Morphology Matters

Hyunji Hayley Park, Katherine J. Zhang, **Coleman Haley**, Kenneth Steimel, Han Liu, & Lane Schwartz

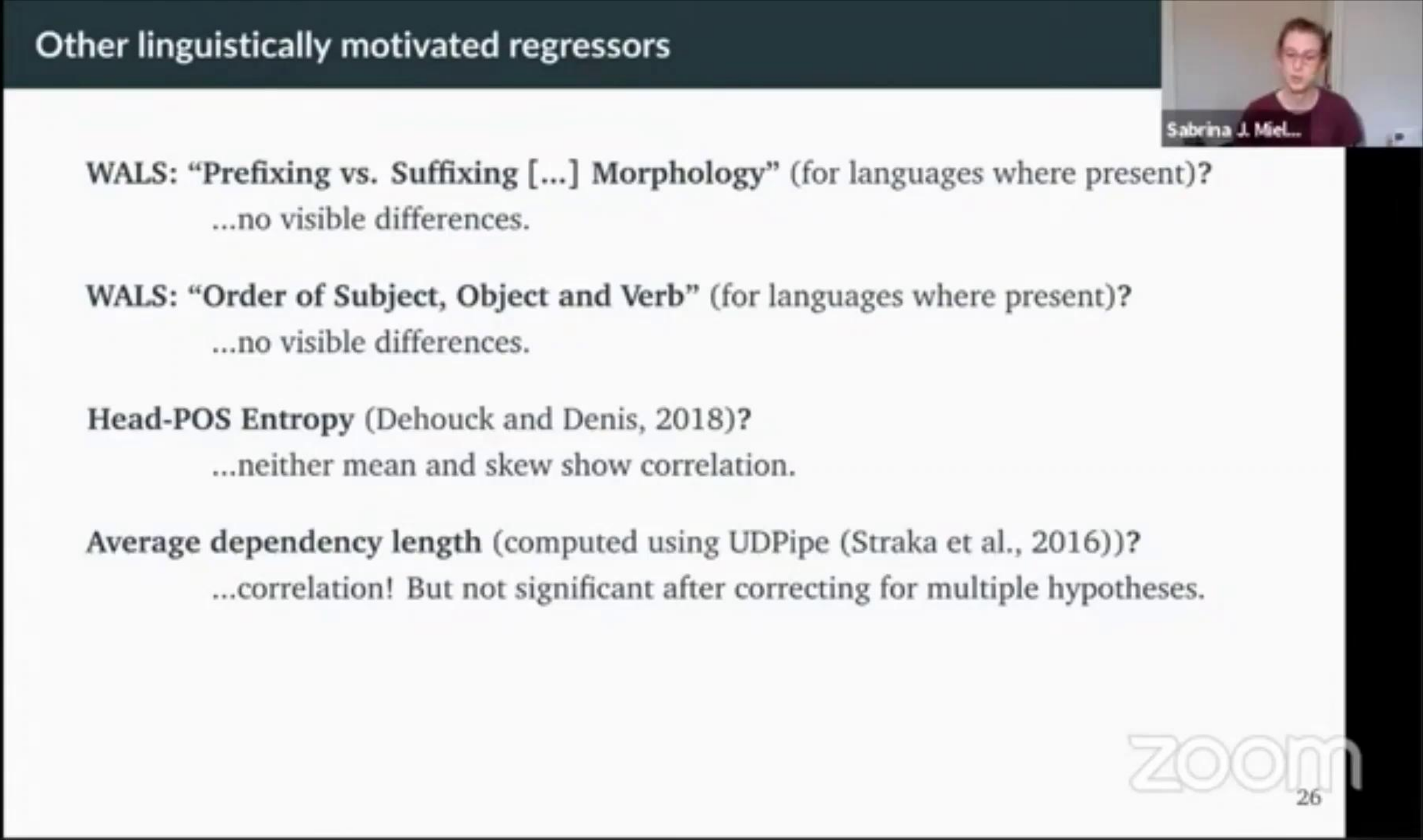
Why this paper exists.

2018 -- Cotterell et al. “Are all languages equally hard to language model?”

- Using Europarl, found correlations between morphological counting complexity and LM performance.

2019 -- Mielke et al. “What Kind of Language Is Hard to Language-Model?”

- Returned to this problem with an additional corpus of 62 bibles. Concluded that morphology was not a major factor in language modelling performance.
- AHHHHHH!



