An efficient communication analysis of morpho-syntactic grammatical features

Francis Mollica (mollicaf@gmail.com)
Charles Kemp (c.kemp@unimelb.edu.au)
School of Psychological Sciences
University of Melbourne, 3010, Australia

Abstract

Grammatical features vary widely across languages and this variation has been studied in detail. The functions of grammatical features, however, are not entirely clear and a number of puzzles remain. For example, why do some languages have rich feature inventories but others have few if any grammatical features? Why do many languages have features that appear to encode semantic information (e.g. animacy) that is already known to the listener? We present a computational framework that addresses questions like these by formalizing one way in which grammatical features aid communication. We use the model to illustrate how morpho-syntactic feature inventories help to solve the problem of communicating semantic structures under cognitive pressures.

Keywords: grammatical features; syntactic typology; information theory

While the extent to which language influences cognition is debated, it is generally accepted that cognitive pressures shape language (e.g., Culbertson, 2012). Consistent with this view, linguists have documented robust similarities (often called universals) across the world's languages (Greenberg, 1957). Yet despite solving the problem of communication under similar cognitive pressures, language evolution in different environments has lead to very different communicative systems. Inspired by the tradition of "competing motivations" (e.g., Haiman, 2010; Hawkins, 2004), we hypothesize that extant languages achieve near-optimal tradeoffs among the cognitive pressures within a given environment. This hypothesis has been formalized using information theory and positively evaluated against readily available samplings of the world's languages (for review see Gibson et al., 2019). Here, we argue that the efficient communication approach is useful for understanding grammatical feature inventories across languages. We describe a normative model for the communication of semantic structures that aims to capture how grammatical features ensure robust communication of semantic dependencies and semantic roles.

Across the languages of the world, word order and grammatical features are the two main strategies used to convey semantic roles and dependencies. Prior work has formalized the tradeoff between complexity and parsing efficiency in order to explain cross-linguistic word order universals (Ferrer-i-Cancho & Solé, 2003; Futrell, Mahowald, & Gibson, 2015; Hahn, Jurafsky, & Futrell, 2020). Our approach relies on a similar formal framework, but can be used to study the extent to which both word order and grammatical features support efficient communication of semantic structure. Because previous work has focused on word order, we focus here on the role of grammatical features.

Grammatical features are often integral parts of the morphological paradigms of a language, concisely conveying highly frequent semantic distinctions and preserving the structure of semantic dependencies across communicative channels. Kibort and Corbett (2008) distinguish two kinds of features: *morpho-semantic* features introduce additional semantic content to the message (e.g., tense and evidentiality) and *morpho-syntactic* features (e.g., gender and case) reflect information about the dependencies between the lexical concepts in the message, but often do not inherently convey new semantic content. The evolution of morpho-semantic grammatical features is well documented in the grammaticalization literature (e.g., Heine, 2017) and is likely a product of core cognitive pressures on a communication system. There is a strong theoretical tradition approaching grammaticalization in terms of frequency of use (Bybee, 2003; Haspelmath, 2019), which has been mirrored in psycholinguistic accounts in terms of productivity and reuse (Hawkins, 2004; O'Donnell, 2015) and production efficiency (Zipf, 1949; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). In both accounts, the frequency of linguistic units reflects communicative need (Anderson & Schooler, 1991), and languages are pressured to *compress*, or reduce the complexity of, highly needed forms to aid language processing. Morpho-semantic features follow the traditional frequency-driven productivity/reuse analyses and predict increased compression as communicative need/experience increases.

In addition to cognitive pressures of productive efficiency, the evolution of morpho-syntactic grammatical features is driven by pressures for robust communication of semantic dependencies (Comrie, 1989; Jäger, 2007). For example, consider the message, CAT FIELD CHASE MOUSE. A rational pragmatic agent, armed with prior information about the component lexical concepts, would reason that the likely interpretation is that the cat was chasing the mouse. Language, however, is used to convey both likely and unlikely events (e.g., a mouse named Jerry chasing a cat named Tom). In addition to lexical concepts, languages must therefore convey semantic dependencies between lexical concepts and their semantic roles. Word order and morpho-syntactic features are both possible strategies for preserving this graph structure.

