An Information-Theoretic Characterization of Morphological Fusion

Neil Rathi
Palo Alto High School
neilrathi@gmail.com

Michael Hahn*
Stanford University
SFB 1102, Saarland University
mhahn2@stanford.edu

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

Abstract

Linguistic typology generally divides synthetic languages into groups based on their morphological fusion (von Humboldt, 1825). However, this measure has long been thought to be best considered a matter of degree (e.g. Greenberg, 1960). We present an information-theoretic measure, called informational fusion, to quantify the degree of fusion of a given set of morphological features in a surface form, which naturally provides such a graded scale. Informational fusion is able to encapsulate not only concatenative, but also nonconcatenative morphological systems (e.g. Arabic), abstracting away from any notions of morpheme segmentation. We then show, on a sample of twenty-one languages, that our measure recapitulates the usual linguistic classifications for concatenative systems, and provides new measures for nonconcatenative ones. We also evaluate the long-standing hypotheses that more frequent forms are more fusional, and that paradigm size anticorrelates with degree of fusion. We do not find evidence for the idea that languages have characteristic levels of fusion; rather, the degree of fusion varies across part-of-speech within languages.

1 Introduction

Traditional morphological typology divides synthetic languages into two distinct groups, agglutinative and fusional (von Humboldt, 1825). **Agglutinative** languages have morphemes which can be separated into identifiable parts corresponding to single features. For example, the Hungarian form *embereknek* can be separated into a root and two suffixes, each of which expresses a single morphological feature: *ember-ek-nek* (person-PL-DAT). On the other hand, **fusional** languages express multiple features in a single morpheme, such as Latin *servīs* (servant-DAT.PL), where the suffix *–īs* indicates the dative and plural simultaneously and cannot be analyzed into parts that individually correspond to the genitive or plural features (Brown, 2010; Plank, 1999).

Linguistic typologists have long recognized that this distinction is more of a spectrum than a categorical distinction, with Greenberg (1960) defining an 'index of agglutination' metric to determine the degree to which a language is agglutinative across its morphological paradigms. Interestingly, the notion appears to be graded even within a language. For example, the Latin adjectival feminine genitive plural suffix is *–ārum*, where the thematic vowel *ā* corresponds weakly to the feminine.

Here, we provide an information-theoretic characterization of the degree of fusion of any given form in a language, naturally providing a graded measure. Our core intuition is that a form which expresses a given set of features can be classified as fusional if it cannot be predicted given the forms for other sets of morphological features (i.e. the "rest of the paradigm"). For example, the Latin ending *–īs* in Table 1 is almost entirely unpredictable from the rest of the paradigm: it does not decompose into parts whose meaning can be determined based on other forms. Therefore, we would say that the degree of fusion of *servīs* is high. On the other hand, the Hungarian *–eknek* in Table 2 is fully predictable based on the deduction that *–ek* corresponds to the plural and *–nek* to the dative, so we would say that *embereknek* would have a low degree of fusion.

Our measure of fusion abstracts away from issues of morpheme segmentation. 'Agglutination' and 'fusion' traditionally refer to the extent to which individual features correspond to individual concatenated morphemes: for example, the Hungarian example is considered agglutinative because the suffix *–nek* for the feature DATIVE is concatenated to the morpheme *–ek* for the feature PLURAL. In contrast, our measure of fusion indicates the extent to which a form may be explained as a result of

*Equal contribution by MH and RF.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10115–10120
November 7–11, 2021.

	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.

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

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

individual morphological *processes* corresponding to features, including nonconcatenative processes such as infixation, vowel alternations, reduplication, etc. Effectively, we measure the extent to which a form cannot be predicted or explained in terms of *any strict subset* of its morphological features. Because our measure abstracts away from the form of the morphological processes involved, we name it **informational fusion**.

Previous work has argued that the idea of 'fusion' conflates (at least) three distinct ideas: phonological fusion (the extent to which morphemes are phonologically merged or interleaved with the root), flexivity (the degree of allomorphy with the root), and exponence or cumulativity (the number of distinct features expressed by an unanalyzable morpheme) (Haspelmath, 2009; Bickel and Nichols, 2013). Informational fusion aligns most closely with the idea of exponence, measuring the extent to which multiple features are expressed by an unanalyzable morphological process.