Other linguistically motivated regressors

WALS: "Prefixing vs. Suffixing [...] Morphology" (for languages where present)?

...no visible differences.

WALS: "Order of Subject, Object and Verb" (for languages where present)?

...no visible differences.

Head-POS Entropy (Dehouck and Denis, 2018)?

...neither mean and skew show correlation.

Average dependency length (computed using UDPipe (Straka et al., 2016))?

...correlation! But not significant after correcting for multiple hypotheses.

What *did* they find, then?

BPE LM performance correlates with type-token ratio…

Hmmmm... haven’t people said that the type-token ratio can be a proxy for morphology? (Kettunen 2014, Bentz et al. 2016)

Mielke et al. only tried 1 WALS morphology feature… and why would prefixing vs. suffixing actually matter?

And, their corpus still has a pretty strong bias to European languages…

Goal: replicate and expand upon Mielke et al. 2019, and show that morphology actually matters (hopefully).

Additionally: How does segmentation method play into this...

1. Creating a new dataset

Starting point: the 106 Bibles/62 languages from Mielke et al. 2019

Removed Esperanto and Klingon: 104 Bibles/60 languages

Added 41 bibles from 32 languages from previous corpora and new scraping. Large addition of polysynthetic langauges.

- Tosk Albanian, Amharic, Zarma, Hebrew, Icelandic, Japanese, Korean, Paite, Slovak, Slovene, Spanish, Swedish, Thai, **Plains Cree, Guarani, East Bolivian Guarani,** Hindi, **East Canadian Inuktitut, Inuinnaqtun,** Kannada, Malayalam, Marathi, **Central Huasteca Nahuatl**, Nepali, **Eastern Huasteca Nahuatl**, Western Persian, Polish, Shona, Telugu, **Toba Qom**, Turkish, **Central Alaskan Yupik, Greenlandic**

Resulting Data: 145 Bibles/92 languages

2. Expert-based measures of morphology

Previous work took 1 of 2 approaches:

1. Take feature values from WALS and call it a day, or
2. Slap a label like “agglutinative”, “fusional” on each language under consideration, call it a day.

Well, the second seems hard to scale, and there are problems with the first...

The WALS problem

WALS data is very incomplete, often wrong.

6 of our languages not in WALS, many with no features for morphology.

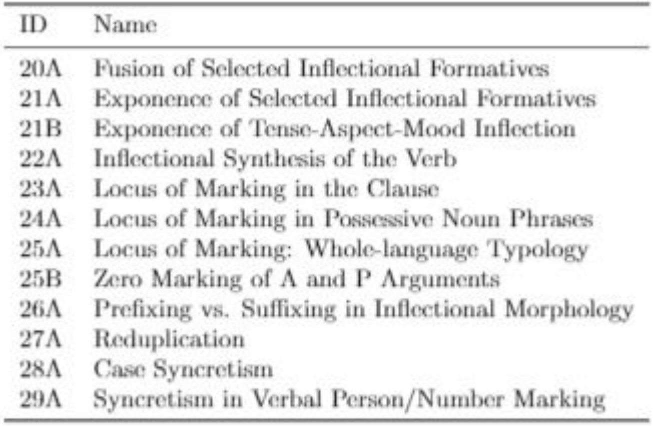
Bentz et al. 2016 concluded lower correlations of other morphological measures to WALS was due to missing data.

Mielke et al. selected “Prefixing vs. Suffixing morphology” primarily due to it being one of the most complete morphological features in WALS

The dream--investigate **every feature** labelled as morphological in WALS (12)

- This way, no judgement calls about what features to include.

So, we cracked open ~20 grammars and filled in and corrected the WALS feature values for many of the languages under consideration. (Huge thanks to Katherine Zhang!)



ID	Name
20A	Fusion of Selected Inflectional Formatives
21A	Exponence of Selected Inflectional Formatives
21B	Exponence of Tense-Aspect-Mood Inflection
22A	Inflectional Synthesis of the Verb
23A	Locus of Marking in the Clause
24A	Locus of Marking in Possessive Noun Phrases
25A	Locus of Marking: Whole-language Typology
25B	Zero Marking of A and P Arguments
26A	Prefixing vs. Suffixing in Inflectional Morphology
27A	Reduplication
28A	Case Syncretism
29A	Syncretism in Verbal Person/Number Marking

Table 1: The 12 morphological features in WALS.