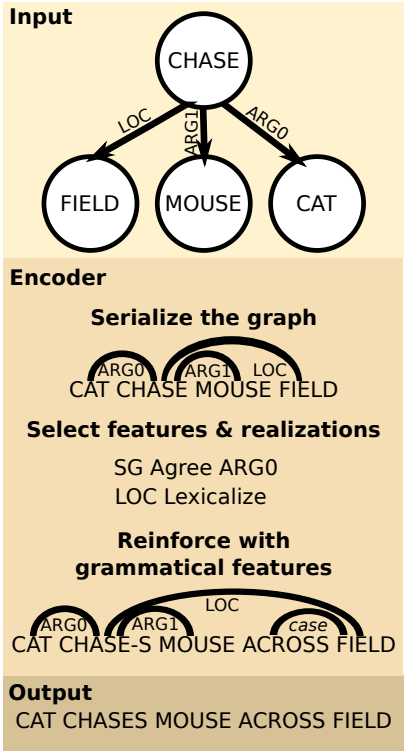

Figure 1: A schematic of a grammatical encoder. As input an AMR graph is serialized and reinforced with grammatical features by an encoding algorithm and returns a linear string.

To make this concrete, consider the Abstract Meaning Representation (AMR; Banarescu et al., 2013) graph at the top of Figure 1. The lexical concepts (nodes) are connected by directed, semantic dependencies (edges), which are each labeled with a semantic role. The edges are unordered but the constraints of human language require the message to be an incremental sequence, which requires serialization of the graph structure. If we assume that all lexical concepts must be realized, there are at least 24 possible serializations, or word orders, that could be used to communicate the input structure in Figure 1. For example, we could serialize a hierarchy by mentioning agents (arg0) before patients (arg1) which in turn are mentioned before adjuncts (e.g., location). While this might work well for *cat chase mouse*, *man painted studio* would still be ambiguous between location (*painting at the studio*) and patient (*painting the studio*) interpretations.

An obvious innovation would be to explicitly reinforce structure by adding to the sentence. For example, in Figure 1, English reinforces both the semantic role and the dependency between CAT and CHASE through agreement—i.e., morphologically tagging the head of the dependency (CHASE) with the number information of the dependent (CAT). Additionally, English reinforces the semantic role of FIELD via government—i.e., inserting a dependent morpheme (*across*), and thus a new dependency, that carries the semantic role of the head (FIELD). Head-marking and dependent-marking strategies are present to varying degrees throughout the languages of the world (Nichols, 1986). Typically, adjunct semantic roles (e.g., instrument and location) are dependency-marked; whereas, quantifiers, delimiters and negation tend to be head-marked (Nichols, 1986).

Adding grammatical markers may support robust communication of meaning, but increases the complexity of a language. The primary goal of our work is to explore how grammatical complexity trades off against success at reconstructing semantic structures. We begin by explaining our normative model, then use it to discuss the properties of both natural languages and hypothetical languages that achieve near-optimal tradeoffs between grammatical complexity and communicative success.

Theoretical framework

Consider a speaker who wishes to convey a semantic structure (e.g. top panel of Figure 1) to a hearer. The language used by speaker and hearer can be formalized as an encoder $\mathcal{L}$ that maps semantic structures to strings of words.

Encoder. Given a graph to be communicated, the encoder decides how to serialize the graph, which grammatical features are added to the resulting sequence, and how these features are realized (Figure 1). We define a distribution over encoders $\mathcal{L}$ by combining terms that correspond to these three steps:

$$P(\mathcal{L}|G) = P(S|G)P(F)P(R|F). \quad (1)$$

The distribution is defined relative to a set G that includes all unique graph structures that the encoder will encounter. Each graph in G has a serialization, and S is the set of all serializations. F is the set of grammatical features used by the encoder, and R specifies the realizations of these features.

The serialization s of each graph g is drawn from a distribution over all possible depth first traversal paths of a graph with the dependency structure of g.[1] $P(s|g)$ is formulated using a Dirichlet-multinomial distribution, which favors reuse of serialization strategies across graphs that share similar dependency structures regardless of semantic role labels.

