Word boundaries and the morphology-syntax trade-off

Pablo Mosteiro
Utrecht University, the Netherlands
p.mosteiro@uu.nl

Damián Blasi
Pompeu Fabra University, Spain
Harvard University, USA
dblasi@fas.harvard.edu

Abstract

This paper investigates the relationship between syntax and morphology in natural languages, focusing on the relation between the amount of information stored by word structure on the one hand, and word order on the other. In previous work, a trade-off between these was observed in a large corpus covering over a thousand languages, suggesting a dynamic 'division of labor' between syntax and morphology, as well as yielding proof for the efficient coding of information in language. In contrast, we find that the trade-off can be explained by differing conventions in orthographic word boundaries. We do so by redefining word boundaries within languages either by increasing or decreasing the domain of wordhood implied by orthographic words. Namely, we paste frequent word-pairs together and split words into their frequently occurring component parts. These interventions yield the same trade-off within languages across word domains as what is observed across languages in the orthographic word domain. This allows us to conclude that the original claims on syntax-morphology trade-offs were spurious and that, more importantly, there does not seem to exist a privileged wordhood domain where within- and across-word regularities yield an optimal or optimized amount of information.

1 Introduction

Few taxonomic distinctions in the study of language are as storied as that of 'syntax' and 'morphology', in spite of the numerous conceptual and technical obstacles in defining them. Glossing over theory-specific approaches to these two levels of description, as a first approximation syntax can be regarded as the study of combinations *between* words into grammatical phrases and sentences, whereas morphology is the study of processes that hold *within* words. Thus, linguistic phenomena such as word order, phrasal and constituency structure fall within syntax, whereas inflectional paradigms and allomorphy are uncontroversially assigned to morphology.[1] Whether these two levels involve truly different linguistic processes is a matter of controversy (Tallman and Auderset, 2023), and the descriptions of many linguistic phenomena, such as noun incorporation, seem to sit right between the two. Many attempts to distinguish between syntax and morphology need first to tackle the challenge of embracing some definition of wordhood, which is no less complex a task. Circularity of definition (e.g. by defining 'word' as the maximal domain of morphological processes), complicated cases (can clitics be words? should collocations be treated as words?) and reliance on phonological, morphosyntactic, psycholinguistic, etc. criteria have led to a seemingly inescapable situation where only a few solutions are available. According to Haspelmath (2023), we can either (1) drop the term 'word' altogether, perhaps along with the syntax-morphology distinction, (2) ignore the problem with the definition and hope that our results will be robust regardless of minutiae with the definition, (3) regard certain words as prototypical, or (4) come up with a potentially awkward and unpractical technical definition that covers much of the effective uses of the term.

Yet some patterns in language seem to provide independent support to the morphology-syntax divide (and with it, perthaps, to the notion of a 'true' wordhood). One such pattern is the celebratred trade-off between the amount of information conveyed by morphology (word structure) versus syntax (word order) (Crystal, 2010), i.e. the notion that if one system is flexible and arbitrary in its rules, the other will compensate through the enrichment

[1] We acknowledge that the definitions of syntax and morphology are more nuanced than those provided here. We merely chose simple definitions because those are the ones that give rise to the operationalizations we employed of word-order information and word-structure information. These come from Koplenig et al. (2017) and will be discussed in Section 2.

Proceedings of the New Horizons in Computational Linguistics for Religious Texts (Coling-Rel), pages 86–93
January 19, 2025.

of its relevant rules.

A statistical corpus study found the trade-off to be observable across many languages (Koplenig et al., 2017), using the Parallel Bible Corpus (Mayer and Cysouw, 2014): languages seem to rely more on word order or more on word structure in order to convey information.

In the present study, we address the following research questions: *Can the morphology-syntax trade-off be explained by orthographic conventions?* In other words, if the word boundaries are re-defined, does a language now distribute information differently across morphology and syntax? And *do languages* optimise *the amount of information conveyed by the sum of morphology and syntax?* In other words, how much redundancy is there in the information conveyed by morphology and syntax?