3. Corpus-based measures

TTR/MATTR

- Number of distinct *types* divided by number of *tokens*
- Higher is more complex
- MATTR: average over a moving window -- comparable across lengths
- Kettunen 2014 -- strong correlations with number of distinct noun forms, weak correlation with verb synthesis
- Bentz et al. 2016 -- correlated with word entropy, word alignment, WALS features, within-word entropy.
- Useful because it can be compared with all languages

Mean Length of Word -- expected to be longer in morphologically complex languages, though this interacts with writing system.

4. Segmentation methods

Looking ahead--if morphology affects language modelling, does segmentation matter for this?

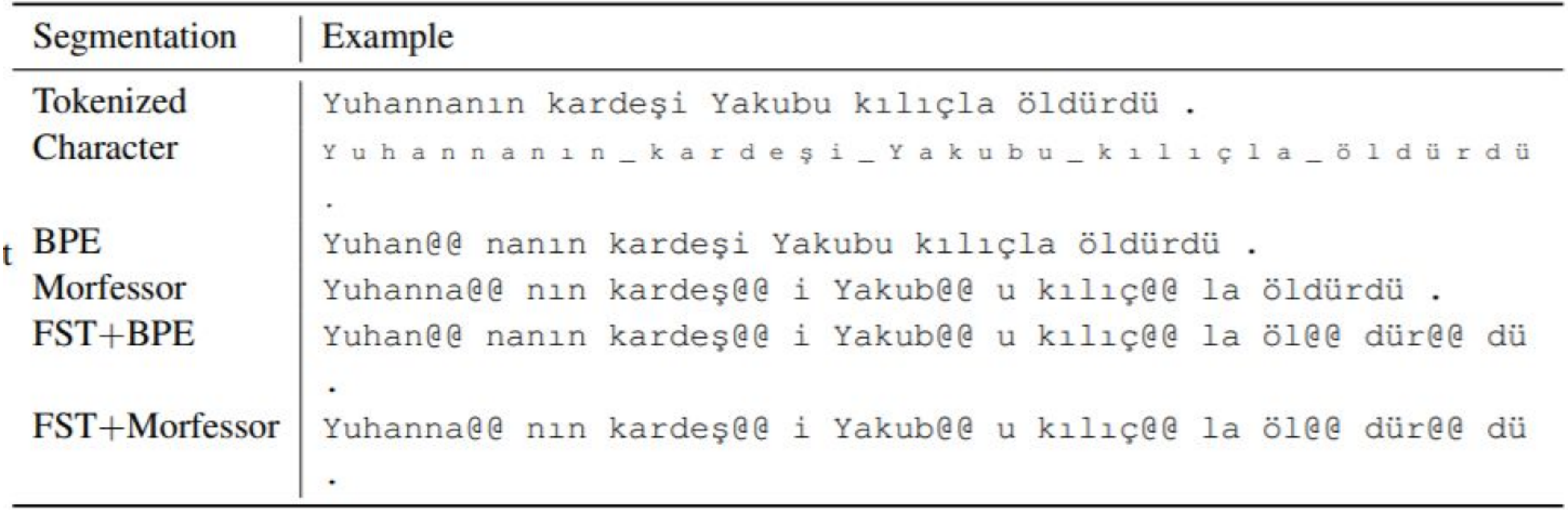
Maybe segmentations that are more closely aligned with morphology will do better.

So in addition to Mielke et al.’s character and BPE, we tried:

- Morfessor (default settings)
- FST+BPE
- FST+Morfessor

With BPE -- we used 0.4 x types just like Mielke et al. did.

- Does this invalidate TTR correlations? *No*.



t

Segmentation	Example
Tokenized	Yuhannanın kardeşi Yakubu kılıçla öldürdü .
Character	Y u h a n n a n ı n _ k a r d e ş i _ Y a k u b u _ k ı l ı ç l a _ ö l d ü r d ü .
BPE	Yuhan@@ nanın kardeşi Yakubu kılıçla öldürdü .
Morfessor	Yuhanna@@ nın kardeş@@ i Yakub@@ u kılıç@@ la öldürdü .
FST+BPE	Yuhan@@ nanın kardeş@@ i Yakub@@ u kılıç@@ la öl@@ dür@@ dü .
FST+Morfessor	Yuhanna@@ nın kardeş@@ i Yakub@@ u kılıç@@ la öl@@ dür@@ dü .

FST + X

For a morphologically diverse subset of 7 languages, we tried out a simple segmentation method incorporating FST morphological analysis.