While serializations can differ for each unique graph, the set of grammatical features F is a global property of the encoder. The features in G and their realizations R are sampled from a probabilistic context free grammar (PCFG), with uniform probability over production rules. The PCFG flips a coin to decide if there will be features or not. If there will be features, then each time a feature is determined, the PCFG flips a coin to decide whether or not this is the last feature. The PCFG contains rules for eleven features: number (single SG, plural PL), gender (female F, male M), and case (nominative NOM, accusative ACC, duration DUR, benefactive BEN, instrument INST, locative LOC, manner).

Each feature has three possible realizations. First, a feature can be expressed implicitly and assigned to a given word. For example, in Spanish, gender is implicitly encoded in some nouns (e.g., *madre*) with no overt linguistic markings. Second, a feature can be lexicalized. For example, in English

[1] We also implemented a version which allows for cross-serial dependencies and found that it produced qualitatively similar results.

the definiteness of a noun is encoded in a determiner. Typically determiners proceed the noun they modify, but as we are building a model of all possible encoders, each lexicalized feature may consistently appear either before or after the lexical concept it modifies. Third, a grammatical feature can be expressed via agreement. In agreement, the grammatical features of a dependent of a given thematic relation are typically expressed on the head by explicit marking (e.g., verb conjugation); however, we do not make directional restrictions.

The distribution over encoders in Equation 1 has two important qualitative properties. First, the $P(S|G)$ term favors encoders with consistent serialization patterns. For example, an encoder that generates *hero eats cereal* and *hero eats outside* would be preferred to an encoder that generates *hero cereal eat* and *hero eats outside* because the former encoder uses the same traversal pattern for both graphs. Second, the $P(F)$ term favors encoders that use smaller numbers of features. Intuitively, then, the distribution $P(\mathcal{L}|G)$ is inversely related to the complexity of L, and we will treat $-\log P(\mathcal{L}|G)$ as a measure of complexity.

Decoder. For convenience, we use an existing model of parsing as our decoder. We use the modified MST dependency parser (Le & Zuidema, 2015) because it builds in fewer assumptions about language than alternatives (e.g., Klein & Manning, 2004), and therefore can be used to evaluate both attested and hypothetical languages. Future work should investigate how psychologically motivated parsers and noise models influence decoding (for a start see Futrell & Levy, 2017).

Efficient communication. As formulated above, the speaker uses the encoder $\mathcal{L}$ to deterministically generate a message m based on the graph g to be conveyed. We assume that this message is transmitted without error, and upon receiving it the hearer computes a distribution $P(g'|m,\mathcal{L})$ over possible graphs. Communication succeeds to the extent that the hearer distribution assigns high probability to the original graph g. More formally, the distortion associated with the interaction is

$$d(g|\mathcal{L}) = -\log(P(g|m,\mathcal{L}). \quad (2)$$

Because the speaker is certain about the graph g to be conveyed, Equation 2 is equivalent to the KL divergence between the representations of the speaker and hearer. In expectation over all possible graphs, our distortion measure becomes the cross-entropy or information loss for having used encoder $\mathcal{L}$.

Our distortion measure trades off against complexity—intuitively, as $\mathcal{L}$ becomes more complex by reinforcing structure, reconstructive distortion should drop. $\mathcal{L}$ achieves an *efficient* tradeoff between the two if no other encoder performs better along both dimensions.

Our approach is closely related conceptually but formally distinct from previous formulations of communicative efficiency (Zaslavsky, Kemp, Regier, & Tishby, 2018; Hahn et al., 2020) based on Rate-Distortion theory (Berger, 2003) or the closely-related Information Bottleneck (Tishby, Pereira, & Bialek, 2000). We use the same distortion measure as this previous work, but formalize complexity in terms of Kolmogorov complexity (similar to Steinert-Threlkeld, 2020) rather than mutual information (as in Zaslavsky et al., 2018) or entropy (as in Hahn et al., 2020; Ferrer-i-Cancho & Solé, 2003). Kolmogorov complexity is the length of the shortest algorithm that produces an object, and our complexity measure $-\log P(\mathcal{L}|G)$ can be interpreted as the number of bits required to specify an encoder $\mathcal{L}$. Our decision is motivated by our assumption that algorithmic complexity will better reflect interpretable properties of the encoder than does mutual information. Characterizing an encoder as a probability distribution (as required for mutual information) tells us little about how linguistic representations affect compression or reconstruction; whereas, characterizing an encoder as an algorithm for mapping structures to strings (as required by Kolmogorov complexity) forces us to think about how this mapping is actually carried out.

