Confounding Factors in Relating Model Performance to Morphology

Wessel Poelman* and **Thomas Bauwens*** and **Miryam de Lhoneux**
L^AG_OM·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 (Cotterell et al., 2018; Gerz et al., 2018a; Park et al., 2021, *inter alia*). We argue this conflicting evidence is due to confounding factors in experimental setups, making it hard to compare results and draw conclusions. We identify such factors in analyses trying to answer the question of *whether, and how, morphology relates to language modeling.* Next, we re-assess three hypotheses by Arnett and Bergen (2025) for why modeling agglutinative languages results in higher perplexities than fusional languages: they look at morphological alignment of tokenization, tokenization efficiency, and dataset size. We show that each conclusion includes confounding factors and suggest methodological improvements. Finally, 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 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 of approach is morphological complexity (Gerz et al., 2018a; Park et al., 2021): if the internal structure of words is more unpredictable in one language than another according to some standard, 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; Amrhein and Sennrich, 2021). *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. There is particular focus in NLP to contrast ALs and FLs. Results have been mixed, with some evidence pointing to ALs being harder to model than FLs (*e.g.,* Cotterell et al., 2018; Gerz et al., 2018a,b) whereas others have shown that there is no difference between the two (*e.g.,* Mielke et al., 2019; Arnett and Bergen, 2025).

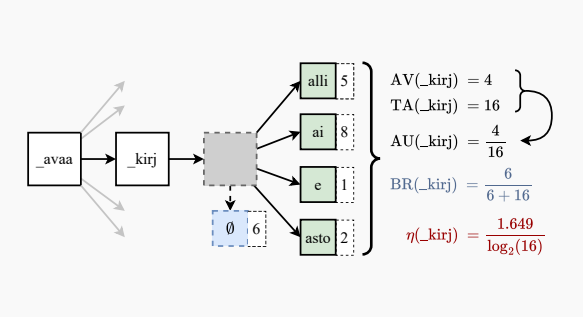


Figure 1 – Computation of our tokenizer-based gradient proxies of morphology (§ 5) for the right accessors of a Finnish subword `_kirj`: *accessor variety* (AV), *total accessors* (TA), *uniqueness* (AU), and *entropic efficiency* (η). The metrics are computed in fixed-size windows for each subword in the vocabulary, to mimic MATTR by Covington and McFall (2010). Our metrics better capture the relation between morphology and tokenization compared to word-based or unigram evaluation metrics.

Unfortunately, such analyses often introduce confounding factors when studying *whether, and how, morphology relates to language modeling.* A recent, award-winning entry in the list of evidence is a study by Arnett and Bergen (2025), who propose three hypotheses for why there might be a gap in perplexity (PPL) of monolingual causal LMs (CLMs) trained on ALs versus FLs:

* Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7262–7287
November 4-9, 2025

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

It is sometimes wrongly assumed that subword tokenization segments words into morphemes (Bostrom and Durrett, 2020) and the effects of this on tokenization and language modeling are unclear, hence **H1**. Intrinsic tokenizer evaluations, such as measuring vocabulary distributions, could reveal differences between languages that might lead us to an explanation for the performance gap. Combining these corpus-based metrics with insights from morphology might prove useful, thus **H2**. And lastly, **H3** is less directly related to morphology, but provides a good alternative to "just" word formation strategies as an explanation. Hypotheses like these are proposed regularly in such analyses and Arnett and Bergen summarize them well.

Ultimately, we outline what experimental conditions and metrics are necessary to reliably answer the aforementioned question. Our contributions:

- 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.
- We take the aforementioned hypotheses, identify confounding factors, and suggest methodological improvements to control for them.
- We propose predicting CLM difficulty with the variety and entropic efficiency of neighboring tokens, and find that they are proxies for morphological complexity.

2 Related Work

Language Modeling and Morphology. The relation between morphology and language models is mainly approached in two ways. The first asks *whether certain languages are harder to model* (Cotterell et al., 2018; Mielke et al., 2019; Park et al., 2021; Gerz et al., 2018a,b, *inter alia*). Studies taking this approach emphasize the use of parallel data, focus on model evaluations, and try to keep most experimental variables constant, except for the languages investigated. The second way asks *how aligned model architectures and tokenizers are with morphology* (Bostrom and Durrett, 2020; Amrhein and Sennrich, 2021; Bauwens and Delobelle, 2024; Limisiewicz et al., 2024, *inter alia*). Larger, unaligned corpora are often used, tokenizers are the main evaluation target (using reference lexicons or segmentations), and attention is paid to specific word-formation mechanisms (e.g. compounding).

These approaches can be at odds with each other; parallel corpora are generally quite small and specific treatments introduce potential confounds. The previously mentioned study by Arnett and Bergen (2025) attempts to combine these perspectives. We discuss their findings in §4.

Morphological Complexity. The "complexity" of a language can refer to many aspects of it (Sampson et al., 2009). Since a CLM builds words by predicting tokens, we care about the complexity of word formation, i.e. morphology.

Complexity metrics can be corpus statistics that measure how a language uses a vocabulary of characters/subwords/words. These include *mean word length (MWL)*, *type-token ratio (TTR)*, and *moving-average TTR (MATTR)* (Kettunen, 2014; Bentz et al., 2016; Park et al., 2021; Çöltekin and Rama, 2023). These metrics somewhat relate to morphology since ALs tend to have more unique words that use more morphemes, resulting in higher values for the aforementioned metrics. They have been regularly used as gradient proxies of morphology.

Other metrics are corpus-agnostic and make use of expert annotations or grammars; in their simplest form, they may be binary typological groupings (Bickel and Nichols, 2013). One can also quantify the richness of paradigms. Either by counting the number of paradigms or by counting how many word forms paradigms can produce on average (both referred to as *E-complexity*), or by calculating how predictable other forms of a paradigm become once one is known (conditional entropy, known as *I-complexity*; Ackerman and Malouf, 2013; Cotterell et al., 2018). These metrics do not reflect how a vocabulary encodes a text, which is what happens in language modeling.

In §5, we introduce metrics combining both perspectives: they are *corpus-based* to have a direct connection with tokenization and avoid the need for experts, but, unlike for instance MWL or TTR, look at relations *between*[1] tokens rather than tokens in isolation. Additionally, by using *tokens* instead of *words*, we calculate the proxy using the same units a language model would generally use.

[1]Just as I-complexity provides more nuance than E-complexity by relating suffices *to each other*.

3 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 to justify this? Coarse morphological groupings hide potentially relevant information (Table 5).
3. **Tokenization algorithm**: What subword tokenization algorithm is used? What are its hyperparameters? The vocabulary of units and segmentation choice both influence the entire pipeline and conclusions drawn. ULM (Kudo, 2018) is more morphologically aligned than BPE (Sennrich et al., 2016) for e.g. English and Japanese (Bostrom and Durrett, 2020).
4. **Vocabulary size vs. data size**: How does the amount of subword types relate to the amount of training data? If a tokenizer's training corpus is too small relative to the vocabulary size, it will partially store a long tail of corpus-specific strings (Reddy et al., 2025). If the language model's corpus is too small relative to the vocabulary size, it will have poorly trained embeddings (Rumbelow and Watkins, 2023).[2]
5. **Corpus domain**: Are tokenizers and models trained on the same data? Are datasets comparable across languages (ideally multi-parallel or similar amounts of data)?
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? In §4.4, we argue against comparing monolingual PPLs of different tokenizers and test texts.

These factors show a way *towards* an ideal experimental setup.[3] Practically, one must work *backwards* from this ideal to a feasible setup in terms of data and analysis. We now re-assess the hypotheses of Arnett and Bergen and outline broader concerns.

[2]Empirically, Ding et al. (2019) found that transformer models prefer smaller vocabularies than often used.

[3]There are more factors not directly relevant, see §6.

4 Re-assessment of Hypotheses

4.1 H1: Morphological Misalignment

A segmentation of a word is said to be morphologically aligned when the splits it places match the boundaries between the morphs (i.e. the visible parts of morphemes) that make up the word (Kurimo et al., 2006). Misalignment means oversegmentation of a morph and/or a token containing characters from more than one morph.

The argument for better morphological alignment causing better language modeling is that models cannot see characters. When a morph's characters are scattered across tokens, a model may struggle to know from their embeddings that the morph is there, potentially missing its semantics.

MorphScore. Micro-averaged precision, recall and F_1 are established metrics for measuring morphological boundary recognition (Kurimo et al., 2006; Grönroos et al., 2014; Bostrom and Durrett, 2020, *inter alia*). Arnett and Bergen introduce a new metric, MorphScore, with associated datasets. The MorphScore datasets consist of one morpheme boundary per word, namely the boundary between a stem and its suffix(es). The MorphScore metric considers micro-averaged recall of these boundaries in two modes: one in which words that appear in both the test set and the tokenizer's vocabulary are left untested, and another in which those words are always counted as correct even if the segmentation does not recall the stem-suffix boundary. Table 2 shows why tracking only recall, and why only considering the stem-suffix boundary, does not accurately judge morphological 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

Table 2 – Examples of what is being evaluated by full-alignment (F-F_1) and MorphScore (MS). Full-alignment refers to evaluating a tokenizer on all morpheme boundaries. MorphScore evaluates the recall of stem-suffix boundaries. Stems are marked green, other colors indicate which characters belong to a refrence morpheme.

Ignoring suffix-suffix boundaries is problematic when relating ALs, FLs, and LMs. Consider a fusional word of the form *aaaaa/bc* and an agglutinated word of the form *wwwww/xx/yy/zz*.

	Full			Full$_{\geq 3}$ (stem-suffix)			Full$_{\geq 3}$ (suffix-suffix)			MorphScore		
	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1	Pr	**Re**	F_1
German	28.27	61.64	38.76	31.85	88.68	46.87	6.43	17.89	9.46			
English	29.65	62.44	40.21	32.88	86.18	47.60	17.85	44.82	25.53	65.88	20.85	31.67
Polish	23.91	41.29	30.28	32.79	57.14	41.67	31.15	52.78	39.18			
Swedish	43.81	42.99	43.40	27.58	48.58	35.19	23.44	34.60	27.95			
Russian	25.32	30.78	27.78	19.71	44.33	27.29	9.22	15.13	11.46			
Catalan	29.58	32.37	30.92	23.40	39.58	29.42	11.60	19.62	14.58			
Spanish	27.08	41.46	32.76	13.65	38.18	20.11	17.33	47.82	25.44	52.50	34.15	41.38
Czech	20.98	35.31	26.32	14.24	27.44	18.75	17.69	34.06	23.29			
French	23.51	41.89	30.12	14.89	25.88	18.91	19.43	27.39	22.73			
Portuguese	12.43	22.39	15.98	9.52	14.75	11.57	10.78	12.54	11.60			
Hungarian	47.22	69.58	56.26	35.17	71.20	47.08	21.85	44.21	29.24	57.59	43.05	49.27
Finnish	19.43	35.35	25.08	14.10	37.20	20.45	7.52	19.85	10.91			
Turkish	74.34	34.25	46.90	32.50	33.10	32.80	49.58	28.31	36.04	23.18	48.27	31.32

Table 1 – Morphological boundary recognition. Full segmentations are from MorphyNet (Batsuren et al., 2021) and MorphoChallenge (Kurimo et al., 2010, only Turkish). The MorphScore data is from Arnett and Bergen (2025). The second and third columns ("stem-suffix" and "suffix-suffix") correspond respectively to testing only the one stem-suffix boundary of a word (mimicking how MorphScore works), and testing all boundaries except for that one (for which at least 3 morphemes are needed). The top languages are considered fusional, the bottom agglutinative. Word counts are in §C.

First, both words have a stem-suffix boundary, but the odds of the stem and suffix sticking together are slightly lower in ALs, since the suffix morphs might already form a bigger token as in *wwwww*/*xxyy*/*zz*, which Arnett and Bergen also observe. Second, missing a stem-suffix boundary produces highly specific tokens for both, as in *aaaaab*/*c* or *aaaa*/*abc*, and *wwwwwx*/*x*/*yy*/*zz* or *wwww*/*wxx*/*yy*/*zz*. Third, 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. Finally, misses can cascade, as in *wwwww*/*xxy*/*yz*/*z* or *wwwww*/*xxy*/*yzz*; the three suffix morphemes will be harder for a model to piece together from the embeddings of two tokens.[4] In short: stem-suffix boundaries are not explanatory for ALs underperforming to FLs. Losing one boundary in ALs may cause the same performance hit as losing many more boundaries in FLs, making morphological alignment a poor predictor for language modeling.

Using MorphScore, Arnett and Bergen find that tokenization for ALs is more aligned (higher stem-suffix recall) than for FLs. This is based on averages across MorphScore's 22 languages, evenly divided between ALs and FLs. Yet, this average hides notable inconsistencies: English (FL) has the second-highest MorphScore of all languages, and five FLs have a higher MorphScore than Turkish (AL). The conclusion that ALs are more aligned than FLs is partially caused by this averaging across groups. We get back to this grouping issue in §5.

[4]This is what happens in Table 3 of Ataman et al. (2017).

