Title
Explaining patterns of fusion in morphological paradigms using the memory--surprisal tradeoff

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors
Rathi, Neil
Hahn, Michael
Futrell, Richard

Publication Date
2022

Peer reviewed

Explaining patterns of fusion in morphological paradigms using the memory–surprisal tradeoff

Neil Rathi[1], **Michael Hahn**[2], **and Richard Futrell**[3]

[1]Palo Alto High School, neilrathi@gmail.com
[2]Department of Linguistics, Stanford University; SFB 1102, Saarland University, mhahn2@stanford.edu
[3]Department of Language Science, University of California, Irvine, rfutrell@uci.edu

Abstract

Languages often express grammatical information through inflectional morphology, in which grammatical features are grouped into strings of morphemes. In this work, we propose that cross-linguistic generalizations about morphological fusion, in which multiple features are expressed through one morpheme, can be explained in part by optimization of processing efficiency, as formalized using the memory–surprisal tradeoff of Hahn, Degen, and Futrell (2021). We show in a toy setting that fusion of highly informative neighboring morphemes can lead to greater processing efficiency under our processing model. Next, based on paradigm and frequency data from four languages, we consider both total fusion and gradable fusion using empirical measures developed by Rathi, Hahn, and Futrell (2021), and find that the degree of fusion is predicted by closeness of optimal morpheme ordering as determined by optimization of processing efficiency. Finally, we show that optimization of processing efficiency can successfully predict typological patterns involving suppletion.

Keywords: language processing; morphology; information theory; memory–surprisal tradeoff

Introduction

In human language, grammatical information is often expressed via **inflectional morphology**, in which a set of **grammatical features** are encoded in wordforms. For example, in the English word *walked*, the suffix *–ed* expresses the tense feature PAST; we say that the word consists of two **morphemes**, the root *walk* plus the suffix *–ed*.

Morphological systems have long held a fascination for the field of linguistics, because they differ widely across languages, while also showing clear universal tendencies. They are often the target of theoretical models aiming to quantify their complexity (for example, del Prado Martín, Kostić, & Baayen, 2004; Baerman, Brown, & Corbett, 2015; Bentz, Ruzsics, Koplenig, & Samardžić, 2016) and to explain their properties in terms of constraints on complexity (for example, Ackerman & Malouf, 2013; Cotterell, Kirov, Hulden, & Eisner, 2019).

In this work, we focus on cross-linguistic generalizations about how information is packaged into morphemes, and in particular cases where multiple features are expressed simultaneously in a single morpheme—a phenomenon called **fusion**.[1]

We argue that key properties of fusion can be explained in terms of a recently-introduced theory of linguistic complexity called the **memory–surprisal tradeoff** (Hahn et al., 2021), which is based in information theory and models of online language processing. Previous work has argued that the memory–surprisal tradeoff can explain certain aspects of the ordering of morphemes within words (Hahn et al., 2021; Hahn, Mathew, & Degen, 2022).

Intuitively, the memory–surprisal tradeoff measures how the predictability of a form trades off with the memory resources required for processing it. We say a morphological system is **efficient** when it allows for favorable tradeoffs. We propose that cross-linguistic patterns of fusion arise from a pressure for efficiency in the sense of the memory–surprisal tradeoff—i.e. that the ETH can predict which features tend to be fused—and provide empirical evidence to support this hypothesis.

First, we present a model simulation showing how fusion is related to our notion of efficiency. We demonstrate that morphological systems are efficient when highly correlated features are expressed simultaneously as one morpheme; i.e. when they are **polyexponent**.

Next, we focus on **informational fusion** (Rathi et al., 2021), a graded measure of the extent to which a set of features are expressed together in an unanalyzable morpheme (Brown, 2010; Bickel, 2001). We show that the memory–surprisal tradeoff can accurately explain which pairs of features have higher levels of informational fusion. Furthermore, across languages, we are able to predict which sets of three features will be polyexponent, such as tense–aspect–mood (TAM) markers.

Finally, we consider the case of **suppletion**, in which the form of a root changes unpredictably based on grammatical features (Veselinova, 2013). We treat suppletion as fusion of a grammatical feature with the root, and show that cross-linguistic patterns of which features are suppletive can be predicted via the memory–surprisal tradeoff on a sample of 17 languages.

The remainder of the paper is structured as follows. First, we introduce the memory–surprisal tradeoff as a theory of linguistic complexity, how it can be calculated from linguistic datasets, and what properties we generally expect from linguistic systems that optimize the tradeoff. Next, we present our simulations and computational studies in four experiments; these sections also review the relevant linguistic phenomena. In the conclusion we discuss the implications of our results.