Some FSTs were analyzers, some had existing segmenters.

If the FST was an analyzer, we removed the morpheme boundary cleanup rules and used some fancy manipulation to get the FST to map from surface to segmentation.

When there were multiple segmentations, we chose the output with the fewest segments > 1 segment.

When a word wasn’t analyzed by the FST, we used BPE or Morfessor trained on the corpus to segment it.

Models & Metrics

Salesforce LSTM(!) LM, hyperparameters from Mielke et al. 2019

Use surprisal per verse, which has been argued to be comparable across languages and segmentations.

5. Results: WALS

We use the Kruskal-Wallis test (one-way ANOVA on ranks)

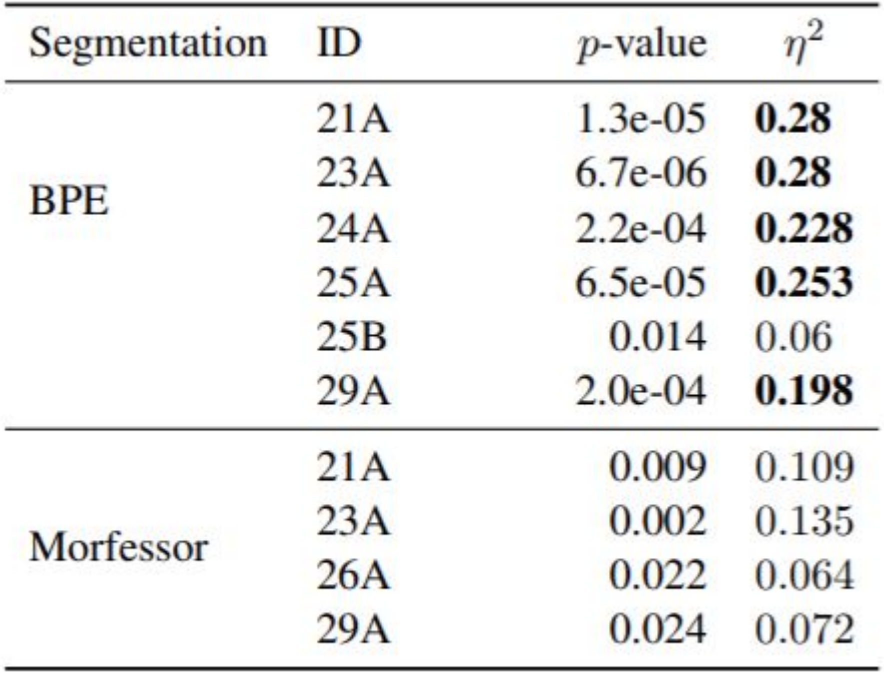
We use it because it is non-parametric and the surprisals were not normally distributed.

Character models: no correlation with WALS features

BPE models: 6/12 features have significant correlations! Mostly large effect size

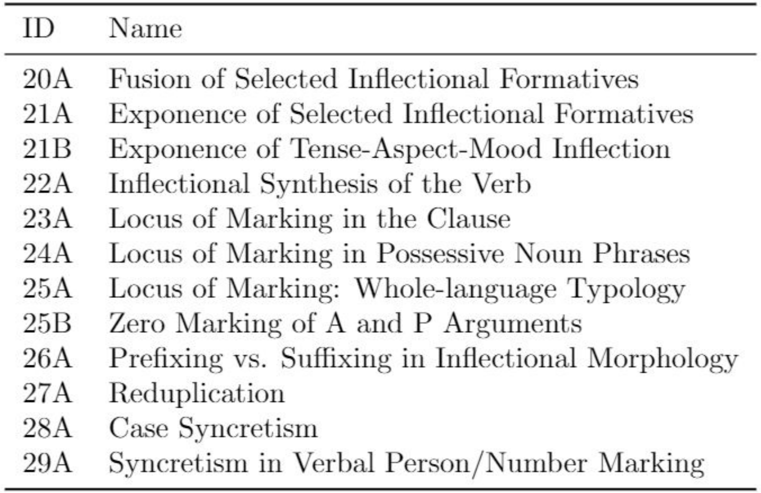
- Note that prefixing vs. suffixing was not significant, replicating Mielke et al.