Full Alignment. We now use inflectional and derivational segmentations from MorphyNet (Batsuren et al., 2021) to measure full alignment. Additionally, we create two datasets from words with at least three morphemes: one with only the stem-suffix boundary, the other with only the remaining suffix-suffix boundari(es).

The tokenizers Arnett and Bergen use for **H1** are from Chang et al. (2024a), which are not openly available. These are monolingual ULM[5] tokenizers trained on 10k randomly sampled "lines", with a vocabulary size of 32k. Instead, we use the monolingual tokenizers from the *Goldfish* suite of models (Chang et al., 2024b); these also use ULM, but they are trained on more data (1 GiB[6] sampled from a pool of several datasets) and with a vocabulary size of 50k. Table 1 shows the results of full alignment and compares to MorphScore where available.[7]

Findings. To reliably conclude whether tokenization for ALs is more aligned than FLs, and what it implies for LMs, one would ideally use full reference segmentations for a large set of languages and evaluate various tokenization methods.

The data we have limits us to MorphScore, containing 22 languages with about 100 to 2000 exam-

[5]Chang et al. (2024a,b); Arnett and Bergen (2025) all refer to these as "SentencePiece tokenizers", but SentencePiece itself is a software package (Kudo and Richardson, 2018), where the user chooses between ULM (Kudo, 2018) or BPE (Sennrich et al., 2016), resulting in different tokenizers.

[6]This 1 GiB is "byte-premium-adjusted", see §4.3.

[7]We added precision and F_1 for MorphScore's boundaries, but it is technically only the **Re** column.

ples each, MorphyNet, containing 13 languages[8] ranging from 100k to over 1 million examples, and MorphoChallenge with about 1000 examples (§C).

The "stem-suffix" and "suffix-suffix" columns in Table 1 show that *within* a language, suffix-suffix boundaries are missed much more than stem-suffix boundaries in almost all cases, but at an unpredictable rate. Thus, only checking alignment for stem-suffix boundaries does not allow assessing the alignment of the boundaries that possibly matter even more (as discussed at the start of this section).

Comparing scores *across* languages shows neither grouping consistently having a higher F_1.

4.2 H2: Tokenization Efficiency

Whereas alignment refers to morphological correspondence between tokens and morphs, efficiency refers to whether a tokenizer optimally uses its allocated vocabulary for encoding a text.

Corpus token count (*CTC,* Schmidt et al., 2024) and *Rényi efficiency* (*RE,* Zouhar et al., 2023) have been proposed to quantify this. CTC measures how many tokens are needed to encode a given text. Despite sometimes claimed to measure "compression", it cannot be compared across texts and languages unless (at least) normalized by the length of the source text (yielding the inverse of *mean token length (MTL)*). RE quantifies the uniformity of the token distribution for a certain text, as measured by its entropy H_α normalized by its maximally achievable entropy H_0. High entropy means the distribution is uniform (flat), low entropy means it is skewed with very frequent and very rare tokens (Zipf, 1949). High entropy is desirable since it means the whole vocabulary receives training data, rather than overloading a small number of types.

ALs have a smaller affix inventory than FLs because more are used when forming words, so no affix suffers from information scarcity. Arnett and Bergen compute RE and CTC on the multi-parallel FLORES-200 dataset (Team NLLB et al., 2022) and indeed find higher RE for tokenizers in ALs, although they conclude that this is undesirable. They find little connection to PPL (surprisingly, since RE, CTC, and PPL were all higher for ALs) for the monolingual CLMs by Chang et al. (2024a) in 36 ALs and 16 FLs.[9] In §5, we find that in larger corpora, there are minimal differences between the REs of FLs (e.g. Romanian) and ALs (e.g. Finnish); we argue that CLMs are more affected by the distribution of token bigrams than token unigrams.

Findings. To conclude that ALs have worse tokenizer efficiency and that this impacts LMs, an ideal experiment would have parallel training and testing data for a large set of languages, and would relate multiple intrinsic metrics (e.g. alignment and efficiency) to LM metrics (e.g. language characteristics (Meister and Cotterell, 2021) and downstream performance) that are comparable, with no unnecessary grouping. All this is a large-scale effort and outside the scope of the current paper.

The conclusions by Arnett and Bergen are arguably not reliable due to comparing monolingual PPLs (§4.4) and coarse, unbalanced groupings (§5 and Table 3a). Their claim that higher entropy is undesirable also conflicts with Zouhar et al. (2023).

4.3 H3: Dataset Size

More data generally results in better LMs (Kaplan et al., 2020; Bousquet et al., 2022). When training models on non-parallel monolingual corpora, each model should be supplied with the same amount of information. Neither the corpus character count (CCC), token count (CTC), or word count (CWC) reliably quantify this due to being confounded by morphology and tokenization. The *corpus sentence count (CSC)* is less confounded, if sentence boundaries can be found (Minixhofer et al., 2023).

Byte-premiums. To test whether models for ALs were just trained on less data, Arnett and Bergen compare PPLs computed for the previously mentioned monolingual *Goldfish* models, whose training data was scaled per language to reflect its *byte-premium (BP)*. BPs were introduced by Arnett et al. (2024) to measure how many extra bytes are needed to encode parallel texts in UTF-8 compared to English (due to e.g. script or diacritics). For the Goldfish models, using 10 MiB of English text as reference, a language requiring $3\times$ more bytes to represent got 30 MiB of (non-parallel) training data. The comparison reported 154 languages,[10] grouped into 85 ALs and 64 FLs.

Arnett and Bergen observe that the previously seen PPL gap between FLs and ALs shrinks to have a p-value of 7.7% when using the Goldfish models, whose datasets controlled for BP. They conclude from this that there is no longer a performance gap and that BP explains this compared to the original

[8] There are 15 total, but manual inspection showed questionable labels for Italian and Mongolian.

[9] The reported number is 63, the actual 53: Table 3a.

[10] The actual number is 149, see Table 3a.

Experiment	|L|	ALs	FLs	|V|	Tokenizer Data	Metric
H1: Alignment	22	11	11	32k	10k lines	MorphScore, PPL*
H2: Efficiency	63 (53†)	37‡	16	32k	10k lines	CTC, RE, PPL*
H3: Data Size	154 (149†)	85‡	64	50k	100 MiB	PPL*

(a) Language-script pairs, morphological groupings, tokenizers, and metrics per hypothesis.

Hypotheses	|L|
H1 ∩ **H2**	3
H1 ∩ **H3**	22
H2 ∩ **H3**	52
H1 ∩ **H2** ∩ **H3**	3
H1 ∪ **H2** ∪ **H3**	145

(b) Language overlap.

Table 3 – Experimental conditions in Arnett and Bergen (2025). Multiple experimental variables change per hypothesis, making it impossible to know what caused observed effects. † means the reported number is incorrect due to null values, which are silently dropped in R. ‡ means the grouping contains languages that are included twice, but written in different scripts. **H2** has 52 and **H3** 145 unique languages. All duplicate language-script combinations are ALs. We report *languages*, not language-script combinations in Table 3b. *__H1__ & **H2** are from (Chang et al., 2024a), **H3** from (Chang et al., 2024b).

experiments that showed a larger gap (statistical discussion: §B). Yet, the previous experiments used different vocabulary sizes, training data, and languages (see Table 3); all these confounding changes are potential causes for the observed effect.

> *The results show that after taking into account byte premiums, there is no difference in performance according to morphological typology. [...] This suggests, therefore, that differences that seemed to be driven by morphological typology are actually being driven by disparities in dataset size measurement. – Arnett and Bergen*

Findings. We agree that BP is an interesting alternative to CSC, but by scaling the datasets using BP, changing the tokenizers, the model sizes, and the languages studied, it becomes impossible determine the effect of BP. To isolate it, one could do a paired t-test between pairs of LMs that share their language, tokenizer, architecture, and test set (preferably part of a fully parallel corpus), with one model trained on e.g. 1×10 MiB and the other on $\text{BP}_L \times 10$ MiB of data. We would also need a fair metric to assess LM performance across languages (see §4.4). Concluding that BP explains away morphology is not possible otherwise. A p-value of 7.7% implies that if there truly was no difference between the PPLs of BP-adjusted ALs and FLs (the null hypothesis), gaps as high as the one found for the Goldfish models would only occur in 1 out of every 13 repeats of the experiment. We outline additional issues regarding experimental setups and hypothesis testing in §B.1.

4.4 Recap

In Table 3a, we list the experimental variables per hypothesis of Arnett and Bergen, and in Table 3b which languages they each share. On top of confounding factors (see §3), only three languages are present in all three hypotheses, meaning conclusions are predominantly drawn about *different* languages, bringing into question their reliability.

Perplexities Across Languages. To answer the central question, a metric to quantify language modeling performance is needed. Variations[11] of *perplexity (PPL)* are often used which measure if a model assigns low probabilities to each next token in a test sequence (Cotterell et al., 2018; Gerz et al., 2018a; Mielke et al., 2019; Park et al., 2021; Wan, 2022; Chang et al., 2024b). As shown, experimental setups commonly use *monolingual* models evaluated on a test set in the model's language. One model achieving a lower PPL than another would indicate that it is "better".

Comparing PPLs between monolingual models is not straightforward without strong assumptions. Comparing PPLs of different *models* with the same tokenizer and test set (Chang et al., 2024a) is valid. Comparing PPLs of different *segmentations* of the same test text can be compared after rescaling to a shared underlying unit like characters (Mielke, 2019; Bauwens, 2024). Yet, when comparing PPLs from different models targeting different test sets in different languages, not only does the segmentation change,[12] but also the underlying distribution of the test set – even with parallel texts, see e.g. Table 4. The argument made in favor of this comparison is that with parallel texts, models are capturing "semantic information" across languages, which is what we would ideally use. However, even after transforming PPL into negative log-likelihood, bits-per-character, or relative metrics such as bits-per-English-character, we are still comparing different distributions (texts) using different segmentations. Grouping monolingual PPLs into morphological categories and performing analyses on them is based on the assumption that these PPLs are drawn from the same distribution.[13] We argue for future

[11] E.g. negative log-likelihood (NLL) or log-perplexity, both monotone transformations and thus interchangeable.

[12] Assuming we are using subword tokenization.

[13] Even if we assume they are comparable, we have to deal

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

Table 4 – Two valid parallel English sentences for a Spanish sentence (same semantic information) with hypothetical PPLs. If we (arbitrarily) select the first parallel English option, suddenly model A is "better" than B, and vice-versa for the second option.

research on comparable and informative intrinsic language modeling metrics if we want to find a reliable answer to the central question.

5 Gradient View of Languages

Coarse morphological groupings are useful to talk about general tendencies, not for answering the central question. Reporting averages over these groupings hides individual language characteristics for both morphological phenomena (fusional–agglutinative scale) and performance differences.

CLMs predict a token after a prior sequence. Intuitively, this is easier if there are fewer valid options to choose from; we can quantify this by measuring per context (the current token) how many possible follow-ups there exist for it in a corpus. Such token *bigram* metrics will be more informative than *unigram* metrics like TTR.[14]

Accessor Variety. 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 or word. Feng et al. (2004) coined *accessor variety (AV)* as the minimum of predecessor and successor variety. Wu and Zhao (2018) applied this to subword tokens to learn BPE merges. We use ULM tokens.

Formally, let V be a subword vocabulary, $t_1, t_2 \in V$, and $f(t_1, t_2)$ be the amount of times a token of type t_1 is followed immediately by a token of type t_2 in a corpus. The sets

$$\begin{aligned} \mathcal{A}_L(t) &= \{t' \in V \mid f(t', t) > 0\} \\ \mathcal{A}_R(t) &= \{t' \in V \mid f(t, t') > 0\} \end{aligned} \tag{1}$$

are respectively the *left accessors* (predecessors) and *right accessors* (successors) of $t \in V$. We similarly define a *left AV* and *right AV*:

$$\mathrm{AV}_L(t) = |\mathcal{A}_L(t)| \qquad \mathrm{AV}_R(t) = |\mathcal{A}_R(t)|. \tag{2}$$

Since AV is bounded by the *total accessors (TA)*

$$\begin{aligned} \mathrm{TA}_L(t) &= \sum_{t' \in \mathcal{A}_L(t)} f(t', t) \\ \mathrm{TA}_R(t) &= \sum_{t' \in \mathcal{A}_R(t)} f(t, t'), \end{aligned} \tag{3}$$

it can be confined to a fixed range, which we denote as *accessor uniqueness (AU)*:

$$\mathrm{AU}_L(t) = \frac{\mathrm{AV}_L(t)}{\mathrm{TA}_L(t)} \qquad \mathrm{AU}_R(t) = \frac{\mathrm{AV}_R(t)}{\mathrm{TA}_R(t)}. \tag{4}$$

AU is analogous to TTR, except for token *bigrams*. TTR has been criticized for its dependency on corpus size, which can be relieved by computing it in fixed-size windows and averaging across those (Covington and McFall, 2010). Thus, for $\mathrm{AV}(t)$ and $\mathrm{AU}(t)$, each t keeps a 1000-accessor window.

