# How Can We Relate Language Modeling to Morphology?

**Wessel Poelman*** and **Thomas Bauwens*** and **Miryam de Lhoneux**
LAGOM·NLP, Department of Computer Science, KU Leuven
firstname.lastname@kuleuven.be

## Abstract

The extent to which individual language characteristics influence tokenization and language modeling is an open question. Differences in morphological systems have been suggested as both unimportant and crucial to consider (e.g., Cotterell et al., 2018; Park et al., 2021; Arnett and Bergen, 2025). We argue this conflicting evidence is due to confounding factors in experimental setups, making it hard to compare results and draw conclusions. We identify confounding factors in analyses trying to answer the question of *whether, and how, morphology relates to language modeling.* Next, we introduce token bigram metrics as an intrinsic way to predict the difficulty of causal language modeling, and find that they are *gradient proxies* for morphological complexity that do not require expert annotation. Ultimately, we outline necessities to reliably answer whether, and how, morphology relates to language modeling.[1]

## 1 Introduction

Are certain languages *inherently* easier or harder to model (Cotterell et al., 2018; Mielke et al., 2019)? The interplay between language modeling and individual differences among languages is an open problem. One angle it can be approached from is morphological complexity (Gerz et al., 2018a; Park et al., 2021): if in one language the internal structure of words is more unpredictable according to some standard than another, then perhaps language models (LMs) have a harder time learning to predict text in that language.

Morphological systems are widely recognized as being gradient, but coarse groupings are often used, especially in NLP (Oncevay et al., 2022). *Agglutinative* languages (ALs) tend to add one grammatical feature to a word with each added morpheme, resulting in long words with many morphemes. *Fusional* languages (FLs) tend to express information through inflection, where a single morpheme can express multiple features, resulting in shorter words with fewer morphemes. Results contrasting ALs and FLs have been mixed, with some evidence pointing to ALs being harder to model than FLs (e.g., Gerz et al., 2018b) whereas others have shown that there is no difference between the two groupings (e.g., Arnett and Bergen, 2025).

We outline what experimental conditions and metrics are necessary to reliably answer whether, and how, morphology relates to language modeling. Our contributions: (1) We list confounding factors that have to be taken into account when attempting to answer the central question above. They can be seen as criteria for an "ideal" experiment. (2) We propose predicting CLM difficulty with the variety and entropic efficiency of neighboring tokens, and find they are proxies for morphological complexity.

## 2 Confounding Factors

It is not obvious how morphology impacts language modeling. What is clear is that research that seeks to draw reliable conclusions relating the two must control for the following confounding factors:

1. **Languages**: What set of languages is under consideration? If multiple hypotheses are tested, that set should ideally stay constant.
2. **Grouping**: If results/languages are grouped, is there enough in-group agreement?
3. **Tokenization algorithm**: What subword tokenization algorithm is used? What are its hyperparameters?
4. **Vocabulary size vs. data size**: How does the amount of subword types relate to the amount of training data?
5. **Corpus domain**: Are tokenizers and models trained on the same data? Are datasets comparable across languages (ideally, multi-parallel), or made to be so?

* Equal contribution.

[1] This an extended abstract of Poelman et al. (2025) which is accepted at the EMNLP 2025 main conference.



*Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 196–198

| | | Token Bigrams | | | | Token Unigrams | | | Words | |
|---|---|---|---|---|---|---|---|---|---|---|
| Language | Grouping* | AV | $\eta$ (↓) | AU | LR | MATTR | MTL | RE | $\mathcal{S}$ | MWL |
| English | Fusional | 2.12 | 15.92 | 61.08 | 59.29 | 31.78 | 4.89 | 36.68 | 9.27 | 5.54 |
| French | Fusional | 2.39 | 19.11 | 57.77 | 51.55 | 34.27 | 5.08 | 40.30 | 2.30 | 5.91 |
| Dutch | Fusional | 3.33 | 20.75 | 60.61 | 43.60 | 33.85 | 5.17 | 37.83 | 8.36 | 6.01 |
| Portuguese | Fusional | 3.06 | 21.31 | 52.64 | 51.49 | 35.38 | 4.91 | 36.38 | 10.64 | 5.79 |
| Spanish | Fusional | 2.95 | 22.70 | 56.97 | 52.62 | 33.85 | 5.05 | 36.16 | 9.05 | 5.72 |
| Danish | Fusional | 3.84 | 24.12 | 57.44 | 38.71 | 33.32 | 4.78 | 35.53 | 11.91 | 5.82 |
| Bulgarian | Fusional | 3.37 | 24.12 | 52.91 | 40.74 | 36.37 | 4.86 | 34.88 | 12.21 | 5.97 |
| Swedish | Fusional | 3.84 | 24.18 | 57.29 | 35.71 | 35.90 | 5.11 | 39.79 | 8.73 | 6.10 |
| Greek | Fusional | 4.20 | 24.48 | 51.62 | 46.81 | 38.71 | 5.11 | 37.44 | 10.35 | 6.15 |
| Romanian | Fusional | 3.12 | 25.09 | 51.81 | 51.01 | 37.80 | 5.04 | 36.98 | 10.52 | 5.95 |
| German | Fusional | 4.04 | 26.33 | 57.29 | 33.66 | 35.83 | 5.28 | 35.14 | 12.12 | 6.52 |
| Italian | Fusional | 3.65 | 27.10 | 61.54 | 59.88 | 37.56 | 5.22 | 38.85 | 9.39 | 6.21 |
| Latvian | Fusional | 4.45 | 28.07 | 50.99 | 43.81 | 41.75 | 5.00 | 32.29 | 15.76 | 6.41 |
| Czech | Fusional | 4.58 | 30.07 | 50.71 | 41.32 | 43.06 | 4.70 | 35.15 | 13.67 | 6.01 |
| Polish | Fusional | 4.74 | 30.85 | 50.61 | 43.80 | 44.51 | 5.25 | 35.76 | 12.75 | 6.68 |
| Slovak | Fusional | 4.70 | 31.12 | 51.43 | 44.68 | 43.04 | 4.82 | 34.91 | 13.39 | 6.13 |
| Slovenian | Fusional | 4.09 | 32.04 | 52.85 | 48.35 | 40.42 | 4.77 | 33.74 | 13.66 | 5.88 |
| Lithuanian | Fusional | 6.26 | 33.62 | 52.82 | 44.35 | 44.11 | 5.00 | 32.26 | 16.58 | 6.61 |
| Finnish | Agglutinative | 7.14 | 36.83 | 55.05 | 28.95 | 45.72 | 5.37 | 34.60 | 16.23 | 7.78 |
| Hungarian | Agglutinative | 6.69 | 39.11 | 56.24 | 31.37 | 41.73 | 5.05 | 34.10 | 14.63 | 6.78 |
| Estonian | Agglutinative | 6.27 | 40.31 | 55.89 | 34.39 | 43.66 | 5.22 | 34.58 | 14.87 | 6.96 |

