Confounding Factors in Relating Model Performance to Morphology

EMNLP 2025

Wessel Poelman* Thomas Bauwens* Miryam de Lhoneux

LAGoM-NLP, Department of Computer Science, KU Leuven

November 5, 2025

Are certain human languages easier or harder to model?

Conditional Language Models: *_wel come _every one*
Morphology: gather {s, ed, ing}

Language characteristics $\leftrightarrow$ **CLMs**

Morphological Complexity

- **Fusional**: aaaaa/bc
  inflection; one morpheme multiple features; shorter words; fewer morphemes
- **Agglutinative**: wwwww/xx/yy/zz
  one feature per morpheme; longer words; many morphemes
- ...

"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." – Sapir (1921)

Language characteristics $\leftrightarrow$ CLMs

- **Languages**
- **Grouping**
- **Tokenization algorithm**
- **Vocabulary size vs. data size**
- **Corpus**
- **Performance indicator**

Confounding Factors: affect what is *measured* and the *conclusions*.

Ideal $\rightarrow$ **feasible**

Hypotheses: ALs $<<$ FLs?

Award-winning[1] research from Arnett & Bergen (2025): hypotheses.

[1] Best Paper Award at COLING 2025.

Hypthesis 1: Subword tokenization is less morphologically aligned for ALs

MorphScore: **recall** stem-suffix boundaries

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

Hypothesis 2: Subword vocabularies are used more inefficiently for ALs

Rényi Efficiency (RE): H_α/H_0

FLORES-200 (2k lines) $\rightarrow$ higher RE for ALs

Not seen on larger corpora (EuroParl or FineWeb-2: 200k+ lines)

Unigrams and **morphological complexity**?

Hypothesis 3: Less training data is available for ALs

L_{JA}: 150 UTF-8 bytes

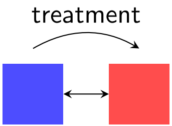
L_{EN}: 100 UTF-8 bytes

L_{JA} has a $1.5\times$ byte premium (Arnett et al., 2024)

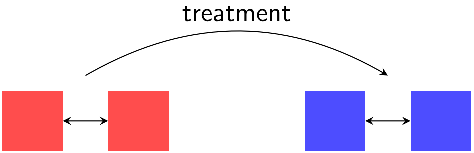
Scale L_{JA} data $\times 1.5 \rightarrow$ decrease performance gap[2]

[2] $p = 0.07$

Hypothesis 3: Less training data is available for ALs



(a) Hypothesis test of difference



(b) Difference of hypothesis tests

Confounding Factors: Round Two

- **Languages:** $H1 \cap H2 \cap H3 = 3$
- **Grouping:** Sapir (1921)
- **Tokenization algorithm:** Consistent*
- $|V|$ **vs. data size:** $^*(H1 = H2) \neq H3$
- **Corpus:** $^*(H1 = H2) \neq H3$
- **Performance indicator:** MorphScore, PPL, CTC, RE, ...

How can we be certain what caused observed effects?

Gradient measure of morphological complexity; relevant to **CLMs**

Accessor Variety

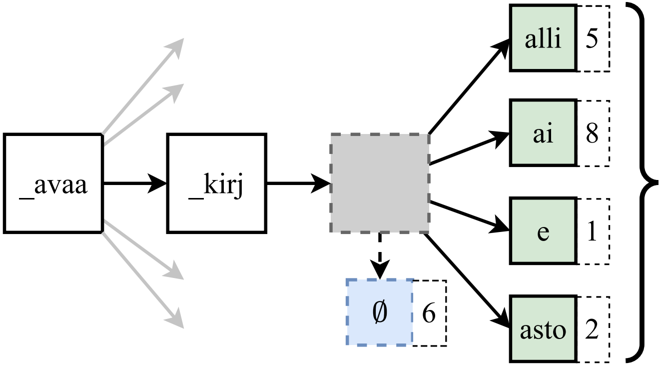
Old idea (Harris, 1955):
count predecessors and successors; unusual spikes → morpheme or word

Feng et al. (2004): Accessor Variety → minimum predecessor and successor variety

Apply to tokenizer **vocabulary**!

Sliding window (like MATTR) → text size.