Since each type has a *distribution* of accessors on its left and right, we measure their *Shannon efficiency*: how close they are to a uniform distribution. For the right accessors, this is

$$\eta_R(t) = \frac{1}{\hat{H}_0^R} \sum_{t' \in \mathcal{A}_R(t)} \frac{f(t, t')}{\mathrm{TA}_R(t)} \log_2 \frac{f(t, t')}{\mathrm{TA}_R(t)} \tag{5}$$

where $\hat{H}_0^R(t) = \log_2 \min\{|\mathrm{dom}\, \mathcal{A}_L|, \mathrm{TA}_R(t)\}$ is the maximally achievable entropy with $\mathrm{TA}_R(t)$ samples taken from the $|\mathrm{dom}\, \mathcal{A}_L| \leq |V|$ right accessors that appear in the corpus. Figure 1 shows a diagram of the above metrics. In what follows, we filter out types with little to no accessors (see § A).

Finally, there are two ways of applying the bigram metrics: either to characterize *morphological complexity* or *data difficulty*. For the former, we look at "intra-word" tokens, meaning we pretokenize our input and subsequently tokenize the pretokens. This is what is Feng et al. (2004) do. Alternatively, we can forego the pretokenization step and calculate AV directly. The former gives us an idea about the token-to-token ambiguity within pretokens (closer to morphology), the latter gives an idea of the token-to-token ambiguity without directly relating to words or pretokens (closer to the data). See § A for more details.

Multi-Parallel Results. In Table 5, we calculate our metrics on a multi-parallel aligned subset of

with outliers. Perplexity ranges from 1 to ∞, hindering robust, direct comparisons. Indeed, the results by Arnett and Bergen contain clear outliers (see § B.4). The mere existence of these outliers also shows byte-premiums are not the ultimate solution to performance disparities across languages.

[14] Imagine a text that enumerates the alphabet. Its TTR is maximal, yet a CLM can deterministically reproduce it.

		Token Bigrams				Token Unigrams			Words	
Language	Grouping*	AV	η (↓)	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 5 – We propose to use gradient proxies of morphology that operate on token *bigrams* (Figure 1) within "words" (pretokens): the variety of a type's accessors (AV), their uniqueness (AU), and the Shannon efficiency of their distribution (η). We report averages over types in the tokenizer's vocabulary that appear at least once and were not filtered (see § A); the fraction of types excluded from each average is its lexicalization ratio (LR). We also give existing metrics operating on token *unigrams*: moving-average type-token-ratio (MATTR), micro-average characters per token (mean token length; MTL), 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 the same tokenizers as Table 1. *Groupings taken from Arnett and Bergen (2025). The gradient in the columns ranges from its minimum to maximum and are intended to highlight how the metrics differ. We sort by η. For AU and LR, the top three are highlighted yellow, the bottom three orange. For visual clarity, all metrics except for AV, MTL, and MWL are multiplied by 100.

the EuroParl (Koehn, 2005) corpus. The alignment is for the sake of removing confounds of data sizes and domains, not to draw conclusions based on the parallel meaning (c.f. §4.4). In the next section, we loosen this restriction and expand our language set.

Table 5 shows that 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 compounding languages (e.g. German and Danish) as opposed to the lower ones (e.g. English and Romanian). The shape of the accessor distribution as summarized by η follows the same trend, being higher (more uniform) for ALs. These results for AV and η are both crucial in light of our hypothesis above, i.e. that the difficulty of causal language modeling, and hence the source of higher PPL, is having *more* and *more equally likely follow-up options* at each token. This is what AV and η measure. Therefore, if the hypothesis is correct, then higher AV and η are causally linked to higher PPL, thus explaining the ALs and FLs gap.

The word-based metrics recover the groupings somewhat, but are less directly related to CLMs, unless they also use words instead of subword tokens. Additionally, we have to define reference *words* which is another potential confound. The token unigram metrics look rather even across the languages in EuroParl, showing less correspondence with the other metrics. 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.[15]

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

Expanded Results. When we loosen the restriction of multi-parallelism, we can expand our language coverage. We use data from FineWeb 1 & 2 (Penedo et al., 2024, 2025). Our selection is based on (1) the language has to have a 1 GiB Goldfish tokenizer and (2) it has to have 200k lines of available data. All languages thus have similar *amounts* of data for both datasets. Figure 2 shows the combined results.

[15]EuroParl has 211k multi-parallel lines with Italian as a pivot, FLORES-200 has about 2k when combining the dev and test splits, which are not available for all languages; full multi-parallel alignment results in about 1k lines.

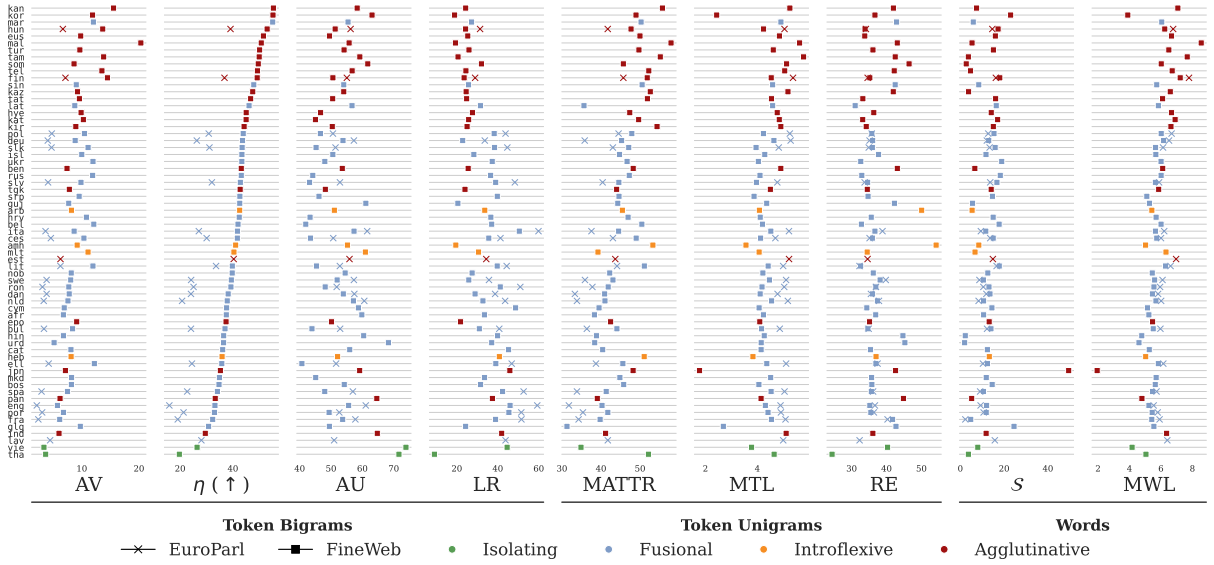


Figure 2 – Metrics across EuroParl and FineWeb. The bigram metrics are calculated within pretokens, Figure 9 contains results without pretokens. Similar to Table 5, we sort by η to show how (dis)similar the metrics are from existing metrics. EuroParl (EP) contains 21 languages; FineWeb (FW) 63. FW $\cap$ EP = 19; FW $\cup$ EP = 65. Full results in table form are in §D. The added coarse groupings are *Isolating* languages, which tend to have little to no inflection and use few morphemes per word, and *Introflective* (or non-concatinative) languages, which modify roots and tend to use little to no morphemes.

Starting with AV and η, we see a clear domain influence, with EuroParl always having lower values compared to FineWeb. The opposite is true for MATTR, suggesting that EuroParl is more lexically dense than web texts, but in a repeating fashion. The expanded language coverage shows even more clearly that coarse groupings hide information. While the top languages are agglutinative, and the bottom ones isolating, the middle shows the need for a gradient proxy. Perhaps unsurprisingly, MWL is another decent proxy for morphological complexity, but as mentioned, LMs do not tend to use words. Language modeling difficulty is tokenizer-dependent. Similarly, token unigram metrics measure noticeably different phenomena, and, as discussed in §4.2, these are less important for relating morphology to language modeling compared to the bigram metrics.

Morphology or Data. We have shown results *with* pretokenization given the link with morphology (requiring "words"). Results *without* (Table 10) also show ALs near the top, but especially languages using long words written in systems with large character inventories (e.g. abugidas).

This relates to **H3** by Arnett and Bergen (2025): some writing systems require more UTF-8 bytes to encode and more data mitigates this. However, characteristics of written language are *part of* writing systems. Or, as put by Gorman and Sproat (2023) *"A writing system is, at its base, a linguistic analysis of the language it is used to write."* Arnett and Bergen's conclusion could be explained in opposite terms: since morphology is encoded in writing systems, we need to account for it (byte-premiums).

6 Conclusion

We identify confounding factors to consider in order to reliably answer the question of *whether, and how, morphology relates to language modeling*. These factors imply "ideal" experiments, from which to work backwards to what is feasible.

We re-evaluate three hypotheses from Arnett and Bergen (2025) for why there might be a causal language modeling performance gap between agglutinative and fusional languages: morphological alignment of tokenization, tokenization efficiency, and dataset size. We show recall of stem-suffix boundaries (MorphScore) is not full alignment and outline how alignment relates to LMs (§4.1). We agree and re-confirm that token unigram metrics are poor explanations for the gap (§4.2). We disagree with the conclusion that dataset size explains away modeling difficulty caused by morphology and suggest methodological improvements (§4.3).

Finally, we introduce token bigram metrics (accessor variety and entropic efficiency) that quantify the ambiguity language models face, and show they are gradient proxies of morphology, providing a new hypothesis for why causal language models might struggle more with agglutination.

Limitations

Additional Factors. The confounding factors we discuss especially relate to the question of *whether, and how, morphology relates to language modeling*. We acknowledge there are many more confounding factors, such as architecture choices, domain, data quality, translation effects of parallel data, among others. We do not discuss these since, while all important, they should ideally stay fixed when trying to answer the central question.

Segmentation Availability. Our full alignment analysis from §4.1 relies on high quality reference segmentations. These are rare and their language coverage is quite limited, which prevents us from making broad conclusions.

Text Features and Morphology. Our metrics from §5 are *not* metrics for morphological complexity as understood in the linguistics literature (E- and I-complexity, see §2). Instead, they are proxies for morphological phenomena as seen *through the lens of* a particular tokenizer (here, monolingual ULM tokenizers) over a particular corpus (here, EuroParl or FineWeb). The assumption for this to work – but empirically, this seems to be correct – is that patterns in how tokenizers construct words mimic patterns of how words are constructed from morphological systems.

Dataset Size. We do not compute our metrics on the corpus for which Arnett and Bergen have PPL values, i.e. FLORES-200 (Team NLLB et al., 2022), since we found it too small to get stable results, see footnote 15. EuroParl is significantly larger and showed more stable metrics, which is what we use as our multi-parallel corpus. For our non-parallel corpus, but using a consistent number of *lines*, we use FineWeb. We make sure the number of lines between EuroParl and FineWeb are comparable (∼200k).

Morphological Groupings. There are many ways to characterize morphological systems of languages. The course groupings are commonly used, but have also long been criticized, such as by Sapir (1921), who notes: *"In any case it is very difficult to assign all known languages to one or other of these groups, the more so as they are not mutually exclusive."* Our use of these groupings is purely for the sake of comparison with previous work.

Acknowledgments

We thank Kushal Tatariya for suggestions, comments, and proofreading an earlier draft of this work. We also thank Catherine Arnett for answering our questions and making the analysis scripts openly available. Finally, we thank the anonymous reviewers for their suggestions. 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

Farrell Ackerman and Robert Malouf. 2013. Morphological Organization: The Low Conditional Entropy Conjecture. *Language*, 89(3):429–464.

Chantal Amrhein and Rico Sennrich. 2021. How Suitable Are Subword Segmentation Strategies for Translating Non-Concatenative Morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705. Association for Computational Linguistics.

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.

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024. A Bit of a Problem: Measurement Disparities in Dataset Sizes across Languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 1–9. ELRA and ICCL.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331–342.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: A Large Multilingual Database of Derivational and Inflectional Morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48. Association for Computational Linguistics.

Thomas Bauwens. 2024. Bits-per-character and its relation to perplexity. Blog post.

Thomas Bauwens and Pieter Delobelle. 2024. BPE-knockout: Pruning Pre-existing BPE Tokenisers with Backwards-compatible Morphological Semi-supervision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5810–5832. Association for Computational Linguistics.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 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. The COLING 2016 Organizing Committee.

Balthasar Bickel and Johanna Nichols. 2013. Fusion of selected inflectional formatives (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Kaj Bostrom and Greg Durrett. 2020. Byte Pair Encoding is Suboptimal for Language Model Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624. Association for Computational Linguistics.

Olivier J. Bousquet, Amit Daniely, Haim Kaplan, Yishay Mansour, Shay Moran, and Uri Stemmer. 2022. Monotone Learning. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 842–866. PMLR. ISSN: 2640-3498.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024a. When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096. Association for Computational Linguistics.

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

Çağrı Çöltekin and Taraka Rama. 2023. What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*, 9(s1):27–43.

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.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

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.

Kyle Gorman and Richard Sproat. 2023. Myths about Writing Systems in Speech & Language Technology. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 1–5. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. ArXiv:2001.08361 [cs].

Kimmo Kettunen. 2014. Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

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

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. Association for Computational Linguistics.

Mikko Kurimo, Mathias Creutz, Matti Varjokallio, Ebru Arisoy, and Murat Saraclar. 2006. Unsupervised segmentation of words into morphemes – Challenge 2005 An Introduction and Evaluation Report. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*, pages 1–11.