Method

To evaluate the model we constructed a minimal corpus of AMR graphs with nodes reflecting the following structures: 1) Agent Verb Theme Location—as shown in Figure 1, 2) Agent Verb Instrument, 3) Agent Verb Theme, 4) Agent Verb Benefactor, 5) Agent Verb Duration, 6) Agent Verb Theme Benefactor, 7) Agent Verb Location Duration, 8) Agent Verb Location, 9) Agent Verb Theme Manner and 10) Agent Verb Theme Instrument. Each lexical concept was manually annotated for part of speech and grammatical feature. All possible assignments for arbitrary grammatical features like gender and number were enumerated resulting in 352 graphs.[2]

For an initial analysis of this trade-off amongst extant languages, seed sentences for each graph in the corpus were translated from English into Estonian, Japanese, Korean, Spanish and Russian and the encoding function for each language (Table 1) was specified by hand. As anchor points for comparison, we also include three encoders which implicitly express typically head-marked features (gender, number), typically dependent-marked features (case) or all features. The word order of these anchor points was fixed to Spanish/English word order. For a subsequent analysis, we sampled 450 counterfactual encoders from $P(\mathcal{L}|G)$ using MCMC methods. We ran one chain initialized on each of the encoders in Table 1 for 100 steps, retaining the top 50 encoders from each chain. As a result, these encoders are conditioned on the graphs in our corpus and therefore optimized for encoding these graphs; whereas, the encoders of extant languages might appear sub-optimal on our corpus G even if they are optimal on a more naturalistic corpus.

To calculate the listener's surprisal for our distortion metric, we use the encoder to serialize our corpus of semantic

[2]In ongoing work we are annotating verb classes (Levin, 1993) to create a more extensive corpus. Preliminary analyses suggest that most unique structures for English verb classes are present in this corpus.

Encoder	Agree	Express	Lexicalize
Dependent	-	ACC BEN DUR INST LOC MANNER NOM	-
Head	-	F M SG PL	-
All	-	F M SG PL ACC BEN INST LOC MANNER NOM	-
English	SG_{arg0}	PL	BEN DUR INST LOC
Estonian	SG_{arg0}	SG PL ACC BEN INST LOC NOM	-
Japanese	-	-	ACC BEN INST LOC NOM
Korean	-	ACC INST LOC NOM	BEN DUR
Russian	SG_{arg0}	F M SG PL ACC INST NOM	BEN LOC
Spanish	SG_{arg0} PL_{arg0}	F M SG PL	BEN DUR INST LOC

Table 1: Features and realization for each encoder in Figure 2.

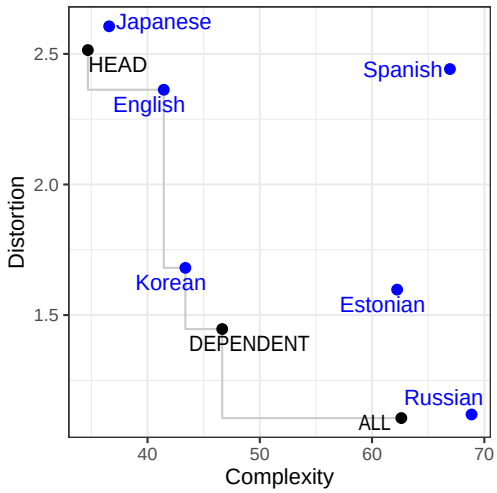

Figure 2: Tradeoff between distortion and complexity for extant languages. The anchor points are encoders that use implicit realization only (i.e. no agreement or lexicalization) and include typically head-marked features only (HEAD), typically dependent-marked features only (DEPENDENT) or all features (ALL).

graphs into messages, which we then use to train the modified MST dependency parser (Le & Zuidema, 2015). For each message in the corpus, we construct a probability distribution over each dependency by sampling the top 1000 most likely parses and normalizing their parser score with additive smoothing ($\alpha = 1e\text{-}5$). To predict the semantic roles, we used a naive Bayes classifier with smoothing ($\alpha = 1e\text{-}5$) using part of speech, grammatical features and agreement of these features across head and dependent as predictors.