[1]The term 'fusion' is used with different senses in the linguistic literature (Plank, 1999; Brown, 2010). The two major senses are phonological fusion—when two morphemes appear to be merged together because of the action of phonological rules—and polyexponence: when a single morpheme expresses multiple features. Our usage is more aligned with the latter sense.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

The Memory–Surprisal Tradeoff

The memory–surprisal tradeoff describes the complexity of incremental language processing in terms of two factors: (1) the difficulty of predicting upcoming material, and (2) the difficulty of maintaining memory representations of past material. These two factors trade off: a high-fidelity representation of past material enables accurate predictions about future material, but at the cost of higher investment of memory resources. The **Efficient Tradeoff Hypothesis** (ETH; Hahn et al., 2021) holds that languages are structured so that favorable tradeoffs of these two factors are possible: that is, so that upcoming material is highly predictable even given very little information stored in memory.

The memory–surprisal tradeoff is grounded in the Surprisal theory of online language comprehension (Hale, 2001; Levy, 2008), which holds that the processing difficulty of word (or any linguistic unit) w_t in context $w_1 \dots w_{t-1}$ is proportional to its **surprisal**, which is the negative log probability of the word in context:

$$S = -\log p(w_t \mid w_1 \cdots w_{t-1}). \tag{1}$$

Surprisal, as measured by n-gram, PCFG, or neural models, has been shown to accurately predict word-by-word reading times (Demberg & Keller, 2008; Smith & Levy, 2013; Goodkind & Bicknell, 2018; Wilcox, Gauthier, Hu, Qian, & Levy, 2020; Frank & Ernst, 2019; Rathi, 2021) (but see van Schijndel & Linzen, 2021). Within this paradigm, Futrell, Gibson, and Levy (2020) argue that the context accessible to the processor is best thought of as a *lossy* memory representation m_t of the true context:

$$S_M = -\log p(w_t \mid m_t), \tag{2}$$

where the memory representation $m_t = M(w_1 \dots w_{t-1})$ is given by some memory encoding function M.

The core idea of the memory–surprisal tradeoff is that with more information in the memory m_t, the average surprisal achieved will be lower. That is, more precise memory leads to more precise predictions and lower processing difficulty. We can quantify the average amount of information stored in the memory state as H_M, the entropy of the memory state:

$$H_M = -\sum_{m_t} p(m_t) \log p(m_t). \tag{3}$$

The **memory–surprisal tradeoff curve** quantifies the lowest achievable average surprisal S_M given a particular average amount of information stored in memory H_M. It is a form of the 'predictive information bottleneck' curve studied in information theory (Still, 2014). With a steeper curve, lower processing difficulty can be achieved with less memory resources, i.e. while storing less information in memory. Thus, a steeper curve is more efficient.

Given this setting, Hahn et al. (2021) formalize the ETH as follows: "the order of elements in natural language is characterized by a distinctively steeper memory–surprisal tradeoff curve compared to other possible orders." For example, in Figure 1, Language A has a steeper curve than Language B and would thus be more efficient, as we would expect a natural language to be. Hahn et al. (2021) furthermore show that an efficient tradeoff is generally achieved when languages follow the **information locality principle** (Futrell, 2019), which states that atomic units that predict each another will be close in linear order (cf Behaghel, 1930). It can be formalized by measuring predictivity using **mutual information** (MI), an information-theoretic measure of statistical dependence.

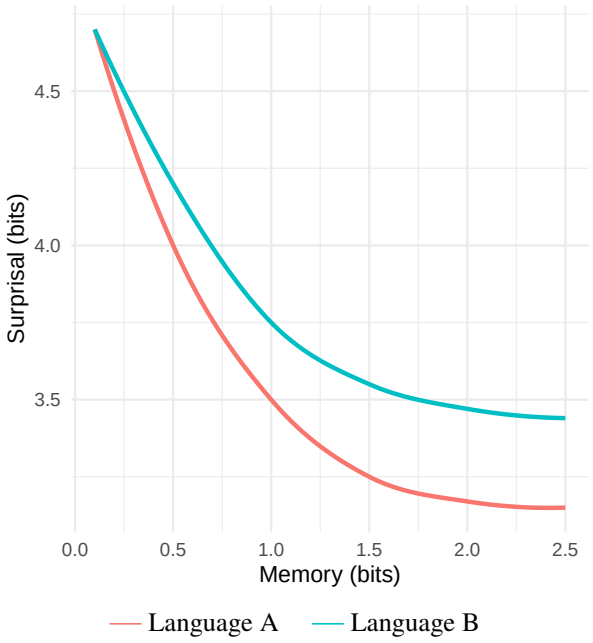

Figure 1: Sample memory–surprisal tradeoff curves of two hypothetical languages, Language A and Language B. The curve for Language A is steeper, and thus we would say it is more efficient in terms of the cognitive resources required.