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 2010. Morpho Challenge 2005-2010: Evaluations and Results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95. Association for Computational Linguistics.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-Driven Byte Encoding for Better and Fairer Multilingual Language Modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076. Association for Computational Linguistics.

Paul McCann. 2020. Fugashi, a Tool for Tokenizing Japanese in Python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51. Association for Computational Linguistics.

Clara Meister and Ryan Cotterell. 2021. Language Model Evaluation Beyond Perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339. Association for Computational Linguistics.

Sabrina J. Mielke. 2019. Can you compare perplexity across different segmentations? Blog post.

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.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. 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.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. ArXiv:2506.20920 [cs].

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai Natural Language Processing in Python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36. Association for Computational Linguistics.

Varshini Reddy, Craig W. Schmidt, Yuval Pinter, and Chris Tanner. 2025. How Much is Enough? The Diminishing Returns of Tokenization Training Data. ArXiv:2502.20273 [cs].

Peter J. Rousseeuw and Mia Hubert. 2011. Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, 1(1):73–79.

Jessica Rumbelow and Matthew Watkins. 2023. SolidGoldMagikarp (plus, prompt generation). Blog post.

Geoffrey Sampson, David Gil, and Peter Trudgill, editors. 2009. *Language Complexity as an Evolving Variable*. Oxford Studies in the Evolution of Language. Oxford University Press.

Edward Sapir. 1921. *Language: An Introduction to the Study of Speech*. Harcourt, Brace.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

George A. F. Seber and Alan J. Lee. 2003. Straight-Line Regression. In *Linear Regression Analysis*, pages 139–163. John Wiley & Sons, Ltd.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. ArXiv:2207.04672 [cs].

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA).

Ada Wan. 2022. Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling. In *International Conference on Learning Representations*.

B. L. Welch. 1947. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35. Publisher: [Oxford University Press, Biometrika Trust].

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.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, Massachusetts.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the Noiseless Channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207. Association for Computational Linguistics.

A Pretokenization and Filtering

A.1 Pretokenization

For the word-based approaches discussed in §5, we preprocess sentences by splitting them on spaces and punctuation. For languages where this pretokenization step is problematic, Japanese and Thai in our case, we use a dedicated word segmenter (McCann, 2020; Phatthiyaphaibun et al., 2023).

Tokenization for our bigram metrics happens within pretokens, and as in Feng et al. (2004), tokens of neighboring pretokens *cannot see each other*. Instead, the first token sees a "dummy" token to its left that always counts as a unique accessor no matter how many times it has been seen. The same is true for the last token.[16]

Unlike Feng et al. (2004), we do not include these dummy accessors in $\mathcal{A}_L(t)$ and $\mathcal{A}_R(t)$. Instead, we count them separately as $b_L(t)$ and $b_R(t)$, resp. the amount of times a type t occurs as the first and last token of a word. Note that $f(t) = \mathrm{TA}_L(t) + b_L(t) = \mathrm{TA}_R(t) + b_R(t)$. The fraction of dummy accessors is the *boundary ratio (BR)*:

$$\mathrm{BR}_L(t) = \frac{b_L}{\mathrm{TA}_L(t) + b_L} \quad \mathrm{BR}_R(t) = \frac{b_R}{\mathrm{TA}_R(t) + b_R}. \tag{6}$$

Lastly, we also include results of our bigram metrics without pretokenization in §D.

A.2 Filtering

Per language, after tokenizing the dataset and counting accessors, we retroactively apply two filtering steps to the vocabulary (modifying the counts appropriately) with the goal of reducing noise in the statistics computed from them.

First, we filter out all types that contain at least one character whose value for the Unicode character property *general category*[17] is either *Punctuation* or *Digit*. In short: we filter out all types matching the regular expression

```
.*(\p{Punct}|\p{Digit}).*
```

The resulting vocabulary V' is assumed to come entirely from the language's lexicon for that corpus.

Since we are interested in the distribution of tokens (i.e. non-boundaries) around each type (the tokenization equivalent of morphology), we further exclude all types which are almost never accessed by other types and thus mostly by dummies, i.e.

$$\mathbb{K} = \{t \in V' \mid \min\{\mathrm{BR}_L(t), \mathrm{BR}_R(t)\} \geq 0.95\}. \tag{7}$$

These types could be said to be *lexicalized* by the tokenizer and have such sparse or empty accessor distributions that including summary metrics for those distributions would merely be noise.

We then call the fraction of types excluded from the vocabulary its *lexicalization ratio (LR)*:

$$\mathrm{LR} = \frac{|\mathbb{K}|}{|V'|}. \tag{8}$$

B Statistical Sidenotes

We want to address some of the consequences of particular decisions made by Arnett and Bergen (2025) in their statistical analyses. Specifically:

- Designing their study as a difference of hypothesis tests rather than a hypothesis test of a difference (§B.1);
- Defining the null hypothesis and insignificance as the success of a treatment, meaning the p-value does not express whether the result is uncommon in the control;
- Using different statistical tests to prove the existence of the gap between FLs and ALs versus proving its disappearance (§B.2);
- Lacking Bonferroni correction, raising the chances of encountering at least one significant hypothesis test (§B.3);
- Assuming PPL has no outliers or is not normally distributed, either way invalidating hypothesis tests and correlations (§B.4);
- Duplicating measurements for CTC and RE, arbitrarily lowering p-value (§B.5);
- Large, skewed predictor distribution in regression, causing highly significant but highly unpredictive regression coefficients (§B.6);
- Suggesting that causation implies correlation (§B.7).

B.1 Effect of designing the experiments as a *difference of hypothesis tests*, rather than a *hypothesis test of differences*

In essence, what Arnett and Bergen study is the effect of a treatment on a group of subjects.

[16] The idea is that there is such high flexibility in what lies beyond a word boundary that one can work with the upper limit that there is always a different accessor there. This higher flexibility is confirmed in Table 10.

[17] unicode.org/versions/Unicode17.0.0/core-spec/chapter-4/#G124142

B.1.1 Hypothesis test of differences

Conventionally, such studies are laid out according to the following recipe:

1. Compute a statistic S on the group before the treatment: s_{before}.
2. Apply the treatment.[18]
3. Compute the same statistic S on the group after the treatment: s_{after}.
4. Assuming the treatment has no effect (H_0), formulate a hypothesis test for the statistic such that the more effect the treatment has, the more unlikely s_{after} would appear if characterizing the population by s_{before}. Compute the p-value, i.e. the probability of S taking on all values even more unlikely than s_{after}.
5. Conclude that the treatment has significant effect (H_1) if p is lower than a prespecified threshold α. The probability that this conclusion is incorrect, is p.

To study what causes the disparity in PPL between FLs and ALs, this template might look like:

- The group of subjects are a set of languages.
- The statistic S is the *gap in average PPL*, $\bar{X}_1 - \bar{X}_2$, between the FLs and ALs.
- The treatment is different for **H1**, **H2**, and **H3**. Each of them tries to "explain the gap", which should be designed as a treatment that could make the gap disappear. For each language:
  - **H1**: train 1 model with a more and 1 with a less morphologically aligned tokenizer;
  - **H2**: train 1 model with a more and 1 with a less compressive tokenizer;
  - **H3**: train 1 model with and 1 without byte-premium-scaled training data;

  here, "more" and "less" can be measured continuously using respectively alignment F_1 (or MorphScore), CTC (or Rényi efficiency), and byte-premiums.
- The hypothesis test would measure whether the *gap has decreased significantly*. This calls for a *one-sided* hypothesis test: assigning FLs and ALs to the subscripts 1 and 2 such that $\Delta_{\text{before}} = \bar{X}_{1,\text{before}} - \bar{X}_{2,\text{before}} > 0$, a significant treatment would make the difference $\Delta_{\text{after}} = \bar{X}_{1,\text{after}} - \bar{X}_{2,\text{after}}$ move significantly far away to the left of Δ_{before} past some value $\Delta_\alpha < \Delta_{\text{before}}$, as shown in Figure 3. (That is: either the gap becomes smaller in absolute value, or the sign flips and the absolute value can be anything.)

[18] Applying the treatment to a second group (the first being a control) is also possible, but less practical here.

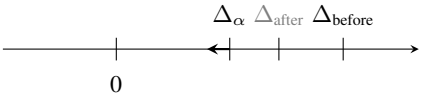


Figure 3 – One-sided hypothesis test for a significant reduction in an initially positive difference. Everything left of Δ_α is significant.

Because it is assumed $\bar{X}_{i,\text{before}}$ and $\bar{X}_{i,\text{after}}$ are normally distributed, so are Δ_{before} and Δ_{after} and thus $Y = \Delta_{\text{before}} - \Delta_{\text{after}}$ as well, with the sum of the variances of the four means as its variance:

$$S_Y^2 = \frac{S_{1,\text{before}}^2}{n_{1,\text{before}}} + \frac{S_{2,\text{before}}^2}{n_{2,\text{before}}} + \frac{S_{1,\text{after}}^2}{n_{1,\text{after}}} + \frac{S_{2,\text{after}}^2}{n_{2,\text{after}}}. \quad (9)$$

This means Y satisfies the conditions described by Welch (1947) for $Y/S_Y \sim t(\nu)$ (with ν given by Welch (1947, Eq. 28), a Welch-Satterthwaite equation). Therefore, the hypothesis test

$$H_1 \text{ if } \frac{\Delta_{\text{before}} - \Delta_{\text{after}}}{S_Y} > t_{1-\alpha,\nu} \quad (10)$$

works. That is, when the expected gap between the average PPL for FLs and ALs after a treatment is *not* below the expected gap before, the measured gap Δ_{after} will fall below

$$\Delta_\alpha = \Delta_{\text{before}} - t_{1-\alpha,\nu}\, S_Y \quad (11)$$

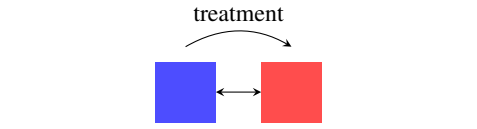
only with probability α. In short, if we use the fact that $\Delta_{\text{after}} < \Delta_\alpha$ as the decision rule to decide that $\mathbb{E}[\Delta_{\text{after}}] < \mathbb{E}[\Delta_{\text{before}}]$, we are incorrect with probability α when actually $\mathbb{E}[\Delta_{\text{after}}] \geq \mathbb{E}[\Delta_{\text{before}}]$.

The effectiveness of the treatment can additionally be assessed by just comparing $\bar{X}_{1,\text{before}}$ to $\bar{X}_{1,\text{after}}$ using a usual Welch t-test, or comparing $\bar{X}_{2,\text{before}}$ to $\bar{X}_{2,\text{after}}$. However, this does not say anything about what the treatment does to the gap *between* the averages; it could be that both averages drop significantly after treatment, but by the same amount, and hence the gap does not change.

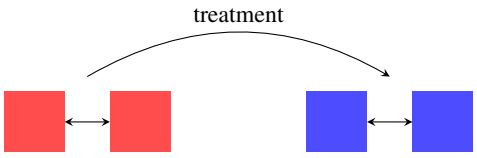
B.1.2 Difference of hypothesis tests

Arnett and Bergen (2025) do not follow the above template. In particular, rather than comparing difference statistics as per above, they compare averages;

since there are four averages, they do *two* hypothesis tests for determining if a treatment is significant: one[19] comparing $\bar{X}_{1,\text{before}}$ to $\bar{X}_{2,\text{before}}$, and another comparing $\bar{X}_{1,\text{after}}$ to $\bar{X}_{2,\text{after}}$. A treatment is deemed significant if it causes unequal decisions between the two tests, and in particular, because the first test is significant, a treatment is deemed significant if the second test is *not* significant. Figure 4 symbolically represents the difference with the above test.



(a) Hypothesis test of difference



(b) Difference of hypothesis tests

Figure 4 – The two experimental setups discussed in §B.1. Each box is a statistic. Each double arrow is a hypothesis test. The red boxes are the values discussed in the text to be desired as causing a significant hypothesis test (i.e. rejecting H_0) when the treatment *is* effective.

Equivalently, these two tests compare Δ_{before} and Δ_{after} to 0 rather than comparing them to each other, respectively testing

$$H_1 \text{ if } \left| \frac{\Delta_{\text{before}} - 0}{S_{\Delta_{\text{before}}}} \right| > t_{1-\alpha/2,\nu} \tag{12}$$

and

$$H_1 \text{ if } \left| \frac{\Delta_{\text{after}} - 0}{S_{\Delta_{\text{after}}}} \right| > t_{1-\alpha/2,\nu}. \tag{13}$$

This system of hypothesis tests is inappropriate for several reasons. Firstly, despite the sign of the gap being known to the tester, the second test is two-sided: therefore, if a treatment is so good at closing the gap between FLs and ALs that in fact ALs become *easier* to model than FLs, this test will conclude that the treatment "cannot explain the gap" if the effect is large enough (because the absolute gap will be bigger). Secondly, these tests do not consider the size of the decrease in the gap, but simply the size of the gap itself. That means that if a treatment causes a tiny decrease in the gap yet pushes it just past the $p = \alpha$ threshold, this treatment will be said to "explain the gap" despite its small effect.