Results

Figure 2 plots the encoders for extant languages and our anchor points according to our metrics for complexity and distortion. Optimal solutions should be simple and have low distortion and therefore lie towards the origin. Looking at the anchor points, when typically head-marked features are inherently expressed they result in high distortion. This is unsurprising because it is rare that head-marked features *inherently* carry semantic information about structure; instead, they carry information about structure through agreement. On the other hand, expressing typically dependent-marked features increases complexity (as there are more case features) but greatly reduces distortion. The results for encoders corresponding to the six natural languages reveal two points of interest. First, by comparing encoders with the most lexicalization (English, Japanese and Spanish) to the rest, we see that lexicalizing increases distortion, presumably because this introduces additional dependencies that the listener must correctly resolve. Second, we see that encoders expressing multiple typically head-marked features (Estonian, Spanish and Russian) increase in complexity without decreasing in distortion.

There are several possible reasons why our analysis does not find that typically head-marked features reduce distortion. First, our corpus may lack the complex structures that are usually reinforced by head-marking. For example, our corpus has no complex noun phrases, no coordinate, subordinate or relative clauses, and no discourse structure. Another possibility is that these features convey little information about semantic structures when isolated from inflectional paradigms, which are beyond the scope of our initial analysis. Alternatively, it might be that by using all possible combinations of gender and number in our corpus, we have removed correlations between structure and feature assignment that might appear in naturalistic environments. Of course, it is possible that these features do not convey structural information at all (see Dye, Milin, Futrell, & Ramscar, 2017, for one such account of gender). Our general framework can be used to investigate these hypotheses but doing so will require the development of new annotated corpora, containing both the semantic structures to be conveyed and tags for both extant and potential grammatical features.

To control for the limitations of our corpus, we compared simulated encoders optimized for this corpus (left panel of Figure 3) to typological generalizations about head- and dependent-marking of extant languages (Nichols, 1986). The right panel of Figure 3 shows how our simulated encoders realize grammatical features. Note first that 54% of the encoders make use of both head-marking via agreement (blue circles) and dependent-marking via lexicalization (red squares) similar to attested languages. Further in line with typological data, grammatical features associated with adjuncts (BEN DUR INST LOC MANNER) were more likely to be dependency marked via lexicalization (49% of realizations) than marked via agreement (1% of realizations); whereas, the other grammatical features are more likely to be marked via agreement (24% of realizations) than lexicalization (5% of realizations). Note, however, that agreement is rarely used along the Pareto front in line with our analysis of extant lan-