Morfessor: 4/12 features have significant correlations. Small effect size



Segmentation	ID	p-value	η^2
BPE	21A	1.3e-05	**0.28**
	23A	6.7e-06	**0.28**
	24A	2.2e-04	**0.228**
	25A	6.5e-05	**0.253**
	25B	0.014	0.06
	29A	2.0e-04	**0.198**
Morfessor	21A	0.009	0.109
	23A	0.002	0.135
	26A	0.022	0.064
	29A	0.024	0.072

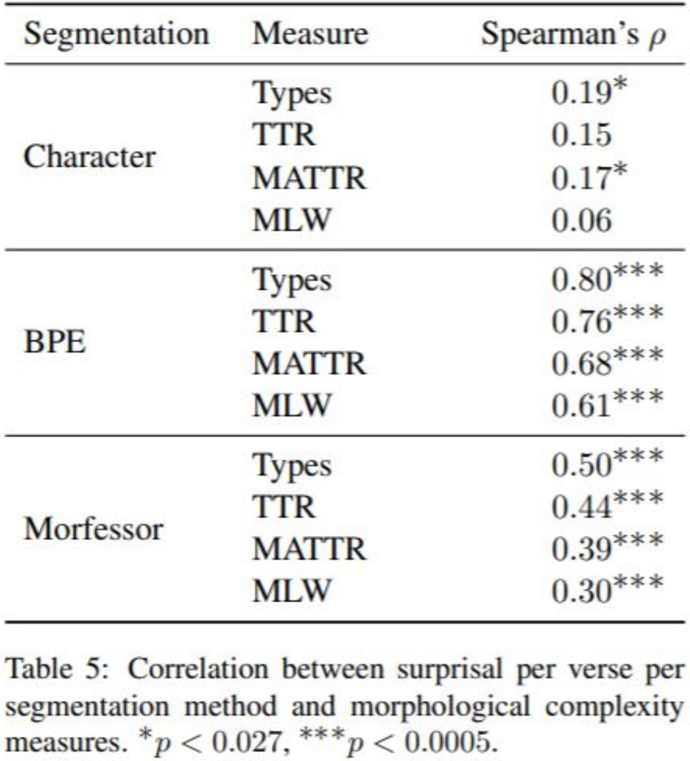
Table 4: p-values and effect sizes of WALS features that showed significant effect on surprisal per verse. Large effect sizes (≥ 0.14) are in bold.



ID	Name
20A	Fusion of Selected Inflectional Formatives
21A	Exponence of Selected Inflectional Formatives
21B	Exponence of Tense-Aspect-Mood Inflection
22A	Inflectional Synthesis of the Verb
23A	Locus of Marking in the Clause
24A	Locus of Marking in Possessive Noun Phrases
25A	Locus of Marking: Whole-language Typology
25B	Zero Marking of A and P Arguments
26A	Prefixing vs. Suffixing in Inflectional Morphology
27A	Reduplication
28A	Case Syncretism
29A	Syncretism in Verbal Person/Number Marking

Table 1: The 12 morphological features in WALS.

6. Results: Corpus based



Segmentation	Measure	Spearman's ρ
Character	Types	0.19^{*}
	TTR	0.15
	MATTR	0.17^{*}
	MLW	0.06
BPE	Types	0.80^{***}
	TTR	0.76^{***}
	MATTR	0.68^{***}
	MLW	0.61^{***}
Morfessor	Types	0.50^{***}
	TTR	0.44^{***}
	MATTR	0.39^{***}
	MLW	0.30^{***}

Table 5: Correlation between surprisal per verse per segmentation method and morphological complexity measures. $^{*}p < 0.027$, $^{***}p < 0.0005$.

“Wow! Seems like character is the way to go!”

