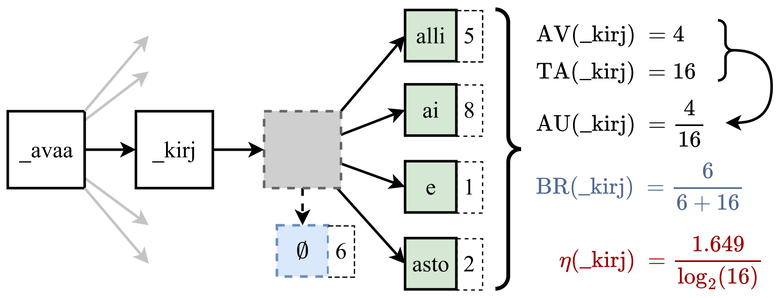
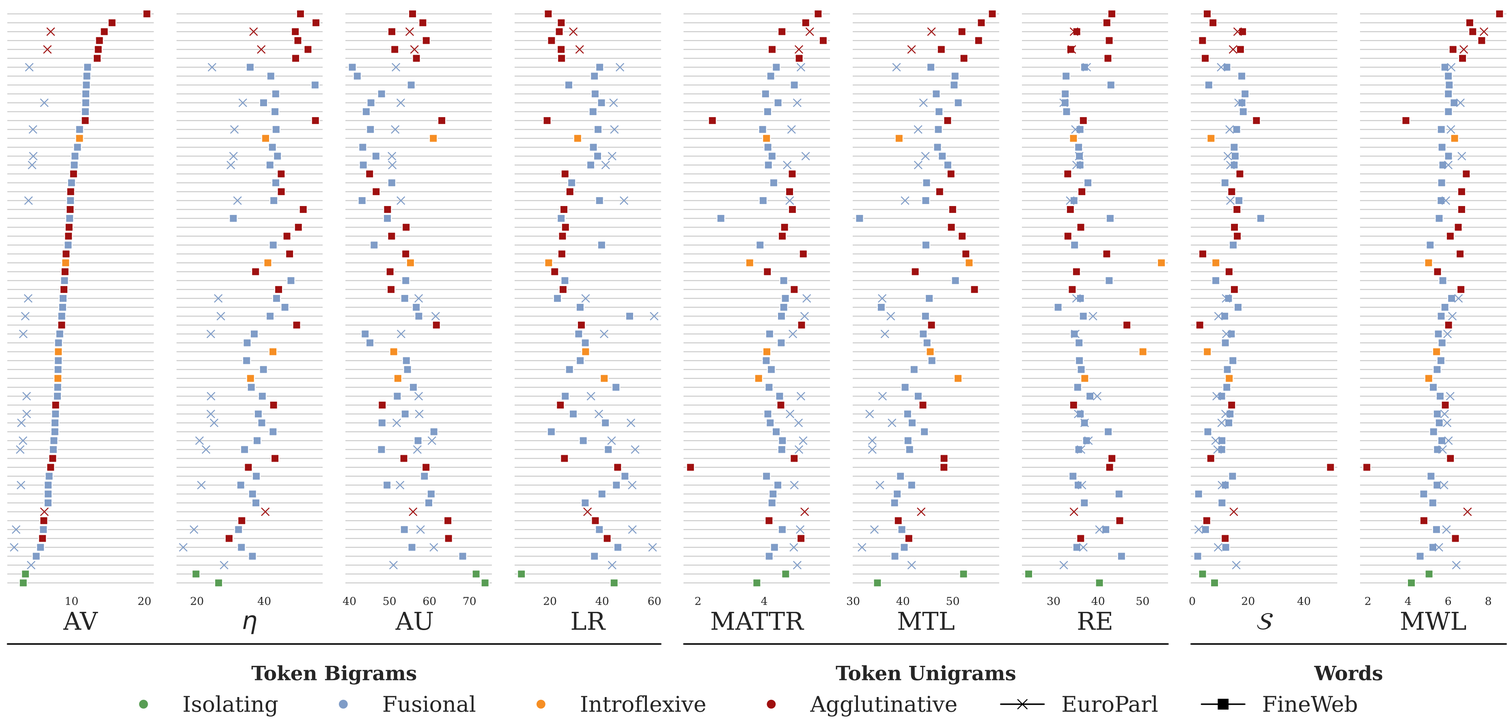
Relating Language Modeling Performance to Morphology

Wessel Poelman* Thomas Bauwens* & Miryam de Lhoneux
LAGoM-NLP, Department of Computer Science, KU Leuven
Empirical Methods in Natural Language Processing 2025





1. Background

- The relation between **intrinsic language differences** and **language modeling** remains an open question.
- **Morphology** has been suggested both unimportant and crucial.
- Existing experimental setups contain **confounding factors** that make it hard to draw reliable conclusions.
- We identify these factors, suggest methodological improvements, and introduce corpus-based token bigram metrics as a gradient proxy for morphological complexity.