In the remainder of the paper, we formally state our fusion measure and describe its implementation and estimation from data (Section 2), and then evaluate our measure's ability to capture linguistic intuitions and use it to test linguistic hypotheses (Section 3). Section 4 concludes.

2 Definition of Informational Fusion

2.1 Preliminaries

Adopting the framework of Wu et al. (2019), we consider a **word** to be a triple of a **lexeme** ℓ, a **feature combination** or **slot** σ, and a **surface form** w. The lexeme is a string that captures an abstract notion, which is then split into slots σ containing information about the inflection. For example, a slot σ may consist of $\langle$GEN, PL$\rangle$ for a genitive plural form. A **paradigm** is a mapping from lexemes and slots to surface forms. For example, Table 1 provides the paradigm for the Latin lexeme *serv*. The form *servōrum* would be defined as a triple (ℓ = *serv*, $\sigma = \langle$GEN, PL$\rangle$, w = *servōrum*), such that (ℓ, σ) is mapped to w according to the Latin nominal paradigm.

2.2 Informational fusion

We define the informational fusion ϕ of a given surface form w with feature combination σ and lexeme ℓ by taking the surprisal of the surface form given the "rest of the paradigm":

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

where $\mathcal{L}_{-\sigma}$ indicates the language $\mathcal{L}$ without any forms with feature combination σ, and the **predictive model** $p(\cdot \mid \mathcal{L}_{-\sigma}, \sigma, \ell)$ is a conditional probability distribution on forms w given features σ and lexemes ℓ, which is based only on data from $\mathcal{L}_{-\sigma}$.

Informational fusion 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 feature combination σ from the data used to train the predictive model, instead of the lemma ℓ. For example, the informational fusion of *servōrum* would be its negative log probability given every other surface form w in the language outside of those that share $\sigma = \langle$GEN, PL$\rangle$.

If a surface form w is entirely predictable from the paradigm, then it will have an informational fusion of 0, while if it is entirely unpredictable, its informational fusion will be high. A form like *servōrum* is highly unpredictable from the Latin paradigm, so it should have high fusion, while *embereknek* would have low fusion in Hungarian.

To handle syncretism, as in Wu et al. (2019) we "collapse" identical forms into one slot, such that during training of the predictive model, the model does not have access to any syncretic forms. Therefore, with serv.ABL.SG in the table above, the

	lat	hun	tur	que	fra	fro	por	rus	spa	ita	ara	deu	xcl
Overall	9.84	8.19	2.22	0.67	9.32	6.94	6.26	21.41	7.50	10.26	8.27	4.35	11.62
Nouns	13.36	4.73	2.22	0.67				14.88			6.15	4.16	9.46
Verbs	6.37	10.36			9.32	6.94	6.26	25.32	7.50	10.26	2.17	4.41	8.57
Adjectives	20.25										23.97		14.67

	hye	klr	ell	ces	pol	fin	mkd	hbs
Overall	2.63	1.33	10.79	14.79	18.63	7.02	4.73	3.67
Nouns	1.88		29.61	9.93	15.80	6.78	6.26	12.66
Verbs	3.50	1.33	27.90		2.72	2.47	4.58	9.02
Adjectives	1.88		6.58	16.73	23.51	8.53		2.31

Table 3: Average informational fusion across forms in each language, indicated here and elsewhere by three-letter codes (ISO 639-3:2007). Empty cells represent parts of speech with a lack of training data.

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.

2.3 Implementation

We estimate ϕ from paradigm data for 21 languages drawn from UniMorph (Sylak-Glassman, 2016). For Arabic data, we used a transliteration with the ALA-LC standard.[1] All other languages used had separable characters, and thus did not require romanization.