--a wrong person

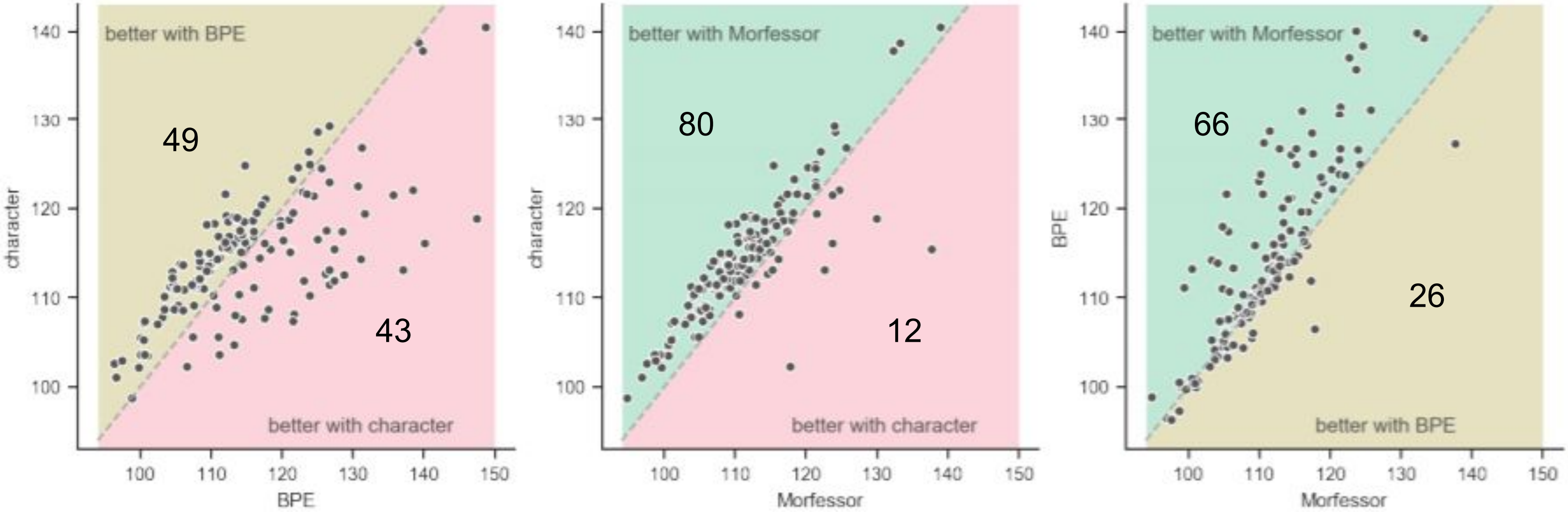
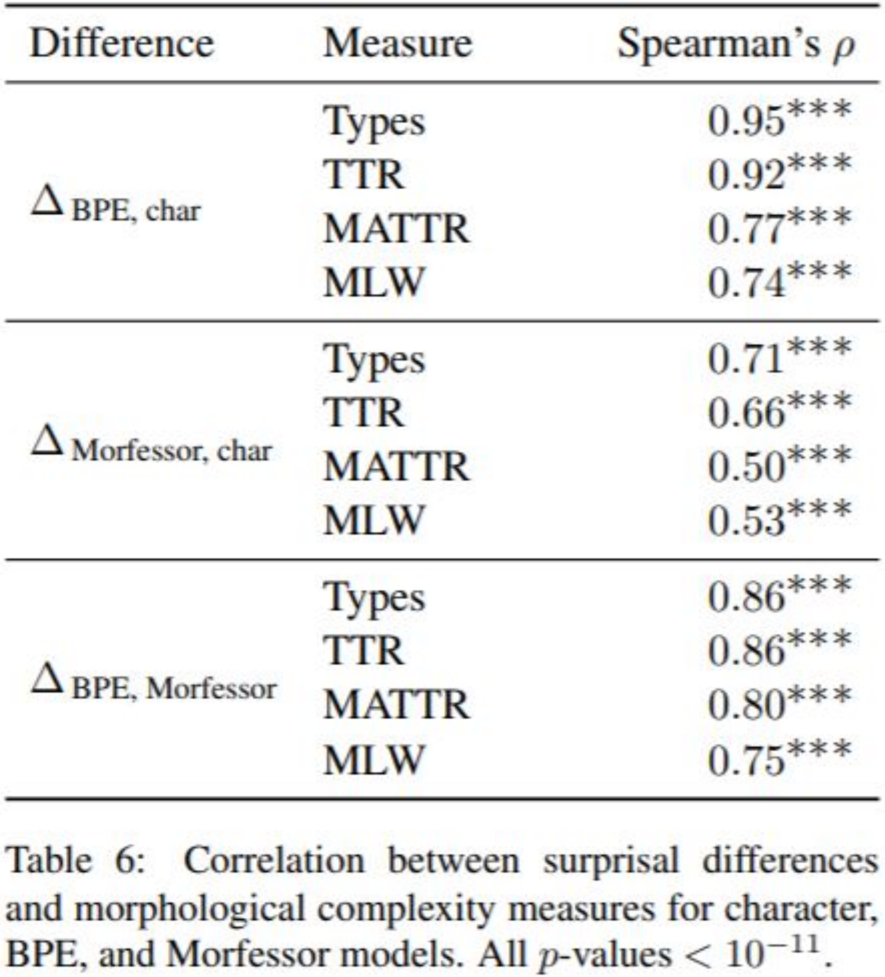


Figure 1: Pairwise comparisons of surprisal per verse values for character, BPE, and Morfessor models. For the majority of the languages, Morfessor segmentation resulted in lower surprisal per verse than character or BPE segmentation.