We find that the morphology-syntax trade-off can be reproduced by manipulations of the word boundaries. In a single language, changes to the word boundaries causes the information distribution to change in the word-order/word-structure plane. Our contribution is to show that specific previous evidence for the morphology-syntax trade-off (Koplenig et al., 2017) can be reproduced by manipulations of word boundaries, and that therefore this evidence should not be considered as supporting the claim that morphology and syntax are separate cognitive processes.

2 Related Work

Using mathematical tools to quantify the amount of information conveyed by sequences of symbols starts with the seminal work of Shannon (1948). The metric *entropy* as defined therein has been widely used to compute the information content in sequences of language (Arora et al., 2022; Bentz et al., 2022; Gutierrez-Vasques et al., 2021; Ferrer-i Cancho and Martín, 2011; Jaeger, 2010). A popular approach to computing Shannon's entropy is to *plug in* empirical probabilities into the formula. However, this underestimates the entropy (Miller, 1955). Several corrections to Shannon's entropy were proposed to mitigate this (Arora et al., 2022). However, those corrected formulas still depend crucially on estimating the probabilities of text sequences. Doing this empirically is already unreliable for sequences of length five (Schürmann and Grassberger, 1996). To avoid this problem altogether, we followed previous work on estimating Shannon's entropy using the key insight from a compression algorithm (Kontoyiannis et al., 1998).

This estimation method was previously used to estimate the entropy per word in books in multiple languages, which led to the proposal of a linguistic universal: the amount of information per word that is encoded by word ordering is the same across all languages (Montemurro and Zanette, 2011). This study used different texts for different languages. One way to make the study more robust is to use a parallel corpus such as the Parallel Bible Corpus (Mayer and Cysouw, 2014). Using this corpus and the compression-algorithm-based entropy estimation method, Bentz et al. (2017) confirmed the finding of the linguistic universal.

Along these same lines, Koplenig et al. (2017) studied the trade-off between word order information and word structure information (both in bytes per character) using the Parallel Bible Corpus. The word-order information and word-structure information are operationalizations of the information contained in syntax and morphology, respectively. As mentioned previously, their study is cross-language, and in the present work we extend it by analyzing each language individually, varying the amount of common word-pairs that are pasted together and words that are split into component parts.

Gibson et al. (2019) have reviewed the various ways in which the question of a morphology-syntax trade-off has been studied. They conclude that evidence for an efficient trade-off between these quantities puts pressure on the theories of the evolutionary origins of language. They also suggest that cognitive processes could be associated with the different ways in which we communicate information, but they do not claim any causal relationships.

On the machine-learning end, Abdou et al. (2022) have found that language models that presumably produce state-of-the-art results using shuffled sentences (Sinha et al., 2021) are actually employing sub-word information (e.g., morphology). They stress the importance of word-order information in language.

As for the observation by Koplenig et al. (2017) that languages tend to *optimize* their position along the morphology-syntax trade-off, Jaeger (2010) proposed the principle of Uniform Information Density: language production is affected by a preference to distribute information uniformly across the linguistic signal.

3 Data

We use the Parallel Bible Corpus (Mayer and Cysouw, 2014). It contains 2000 translations[2] of the Bible in 1460 languages in a verse-aligned parallel structure, covering over 40 language families from the Americas, Europe, Africa, Asia and Oceania. Each translation is tokenized and Unicode-normalized, with spaces inserted between words and both punctuation marks and non-alphabetic symbols. We follow the same pre-processing steps as Koplenig et al. (2017). Namely, we lowercase all text following the Unicode Standard (The Unicode Consortium, 2022) using the Python `str.lower` method. We then split each bible translation into different books of the bible, treating each book as a different text sample. We focus on the same six books of the New Testament studied by Koplenig et al. (2017): the four Gospels (Matthew, Mark, Luke, John), the Book of Acts and the Book of Revelation. Restricting our dataset to translations that contain at least one verse of at least one of the aforementioned books leaves 1962 bible translations in 1444 languages. This dataset is appropriate to answer our research question because it is available in many languages across multiple families, so that any findings cannot be ascribed to specific features of a given language.