Lastly, in both tests, the hypotheses are inverted: the gap between FLs and ALs is assumed to *not exist* in both cases when doing the t-test, despite the point of the preliminary analysis being that there *is* a gap. For the second test, because the *absence* of a gap is the desired result of the treatments, this means that the null hypothesis is that the treatment *works*. While this would be a straightforward way to set up a hypothesis test comparing averages (if one chooses to use it) and while there is no rule saying the null hypothesis should be the status quo, extreme care should be taken interpreting p-values and the role of the significance level α: the latter is no longer the false-positive rate of the treatment (the chosen, guaranteed, small probability that a treatment is detected as reducing the gap, given the ground truth that the gap stays) but the false-negative rate (the chosen, guaranteed, small probability that a treatment is detected as not reducing the gap, given the ground truth that the gap is gone). Only one of these rates can be controlled (by setting α). Since the status quo is that there is a gap, the assumption should be that this is the ground truth. Therefore, it should also be assumed problematic to fix the probability conditioned on the opposite of this ground truth.

Normally, the lower a p-value, the more confidence we have that we are not mistaken in thinking the treatment works. With every decreasing order of magnitude, a p-value gives higher confidence; in principle, the p value can keep approximating 0 indefinitely. When swapping hypotheses, however, a p-value *higher* than α is seen as giving confidence in the treatment, but it is unknown to what degree. What a confidence of $p = 5\%$ or $p = 10\%$ or $p = 90\%$ tells about the treatment is unclear.

This particular setup also allows interpreting the conclusion to always favor the treatment. Normally, decreasing α makes it harder for p-values of treatments to be considered significant. When swapping hypotheses, this stricter α (5%, 1%, 0.001%, ...) causes almost all p-values to accept the null hypothesis (the treatment working).

Let us now consider the p-value that is found, and what is concluded about it:

[19] Note that this test is not actually reported in the paper.

> *The Goldfish models exhibit numerically higher perplexity for agglutinative ($M = 143.62$) than fusional languages ($M = 132.63$), but this difference is not statistically significant ($t(137.36) = 1.180$, $p = 0.077$). Therefore, after taking byte premiums into effect, the Goldfish models do not exhibit the same performance gap that was demonstrated in previous research and in Section 3 above.*
> – *Arnett and Bergen*

The meaning of this 7.7% is that *if* there is truly no gap, averages even further apart than this would rarely occur – 7.7% of all re-executions of the experiment, or 1 in every 13. Yet, this rarity is used to *reject* that there is a gap. This is correct according to the hypothesis test as formulated, but is not a strong case.

It is in fact impossible to conclude, based on this p-value, that "the Goldfish models do not exhibit the same performance gap that was demonstrated". Consider the situation where the ground truth – which is unknown – is that there *is* a gap. In that case, the t-test statistic is drawn from an unknown distribution, and it could be that it has a *higher p-value* (i.e. it is less out-of-place) in this ground truth distribution. One possible scenario where this happens is sketched in Figure 5. Here, the fact that p was bigger than α is irrelevant, as it is still smaller than the p-value of the alternative hypothesis (the existence of a gap, i.e. the treatment not working).

B.2 Effect of using hypothesis tests of different type to measure before and after

Across **H1**, **H2**, and **H3**, the only treatment that is deemed significant is **H3**. This is concluded using the inverted t-test discussed above. However, this t-test comparing mean PPLs of FLs and ALs is never actually reported for the "before" case, despite the data being available to do so. Instead, several hypothesis tests for regression coefficients predicting PPL from morphological type are given. From what we could gather, the pre-analysis by Arnett and Bergen (Section 3) is intended to serve as stand-ins for the missing t-test.

In the quote above, it was stated that

> *(...) the Goldfish models do not exhibit the same performance gap that was demonstrated (...) in Section 3 above.*

Yet, section 3 reads:

> *This section describes three analyses that show lower performance for agglutinative languages.*
> *(...) there is still a significant effect of morphological type, where agglutinative languages had higher perplexities than fusional languages.*
> *(...) there is still a significant effect of morphological type, where fusional languages show better performance than agglutinative languages.*
> *(...) morphological type still explains additional variance ($\chi^2(3) = 3.3324$, $p = 0.02$).*
> *(...) we found a robust performance gap between agglutinative languages and fusional languages.*

All that is quantified is how predictive morphological type is in a linear regression with PPL as response, but this has no bearing on the size of the gap. Presumably the statements comparing FLs to ALs are about their average PPLs, but no size nor significance is stated.

The lack of a baseline t-test analogous to the one in **H3** also means that when α was chosen for the latter, there was no precedent set from a previous test, so it could have been set at the time the p-value was generated, as mentioned above.

B.3 Effect of many hypothesis tests

All main hypotheses (i.e. the three treatments **H1**, **H2**, **H3**) are studied with multiple significance tests (8, 4, and 2, respectively). If it so happens that the ground truth is that no alternative hypotheses hold, then the probability of finding a significant result in a set of m tests is

$$\begin{aligned} P(\text{any test sig.} \mid H_0) &= \sum_{i=1}^{m} P(\text{test } i \text{ sig} \mid H_0) \\ &= \alpha + \ldots + \alpha \\ &= \alpha \cdot m \end{aligned} \tag{14}$$

assuming each test uses the same significance level α. Thus, the odds of seeing rare (significant) events across the study increases proportional to the amount m of tests done, known as the *multiple comparisons problem*. A simple way to mitigate this is to adjust the significance level using a (Holm-)Bonferroni correction (Holm, 1979). In its most basic form, it replaces α by $\alpha' = \alpha/m$.

Note that although p-values and α guarantees are concerned with repeated computation of a specific statistic from a specific distribution, the Bonferroni correction should be applied even if each individual test is only performed once, by the above equation. (Nevertheless, some tests by Arnett and Bergen are partially about overlapping data, meaning their test decisions are not independent.)

B.4 Effect of large values on mean and t-test

In **H3**, the means of the PPL distributions of FLs and ALs are computed and compared using a Welch t-test. In Figure 6, we see that although most of the models trained by Arnett and Bergen achieve a PPL between 110 and 140, there are several much larger

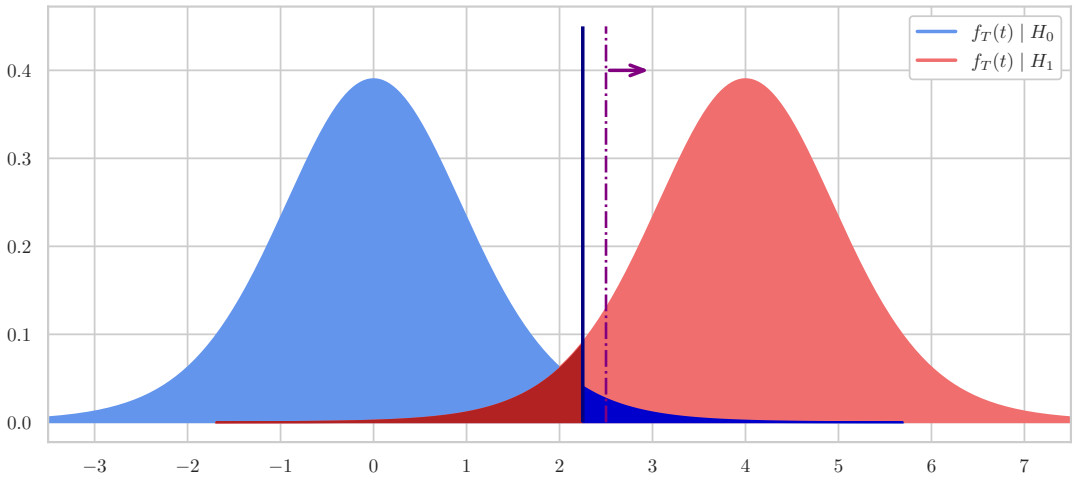


Figure 5 – Distribution of a T-test statistic under the null hypothesis (e.g. "no gap; treatment worked") and the alternative hypothesis (e.g. "gap; treatment did not work"). The purple line indicates the hypothesis threshold $t_{\alpha,\nu}$ under H_0. The blue line indicates a value of T which is not significant (it is to the left of t_α, or equivalently, its p-value under the null hypothesis – the dark blue area – is bigger than α), yet it would be more likely under H_1 (the dark red area is bigger than the dark blue area) despite causing H_1 to be rejected.

values, namely four that lie above 300. Note also that one morphological grouping has 1 such value (FLs) and the other has 3 (ALs).

If we assume it unproblematic to compare and aggregate PPL values across languages (despite our objections in §4.4). Even then, the existence of such much larger values is always problematic for the statistical tools used: either the tools break down due to the outlying values, or the tools are inapplicable due to a distribution mismatch.

B.4.1 If large values are outliers

If the large values are considered outliers of an otherwise normally distributed PPL, then they should be filtered out to not distort downstream statistics.

The *robustness* of a statistical estimator is the ability of its value to withstand perturbation due to outliers. *Breakdown* happens when the value can be changed arbitrarily much by replacing some amount of samples in the dataset by outliers (Rousseeuw and Hubert, 2011). In particular, the most basic sum-based estimators break down with just 1 outlying sample: the sample *mean* and corrected sample *(co)variance* are respectively

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{15}$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{16}$$

$$S_X^2 = S_{XX} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{17}$$

from which other tools for drawing conclusions are derived, such as *correlation* and the *hypothesis t-test statistic*

$$R_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}\,S_{XY}}} \tag{18}$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_T} \tag{19}$$

with various definitions of S_T all based on S_X^2.

Example. A list of samples $x = [1, 2, 3, 4, k]$ has median $\tilde{x} = 3$ for any k between 3 and $+\infty$. Meanwhile, the mean $\bar{x} = (10 + k)/5$ is 3 for $k = 5$, 30 for $k = 140$ and 300 for $k = 1490$. Its correlation with a list of samples $y = [10, 20, 30, 40, 50]$ is $r_{xy} = 1.00$ when $k = 5$, but drops to $r_{xy} = 0.89$ when $k = 10$, drops to $r_{xy} = 0.80$ when $k = 20$, and drops to $r_{xy} = 0.72$ when $k = 100$.

Arnett and Bergen compare *means* using a *Welch t-test* for the PPLs in **H3**, and compute *correlation* between PPL and CTC in **H2**. All three methods are not robust, meaning the results are distorted due to the outlying PPLs.

B.4.2 If large values are expected

If the large values are not considered outliers, then the assumption is that the distribution from which PPL values are drawn is one with a large right tail, and thus not a normal distribution. The consequence is that both the p-value and the acceptance of the null hypothesis, resulting from the Welch t-test, are invalidated.

When doing a two-sided t-test using the statistic in Eq. 19, its value is compared to the quantiles $t_{\alpha/2,\nu}$ and $t_{1-\alpha/2,\nu}$ of Student's t-distribution with ν degrees of freedom. The statistic is considered significant at significance level α (e.g. 1%) if it is even less than the former or even more than the latter. That is when the null hypothesis is rejected. The meaning of "at significance level α" is that under the null hypothesis, if the experiment were repeated many times on the same population the original sample set came from, the *fraction of experiments* that would turn out significant is α. In other words, the probability that the test statistic T falls outside $[t_{\alpha/2,\nu}\,;\,t_{1-\alpha/2,\nu}]$ is α. Since we reject the null hypothesis when this occurs, put differently, α is the probability of a type-I error. The only reason we can be certain of this probability is that the quantile $t_{\alpha/2,\nu}$ is exactly the value for q for which

$$P(T \leq q) = \alpha/2 \tag{20}$$

holds, because $T \sim t(\nu)$ and therefore we know

$$P(T \leq q) = \mathrm{CDF}_{t(\nu)}(q) \tag{21}$$

and only then can Eq. 21 inform Eq. 20 to get

$$\begin{aligned} \mathrm{CDF}_{t(\nu)}(q) &= \alpha/2 \\ q &= \mathrm{CDF}^{-1}_{t(\nu)}(\alpha/2) \\ &\equiv t_{\alpha/2,\,\nu} \end{aligned} \tag{22}$$

When $\nexists \mu, \sigma : X_i \sim \mathcal{N}(\mu, \sigma^2)$, we know Eq. 15 is no longer normally distributed, which, since it is used in Eq. 19, means $\nexists \nu : T \sim t(\nu)$, and thus Eq. 21 no longer holds. That means we do not know for which q Eq. 20 holds. Conversely, using $q = t_{\alpha/2,\nu}$, we do not know what significance level we are working with, so we do not know what the probability for a type-I error is.

B.4.3 Conclusion

As a first step for comparing PPLs, either the PPL samples should be made normally distributed through some transform, and/or the outliers should be filtered out (e.g. using robust statistics) to draw reliable conclusions from the data.

B.5 Effect of duplicating data in t-test

In **H2**, a Welch t-test is done to compare the mean CTC of tokenizers for FLs (n_1) versus ALs (n_2), and the same thing is done for the mean Rényi efficiency. A significant difference is found for the latter and a nearly significant difference for the former.