For the predictive model, we use an LSTM seq2seq model with attention (Sutskever et al., 2014; Kann and Schütze, 2016; Bahdanau et al., 2016). The LSTM takes the feature combination σ, POS tag, and lemma ℓ (in characters) as input, producing the form w in characters as output. The input is represented as a string: for example, for a noun with $\sigma = \langle \text{GEN}, \text{PL} \rangle$ and $\ell = \text{serv}$, the input string is `s e r v N GEN PL`, and the target output string is `s e r v ō r u m`. We then estimate the surprisal of the form as:

$$-\log p(w \mid \ell, \sigma) = -\sum_t \log p_\theta(w_t \mid w_{<t}, \ell, \sigma),$$

where θ represents the LSTM parameters, summing over the characters in the form w. For each language and part-of-speech, for each $\sigma \in \mathcal{L}$, we train a separate LSTM on $\mathcal{L}_{-\sigma}$.[2]

Models were not used if the average cross-entropy loss on the final epoch exceeded 0.1. We found a highly bimodal distribution in final loss, such that nearly all models had either very low ($\sim$0.05) or very high ($>$0.4) loss, with high loss corresponding to feature combinations with little training data. We did not observe a systematic relationship between data size and estimates of ϕ.

[1] https://github.com/MTG/ArabicTransliterator

[2] We used batch size 512, embedding dimension 128, and learning rate 0.001, and trained for 10 passes through the training data with early stopping.

3 Results and Discussion

Here we study whether our fusion measure recapitulates the familiar classifications for selected languages, and study whether it covaries systematically with paradigm size and form frequency, testing linguistic hypotheses.

3.1 Basic results

Average fusion scores for paradigms from 21 languages are shown in Table 3 and Figure 1. The scores are largely consistent with typological classifications. We observed that overall, the languages with lowest average fusion were Turkish and Quechua, whose paradigms are usually classified as agglutinative or monoexponential, while the most fused languages were Greek, Russian, Polish, and Czech, again consistent with typical classifications (Bickel and Nichols, 2013). We also observe clustering based on language family. The Slavic languages as a whole appear to have roughly equal fusion levels, and the same was true for the Romance languages. While these were the only families with more than two languages, the results are suggestive for our measure as an indicator of typological relationships.

We find that fusion differs substantially by part of speech even within languages. For example, Latin and Arabic verbs have much lower fusion than their nominal and adjectival counterparts. This result is in line with Haspelmath (2009)'s arguments against the 'Agglutination Hypothesis.'

Some of the more surprising results shed light

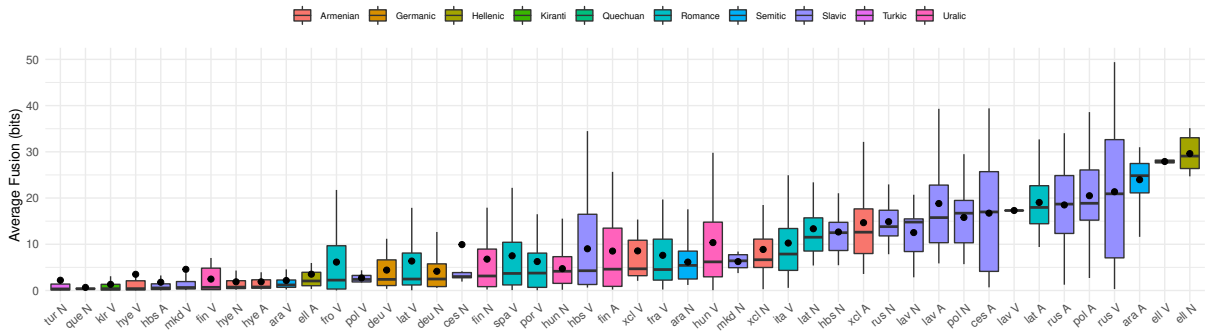

Figure 1: Boxplots of mean informational fusion values by part-of-speech and language. Middle line indicates median fusion; dot indicates mean fusion; colors indicate language family. N = nouns, V = verbs, A = adjectives.

on the nature of informational fusion. For example, the low level of fusion for Latin verbs contrasts with the typical classification of Latin as fusional, but the result is intuitive upon inspection. For instance, the verb form *impugnābāmur* can be segmented into *impugnā–bā–mu–r*, where *bā* represents the feature IMPERFECT, *mu* represents 1.PL, and *r* represents PASSIVE (Bennett, 1994). These parts combine predictably, yielding a correspondingly low fusion of 0.35 for this form.