While the ETH was defined in previous work as a theory of word and morpheme order, we use it here as a theory of how grammatical information is packaged into morphemes, that is, a theory of *why* multiple features might be expressed in a single morpheme, and *which* features are likely to be fused in this way. We hypothesize that attested morphological systems (conceived of generally as mappings from sets of features to wordforms) achieve more favorable memory–surprisal tradeoffs than alternative systems. The information locality principle carries over into this domain in a modified form: we generally predict that features are likely to be expressed by a single morpheme when they statistically depend on each other.

Our general approach, therefore, is to calculate the memory–surprisal tradeoff for attested morphological systems and to compare against other possible systems. We hypothesize that the attested systems will enable more efficient tradeoffs, quantified as the area under the curve of the memory–surprisal tradeoff.

Calculation of the memory–surprisal tradeoff

It is straightforward to calculate a lower bound on memory–surprisal tradeoff curve from corpus data. In particular, Hahn et al. (2021) demonstrate that a lower bound on the curve can be calculated by fitting a series of incremental language models that use successively more context to predict upcoming material. To calculate the lower bound, first we need a quantity

$$I_t = S_{t-1} - S_t, \tag{4}$$

where S_t is the average surprisal of a language model which sees a context of size t, for example an n-gram model which sees only the t previous words. The quantity I_t is the mutual information between words at a distance of t words. Then for all timescales T, for any memory entropy satisfying

$$H_M \leq \sum_{t=1}^{T} tI_t, \tag{5}$$

the average surprisal is lower-bounded as

$$S_M \geq S + \sum_{t=T+1}^{\infty} I_t. \tag{6}$$

The lower bound on the whole curve can be computed by sweeping out values of $T = 0, 1, 2, \ldots$.[2]

For our purposes and given datasets available to us, we measure the memory–surprisal tradeoff at the level of abstract sequences of morphemes. Forms are represented as a sequence consisting a root and a series of affixes: for example, the English form *goes* is represented as ROOT-3-SG-PRS, indicating it expresses the features of third person, singular, and present tense. This representation abstracts away from any ambiguity that might exist in wordforms, and also abstracts away from the length of individual morphemes.

Estimation of optimal orderings

We estimate optimized orderings by optimizing for the area under the memory–surprisal tradeoff curve (AUC). Each ordering corresponds to a potential curve, with the lowest AUC curve being the most "efficient." We use a modified version of the hill climbing method of Gildea and Jaeger (2015): weights in $[0, 1]$ are initially randomly assigned to each feature, and then are iteratively adjusted to reduce AUC. For each iteration, we randomly select one feature, and determine the AUC for each way of ordering it with respect to two other morphemes; we then change the weights to the lowest AUC ordering. We optimize this approximately (only guaranteeing convergence to local rather than absolute minima) by restricting this calculation to morphemes that occurred at least 10 times in the dataset for 95% of iterations, and to 10% of possible orderings for each iteration. This value converged after a few hundred iterations; we iterated 1,000 times for each language.

[2] In the experiments below, we determine I_t using n-gram models with Kneser-Ney smoothing trained on a training set, and estimate average surprisal S_t as cross-entropy on a test set. To mitigate overfitting with larger values of t, we estimate $\hat{S}_t$ as $\min_{t \leq n} S_t$, where S_t is the cross-entropy on the t'th order Markov model.

Experiment 1: Polyexponence Model Simulation

We first discuss how optimization of the memory-surprisal tradeoff could in principle lead to fusion of features that are informative about each other. In this section, by 'fusion,' we refer to polyexponence, where multiple features are expressed in a single morpheme in a way that cannot be decomposed.

To compare the optimality of paradigms with different kinds of fusion, we simulated a language where forms express the features Past/Present Tense, 2nd/3rd Person, and Singular/Plural Number. In order to induce nontrivial probabilistic structure, we (arbitrarily) assigned higher frequency to 2nd Singular and 3rd Plural features than to others; our conclusions do not depend on this particular choice. Thus, in the distribution over features in this simulated language, there is high mutual information between person and number, and low mutual information between tense and person/number. Using this distribution over features, we simulated two languages: in the **Low MI** language, Tense and Person are fused into a single morpheme; in the **High MI** language, Person and Number are fused.