**Table 1** – We propose to use gradient proxies of morphology that operate on token *bigrams*: the variety of a type's accessors (AV), their uniqueness (AU), and the Shannon efficiency of their distribution ($\eta$). We report averages over types in the tokenizer's vocabulary that appear at least once and were not filtered; the fraction of types excluded from each average is its lexicalization ratio (LR). We also give existing metrics operating on token *unigrams*: micro-average characters per token (MTL), moving-average type-token-ratio (MATTR), and Rényi efficiency (RE). Last are word-based metrics: tokens per character averaged per word ($\mathcal{S}$) and mean word length (MWL). All metrics are calculated on EuroParl (Koehn, 2005) using monolingual tokenizers from the Goldfish suite of models (Chang et al., 2024). *Groupings taken from Arnett and Bergen (2025). The gradient in the columns ranges from its minimum to maximum and are intended to highlight how well a metric corresponds with the "Grouping" column. For AU and LR, the top three are highlighted yellow, the bottom orange.

6. **Performance indicator**: What metric is used to evaluate and compare tokenizers and models across languages? Is the setup monolingual or multilingual? Is the metric comparable between any two languages?

These factors show a way *towards* an ideal experimental setup. Practically, one must work *backwards* from this to a feasible setup.

## 3 Accessor Variety

We need a reliable proxy for morphological complexity. Harris (1955) first suggested to count the variety of predecessor and successor units of a given string, where unusual spikes would imply the string's edges delineated something meaningful like a morpheme. Feng et al. (2004) coined *accessor variety (AV)* as the minimum of predecessor and successor variety. Wu and Zhao (2018) applied this to learn BPE merges. We use ULM tokens.

In Table 1, we calculate our metrics on a multi-parallel aligned subset of EuroParl (Koehn, 2005). AV recovers the coarse groupings, with ALs having the highest AV. Additionally, within FLs, a more fine-grained view of morphological complexity is revealed. For instance, higher AV values point to languages using compounding (e.g., German vs English). The shape of the accessor distribution ($\eta$) follows the same trend, being higher (more uniform) for ALs. These results for AV and $\eta$ suggest that the difficulty of causal language modeling, and hence higher PPLs regardless of models, is having *more* and *more equally likely follow-up options* at each token. This is what AV and $\eta$ measure.

The word-based metrics recover the groupings somewhat, but are less reliable for CLMs, unless those models also use words instead of subword tokens. The token unigram metrics MTL, RE, and MATTR look rather even across the languages in EuroParl. Since these estimators become more accurate with more data, their low variance calls into question higher-variance results computed for much smaller corpora like FLORES-200.

Lastly, AV operates on *tokens*, which means it's applicable to other units. For character- or byte-level tokenizers, AV can still provide an estimate of the degree of choice of accessors for a given type.

*In the full paper, we discuss hypotheses of other papers, present results for a larger set of languages, and suggest general methodological improvements for future investigations.*

## Acknowledgments

WP and TB are funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096. The computational resources and services used were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI.

## References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual Language Models for 350 Languages. ArXiv:2408.10441 [cs].

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are All Languages Equally Hard to Language-Model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541. Association for Computational Linguistics.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1):75–93.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018a. Language Modeling for Morphologically Rich Languages: Character-Aware Modeling for Word-Level Prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018b. On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327. Association for Computational Linguistics.

Zellig S. Harris. 1955. From Phoneme to Morpheme. *Language*, 31(2):190–222. Publisher: Linguistic Society of America.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What Kind of Language Is Hard to Language-Model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989. Association for Computational Linguistics.

Arturo Oncevay, 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.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology Matters: A Multilingual Language Modeling Analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Wessel Poelman, Thomas Bauwens, and Miryam de Lhoneux. 2025. Confounding Factors in Relating Model Performance to Morphology. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yingting Wu and Hai Zhao. 2018. Finding Better Subword Segmentation for Neural Machine Translation. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Lecture Notes in Computer Science, pages 53–64, Cham. Springer International Publishing.