2. Hypotheses

Starting point: hypotheses and conclusions from award winning paper by Arnett & Bergen (2025):

H1: Subword tokenization is less morphologically aligned for agglutinative languages.
H2: Subword vocabularies are used more inefficiently for agglutinative languages ("worse quality").
H3: Less training data is available for agglutinative languages.

- **H1:** MorphScore; evaluate recall of stem-suffix boundaries.
- **H2:** Relate PPLs of monolingual models to unigram metrics.
- **H3:** Relate PPLs of monolingual models to data-size adjusted monolingual models based on the language's UTF-8 encoding compared to English.

Ultimate conclusion: morphology is not a factor, just text encoding and dataset scaling.

3. Confounding Factors

1. **Languages**: What set of languages is under consideration?
2. **Grouping**: If results/languages are grouped, is there enough in-group agreement to justify this?
3. **Tokenization**: What subword tokenization algorithm is used? The vocabulary and segmentation are both crucial.
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?
6. **Performance indicator**: What metric is used to evaluate and compare tokenizers and models across languages?

Controlling for these factors can be seen as criteria for the "ideal" experiment. One has to work *backwards* from this to something feasible in terms of data and analysis.

4. Tokenizer Alignment

Segmentation	MS	F-F_1
gathered → gather/ed	1	1.0
gathered → gathere/d	0	0.0
gathered → g/a/t/h/e/r/e/d	1	0.25
arabaları → araba/lar/ı	1	1.0
arabaları → araba/ları	1	0.5
arabaları → arabalar/ı	0	0.5

MorphScore vs full alignment.

Consider a fusional word of the form *aaaaa*/*bc* and an agglutinated word of the form *wwwww*/*xx*/*yy*/*zz*.

1. Both have a stem-suffix boundary, but the odds of the stem and suffix sticking together is lower in agglutinative languages, since there is a chance that the suffix morphs already form a bigger token: *wwwww*/*xxyy*/*zz*,
2. Missing a stem-suffix boundary produces highly specific tokens for both: *aaaaab*/*c* or *aaaa*/*abc*, and *wwwwwx*/*x*/*yy*/*zz* or *wwww*/*wxx*/*yy*/*zz*.
3. Missing an agglutinative suffix-suffix boundary is potentially much worse: in *wwwww*/*xxy*/*y*/*zz*, the *yy* morph has lost half its length, making it potentially meaningless.
4. Misses can cascade: *wwwww*/*xxy*/*yz*/*z* or *wwwww*/*xxy*/*yzz*; the three morphemes will be harder for a model to piece together from the embeddings of two tokens.

Morphological alignment is a bad predictor for language modeling.

5. Metrics

Morphological Complexity

Unigram-based

- Moving average type-token ratio (MATTR)
- Mean token length (MTL)
- Rényi Efficiency (RE)

Word-based

- Tokens per character averaged per word ($\mathcal{S}$)
- Mean word length (MWL)

Language Modeling Performance

Model	Sequence	PPL
A	Sabe _ jugar _ al _ ajedrez	20
B	Do _ you _ know _ how _ to _ play _ chess	22
B	Can _ you _ play _ chess	18

Which model is "better" is arbitrary.

6. Accessor Variety

- Existing metrics are based on **unigrams** or **words**.
- Unigrams are not measuring complexity.
- Language models do not use words.

Bigrams! Idea stems from Harris (1955) and Feng et al. (2004). Results in gradient, corpus-based view of morphological complexity.

Proposed Metrics

- Accessor Variety (AV): *tokens following each other per type*
- Entropic efficiency of AV distribution (η)
- Accessor Uniqueness (AU): *how many unique types?*
- Lexicalization Ratio (LR): *rare or memorized types; ignored*

Findings

1. AV and η provide a gradient view of morphological complexity that is closely related to language modeling.
2. Difficulty is probably is caused by having *more* and *more equally likely follow-up* options at each timestep. This is what AV and η measure.
3. Coarse groupings like *fusional* or *agglutinative* hide information in these setups.

7. Conclusions

- Morphological alignment is a bad predictor for a performance gap between morphological systems.
- Unigram metrics do not measure language modeling difficulty.
- Dataset size alone cannot explain away morphology.
- Bigram metrics such as AV and η are better proxies for morphological complexity that do not require expert annotations.
- And finally...

Keep confounding factors in mind!

Contact & Acknowledgements

Email LinkedIn Publications

WP and TB are funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096.

KU LEUVEN