According to the information locality principle, the high-MI morphemes should be ordered close to each other in order to optimize the memory–surprisal tradeoff; we hypothesize that, as an analogue for closeness, they should also be fused together. This intuition is borne out in the simulations. Memory–surprisal tradeoff curves for these languages are shown in Figure 2. The High MI language shows a more efficient tradeoff at all but very high memory capacities. This happens because fusion of features that are predictive about each other into a sin-

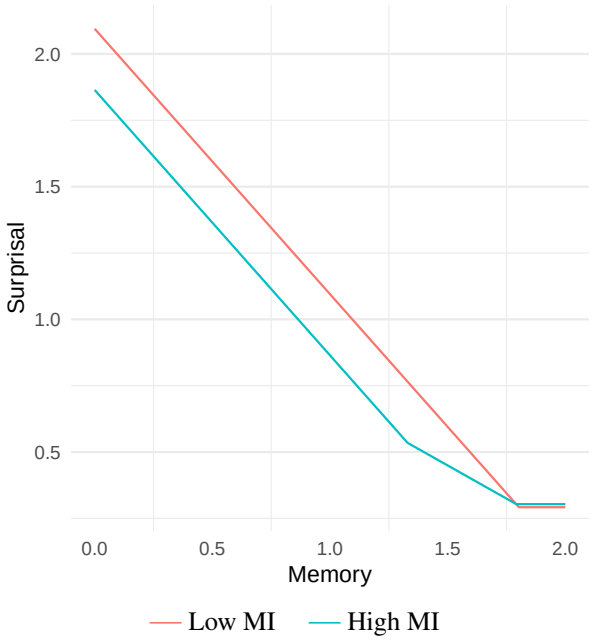

Figure 2: Memory-surprisal tradeoffs with fusion of low or high mutual information pairs in the model simulations in Experiment 1.

gle morpheme reduces the entropy of individual morphemes more than fusion of morphemes that share no information.

Experiment 2: Pairwise Informational Fusion

In the simulation above, we found that fusion of morphological features is most optimal when those features are highly correlated with each other. Here, we explore whether the memory–surprisal tradeoff can more generally predict degrees of morphological fusion in real languages. We use the recently-introduced concept of **informational fusion** as a graded empirical measure of the extent to which multiple features are expressed in a single morpheme. We show that the Efficient Tradeoff Hypothesis can predict the degree of fusion of for pairs of features in four languages, such that features close together in optimal order under the ETH are more likely to be fused.

Informational Fusion Rathi et al. (2021) recently introduced a graded measure of morphological fusion called **informational fusion**. Informational fusion intuitively measures the extent to which a set of features is expressed on a wordform in a way that is unanalyzable: that is, the wordform cannot be decomposed into any morphemes or morphological processes corresponding to subsets of the features in question. For example, the fusion of Latin form *servīs* should be high, since the suffix *–īs* expresses both dative and plural features, while the fusion of Hungarian *embereknek* should be low, since *–eknek* can be split into *–ek* (the plural) and *–nek* (the dative).

More precisely, informational fusion measures the bits of information required to specify a form for a given set of features, beyond what would be required for any subset of the features. Formally, the informational fusion of a surface form w in a language with respect to feature set σ and lemma ℓ is

$$\varphi(w) = -\log p(w \mid \mathcal{L}_{-\sigma}, \sigma, \ell), \tag{7}$$

where $\mathcal{L}_{-\sigma}$ is a dataset consisting of all wordforms in the language along with their features, having removed all occurrences of feature set σ.

If a form is highly predictable based on the rest of the forms in the language, then it will have low informational fusion; on the other hand, if a form is not predictable, it will have high informational fusion. Rathi et al. (2021) show that the measure φ produces numbers that match linguistic intuitions about morphological fusion across a wide variety of languages when the distribution $p(w \mid \mathcal{L}_{-\sigma}, \sigma, \ell)$ is estimated using a neural seq2seq architecture (Sutskever, Vinyals, & Le, 2014).

We are interested in predicting degrees of fusion for *pairs* of features, as opposed to entire slots, so we adapt the idea of informational fusion to study pairs of features in the following way. For a form w with features σ, and a pair of features $f_1, f_2 \in \sigma$, we define the 2-feature informational fusion $\varphi_2(w)$ as

$$\varphi_2(w) = -\log p(w \mid \mathcal{L}_{-(f_1,f_2)}, \sigma, \ell). \tag{8}$$

To get a summary measure for a feature pair f_1, f_2, we calculate the average $\varphi_2(w)$ across forms w expressing those features; we denote this summary measure $\overline{\varphi}_2(f_1, f_2)$. This gives us an average informational fusion value for each pair of features in $\mathcal{L}$. For example, the fusion of the feature pair $(2, \text{PL})$ would be the average surprisal of any 2nd person plural form, conditional on all forms in $\mathcal{L}$ that are not 2nd person plural.[3]

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

Table 1: The second declension Latin noun paradigm for *serv*, 'servant.' Syncretic forms are color-coded.