We remove 4 bible translations because of the presence of a character that cannot be processed by the entropy calculator (which will be described in Section 4).[3] We remove a further 2 bibles because they contain a verse with incorrectly repeated text that leads to mistakes in the entropy calculations.[4] As a result, we have 1956 bible translations in 1442 languages.

4 Methods

We follow most of the methodology employed by Koplenig et al. (2017) to compute word-order and word-structure information, and then apply some manipulations to the word boundaries.

4.1 Word-order and word-structure information

Consider a single book from a single translation of the bible, as described in Section 3, to be a sequence b of N characters. The entropy per symbol H^b is the average amount of information that is needed in order to describe b, per unit character (Shannon, 1948). We estimate entropy using a non-parametric method built upon the Lempel-Ziv compression algorithm (Wyner and Ziv, 1989). This method converges to the entropy at the limit of long texts (Kontoyiannis et al., 1998). The formula for the entropy is

$$H^b = \left[\frac{1}{N} \sum_{i=2}^{N} \frac{l_i}{\log i} \right] \quad (1)$$

where l_i is the length of the shortest substring starting at position i of b that is not also a substring of the part of the book before this position. We use the implementation of this calculation by Koplenig et al. (2017), and we write an independent implementation to verify it[5].

Following Koplenig et al. (2017), we compute the entropy on three variants of the bible books:

1. H^b_{original} is computed on the original book
2. H^b_{order} is computed on a version of the book in which word order has been deliberately destroyed by shuffling all tokens within each verse
3. $H^b_{\text{structure}}$ is computed on a version of the book in which word structure has been deliberately destroyed by replacing every word type in the book by a randomly generated sequence of characters of the same length

This allows us to define $D^b_{\text{order}} = H^b_{\text{order}} - H^b_{\text{original}}$, i.e., the amount of information contained in word ordering; similarly, we define $D^b_{\text{structure}} = H^b_{\text{structure}} - H^b_{\text{original}}$, i.e., the amount of information contained in word structure.

With the setup described, it is possible to compute, for a given book of the bible, the quantities D^b_{order} and $D^b_{\text{structure}}$ for every translation available. We expand the methodology by performing *word-pasting* and *word-splitting* experiments.

[2] The Parallel Bible Corpus is an evolving project. We use the version from 21st October 2021, corresponding to commit c64117d in `git@github.com:cysouw/paralleltext.git`

[3] A solution to this problem is to manually replace the troublesome character by some known character that is not used anywhere in that bible. We leave this for future work.

[4] Nevertheless, we ran the analysis with these 2 bibles included, and we found entirely consistent results.

[5] See `11_validate_bpw.ipynb` in `anonymous.4open.science/r/WordOrderBibles-0F4F`

4.2 Word-pasting experiment

We start with the word-pasting experiment: take the single most commonly occurring pair of words and turn it into a word, then repeat the process iteratively. Given a book of the bible in a given translation b, we define b_{P0} as the version of this book as provided in the Parallel Bible Corpus. We compute $D^{b_{P0}}_{\text{order}}$ and $D^{b_{P0}}_{\text{structure}}$ on b_{P0}. We then find the most common pair of consecutive tokens in the book, and redefine these to be a new word, including the space. For example, if the most common pair of words in a given book is *this book*, we redefine "this book" as a single word. We call this new version b_{P1}. Thereafter, we create the order-destroyed and structure-destroyed versions of the book as defined in the previous section, and obtain new quantities $D^{b_{P1}}_{\text{order}}$ and $D^{b_{P1}}_{\text{structure}}$. We iterate this procedure and obtain, for a given book in a given translation, a sequence of pairs of quantities $(D^{b_{Pi}}_{\text{order}}, D^{b_{Pi}}_{\text{structure}})$, where i is the index of the iteration, i.e., how many times we have redefined the most common token pair as a new token and the P stands for *pasting*. By placing all these dots on a word order versus word structure plot, we can see how the two quantities vary as we redefine the word boundaries in this given language.

4.3 Word-splitting experiment

The previous section explained how we paste common word pairs together and redefine them as tokens. In our word-splitting experiment, the goal is to split words into commonly occurring sub-words. We design our word-splitting experiment in a reverse manner, by starting from characters, and then using Byte-Pair Encoding (BPE) (Gage, 1994) to iteratively paste commonly occurring pairs together.