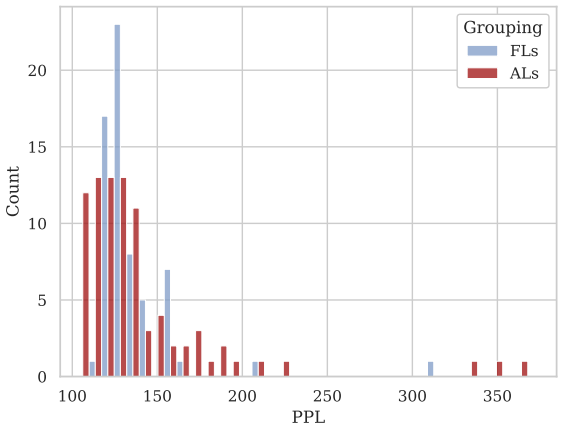


Figure 6 – Distribution of PPL values for the languages used in **H3** in Arnett and Bergen (2025).

However, when looking into the R scripts accompanying the paper, we found that before performing the t-tests, each of the n_1 and n_2 measurements is included $3\times$, i.e. each measurement is represented as three measurements rather than one. We show this in Figure 7, which incudes deduplicated measurements by language codes. The effect of this is an artificially lowered p-value in the hypothesis test, as we show below.

B.5.1 Visual intuition

Given two samples from two different normal distribution, a Welch t-test asks whether the mean of one is located significantly further from the other to conclude that they are not the same. For a one-sided test for whether the second distribution is to the right of the first, this conclusion is drawn when

$$\begin{aligned} H_1 \text{ if } T = \frac{\bar{X}_2 - \bar{X}_1}{S_u} &> t_{1-\alpha,\nu} \\ \bar{X}_2 &> \bar{X}_1 + t_{1-\alpha,\nu} \cdot S_u \end{aligned} \tag{23}$$

Visually, on a horizontal axis representing values of X, we could mark three points to represent the test: $\bar{x}_1$, $\bar{x}_2$, and a threshold which lies $t_{1-\alpha,\nu} \cdot s_u$ to the right of $\bar{x}_1$. If $\bar{x}_2$ falls to the right of this threshold, then it is significantly different from $\bar{x}_1$.

Example. We take two samples from two normally distributed populations with unequal variances, $\mathcal{N}(1, 1^2)$ and $\mathcal{N}(1.4, 1.05^2)$, each of size $n_1 = 25$ and $n_2 = 25$.

In Figure 8, we draw a solid blue line for $\bar{x}_1$ and a solid red line for $\bar{x}_2$. The rest is in light purple: the non-standard t-distribution $f(t) = \bar{x}_1 + s_u \cdot t$ is drawn around $\bar{x}_1$, the $\alpha = 2.5\%$ threshold

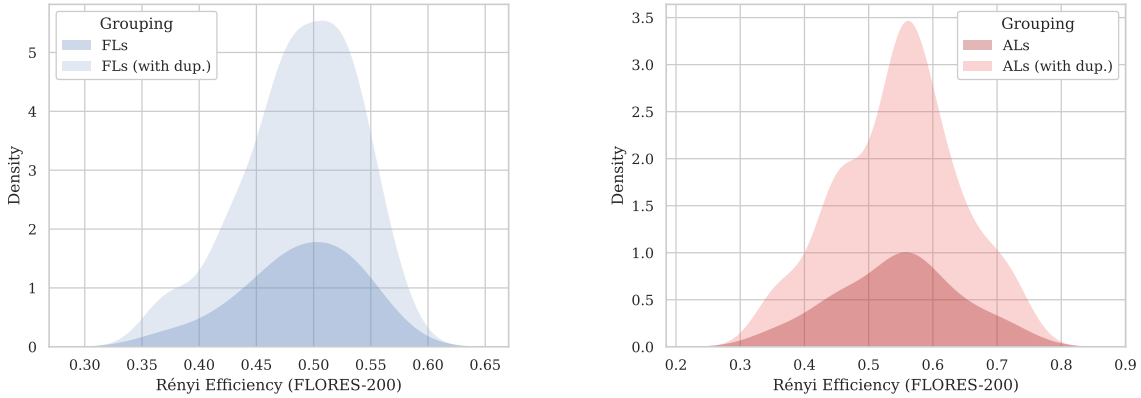


Figure 7 – Sample distributions used for **H2** with and without deduplication. The density scale is such that the *sum* of the areas under both distributions is 1, while keeping each area proportional to the amount of samples it describes.

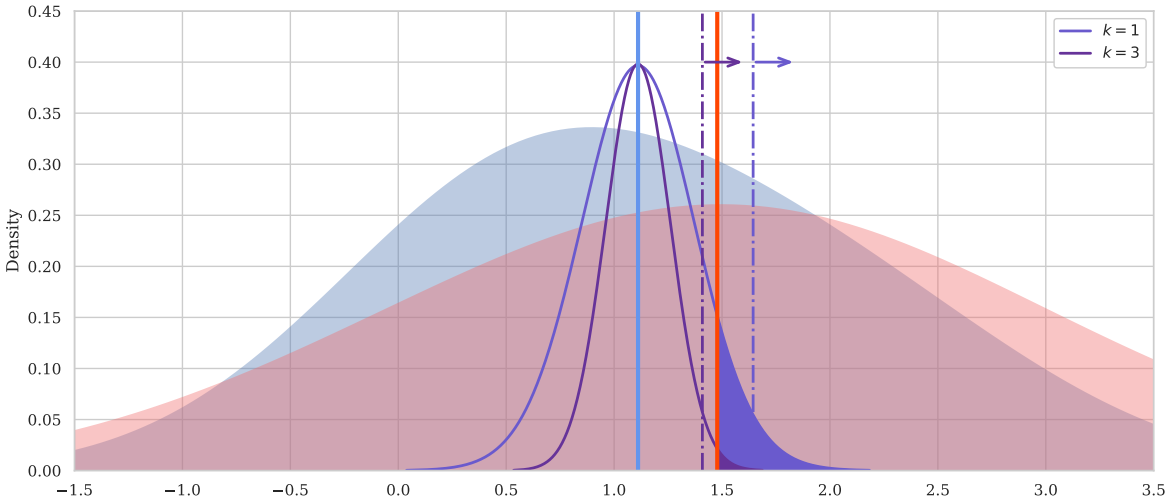


Figure 8 – Example of a Welch t-test between two samples for which the conclusion changes after including samples 3×.

is a dot-dash line at $t_{1-\alpha,\nu} \cdot s_u$ from the center, the p-value of $\bar{x}_2$ is colored as the area under the null-hypothesis distribution past $\bar{x}_2$. Visually, the hypothesis test can be executed either by checking whether the red line is to the right of the dot-dash line ($t > t_{1-\alpha,\nu}$), or whether the colored area is smaller than 2.5% of the total area ($p < \alpha$) under the drawn curve.

We see that the light purple area is large and that the red line stays to the left of the light purple dot-dash line. Thus, we *cannot* conclude from the data that $\bar{x}_2$ is so significantly far from $\bar{x}_1$ that its sample was probably taken from a different population.

On the same figure, we repeat the above drawing procedure but after concatenating the samples to themselves twice, resulting in $n_{1,*} = n_{2,*} = 75$ measurements for the both samples. We use dark purple this time.

Now we see a much smaller purple area and we see that the red line is now *right of* the dot-dash line.[20] The hypothesis test now results in the conclusion that $\bar{x}_2$ *is* significantly far from $\bar{x}_1$ supported by a small p-value below α, despite not having made any more measurements.

B.5.2 Mathematical proof

This hypothesis test is of the form

$$H_0 \text{ if } T = \frac{\bar{X}_1 - \bar{X}_2}{S_u} \in [t_{\alpha/2,\nu}; t_{1-\alpha/2,\nu}]. \quad (24)$$

Using the shorthand $V_i = \frac{S_i^2}{n_i}$ with S_i^2 given by Eq. 17, the unpooled variance estimator S_u^2 is

$$S_u^2 = V_1 + V_2 \quad (25)$$

[20] We will see below that the dot-dash line has moved left for two reasons: $s_{u,*} < s_u$ and $t_{1-\alpha,\nu_*} < t_{1-\alpha,\nu}$.

and ν is given by a Welch–Satterthwaite equation

$$\nu \approx \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \tag{26}$$

Now consider two samples of n_1 and n_2 measurements. When duplicating each measurement by a factor k, the following happens to the above quantities (where "$*$" denotes the new situation):

$$\bar{X}_* = \frac{1}{n_*} \sum_{j=1}^{n_*} X_j = \frac{1}{kn} \sum_{i=1}^{n} k\, X_i = \bar{X} \tag{27}$$

and thus

$$\begin{aligned} S_*^2 &= \frac{1}{n_* - 1} \sum_{j=1}^{n_*} (X_{j,*} - \bar{X}_*)^2 \\ &= \frac{1}{kn - 1} \sum_{i=1}^{n} k (X_i - \bar{X})^2 \\ &= \frac{1}{n - 1/k} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n-1}{n - 1/k} S^2 \\ &\approx S^2 \end{aligned} \tag{28}$$

and thus

$$V_* = \frac{S_*^2}{n_*} = \frac{S_*^2}{kn} \approx \frac{S^2}{kn} = \frac{1}{k} V \tag{29}$$

and thus

$$S_{u,*}^2 = V_{1,*} + V_{2,*} \approx \frac{1}{k}(V_1 + V_2) = \frac{1}{k} S_u^2 \tag{30}$$

and thus

$$T_* = \frac{\bar{X}_{1,*} - \bar{X}_{2,*}}{S_{u,*}} = \frac{\bar{X}_1 - \bar{X}_2}{S_u / \sqrt{k}} = \sqrt{k} \cdot T. \tag{31}$$

Also,

$$\begin{aligned} \nu_* &\approx \frac{(V_{1,*} + V_{2,*})^2}{\frac{V_{1,*}^2}{n_{1,*} - 1} + \frac{V_{2,*}^2}{n_{2,*} - 1}} \\ &\approx \frac{\left(\frac{V_1}{k} + \frac{V_2}{k}\right)^2}{\frac{\left(\frac{V_1}{k}\right)^2}{kn_1 - 1} + \frac{\left(\frac{V_2}{k}\right)^2}{kn_2 - 1}} \\ &= \frac{\frac{1}{k^2}(V_1 + V_2)^2}{\frac{1}{k^3}\left(\frac{V_1^2}{n_1 - 1/k} + \frac{V_2^2}{n_2 - 1/k}\right)} \\ &\approx k \cdot \nu \end{aligned} \tag{32}$$

and thus

$$(t_{1-\alpha/2,\,\nu})_* = t_{1-\alpha/2,\,\nu_*} = t_{1-\alpha/2,\,k\nu}\,. \tag{33}$$

Therefore, we can conclude the following: in the original dataset, for a hypothesis test to turn out significant, the T statistic (proportional to the gap between the means) was required to be so large that

$$H_1 \text{ if } |T| > t_{1-\alpha/2,\,\nu} \tag{34}$$

whereas in the inflated dataset, a significant hypothesis can already reached when

$$\begin{aligned} H_1 \text{ if} \qquad |T_*| &> t_{1-\alpha/2,\,\nu_*} \\ \sqrt{k} \cdot |T| &> t_{1-\alpha/2,\,k\nu} \\ |T| &> \frac{1}{\sqrt{k}} t_{1-\alpha/2,\,k\nu} \end{aligned} \tag{35}$$

which is a much less strict requirement on T since

$$t_{1-\alpha/2,\,\nu} > t_{1-\alpha/2,\,k\nu} > \frac{1}{\sqrt{k}} t_{1-\alpha/2,\,k\nu} \tag{36}$$

and thus even previously insignificant gaps $\bar{x}_1 - \bar{x}_2$ can now be flagged as significant.

B.6 Effect of large sample size on regression t-tests

Most regressions by Arnett and Bergen (2025) are performed on at most a hundred measurements. In **H1**, however, a regression is done with one measurement *per word*:

> *We fit a linear regression with number of tokens per word, word length in characters, and morphological types as predictors for MorphScore. We found that fertility and word length are both negatively correlated with MorphScore ($\chi^2(1) = 61.457$, $p < 0.001$; $\chi^2(1) = 364.03$, $p < 0.001$; respectively); however, the effect sizes were extremely small with an adjusted $R^2 = 0.021$.*

The response is modeled poorly ("small effect size") even though the hypothesis tests are highly significant. The reason for this is that the t-test[21] statistic that checks whether a regression coefficient is significantly different from 0, is proportional to $\sqrt{n}$ (Seber and Lee, 2003, p. 140). The p-values are therefore explained by the sample size of $n = 23\,952$.

The reason why the response is not modeled properly is likely due to heavily skewed predictors: the data published by Arnett and Bergen show that in almost all MorphScore languages, the token

[21] For high n, $t(\nu)^2 \to \chi^2(1)$, so test statistics can either be reported as being T-distributed or χ^2-distributed.

amount M is distributed geometrically ($P(M)$ is geometric) while the word length L is distributed binomially when keeping the token amount fixed ($P(L \mid M)$ is binomial). This means words with few tokens completely swamp words with many tokens, and to the regression, it looks like M is basically a constant while L is unrelated to it and binomially distributed.

Given the negligible modeling performance and the fact that across MorphScore languages, the correlation between $\bar{L} \mid M$ and M is around 77%, the sign of the regression coefficients should not be relied on to infer the signs of the correlations between the predictors and the response.