	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 the Hungarian noun paradigm for *ember*, 'person.' Morphemes are color-coded by feature.

Fusion and the Memory–Surprisal Tradeoff We can construct a memory–surprisal tradeoff curve for any possible permutation of morpheme order. Under the Efficient Tradeoff Hypothesis, we would expect the order that creates the steepest curve to be the order used in natural language. Here, we hypothesize that fusion can be predicted by closeness in optimal order. If a pair of features are close together in optimal order, under the ETH, they should therefore be more fused.

Methods

For all feature pairs f_1, f_2 in each language, we estimate $\overline{\varphi}_2$ using an LSTM sequence-to-sequence model with attention (Sutskever et al., 2014; Kann & Schütze, 2016; Bahdanau, Cho, & Bengio, 2016). We extract data from UniMorph (Sylak-Glassman, Kirov, Yarowsky, & Que, 2015; McCarthy et al., 2020), transliterating Arabic with the ALA-LC standard.

The model takes as input the lemma ℓ, the featureset σ, and the part-of-speech tag, and produces the form w in characters as output. We represent both the input and output as strings. For example, for the Latin form $w =$ *pugnabāmur*, the input would be `p u g n o V 1 PL PST IPFV PASS` and the target string would be `p u g n a b ā m u r`. Then, we

[3] This notion of φ_2 can be generalized such that for any $n \geq 2$, $\varphi_n(w)$ is the informational fusion of n features.

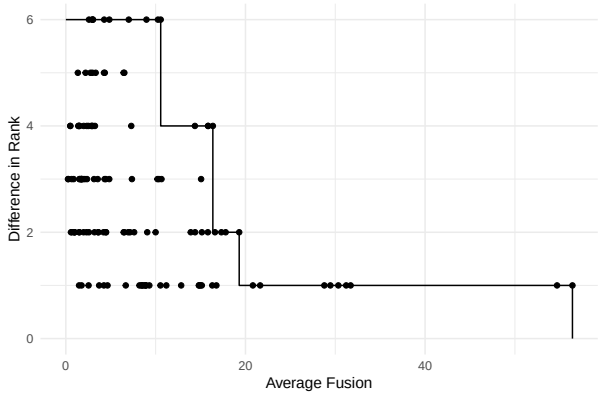

Figure 3A: Arabic ($p < 0.001$)

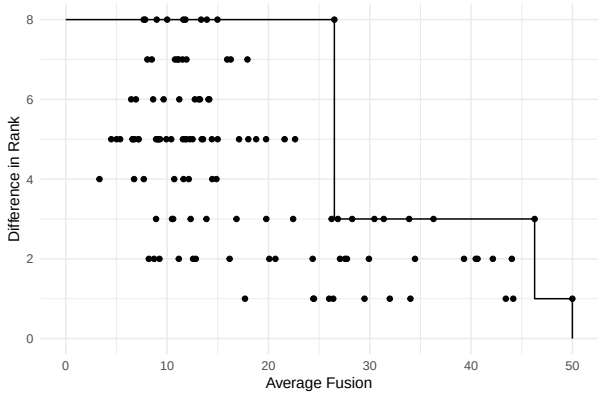

Figure 3B: Latin ($p < 0.001$)

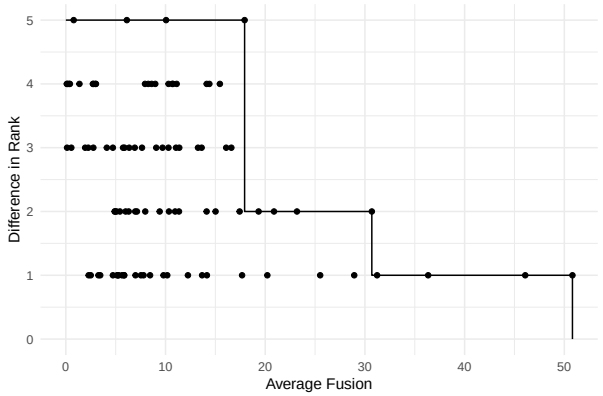

Figure 3C: Spanish ($p < 0.01$)