Given a book of the bible in a given translation b, we define b_{S0} as the version of this book as provided in the Parallel Bible Corpus. We compute $D^{b_{S0}}_{\text{order}}$ and $D^{b_{S0}}_{\text{structure}}$ on b_{S0}. We then train a BPE tokenizer using the Huggingface `BpeTrainer`[6] with a `WhitespaceSplit` tokenizer, which matches the tokenization in the PBC. We give the trainer a maximum vocabulary size of 10 000 or 30 000 words, depending on the bible translation, to ensure that the training reaches completion, i.e., all words in the original text are regenerated from the component characters. We save the training history and then read it backwards, which allows us to create a history of the splitting of the most common word parts.

For each point in the reverse history, we can create the order-destroyed and structure-destroyed versions of the book as defined in Section 4.1, and obtain new quantities $D^{b_{Si}}_{\text{order}}$ and $D^{b_{Si}}_{\text{structure}}$, where i is the index of the iteration, i.e., how many times we have split a token into two component parts, and the S stands for *splitting*. By placing all these dots on a word order versus word structure plot, we can see how the two quantities vary as we redefine the word boundaries in this given language.

4.4 Implementation

For a given book in a given translation, called b, we compute the sequences $\{(D^{b_{Pi}}_{\text{order}}, D^{b_{Pi}}_{\text{structure}})\}$ at 10 points between 0 and 1 000 merges, and at 10 points between 0 and 10 000 merges. We also compute the sequences $\{(D^{b_{Si}}_{\text{order}}, D^{b_{Si}}_{\text{structure}})\}$ at 10 equidistant points between 0 and the maximum number of splits, and at 10 equidistant points between the last two of the aforementioned points.[7] The experiment was carried on in a parallel computing cluster, where each translation was run on a separate CPU. Thanks to the efficiency of the entropy calculator mentioned in Section 4, the entire experiment was run over a few days without requiring GPUs.

We then combine the information from the two experiments on the same plot. The two experiments join at $D^{b_{S0}}_{\text{order}} = D^{b_{P0}}_{\text{order}}$ and $D^{b_{S0}}_{\text{structure}} = D^{b_{P0}}_{\text{structure}}$. This is because both $S0$ and $P0$ are defined as the original books, without any merges or splits, respectively.

To understand our methodology in a different way, in a sense, BPE is doing our word-pasting experiment, but starting from characters.[8] The joining point is where BPE has created the original text, after which we continue the process by pasting words together.

We note that this methodology of splitting and pasting words can naturally generate certain known phenomena at the morphology-syntax interface. Two notable examples are:

[6] https://huggingface.co/docs/tokenizers/api/trainers

[7] To see why the further refinement in some of the areas of the parameter space was necessary, refer to Figure 1.

[8] A perhaps better alternative would be to start by converting all words to phonological forms using a text-to-phonetics converter, and then apply BPE and word-pasting on those phonological forms. Because we are comparing with previous work that operated at the character level, we opted to work at the character level.

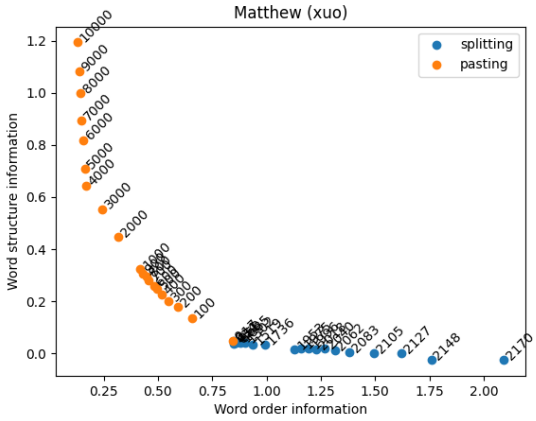

Figure 1: Word-order and word-structure information trade-off for the book of Matthew in the language xuo, in word-pasting and word-splitting experiment. The point labels indicate how many token pairs have been merged or split up to that point.