Accessor Variety



$$\mathrm{AV}(\_\mathrm{kirj}) = 4$$

$$\mathrm{TA}(\_\mathrm{kirj}) = 16$$

$$\mathrm{AU}(\_\mathrm{kirj}) = \frac{4}{16}$$

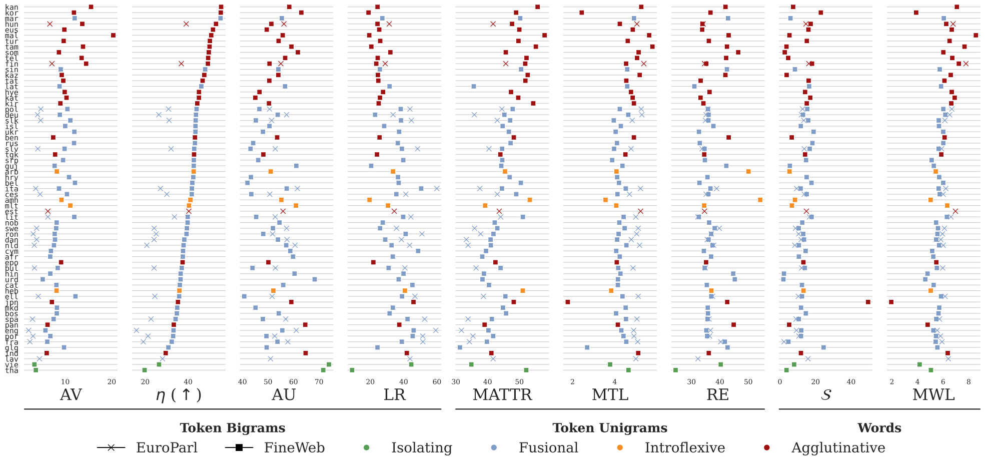
$$\mathrm{BR}(\_\mathrm{kirj}) = \frac{6}{6+16}$$

$$\eta(\_\mathrm{kirj}) = \frac{1.649}{\log_2(16)}$$

Results: Multi-parallel; EuroParl

		Token Bigrams				Token Unigrams			Words	
Language	Grouping*	AV	η ($\downarrow$)	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

Results: Larger Language Set; FineWeb-2



Conclusions

1. Experimental design!
2. AV: bridge insights morphology $\leftrightarrow$ CLMs

Acknowledgements

We thank **Kushal Tatariya** for feedback on a draft of this work,
Catherine Arnett for answering our questions,
and the **Anonymous Reviewers** for surprisingly constructive feedback!

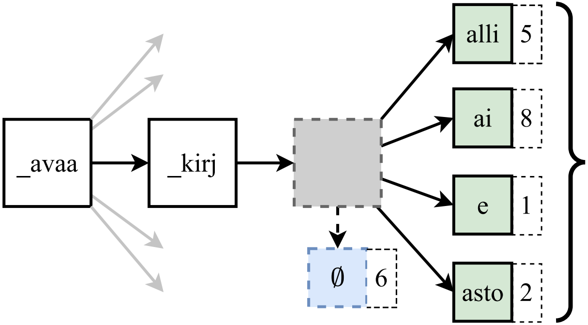
WP and TB are funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096.

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.

Thank you! Questions? Feedback?

`wessel.poelman@kuleuven.be`



$$\mathrm{AV}(\_\mathrm{kirj}) = 4$$

$$\mathrm{TA}(\_\mathrm{kirj}) = 16$$

$$\mathrm{AU}(\_\mathrm{kirj}) = \frac{4}{16}$$

$$\mathrm{BR}(\_\mathrm{kirj}) = \frac{6}{6+16}$$

$$\eta(\_\mathrm{kirj}) = \frac{1.649}{\log_2(16)}$$

Experimental Variables

Experiment	$\lvert L \rvert$	ALs	FLs	$\lvert V \rvert$	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*

Language Coverage

Hypotheses	$\lvert L \rvert$
H1 $\cap$ **H2**	3
H1 $\cap$ **H3**	22
H2 $\cap$ **H3**	52
H1 $\cap$ **H2** $\cap$ **H3**	3
H1 $\cup$ **H2** $\cup$ **H3**	145

PPL Outliers