All correlations are positive.

- Character outperforms BPE and morfessor for very morphologically rich languages still...
- Morfessor outperforms BPE by more on morphologically rich languages.



Difference	Measure	Spearman's ρ
$\Delta_{\text{BPE, char}}$	Types	0.95^{***}
	TTR	0.92^{***}
	MATTR	0.77^{***}
	MLW	0.74^{***}
$\Delta_{\text{Morfessor, char}}$	Types	0.71^{***}
	TTR	0.66^{***}
	MATTR	0.50^{***}
	MLW	0.53^{***}
$\Delta_{\text{BPE, Morfessor}}$	Types	0.86^{***}
	TTR	0.86^{***}
	MATTR	0.80^{***}
	MLW	0.75^{***}

Table 6: Correlation between surprisal differences and morphological complexity measures for character, BPE, and Morfessor models. All p-values $< 10^{-11}$.

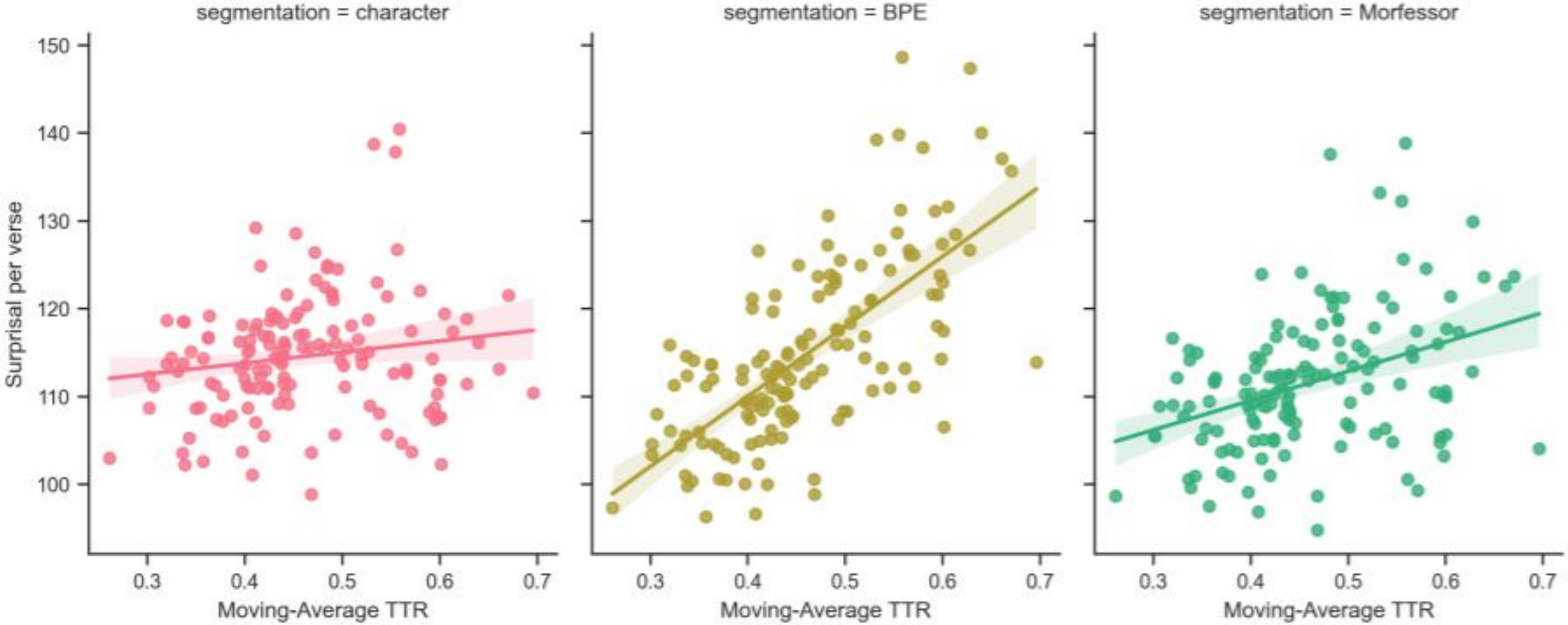


Figure 3: Surprisal per verse plotted against Moving-Average TTR for character, BPE, and Morfessor segmentation methods. Lines indicate the regression estimate with 95% confidence intervals.