- compound nouns are separate words in English, but not in German or Dutch; word-splitting can turn German or Dutch compound nouns into pairs of words; word-pasting can turn English compounds into single words
- agglutinative affixes in Turkish are not considered words. Our splitting methodology can naturally turn long Turkish words into sub-parts.

5 Experimental Results

For every one of the six books considered (the four testaments, plus Acts and Revelation), and for every translation available in the Parallel Bible Corpus, we produce two a plot of $D^{b_i}_{\text{structure}}$ *vs* $D^{b_i}_{\text{order}}$. An example plot is shown on Figure 1, for a single translation-book pair. The results are qualitatively similar for all books considered, for all translations available, and can be found in our repository[9].

In all plots, the datapoint at the center, where the experiments join, corresponds to the original dataset, with no word-pairs pasted. As we paste increasingly more words, the datapoints move towards the top left. As we split increasingly more words, the datapoints move towards the bottom right. In other words, as we paste more common word-pairs, more information is encoded in the word structure, and less information is encoded in the word order; as we split words into more common sub-parts, more information is encoded in the word order, and less information is encoded in the word structure. This suggests that redefining the word boundaries is sufficient to reproduce the word-order vs word-structure trade-off observed previously in the literature.

To evaluate the significance of the correlations, we first join the results of the word-pasting and word-splitting experiments by assigning a negative value to the numbers of word-pairs merged in word-pasting experiments. In this way we can identify every datapoint with a single identifier which, if positive, represents a number of splits and, if negative, represents a number of merges.

Figure 2 is a histogram of the Spearman rank correlation coefficient between the number of splits and the word-order information, for all bible translations and books studied. Because the Spearman correlation coefficient is high and positive, we conclude there is a positive correlation between the number of splits and the amount of information carried by word order. Figure 3 is a histogram of the Spearman rank correlation coefficient between the number of splits and the word-structure information, for all bible translations and books studied. Because the Spearman correlation coefficient is close to -1, we conclude there is a negative correlation between the number of splits and the amount of information carried by word structure. Figure 4 is the histogram of Spearman correlation coefficients between the word-order and word-structure information, for all bible translations and books. Because the Spearman correlation coefficient is close to -1, we conclude there is a negative correlation between the amount of information carried by word structure and the amount of information carried by word order. This is the same observation as was made for a specific bible translation and book by looking at Figure 1.

6 Discussion and Future Work

Our study reveals that the trade-off between morphology and syntax in language observed by Koplenig et al. (2017) can be generated by manipulation of word structure, specifically by joining and splitting words. This finding challenges the notion that the use of morphology or syntax in language necessarily reflects distinct mechanisms for conveying information. Rather, the position of a language on the morphology-syntax trade-off appears to be

[9] https://anonymous.4open.science/r/WordOrderBibles-0F4F

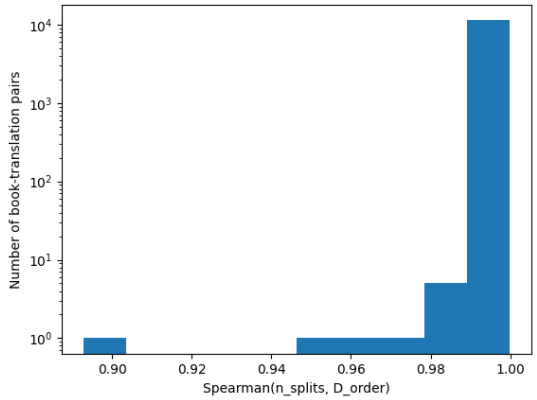

Figure 2: Histogram of the Spearman rank correlation coefficient between word-order information and the number of word-pairs split, for all bible translations and books studied. Word-pasting and word-splitting experiments are joined together by assigning a negative number to the number of word-pairs merged in the word-pasting experiments.