B.7 Causation does not imply correlation

In **H1**, a linear regression is performed that results in a poor model. The conclusion is that the tested predictors cannot explain the response:

> *We fit a linear regression with number of tokens per word, word length in characters, and morphological types as predictors for MorphScore. We found that fertility and word length are both negatively correlated with MorphScore ($\chi^2(1) = 61.457$, $p < 0.001$; $\chi^2(1) = 364.03$, $p < 0.001$; respectively); however, the effect sizes were extremely small with an adjusted $R^2 = 0.021$. Given these small effects, longer words or higher fertility cannot explain the greater than 20% higher MorphScores for agglutinative languages.* – Arnett and Bergen

The underlying assumption here is that *causation implies correlation*: that is, if a predictor explains a response, then it should linearly correlate with it. By the *modus tollens*, given that no correlation (linear relationship) is found – indicated by R^2, not by the p-value of the coefficients – it is assumed that the predictors do not explain the response. But causation does not imply correlation.

Example. We sample $N = 20\,000$ values from a random uniform variable $X \sim U(-1 + \Delta;\, 1 - \Delta)$ with $\Delta = 0.04$ (so, 20 000 random numbers between -0.96 and +1.04). We then square each of them and call the result Y, so that $Y = X^2$. We now perform a linear regression $Y = \beta_0 + \beta_1 X + \varepsilon$. Despite the perfectly deterministic, perfectly explanatory, causal relationship between predictor and response, we get an adjusted $R^2 = 0.021$. (And here too, the coefficient $\beta_1 = 0.076$ is significantly different from 0 with $t(19\,998) = 20.760$, or equivalently $\chi^2(1) \approx 400$, $p < 0.001$.)

C Experimental Setup

Dataset	Link	|L|	Language Coverage (ISO-639-3)
EuroParl (Koehn, 2005; Tiedemann, 2012)	huggingface.co/datasets/Helsinki-NLP/europarl	21	bul, ces, dan, deu, ell, eng, est, fin, fra, hun, ita, lav, lit, nld, pol, por, ron, slk, slv, spa, swe
Morpho Challenge (Kurimo et al., 2010)	morpho.aalto.fi/events/morphochallenge2010/datasets	3	eng,fin,tur
MorphyNet (Batsuren et al., 2021)	github.com/kbatsuren/MorphyNet	15	cat, ces, deu, eng, fin, fra, hbs, hun, ita, mon, pol, por, rus, spa, swe
MorphScore (Arnett and Bergen, 2025)	github.com/catherinearnett/morphscore	22	bul, ceb, ell, eng, eus, gle, guj, hun, hye, ind, isl, jpn, kat, kmr, kor, pes, slv, spa, tam, tur, urd, zul
FineWeb 1 & 2 (Penedo et al., 2024, 2025)	huggingface.co/datasets/HuggingFaceFW/fineweb huggingface.co/datasets/HuggingFaceFW/fineweb-2	63	afr, amh, arb, bel, ben, bos, bul, cat, ces, cym, dan, deu, ell, eng, epo, eus, fin, fra, glg, guj, heb, hin, hrv, hun, hye, ind, isl, ita, jpn, kan, kat, kaz, kir, kor, lat, lit, mal, mar, mkd, mlt, nld, nob, pan, pol, por, ron, rus, sin, slk, slv, som, spa, srp, swe, tam, tat, tel, tgk, tha, tur, ukr, urd, vie
Example sentence Table 4	tatoeba.org/en/sentences/show/13142837	-	-

Table 6 – Datasets used in our analysis. For FineWeb we report the languages we use, not all available languages.

Library	Purpose	Link
TkTkT	Tokenization toolkit.	github.com/bauwenst/TkTkT
MoDeST	Morphological datasets.	github.com/bauwenst/MoDeST
Qwanqwa	Language metadata.	github.com/WPoelman/qwanqwa
Fugashi (McCann, 2020)	Japanese word segmentation.	github.com/polm/fugashi
PyThaiNLP (Phatthiyaphaibun et al., 2023)	Thai word segmentation.	github.com/PyThaiNLP/pythainlp
–	Scripts by Arnett and Bergen (2025).	osf.io/jukzd
–	Scripts used for this paper.	github.com/LAGoM-NLP/ConfoundingFactors

Table 7 – Software used in our analysis.

Language	Source	|Words|
Turkish	MorphoChallenge	1000
Mongolian	MorphyNet	16822
Polish	MorphyNet	58711
Swedish	MorphyNet	94488
Catalan	MorphyNet	121749
German	MorphyNet	186395
Portuguese	MorphyNet	262116
Czech	MorphyNet	326147
French	MorphyNet	383064
English	MorphyNet	414967
Italian	MorphyNet	562273
Russian	MorphyNet	804279
Spanish	MorphyNet	882688
Hungarian	MorphyNet	939067
Finnish	MorphyNet	1629510

Table 8 – Source (MorphoChallenge Kurimo et al. 2010 or MorphyNet Batsuren et al. 2021) and size of the morphological alignment datasets we use. Note that there can be multiple segmentation boundaries per word.

D Full Results

		Token Bigrams				Token Unigrams						Words			
		AV		η		MATTR		MTL		RE		$\mathcal{S}$		MWL	
Language	G*	EP	FW	EP	FW (↓)	EP	FW	EP	FW	EP	FW	EP	FW	EP	FW
Thai	Iso		3.67		19.71		52.16		4.65		24.40		3.69		5.05
Vietnamese	Iso		3.37		26.44		34.88		3.78		40.31		7.98		4.17
Indonesian	Agg		6.01		29.55		41.18		5.11		36.09		11.79		6.36
Galician	Fus		9.74		30.80		31.29		2.69		42.71		24.53		5.55
French	Fus	2.39	6.13	19.11	32.35	34.27	39.78	5.08	4.54	40.30	41.73	2.30	4.75	5.91	5.42
Portuguese	Fus	3.06	6.79	21.31	33.02	35.38	41.75	4.91	4.41	36.38	35.52	10.64	11.83	5.79	5.45
English	Fus	2.12	5.74	15.92	33.20	31.78	40.26	4.89	4.31	36.68	35.22	9.27	11.99	5.54	5.24
Punjabi	Agg		6.19		33.32		39.06		4.14		44.86		5.19		4.79
Spanish	Fus	2.95	7.49	22.70	34.15	33.85	41.34	5.05	4.53	36.16	35.66	9.05	10.56	5.72	5.47
Bosnian	Fus		8.18		34.75		45.81		4.06		35.80		14.55		5.64
Macedonian	Fus		8.21		34.91		44.84		4.51		35.72		11.85		5.70
Japanese	Agg		7.13		35.27		48.22		1.77		42.59		49.50		1.95
Greek	Fus	4.20	12.20	24.48	35.81	38.71	45.60	5.11	4.36	37.44	36.96	10.35	12.39	6.15	5.84
Hebrew	Int		8.13		35.91		51.06		3.83		37.00		13.24		5.03
Catalan	Fus		8.10		36.14		40.43		4.15		35.41		12.37		5.26
Urdu	Fus		5.15		36.48		38.39		4.15		45.26		1.96		4.60
Hindi	Fus		6.79		36.51		38.85		4.27		44.70		2.27		4.78
Bulgarian	Fus	3.37	8.38	24.12	37.02	36.37	44.07	4.86	4.16	34.88	34.66	12.21	13.99	5.97	5.51
Esperanto	Agg		9.10		37.42		42.46		4.09		35.14		13.18		5.47
Afrikaans	Fus		6.78		37.52		38.32		4.23		36.91		10.64		5.23
Welsh	Fus		6.92		37.63		39.50		4.07		34.36		14.43		5.15
Dutch	Fus	3.33	7.57	20.75	37.87	33.85	41.02	5.17	4.55	37.83	37.41	8.36	10.63	6.01	5.68
Danish	Fus	3.84	7.80	24.12	38.22	33.32	40.94	4.78	4.11	35.53	36.00	11.91	13.54	5.82	5.46
Romanian	Fus	3.12	7.73	25.09	39.26	37.80	41.85	5.04	4.18	36.98	36.94	10.52	13.09	5.95	5.56
Swedish	Fus	3.84	8.06	24.18	39.42	35.90	43.06	5.11	4.46	39.79	38.17	8.73	10.54	6.10	5.60
Norwegian Bokmål	Fus		8.15		39.76		42.24		4.21		36.20		12.57		5.45
Lithuanian	Fus	6.26	11.95	33.62	39.79	44.11	51.10	5.00	4.42	32.26	32.59	16.58	17.79	6.61	6.30
Maltese	Int		11.09		40.40		39.23		4.07		34.47		6.72		6.32
Amharic	Int		9.19		41.06		53.28		3.56		54.21		8.45		5.03
Czech	Fus	4.58	10.37	30.07	41.70	43.06	49.01	4.70	4.13	35.15	35.95	13.67	15.02	6.01	5.73
Italian	Fus	3.65	8.65	27.10	41.74	37.56	44.52	5.22	4.52	38.85	36.68	9.39	11.63	6.21	5.65
Belarusian	Fus		12.10		41.95		50.46		4.20		32.81		17.73		6.01
Croatian	Fus		10.81		42.41		46.94		4.11		35.63		15.00		5.69
Standard Arabic	Int		8.19		42.58		45.47		4.08		50.07		5.39		5.42
Gujarati	Fus		7.71		42.61		44.30		4.36		42.27		5.58		5.27
Serbian	Fus		9.53		42.66		44.61		3.87		34.71		14.67		5.10
Tajik	Agg		7.82		42.77		44.01		4.50		34.50		14.11		5.86
Slovenian	Fus	4.09	9.85	32.04	42.84	40.42	44.57	4.77	3.97	33.74	34.59	13.66	16.72	5.88	5.65
Russian	Fus		11.89		43.19		47.24		4.10		32.92		18.26		6.01
Bengali	Agg		7.41		43.19		48.25		4.90		43.10		6.63		6.11
Ukrainian	Fus		11.96		43.40		46.68		4.04		32.62		18.91		6.01
Icelandic	Fus		9.99		43.43		44.74		4.29		37.72		11.73		5.68
Slovak	Fus	4.70	11.10	31.12	43.51	43.04	47.10	4.82	3.95	34.91	35.94	13.39	15.90	6.13	5.66
German	Fus	4.04	8.84	26.33	43.65	35.83	45.28	5.28	4.64	35.14	36.03	12.12	12.98	6.52	6.17
Polish	Fus	4.74	10.47	30.85	43.92	44.51	47.89	5.25	4.24	35.76	35.80	12.75	15.37	6.68	6.02
Kyrgyz	Agg		8.95		44.27		54.35		4.90		34.19		15.08		6.64
Georgian	Agg		10.27		45.03		49.64		4.84		33.18		17.05		6.90
Armenian	Agg		9.88		45.06		47.37		4.77		36.35		14.13		6.67
Latin	Fus		8.77		46.14		35.64		4.59		31.02		16.41		5.84
Tatar	Agg		9.58		46.74		51.88		4.54		33.22		16.12		6.11
Kazakh	Agg		9.25		47.53		52.62		5.18		41.94		3.77		6.59
Sinhala	Fus		9.03		47.93		50.53		4.59		42.48		8.43		5.73
Finnish	Agg	7.14	14.49	36.83	49.22	45.72	51.84	5.37	4.53	34.60	35.22	16.23	18.03	7.78	7.23
Telugu	Agg		13.52		49.33		52.24		5.05		42.21		4.65		6.72
Somali	Agg		8.64		49.63		45.72		5.13		46.47		2.70		6.02
Tamil	Agg		13.82		49.98		55.18		5.78		42.50		3.67		7.67
Turkish	Agg		9.66		50.14		49.71		4.61		36.11		15.11		6.50
Malayalam	Agg		20.34		50.77		57.93		5.63		43.08		5.35		8.56
Basque	Agg		9.81		51.58		49.98		4.85		33.77		16.01		6.67
Hungarian	Agg	6.69	13.66	39.11	52.99	41.73	47.70	5.05	4.24	34.10	33.86	14.63	17.26	6.78	6.24
Marathi	Fus		12.03		55.10		50.27		4.91		42.88		5.92		6.05
Korean	Agg		11.88		55.21		48.97		2.43		36.69		22.98		3.89
Kannada	Agg		15.55		55.37		55.73		5.25		41.97		7.42		7.08
Latvian	Fus	4.45		28.07		41.75		5.00		32.29		15.76		6.41	
Estonian	Agg	6.27		40.31		43.66		5.22		34.58		14.87		6.96	

Table 9 – Full results with pretokenization. EP = EuroParl; FW = FineWeb. All metrics except for AV, MTL, and MWL are multiplied by 100 for visual clarity. *The groupings are taken from Arnett and Bergen.