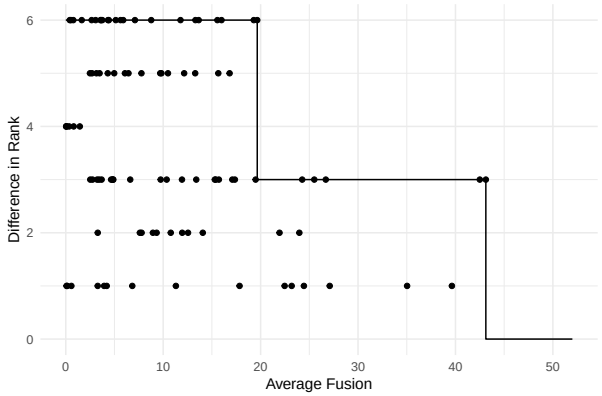

Figure 3D: Portuguese ($p < 0.005$)

Figure 3: Tradeoffs between difference in rank and average fusion. Each point represents two features (f_1, f_2), plotting $(\overline{\varphi}_2(f_1, f_2), R(f_1, f_2))$. Step curve indicates Pareto curve. All tradeoffs are significant ($p < 0.01$) by permutation test for the area under the Pareto curve.

estimate the surprisal of the form as

$$-\log p(w \mid \ell, \sigma) = -\sum_t \log p_\theta(w_t \mid w_{<t}, \ell, \sigma), \qquad (9)$$

where θ refers to the model parameters.[4] We train a model for each feature combination in $\mathcal{L}$.

For each optimized ordering, morphemes are ordered by optimal rank; we average out the ranks of each feature by feature category to determine the optimal *category* ordering, with categories as determined by UD classification. We use categories rather than individual 'fine-grained' features in order to prevent noise in the form of repeated items from interfering. Then, for each pair (f_1, f_2), we compute the difference in rank $R = |r(f_1) - r(f_2)|$ of their categories in the optimized ordering, and plot this against the fusion $\overline{\varphi}_2(f_1, f_2)$.

Results

We trained models for the verbal paradigms of a set of four languages: Arabic, Latin, Spanish, and Portuguese. These languages were chosen based on the size of the datasets available (both from UniMorph and Universal Dependencies) and the variation of their verbal paradigms. Languages with low variation in fusion have little to be explained, and thus were not used. Verbal paradigms were chosen over noun paradigms due to their size and range of degrees of fusion; for example, Latin verbs range from fusional in the present tense to agglutinative in the perfect.[5]

As shown in Figure 3, we find that there is an "empty" upper-right quadrant in all language tested. This means that there are no feature combinations which are both highly fused and far apart in the optimal ordering. To test this for significance, we use a nonparametric permutation test for area under the Pareto curve, as in Cotterell et al. (2019). We stochastically permute y-values of the set of points $\{(x_1, y_1), (x_2, y_2), \dots\}$ to create $\{(x_1, y_{\sigma(1)}), (x_2, y_{\sigma(2)}), \dots\}$. The p-value is the probability that the area under this randomly permuted curve is less than that of the empirical curve; we estimate this using 10,000 permutations; for all languages, we find at least $p < 0.01$. The results support the ability of the ETH to predict morphological

[4] 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. Models were not used if the average cross-entropy loss on the final epoch exceeded 0.1.

[5] We used the Arabic PADT treebank, Latin PROIEL treebank, Spanish AnCora treebank, and Portuguese Bosque treebank.

fusion, going beyond predicting only morpheme order.

Experiment 3: Triple Exponence

Here we study whether the ETH can predict generalizations about when sets of *three* features are expressed together in a morpheme. We refer to this phenomenon as **triple exponence**.

Crosslinguistically, some features tend to be expressed more often than others in this configuration. The most prominent examples are fusion of tense, aspect, and mood (TAM) markers, and the fusion of person, number, and gender (PNG) markers in verbal paradigms. If the ETH is a general predictor of morphological fusion, then we would expect triply-exponential features such as TAM and PNG to be close together in the optimized ordering determined by the memory–surprisal tradeoff.

We evaluated this hypothesis in a set of 15 languages with PNG features and 13 languages with TAM features. For each language, we calculated the optimal ordering of its morphemes. Then we calculated the standard deviation of the three ranks of the relevant features in the optimal ordering. If this standard deviation is low, that means that the three features are close together in the optimal ordering. We normalized this value by the number of feature categories in the language; thus, if a language has a large number of features, chance variation in the distance between features is controlled for. We performed this procedure for all possible combinations of three features, limiting our data to those combinations which had at least 7 languages represented.