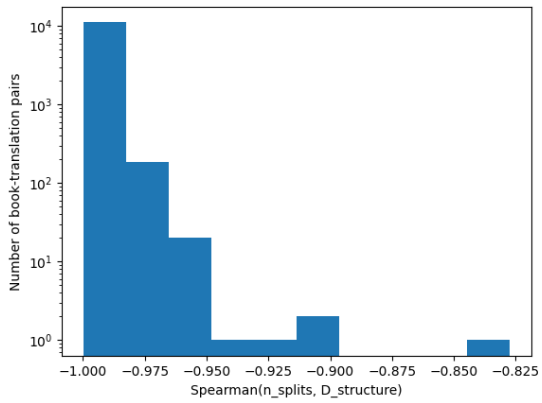

Figure 3: Histogram of the Spearman rank correlation coefficient between word-structure information and the number of word-pairs split, for all bible translations and books studied. Word-pasting and word-splitting experiments are joined together by assigning a negative number to the number of word-pairs merged in the word-pasting experiments.

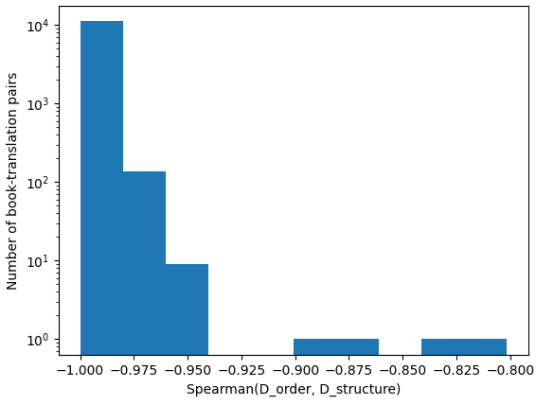

Figure 4: Histogram of the Spearman rank correlation coefficient between word-structure information and word-order information, for all bible translations and books studied. Word-pasting and word-splitting experiments are joined together by assigning a negative number to the number of word-pairs merged in the word-pasting experiments.

determined by how words are constructed. Nevertheless, our results suggest that languages have optimized this trade-off, indicating that a balance between these two mechanisms is preferred. While it may be possible to convey information through either morphology or syntax, the prevalence of the morphology-syntax trade-off across languages suggests that this balance is indeed optimal.

Based on the implications of our findings, there are several avenues for future research that could build upon our work. For example, one possible direction is to investigate whether the balance between morphology and syntax may be subject to change over time. Previous work (Koplenig et al., 2017) has found that the word-order and word-structure information can *evolve* with time; with our methodology, we could verify if this evolution could simply be ascribed to changes in the definition of word boundary. Another potential avenue is to explore the theoretical implications of our findings for our understanding of the relationship between language structure, language use, and cognitive processing (Tallman and Auderset, 2023). Finally, it would be worthwhile to verify the robustness of our methodology by investigating more novel approaches to entropy estimation, such as fine-tuning a pre-trained multi-lingual language model on the Parallel Bible Corpus, and then computing sequence probabilities using this language model.

7 Conclusion

In this paper, we have investigated two research questions concerning the morphology-syntax trade-off. Firstly, we examined whether orthographic conventions, rather than cognitive processes, can account for the trade-off observed by Koplenig et al. (2017). Secondly, we explored whether languages optimise the amount of information conveyed by the sum of morphology and syntax. Our word-pasting and word-splitting experiments showed that a morphology-syntax trade-off can be explained by purely conventional definitions, such as the definition of a word. This would mean that the statistical morphology-syntax trade-off is not necessarily due to a fundamental difference between the cognitive processes responsible for morphology and syntax (Levshina and Moran, 2021). Furthermore, the similarity between the trade-off patterns observed in previous studies and in our experiment suggests that languages do indeed optimise the trade-off between morphology and syntax.

8 Limitations

Like Koplenig et al. (2017), we used six specific books of the bible for which a large number of translations were available, and which were reasonably long for the methodology to work. A natural extension to our work would be to apply the same methodology to all the books of the bible, not just the six books considered here.

Because we restricted our dataset to translations that contain at least one verse of at least one of the aforementioned books, there are some book-translation pairs for which only a single verse or a few verses are available. This is presumably not enough for the entropy estimator we used to approximate the entropy. In future iterations, we shall restrict our analysis only to book-translation pairs for which a sufficient number of verses is available.