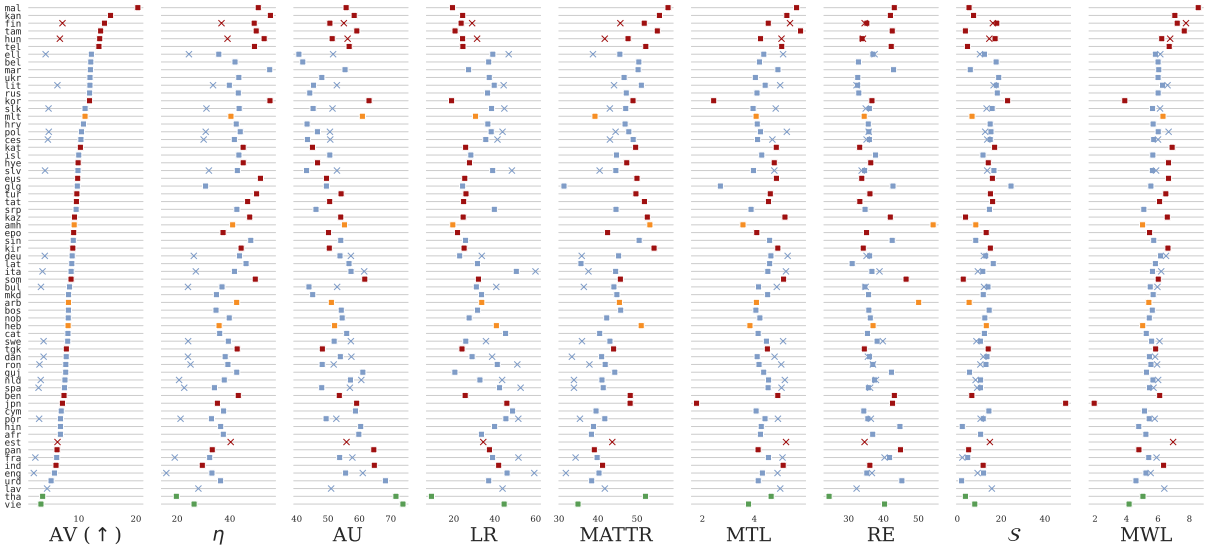
Language	G*	AV EP	AV FW	η EP	η FW (↓)
Thai	Iso		3.67		19.71
Vietnamese	Iso		3.37		26.44
Indonesian	Agg		6.01		29.55
Galician	Fus		9.74		30.80
French	Fus	2.39	6.13	19.11	32.35
Portuguese	Fus	3.06	6.79	21.31	33.02
English	Fus	2.12	5.74	15.92	33.20
Punjabi	Agg		6.19		33.32
Spanish	Fus	2.95	7.49	22.70	34.15
Bosnian	Fus		8.18		34.75
Macedonian	Fus		8.21		34.91
Japanese	Agg		7.13		35.27
Greek	Fus	4.20	12.20	24.48	35.81
Hebrew	Int		8.13		35.91
Catalan	Fus		8.10		36.14
Urdu	Fus		5.15		36.48
Hindi	Fus		6.79		36.51
Bulgarian	Fus	3.37	8.38	24.12	37.02
Esperanto	Agg		9.10		37.42
Afrikaans	Fus		6.78		37.52
Welsh	Fus		6.92		37.63
Dutch	Fus	3.33	7.57	20.75	37.87
Danish	Fus	3.84	7.80	24.12	38.22
Romanian	Fus	3.12	7.73	25.09	39.26
Swedish	Fus	3.84	8.06	24.18	39.42
Norwegian Bokmål	Fus		8.15		39.76
Lithuanian	Fus	6.26	11.95	33.62	39.79
Maltese	Int		11.09		40.40
Amharic	Int		9.19		41.06
Czech	Fus	4.58	10.37	30.07	41.70
Italian	Fus	3.65	8.65	27.10	41.74
Belarusian	Fus		12.10		41.95
Croatian	Fus		10.81		42.41
Standard Arabic	Int		8.19		42.58
Gujarati	Fus		7.71		42.61
Serbian	Fus		9.53		42.66
Tajik	Agg		7.82		42.77
Slovenian	Fus	4.09	9.85	32.04	42.84
Russian	Fus		11.89		43.19
Bengali	Agg		7.41		43.19
Ukrainian	Fus		11.96		43.40
Icelandic	Fus		9.99		43.43
Slovak	Fus	4.70	11.10	31.12	43.51
German	Fus	4.04	8.84	26.33	43.65
Polish	Fus	4.74	10.47	30.85	43.92
Kyrgyz	Agg		8.95		44.27
Georgian	Agg		10.27		45.03
Armenian	Agg		9.88		45.06
Latin	Fus		8.77		46.14
Tatar	Agg		9.58		46.74
Kazakh	Agg		9.25		47.53
Sinhala	Fus		9.03		47.93
Finnish	Agg	7.14	14.49	36.83	49.22
Telugu	Agg		13.52		49.33
Somali	Agg		8.64		49.63
Tamil	Agg		13.82		49.98
Turkish	Agg		9.66		50.14
Malayalam	Agg		20.34		50.77
Basque	Agg		9.81		51.58
Hungarian	Agg	6.69	13.66	39.11	52.99
Marathi	Fus		12.03		55.10
Korean	Agg		11.88		55.21
Kannada	Agg		15.55		55.37
Latvian	Fus	4.45		28.07	
Estonian	Agg	6.27		40.31	

(a) With pretokenization.

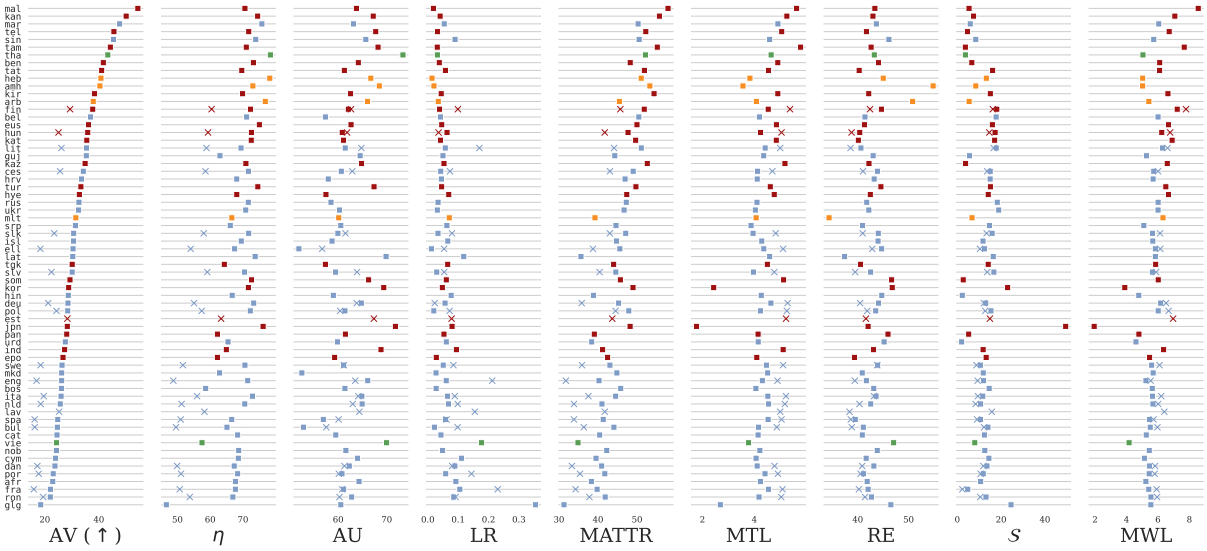
Language	G*	AV EP	AV FW	η EP	η FW (↓)
Galician	Fus		18.50		46.61
Vietnamese	Iso		24.34		57.51
Bosnian	Fus		26.16		58.62
Esperanto	Agg		26.74		62.21
Punjabi	Agg		28.13		62.22
Macedonian	Fus		26.25		62.87
Gujarati	Fus		35.36		62.98
Tajik	Agg		30.14		64.31
Indonesian	Agg		27.34		64.93
Bulgarian	Fus	16.28	24.66	49.56	65.12
Urdu	Fus		27.60		65.45
Serbian	Fus		31.34		66.13
Spanish	Fus	16.25	24.81	50.99	66.57
Maltese	Int		31.49		66.61
Hindi	Fus		28.70		66.71
Romanian	Fus	19.39	22.04	53.77	66.93
Danish	Fus	17.25	23.76	49.90	67.34
Greek	Fus	18.37	30.46	54.04	67.46
French	Fus	15.93	22.08	50.69	67.70
Afrikaans	Fus		22.95		67.71
Croatian	Fus		33.60		68.10
Armenian	Agg		32.80		68.12
Catalan	Fus		24.52		68.38
Portuguese	Fus	17.75	23.11	51.09	68.39
Welsh	Fus		23.98		68.60
Norwegian Bokmål	Fus		24.31		68.70
Lithuanian	Fus	26.16	35.43	58.85	69.42
Icelandic	Fus		30.65		69.50
Tatar	Agg		41.04		69.65
Kyrgyz	Agg		38.40		69.89
Slovenian	Fus	22.47	30.14	59.12	70.49
Dutch	Fus	18.49	25.72	51.26	70.56
Swedish	Fus	18.46	26.38	51.66	70.61
Malayalam	Agg		54.45		70.61
Ukrainian	Fus		32.51		70.85
Kazakh	Agg		34.93		70.91
Tamil	Agg		44.26		71.03
Belarusian	Fus		36.92		71.13
English	Fus	16.95	26.24	48.75	71.46
Russian	Fus		32.64		71.63
Korean	Agg		28.84		71.66
Czech	Fus	25.68	34.26	58.56	71.67
Slovak	Fus	23.50	30.69	58.02	71.75
Telugu	Agg		45.59		71.77
Polish	Fus	24.32	28.49	57.41	72.26
Finnish	Agg	29.35	37.74	60.43	72.32
Georgian	Agg		35.55		72.54
Hungarian	Agg	25.08	35.90	59.34	72.60
Somali	Agg		29.32		72.62
Italian	Fus	19.61	26.03	55.98	72.90
Amharic	Int		40.41		73.03
Bengali	Agg		41.66		73.20
German	Fus	21.25	28.58	55.07	73.30
Latin	Fus		30.36		73.73
Sinhala	Fus		45.41		73.91
Kannada	Agg		50.12		74.50
Turkish	Agg		33.36		74.55
Basque	Agg		36.19		75.02
Marathi	Fus		47.70		75.79
Japanese	Agg		28.33		76.16
Standard Arabic	Int		37.93		76.82
Hebrew	Int		40.78		78.20
Thai	Iso		43.33		78.39
Latvian	Fus	25.30		58.24	
Estonian	Agg	28.36		63.34	

(b) Without pretokenization.

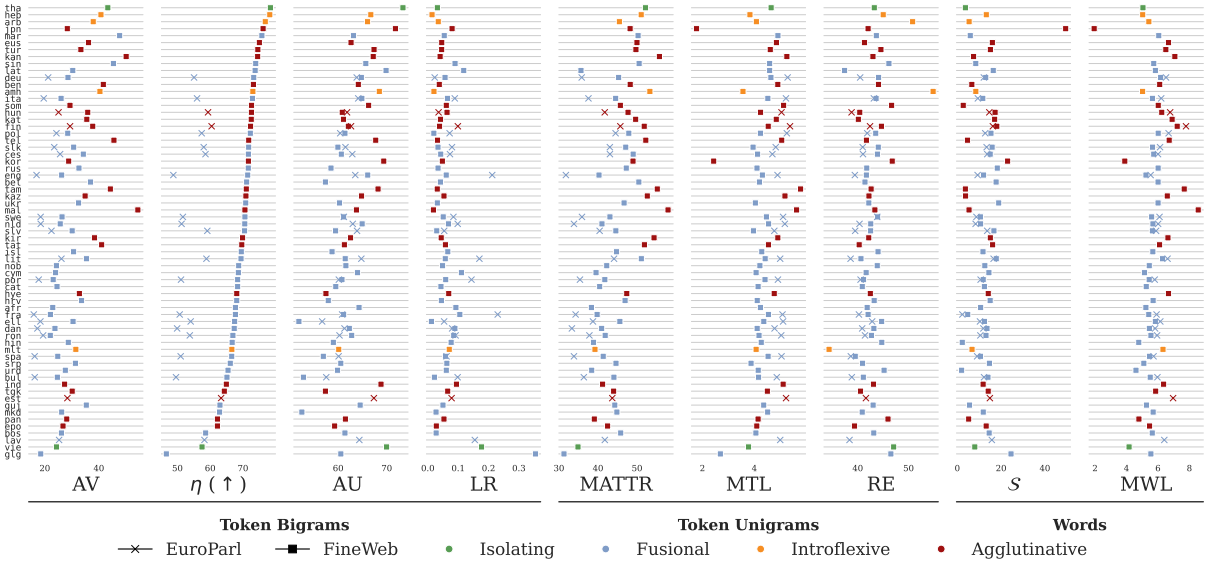
Table 10 – AV and η results with and without pretokenization. EP = EuroParl; FW = FineWeb. The η results are multiplied by 100 for visual clarity. *The groupings are taken from Arnett and Bergen.



(a) Results sorted by AV. The bigram metrics are calculated *with* pretokenization.



(b) Results sorted by AV. The bigram metrics are calculated *without* pretokenization.



(c) Results sorted by η. The bigram metrics are calculated *without* pretokenization.

Figure 9 – Only the sorting and pretokenization for the bigram metrics change, the other results are identical to Figure 2.