Results are shown in Figure 4. Both sets of triply-exponential features had very low normalized standard deviations compared to random sets of three features, with $p < 0.001$ for both TAM and PNG by one-sample t-test. The results indicate that TAM and PNG tend to cluster together in optimized order, confirming the prediction of the ETH.

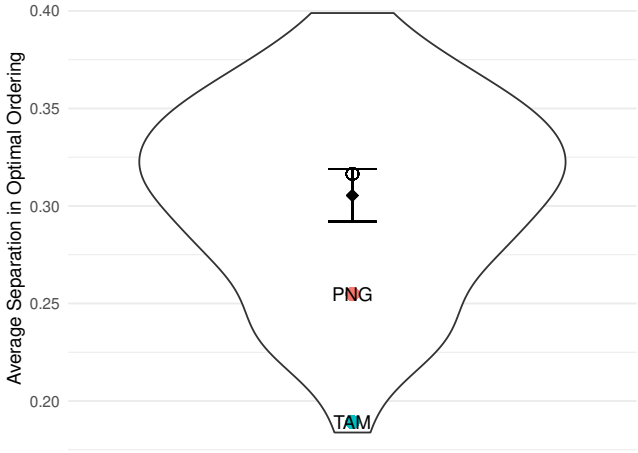

Figure 4: Average separation in optimal ordering (as measured by normalized standard deviation) of three-feature combinations with greater than 7 data points. PNG is colored red, TAM is colored blue; diamond indicates mean, open circle indicates median. Error bars indicate 95% confidence interval.

Experiment 4: Suppletion

Suppletion is a phenomenon in which a given root can take two or more forms when inflected for a particular grammatical feature (Veselinova, 2013). For example, the English root GO is expressed as *go* in the present and future tenses, but as *went* in the past tense. Occasionally, suppletion can be used almost entirely for inflection (e.g. plurality in Dinka; see Ladd, Remijsen, and Manyang (2009)), however, it is usually limited to a few forms. While suppletion is rare, it does exist in the paradigms of a small number of lexemes in some languages and is systematic (Bybee, 1985; Markey, 1985; Aski, 1995; Fertig, 1998).

Cross-linguistically, it is known that some features are more likely to lead to suppletion than others. Notably, in nouns, suppletion to express number is more common that suppletion to express case (Moskal, 2015). Here, we study if this trend can be predicted by the ETH. We treat suppletion as a form of fusion between the feature and the root. Thus, applying the ETH, we would expect features that commonly drive suppletion to be more closely positioned to the root in optimal ordering than those that do not cause suppletion. We generate optimal orderings by feature category using the method described in Experiment 3. We then compare the position of number, which frequently drives suppletion, to case, which does not.

We find that of the 17 languages tested, in 15, number tends to be closer to the root in optimal ordering than case. Thus, we observe that suppletion can be explained by the ETH, with $p < 0.005$ by binomial test (the null hypothesis being that they are equally likely).

Conclusion

We studied cross-linguistic properties of morphological fusion. We found that both fine-grained quantitative measures such as informational fusion, and major cross-linguistic typological generalizations such as the existence of tense–aspect–mood markers, can be explained in terms of optimization of the tradeoff of predictability and memory complexity (Hahn et al., 2021).

Intuitively, the theory predicts that features that are highly correlated with each other in usage are more likely to be fused. In doing so, it provides a possible formal foundation for a recurring intuition about the structure of linguistic systems, according to which aspects of meaning which are 'related' or 'mentally close' are expressed close together or simultaneously (Behaghel, 1930; Bybee, 1985; Givón, 1985).

Because the Efficient Tradeoff Hypothesis is grounded in theories of language processing, our work adds to a growing body of work that suggests that typological generalizations about languages arise out of a need for efficient cognitive processing (Hawkins, 1994; Jaeger & Tily, 2011; Gibson et al., 2019; Hahn, Jurafsky, & Futrell, 2020; Mollica, Bacon, Xu, Regier, & Kemp, 2020).

Acknowledgments

This work benefited from discussion at EMNLP 2021 and 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/morph-order.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Aski, J. (1995). Verbal suppletion: an analysis of Italian, French, and Spanish to go. *Linguistics*, *33*, 403-432.

Baerman, M., Brown, D., & Corbett, G. G. (2015). *Understanding and measuring morphological complexity*. Oxford: Oxford University Press.

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate.*

Behaghel, O. (1930). Zur Wortstellung des Deutschen. *Language*, *6*(4), 29–33.