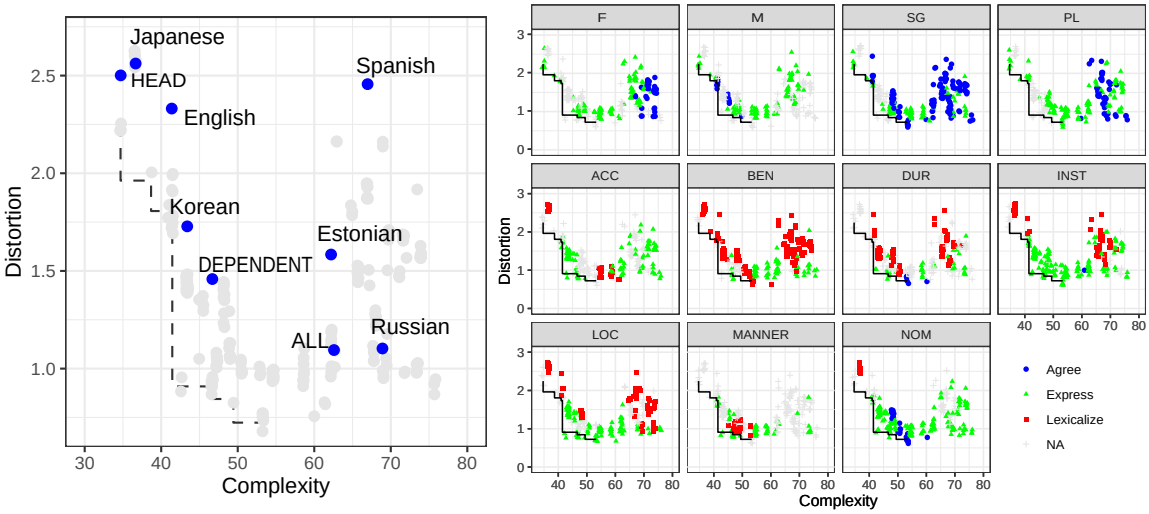

Figure 3: Left: Trade-off between distortion and complexity for simulated encoders. Right: For each simulated encoder, the realization of each grammatical feature is marked via color.

guage encoders.

Despite the limitations of our corpus, this simulation study provides a first glimpse at how individual grammatical features convey information about semantic dependencies and roles. However, the simulations themselves do not provide strong evidence that attested encoders achieve near optimal trade-offs. Another way to test the trade-off is by lesioning grammatical features present in natural language corpora to see if the distortion and complexity associated with an encoder change as predicted by our trade-off. Assuming that an attested encoder lies near the Pareto front, if we decrease complexity by lesioning a grammatical feature then the distortion of the encoder should increase. Our model makes differential predictions for the amount of increase depending upon where the encoder lies along the frontier. For example, we would expect larger increases in distortion after removing features from Korean and smaller increases for Russian because the Pareto front appears steeper by Korean than Russian. Unfortunately, our complexity measure makes it difficult to test these predictions in detail because calculating the complexity of a natural language encoder (and therefore the change in complexity caused by lesioning features) is extremely challenging. We therefore focus on one simple question and ask whether removing morpho-syntactic features from natural languages increases the distortion associated with communicating dependencies. We cannot evaluate communicative efficiency for semantic roles as there is no sufficiently annotated corpus.

Lesion Studies

Given a corpus for a language, we lesion a feature by removing it entirely from the corpus (similar to Attia, Nikolaev, & Elkahky, 2018). We focus here on case and gender. For case, we trained the MST parser with and without grammatical case for a sample of 19 languages in the UD treebank V2.4 (Nivre et al., 2016) from 7 different language families. For gender, we trained the parser on corpora with gender, without gender, and with gender assignments randomly permuted for 11 languages (4 language families) from the same resource. Using 5-fold cross-validation, we compared the expected distortion between a speaker and a hearer attempting to reconstruct the semantic dependencies. If grammatical features aid communication of dependencies, we expect that the natural language encoders will show greater distortion when lesioned than when intact.

The results are given in Table 2. As expected, removing case increased distortion (binomial test $p < 0.05$). Surprisingly, removing gender reduced distortion; yet, arbitrarily assigning gender increased distortion. Binomial tests for gender do not reach significant differences, yet the results align with our previous finding that typically head-marked features did not reduce distortion (Figure 3) and the three hypotheses proposed to explain this finding. The result for lesioning gender completely is consistent with gender either reinforcing a structure not present in our corpus (e.g., discourse) or not reinforcing structure in the absence of a more complex agreement paradigm. In line with our third hypothesis, the swapped gender study suggests that the assignment of gender is non-arbitrary and, thus, perhaps correlated with semantic structure in natural language. From a methodological perspective, the gender lesion results demonstrate that communication is not improved by simply adding features.

