Information-Theoretic Characterization of Morphological Fusion

Neil Rathi
Palo Alto High School
neilrathi@gmail.com

Michael Hahn*
Stanford University
mhahn2@stanford.edu

Richard Futrell*
University of California, Irvine
rfutrell@uci.edu

Traditionally, morphological typology divides synthetic languages into two broad groups (e.g. von Schlegel, 1808; von Humboldt, 1843). Agglutinative languages, such as Turkish, segment morphemes into independent features which can be easily split. On the other hand, fusional languages, such as Latin, "fuse" morphemes together phonologically (Bickel and Nichols, 2013). At the same time, there has long been recognition that the categories "agglutinative" and "fusional" are best thought of as a matter of degree, with Greenberg (1960) developing an "index of agglutination" metric for languages. Here, we propose an information-theoretic definition of the fusion of any given form in a language, which naturally delivers a graded measure of the degree of fusion. We use a sequence-to-sequence model to empirically verify that our measure captures typical linguistic classifications.

Our core intuition is that a form expressing some set of morphological features is fusional if it cannot be predicted on the basis of forms for other sets of morphological features. For example, in the Latin noun paradigm in Table 1, the ending *–ōrum* is almost entirely unpredictable from the rest of the paradigm, and so we would say the degree of fusion for this form is high. On the other hand, in the Hungarian paradigm in Table 2, the dative plural form *embereknek* would be almost entirely predictable based on the observation that plurals are formed with *–ek* and datives with *–nek*; this corresponds to a low degree of fusion.

Our notion of fusion is more general than the traditional linguistic notion, which has to do with whether individual morphemes are in a one-to-many correspondence with morphological features (Brown, 2010; Plank, 1999). By contrast, our approach abstracts away from any questions of morpheme segmentation, and can be applied to non-concatenative morphology and productive alternations, where individual morphemes cannot be reliably identified. At the same time, our measure recapitulates linguistic intuitions for concatenative paradigms.

	SG	PL
NOM	servus	**servī**
GEN	**servī**	servōrum
DAT	**servō**	**servīs**
ACC	servum	servōs
ABL	**servō**	**servīs**
VOC	serve	**servī**

Table 1: Forms of the second declension Latin noun *serv* "servant". Colors represent syncretic forms.

To make this idea more precise, we adopt the morphological descriptive framework of Cotterell et al. (2019). A **word** is defined as a triple of a **lexeme** ℓ, a **feature combination** σ, and a **surface form** w, with a **paradigm** m consisting of a map from slots σ to surface forms w. We define the **fusion** ϕ of a given surface form w with feature combination σ as

$$\phi(w) = -\log P(w \mid \mathcal{L}_{-\sigma}, \sigma, \ell), \quad (1)$$

where $\mathcal{L}_{-\sigma}$ indicates the language $\mathcal{L}$ without any forms with feature combination σ. This is analogous to Wu et al. (2019)'s definition of the irregularity of w as $-\log P(w|\mathcal{L}_{-\ell}, \sigma)$. However, here we remove the combination σ from the language, instead of the lemma ℓ. For example, the fusion of *servōrum* would be its negative log probability given every other surface form w in the language outside of those that share $\sigma = (\text{GEN}, \text{PL})$.

We test this measure empirically across four languages: Latin, Hungarian, Turkish, and Quechua, extracting paradigms from UniMorph (Sylak-Glassman, 2016).[1] We expect Latin to be the most fusional, followed by Hungarian and then Turkish and Quechua (Brown, 2010). Following Kann and

*Equal contribution by MH and RF.

[1] https://unimorph.github.io

	SG	PL
NOM	ember	ember**ek**
ACC	ember**t**	ember**ek**et
DAT	ember**nek**	ember**eknek**
ALL	ember**hez**	ember**ekhez**
ABL	ember**től**	ember**ektől**
...	...	...

Table 2: A subset of forms of the Hungarian noun *ember* "person." Morphemes are color-coded by their meaning.

Schütze (2016) and Cotterell et al. (2019), we use a a long short-term memory sequence-to-sequence (seq2seq) model with attention to model conditional probability (Sutskever et al., 2014; Bahdanau et al., 2016).[2] We train models for each $\sigma \in \mathcal{L}$ on $\mathcal{L}_{-\sigma}$, taking the feature combination σ and lemma ℓ (in characters) as input, and producing the form w in characters. This is a morphological reinflection task, best performed by seq2seq models (Cotterell et al., 2016).

To handle syncretism, as in Wu et al. (2019) we "collapse" identical forms into one slot, such that during training, the model does not have access to any syncretic forms during training. Therefore, with serv.ABL.SG in the table above, the model would not have access to serv.DAT.SG while training. Without this step, the measured fusion of languages such as Latin would be extremely low, because many forms can be predicted from their identical syncretic forms.