Bentz, C., Ruzsics, T., Koplenig, A., & Samardžić, T. (2016, December). A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)* (pp. 142–153). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclanthology.org/W16-4117

Bickel, B. (2001). What is typology?–a short note. *Unpublished paper, University of Leipzig*.

Brown, D. (2010). Morphological Typology. In Jae Jung Song (Ed.), *The Oxford Handbook of Linguistic Typology.* Oxford University Press. Retrieved 2021-03-13, from http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199281251.001.0001/oxfordhb-9780199281251-e-023 doi: 10.1093/oxfordhb/9780199281251.013.0023

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form* (Vol. 9). John Benjamins Publishing.

Cotterell, R., Kirov, C., Hulden, M., & Eisner, J. (2019, March). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, *7*, 327–342. Retrieved from https://www.aclweb.org/anthology/Q19-1021 doi: 10.1162/tacl_a_00271

del Prado Martín, F. M., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1), 1–18.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. doi: DOI: 10.1016/j.cognition.2008.07.008

Fertig, D. (1998). Suppletion, natural morphology, diagrammaticity. *Linguistics*, *36*(6), 1065-1091.

Frank, S. L., & Ernst, P. (2019). Judgements about double-embedded relative clauses differ between languages. *Psychological Research*, *83*(7), 1581–1593.

Futrell, R. (2019, 26 August). Information-theoretic locality properties of natural language. In *Proceedings of the first workshop on quantitative syntax (quasy, syntaxfest 2019)* (pp. 2–15). Paris, France: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-7902

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*, e12814. Retrieved from http://socsci.uci.edu/~rfutrell/papers/futrell2020lossy.pdf

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Gildea, D., & Jaeger, T. F. (2015). Human languages order information efficiently. *arXiv*, *1510.02823*. Retrieved from http://arxiv.org/abs/1510.02823

Givón, T. (1985). Iconicity, isomorphism and non-arbitrary coding in syntax. *Iconicity in syntax*, 187–219.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive modeling and Computational Linguistics (CMCL 2018)* (pp. 10–18). Salt Lake City, UT: Association for Computational Linguistics.

Hahn, M., Degen, J., & Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*.

Hahn, M., Jurafsky, D., & Futrell, R. (2020, February). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *117*(5), 2347–2353. Retrieved 2021-10-08, from http://www.pnas.org/lookup/doi/10.1073/pnas.1910923117 doi: 10.1073/pnas.1910923117

Hahn, M., Mathew, R., & Degen, J. (2022). Morpheme ordering across languages reflects optimization for memory efficiency. *Open Mind: Discoveries in Cognitive Science*. Retrieved from http://stanford.edu/~mhahn2/cgi-bin/files/name_apa7.pdf

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies* (pp. 1–8).

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Jaeger, T. F., & Tily, H. J. (2011). On language 'utility': Pro-

cessing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335.

Kann, K., & Schütze, H. (2016, August). 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)* (pp. 555–560). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-2090 doi: 10.18653/v1/P16-2090

Ladd, D. R., Remijsen, B., & Manyang, C. (2009, 09). On the distinction between regular and irregular inflectional morphology: Evidence from Dinka. *Language*, *85*, 659-670. doi: 10.1353/lan.0.0136

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Markey, T. L. (1985). On suppletion. *Diachronica*, *11*, 51-66.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., ... Yarowsky, D. (2020, May). UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th language resources and evaluation conference* (pp. 3922–3931). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.483

Mollica, F., Bacon, G., Xu, Y., Regier, T., & Kemp, C. (2020). Grammatical marking and the tradeoff between code length and informativeness. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Moskal, B. (2015). Limits on allomorphy: A case study in nominal suppletion. *Linguistic Inquiry*, *46*(2), 363–376.

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

Rathi, N. (2021, June). Dependency locality and neural surprisal as predictors of processing difficulty: Evidence from reading times. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 171–176). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.cmcl-1.21 doi: 10.18653/v1/2021.cmcl-1.21

Rathi, N., Hahn, M., & Futrell, R. (2021, November). An information-theoretic characterization of morphological fusion. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10115–10120). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.793

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Still, S. (2014). Information bottleneck approach to predictive inference. *Entropy*, *16*(2), 968–989.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Sylak-Glassman, J., Kirov, C., Yarowsky, D., & Que, R. (2015, July). A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (pp. 674–680). Beijing, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P15-2111 doi: 10.3115/v1/P15-2111

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12988 doi: https://doi.org/10.1111/cogs.12988

Veselinova, L. N. (2013). Suppletion according to tense and aspect. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from https://wals.info/chapter/79

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.