Furthermore, we applied the analysis independently to each book because by doing so we ensure that all texts within a given analysis have the same content, avoiding the problem whereby a bible translation does not contain all six books. It would be appropriate to combine at least several of the books together and repeat the analysis.

Finally, the PBC is an evolving project, and there are currently more bible translations available than there were at the time of beginning this study. It would be interesting to look at those new bible translations and seeing if the results hold.

On a more fundamental level, the Parallel Bible Corpus consists mostly of translations, not original texts. This means that the individual bibles used might not reflect natural language in those languages (Baets et al., 2020). We believe this is a minor limitation, since we are only exploring the effect of redefining word boundaries on the morphology-syntax trade-off. Furthermore, Kann (2024) has observed that the PBC displays similar word-order statistics to original texts. Still, it would be interesting to re-do our calculations in non-translated corpora. One possible corpus is TeDDi (Moran et al., 2022).

We focused only on demonstrating that there is a correlation between the word-order and word-structure information when performing manipulations on word boundaries in a single translation-book. In a further study, we will analyze whether the functional forms of the word-order vs word-structure distributions match those found by Koplenig et al. (2017) across bible translations.

Acknowledgements

The authors thank Alexander Koplenig and Marcelo Montemurro for providing details regarding previous studies, and Michael Cysouw for providing access to the Parallel Bible Corpus.

Ethical Considerations

In the context of this study, we did not identify any specific ethical considerations that warrant discussion. The research does not involve human subjects, sensitive data, or potentially contentious issues that could raise ethical concerns. We do not find any potential risks associated with this study. We have filled out the Responsible NLP research checklist.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word Order Does Matter and Shuffled Language Models Know It. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919.

Aryaman Arora, Clara Meister, and Ryan Cotterell. 2022. Estimating the Entropy of Linguistic Distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Dublin, Ireland. Association for Computational Linguistics.

Pauline De Baets, Lore Vandevoorde, and Gert De Sutter. 2020. On the usefulness of comparable and parallel corpora for contrastive linguistics. Testing the semantic stability hypothesis. In Renata Enghels, Bart Defrancq, and Marlies Jansegers, editors, *Empirical and Methodological Challenges*, pages 85–126. De Gruyter Mouton, Berlin, Boston.

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).

Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardžić. 2022. Complexity tradeoffs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*.

David Crystal. 2010. *The Cambridge encyclopedia of language*. Cambridge University Press Cambridge.

Ramon Ferrer-i Cancho and Fermín Moscoso del Prado Martín. 2011. Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(12):L12002.

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

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Ximena Gutierrez-Vasques, 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, Online. Association for Computational Linguistics.

Martin Haspelmath. 2023. Defining the word. *Word*, 69(3):283–297.

T. Florian Jaeger. 2010. Redundancy and reduction: speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62. Place: Netherlands.

Amanda Kann. 2024. Massively multilingual token-based typology using the parallel Bible corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11070–11079, Torino, Italia. ELRA and ICCL.

Ioannis Kontoyiannis, Paul Algoet, Yuri Suhov, and Abraham Wyner. 1998. Nonparametric Entropy Estimation for Stationary Processesand Random Fields, with Applications to English Text. *Information Theory, IEEE Transactions on*, 44:1319 – 1327.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-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):1–25. Publisher: Public Library of Science.

Natalia Levshina and Steven Moran. 2021. Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard*, 7(s3):20200081.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

George Miller. 1955. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*.

Marcelo A. Montemurro and Damián H. Zanette. 2011. Universal Entropy of Word Ordering Across Linguistic Families. *PLOS ONE*, 6(5):1–9. Publisher: Public Library of Science.

Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Pelloni, and Tanja Samardzic. 2022. TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1150–1158, Marseille, France. European Language Resources Association.

Thomas Schürmann and Peter Grassberger. 1996. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam JR Tallman and Sandra Auderset. 2023. Measuring and assessing indeterminacy and variation in the morphology-syntax distinction. *Linguistic Typology*, 27(1):113–156.

The Unicode Consortium. 2022. The Unicode Standard. Technical report, Unicode, Inc. Chapter 3: Conformance.

Aaron D Wyner and Jacob Ziv. 1989. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258. Publisher: IEEE.