Discussion

Our goal was to explore how grammatical features trade off complexity and communicative distortion. Our results show that grammatical case is important for communicating structure and influences this trade-off. All natural languages in our sample use case, case is used across the Pareto front in our simulations, and lesions of case in natural languages increased distortion as expected if the languages lay along the Pareto front. Grammatical gender and number, on the other hand, do not appear to significantly influence communicative robustness in our analyses; however, a number of short-

	Intact	No Case	Δ No Case	No Gender	Swap Gender	Δ No Gender	Δ Swap Gender
Basque	13.48	14.14	-0.65	13.50	13.49	-0.02	-0.01
Bulgarian	6.23	6.24	-0.01	6.09	6.26	0.14	-0.03
Chinese	20.27	20.96	-0.69	-	-	-	-
Croatian	17.30	17.45	-0.15	-	-	-	-
Danish	16.93	17.07	-0.13	16.69	16.97	0.25	-0.03
Erzya	10.69	11.02	-0.33	10.64	10.55	0.05	0.14
Estonian	9.20	10.60	-1.40	-	-	-	-
Finnish	6.05	6.92	-0.87	-	-	-	-
Galician	26.99	-	-	27.07	26.99	-0.08	0.008
Hebrew	19.43	19.54	-0.10	19.16	19.52	0.28	-0.08
Hindi	7.88	7.96	-0.08	7.74	7.92	0.14	-0.05
Hungarian	27.17	28.45	-1.28	-	-	-	-
Korean	9.21	9.28	-0.07	-	-	-	-
Latvian	12.20	12.80	-0.59	12.04	12.26	0.16	-0.05
Lithuanian	23.49	-	-	23.55	24.07	-0.05	-0.58
Persian	20.48	20.54	-0.06	-	-	-	-
Serbian	14.98	15.05	-0.07	14.67	15.08	0.32	-0.10
Slovak	4.26	4.33	-0.07	-	-	-	-
Slovenian	9.64	9.79	-0.16	-	-	-	-
Turkish	13.32	13.67	-0.35	-	-	-	-
Urdu	16.45	16.33	0.12	16.02	16.45	0.43	0

Table 2: Our expected distortion measure for each language in our sample. Deltas reflect the difference in scores for Intact and Lesioned languages (negative values denote worse reconstruction for the Lesioned language).

comings of our analysis have been identified and we identified three questions about gender that can be explored in our framework: what if any structural information gender conveys, whether this information is contingent on an inflectional paradigm, and whether this information is contingent on how particular words are assigned to gender classes.

Our goal was not to argue for the optimality of natural languages, but rather to characterize the trade-off involved in communicating semantic structure and to provide a framework that helps to understand the conditions under which grammatical features would provide an optimal solution. More comprehensive tests of the framework will require the development of new annotated corpora, and our current results suggest that this time-intensive step is well worth taking.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psych Sci*, *2*(6), 396–408.

Attia, M., Nikolaev, V., & Elkahky, A. (2018). The morpho-syntactic annotation of animacy for a dependency parser. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 178–186).

Berger, T. (2003). Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*.

Bybee, J. (2003). Mechanisms of change in grammaticization: The role of frequency. *The handbook of historical linguistics*, 602.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago.

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, *6*(5), 310–329.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In *Perspectives on Morphological Organization* (pp. 212–239). Brill.

Ferrer-i-Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*(3), 788–791.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 688–698).

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

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

Greenberg, J. H. (1957). *Order of affixing: a study in general linguistics.* JH Greenberg, Essays in Linguistics, Chicago–London, University of Chicago Press.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*.

Haiman, J. (2010). Competing motivations. In J. J. Song (Ed.), *The Oxford handbook of linguistic typology.*

Haspelmath, M. (2019). Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *to appear*.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press on Demand.

Heine, B. (2017). Grammaticalization. *The handbook of*

historical linguistics, 573–601.

Jäger, G. (2007). Evolutionary game theory and typology: A case study. *Language*, 74–109.

Kibort, A., & Corbett, G. G. (2008). *Grammatical features inventory: Typology of grammatical features.* University of Surrey. doi: http://dx.doi.org/10.15126/SMG.18/1.16

Klein, D., & Manning, C. D. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (p. 478).

Le, P., & Zuidema, W. (2015). Unsupervised dependency parsing: Let's use supervised parsers. *arXiv preprint arXiv:1504.04666*.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Nichols, J. (1986). Head-marking and dependent-marking grammar. *Language*, 56–119.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... others (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659–1666).

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Steinert-Threlkeld, S. (2020). Quantifiers in natural language optimize the simplicity/informativeness trade-off. In *Proceedings of the 22nd Amsterdam Colloquium* (p. 513-522).

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press.