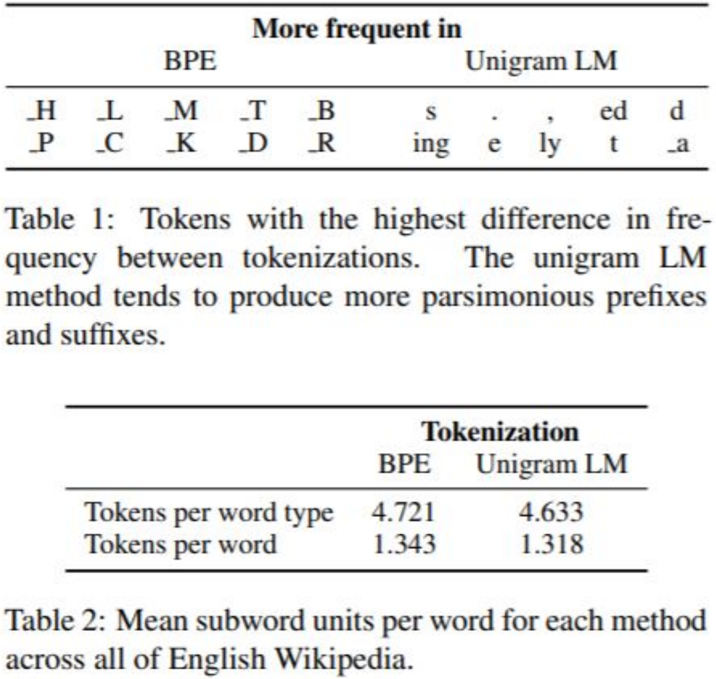
Character out-performs Morfessor for

Amharic, *Egyptian Arabic*, Mandarin, **Central Alaskan Yupik**, *Hebrew*, **Eastern Canadian Inuktitut**, **Inuinnaqtun**, **Greenlandic**, **South Bolivian Quechua**, Telugu, Xhosa

Polysynthetic

Root-and-Pattern

Why might Morfessor be better than BPE?



More frequent in									
BPE					Unigram LM				
_H	_L	_M	_T	_B	s	.	,	ed	d
_P	_C	_K	_D	_R	ing	e	ly	t	_a

Table 1: Tokens with the highest difference in frequency between tokenizations. The unigram LM method tends to produce more parsimonious prefixes and suffixes.

	Tokenization	
	BPE	Unigram LM
Tokens per word type	4.721	4.633
Tokens per word	1.343	1.318

Table 2: Mean subword units per word for each method across all of English Wikipedia.

What about FST segmentations?

Outperform all other methods.

Suggests that segmentation methods that align with morphology may achieve better performance.

Very preliminary results, not enough to do stats to.

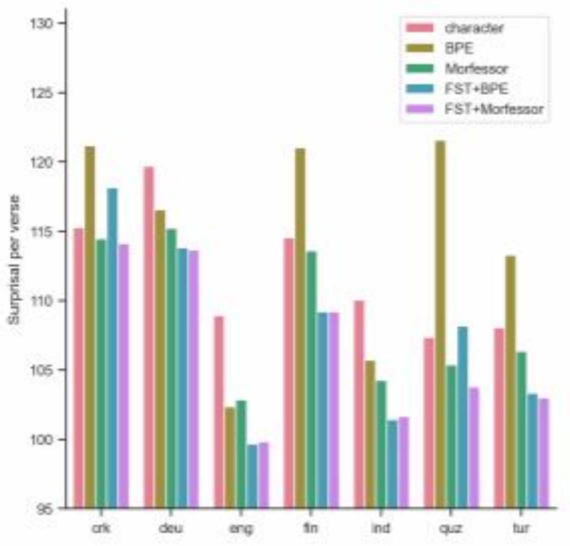


Figure 2: Surprisal per verse per segmentation method including FST segmentation methods. FST+BPE or FST+Morfessor models outperform all other models.

Lessons

1. Your analysis is only as good as your data (and WALS has problems).
2. When considering what linguistic factors might affect NLP, leave your preconceptions at the door.
3. Segmentation matters, isn’t solved.
4. Morphology matters!