Another interesting result is the low level of fusion for Arabic verbs. This result is sensible: although Arabic morphology is highly nonconcatenative, the morphological processes that convey individual features (person, aspect, voice, etc.) are quite regular and compose with each other transparently (Ryding, 2005). This result illustrates how informational fusion abstracts away from the form of the morphological processes.

Some further less anticipated results can be explained as cases of *phonological* fusion. For example, Hungarian, while typically classified as agglutinative, undergoes many regular sound changes across its paradigms, including vowel harmony and vowel coalescence. The latter can be seen in forms such as (ℓ = *gubó*, $\sigma = \langle$AT+ESS, PL$\rangle$, w = *gubóknál*). The suffix for plural is *–ok*, which, when suffixed to a stem ending in *ó*, coalesces with the stem; e.g. *gubó-ok-nál* → *gubóknál* (Szita and Görbe, 2010). As our LSTM learns this phonological process only imperfectly, it falsely predicts *gubóóknál* for this form.

3.2 Covariance with Paradigm Size

Plank (1986) proposed that fusion (in the sense of exponence) limits the number of forms that can exist in a paradigm (i.e. e-complexity: see Ackerman and Malouf, 2013; Cotterell et al., 2019). This hypothesis can be justified cognitively in terms of informational fusion, which indicates the minimum number of bits of information required to store and learn a form. If there is a limit on paradigm complexity in this sense, then paradigms can be either large or highly fusional, but not both.

Figure 2 shows the relationship between average fusion and paradigm size, calculated as the maximum number of forms per lemma in UniMorph. Although there does appear to be a weak negative correlation, it is not robust: we find Spearman's $\rho = -0.30, p = 0.08$. Thus, we do not find support for Plank's hypothesis.

However, we do not take this as strong evidence against the hypothesis, because there is a degree of arbitrariness to measuring paradigm size from datasets such as UniMorph in terms of what

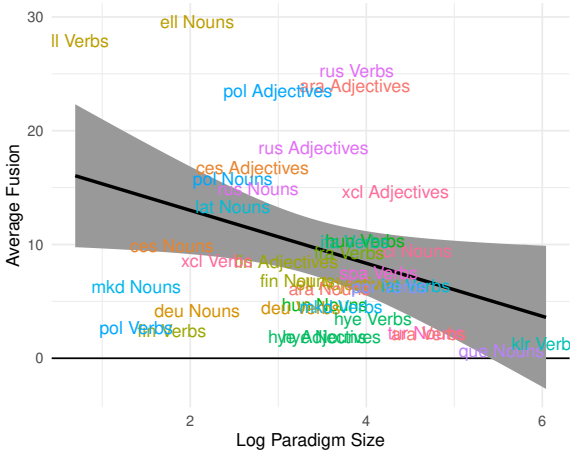

Figure 2: Correlation of log-transformed paradigm size and average informational fusion per paradigm. Text indicates part of speech and language, and datapoints are colored by language.

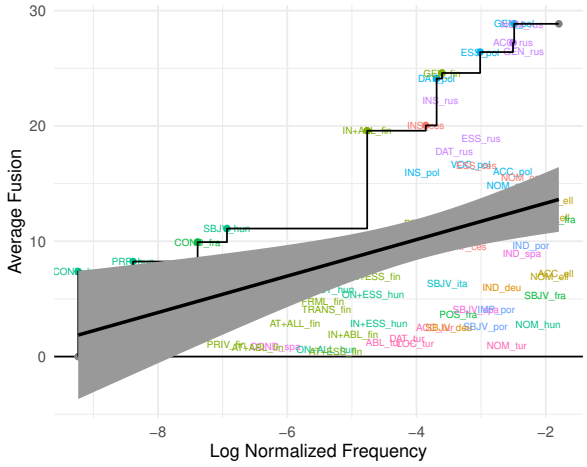

Figure 3: Correlation and tradeoff between frequency and fusion per feature and language. On the x-axis, log normalized frequency of all forms matching a given feature in a given language. On the y-axis, the average informational fusion for those forms. Text indicates feature and language; step curve indicates Pareto curve.

counts as an entry in a paradigm. For example, the Quechua UniMorph dataset includes possessive forms of nouns, while the Hungarian dataset does not, although both languages express possession using suffixes. Differences in measured paradigm size may reflect the choice of what was included in the corpus rather than real linguistic differences.

