocab	H_20	2.96E-06	0.62	47	0.00150
char_vocab	H_0	3.96E-06	0.62	47	0.00200
char_vocab	R_60	4.81E-06	0.61	47	0.00243
char_vocab	H_30	5.67E-06	0.61	47	0.00287
char_vocab	R_20	6.21E-06	0.61	47	0.00314
char_vocab	H_60	1.05E-05	0.59	47	0.00529
char_vocab	R_30	1.12E-05	0.59	47	0.00569
H_60	H_200	3.24E-05	0.57	47	0.01639
char_vocab	H_100	4.30E-05	0.56	47	0.02175
H_200	R_5000	8.54E-05	-0.54	47	0.04322
final_merge	R_350	5.12E-06	-0.61	47	0.00259
final_merge	R_5000	2.45E-06	-0.63	47	0.00124
word_length	R_5000	2.28E-06	-0.63	47	0.00116
word_length	R_350	2.75E-07	-0.67	47	0.00014
H_wordlevel	R_350	6.79E-09	-0.73	47	3.43E-06
H_5000	R_350	5.52E-09	-0.73	47	2.79E-06
H_350	R_wordlevel	2.85E-09	-0.74	47	1.44E-06
H_350	R_350	4.38E-10	-0.76	47	2.22E-07
H_350	R_5000	8.86E-11	-0.78	47	4.48E-08
H_5000	R_wordlevel	4.44E-15	-0.87	47	2.25E-12
H_wordlevel	R_wordlevel	0	-0.91	47	0
H_5000	R_5000	0	-0.94	47	0
H_wordlevel	R_5000	0	-0.96	47	0