Our results, as in Table 3, align with those from traditional approaches to morphological fusion. Latin yielded the highest average surprisals, followed by Hungarian, then Turkish, then Quechua. This differed significantly by part of speech: Latin verbs had lower-than-average surprisals, whereas Hungarian verbs were higher-than-average. Figure 1 shows box plots for each POS and language.

	lat	hun	tur	que
Overall	9.87	8.89	2.34	1.01
Nouns	13.60	5.19	2.34	1.01
Verbs	6.38	11.23		
Adjectives	20.25			

Table 3: Average $\phi(w)$ across forms in each language

[2]We used batch size 512, embedding dimension 128, and learning rate 0.001, trained for 10 epochs with early stopping.

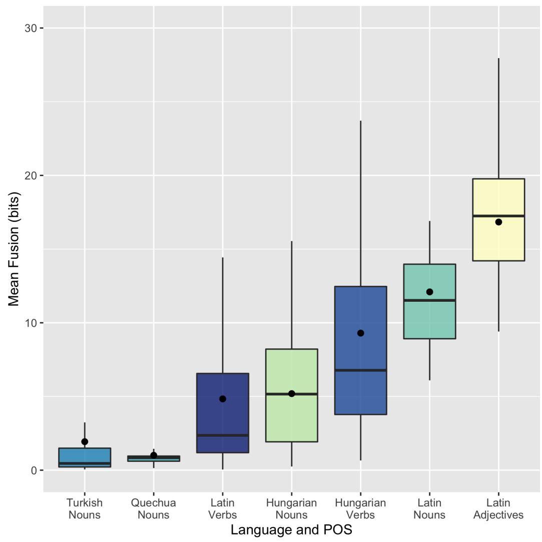

Figure 1: Fusion values by part-of-speech. Middle line indicates median fusion, dot indicates mean.

Some of the higher results for Hungarian and Turkish words comes as a result of phonological changes. For example, the Hungarian adessive plural of *gubó* is *gubóknál*, where *ók* is the plural, and *nál* is the adessive. This has a surprisal of 12.20, as the model instead expects *gubóóknál*, unable to capture the vowel coalescence. This is additionally confirmed by the fact that Quechua, which has few phonological changes, had a lower mean surprisal than Turkish, which features vowel harmony.

Our results for Latin verbs contrast the typical belief that Latin is entirely fusional, but are also intuitive. For example, the verb form *impugnābāmur* can be segmented into *impugnā–bā–mu–r*, where *bā* is IMP, *mu* is 1.PL, and *r* is PASS. This correspondingly has a low surprisal of 0.35.

The differences by POS contradict the traditional "Agglutination Hypothesis," which argues that a language will be equally fusional across all parts of it's morphology (Haspelmath, 2009). Using traditional indices of cumulation, alternation, and suppletion, Haspelmath (2009) showed that the Agglutination Hypothesis does not hold for a selection of 30 languages. Our work therefore corroborates this with an information-theoretic measure of fusion.

In future work, we hope to extend our empirical testing to languages with more diverse morphological systems. Future work will also examine the relationship between fusion and frequency (similarly to Cotterell et al., 2019), as well as the historical development of fusion (see Elsner et al., 2020).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Balthasar Bickel and Johanna Nichols. 2013. Fusion of selected inflectional formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dunstan Brown. 2010. Morphological Typology. In Jae Jung Song, editor, *The Oxford Handbook of Linguistic Typology*, Oxford Handbooks in Linguistics. Oxford University Press.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Micha Elsner, Martha Johnson, Stephanie Antetomaso, and Andrea Sims. 2020. Stop the morphological cycle, I want to get off: Modeling the development of fusion. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 27–37, New York, New York. Association for Computational Linguistics.

Joseph H. Greenberg. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26(3):178–194.

Martin Haspelmath. 2009. An empirical test of the Agglutination Hypothesis. In *Studies in Natural Language and Linguistic Theory*, pages 13–29. Springer Netherlands.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Frans Plank. 1999. Split morphology: how agglutination and flexion mix. *Linguistic Typology*, 3:279–340.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

John Sylak-Glassman. 2016. The composition and use of the Universal Morphological feature schema (UniMorph schema).

Wilhelm von Humboldt. 1843. Über das Entstehen der grammatischen Formen, und ihren Einfluss auf die Ideenentwicklung. In *Band 3*, pages 269–306. De Gruyter.

Friedrich von Schlegel. 1808. *Über die Sprache und die Weisheit der Indier: ein Beitrag zur Begründung der Altertumskunde*. Mohr & Zimmer, Heidelberg. OCLC: 880368442.

Shijie Wu, Ryan Cotterell, and Timothy J. O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Association for Computational Linguistics*.