3.3 Covariance with Form Frequency

We might expect that highly fused forms are also highly frequent in usage. An infrequent but fused form would be unstable, in the sense that language users might forget it in production (defaulting to a more predictable form), or might fail to acquire it in learning. Therefore, here we evaluate the hypothesis that a high degree of informational fusion implies high form frequency; or alternatively, that there is a tradeoff between informational fusion and form frequency.

We test the hypothesis at the level of individual features. We quantify the average fusion of a feature as the average fusion of all forms with that feature, and the frequency of a feature as the total frequency of all tokens expressing that feature in a corpus. Figure 3 shows the relationship between average fusion per feature per language and log feature frequency, estimated from from Wikipedia dumps and normalized by the total number of tokens per Wikipedia corpus. Syncretic forms were removed for this analysis. Average fusion is significantly correlated with frequency (Spearman's $\rho = 0.39, p < 0.001$ by permutation test).

We find an unoccupied quadrant in the data: we do not find features that are *both* infrequent and expressed fusionally. For significance testing, we use a nonparametric permutation test with the area under the Pareto frontier (similarly to Cotterell et al., 2019). The p-value is the probability that a stochastically constructed curve–in which the y-values of the data are randomly permuted–has an "emptier" upper left quadrant, i.e. that the area under the null-hypothesis curve is less than or equal to the area under the empirical curve. This was estimated by permuting the data 10,000 times. We find that the upper-left quadrant is significantly empty ($p < 0.002$), indicating a significant tradeoff between fusion and frequency. This still holds with the cognitive explanation provided above.

4 Conclusion

We introduced an information-theoretic measure of the fusion of a form within a morphological paradigm, called informational fusion. We have shown that informational fusion recapitulates linguists' intuitions and allows for quantitative tests of linguistic hypotheses, including a tradeoff between fusion and frequency. Our work joins a growing body of recent research that aims to operationalize basic linguistic concepts in terms of information theory (Ackerman and Malouf, 2013; Cotterell et al., 2019; Pimentel et al., 2019; Futrell et al., 2019; Mansfield, 2021).

Informational fusion is the extent to which a form cannot be predicted based on any strict subset of its morphological features. As such, it aligns closely with the linguistic notion of the exponence of a form. It can be adapted to provide fusion measures for specific morphemes and features by carefully choosing which features are held out during the training of the predictive model.

Acknowledgments

This work benefited from discussion in the SIGTYP 2021 Workshop. It was supported by NSF Grant #1947307 and an NVIDIA GPU Grant to R.F. All code and data are available at https://github.com/neilrathi/morphological-fusion.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

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

Charles E. Bennett. 1994. *New Latin grammar*. Bolchazy-Carducci Publishers, Wauconda, Ill.

Balthasar Bickel and Johanna Nichols. 2013. Exponence 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.

Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 3–13, Paris, France. 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.

ISO 639-3:2007. 2007. Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages. Standard, International Organization for Standardization, Geneva, CH.

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.

John Mansfield. 2021. The word as a unit of internal predictability. *Linguistics*.

Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

Frans Plank. 1986. Paradigm size, morphological typology, and universal economy. *Folia Linguistica*, 20:29–48.

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

Karin C. Ryding. 2005. *A reference grammar of Modern Standard Arabic*. Cambridge University Press, New York.

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).

Szilvia Szita and Tamás Görbe. 2010. *A practical Hungarian grammar*. Akademiai Kiado, Budapest. OCLC: 935138375.

Wilhelm von Humboldt. 1825. Über das Entstehen der grammatischen Formen und ihren Einfluss auf die Ideenentwicklung. In *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin: Aus den Jahren 1822 und 1823*, pages 401–430. Drückerei der Königlichen Akademie der Wissenschaften, Berlin.

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