An efficient communication account of grammatical features

Francis Mollica (University of Melbourne) & Charles Kemp (University of Melbourne)
mollicaf@gmail.com

Grammatical features vary widely across languages[1] and this variation has been studied in detail. The function of grammatical features, however, is not entirely clear and a number of puzzles remain. 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 speaker? And why are grammatical features encoding other semantic dimensions (e.g. concreteness) never attested? We developed a formal framework that addresses questions like these by formalizing the way in which grammatical features aid communication.

Grammatical features can be separated into morpho-semantic features (e.g. tense, aspect) that provide new semantic information, and morpho-syntactic features (e.g. gender, case) that involve a dependency[2]. Here we focus on morpho-syntactic features and propose that the function of these features is to communicate information about semantic dependencies and semantic roles. It is natural to think that case-markings convey semantic dependencies, but not at all obvious that morpho-syntactic features such as gender can be explained in this way.

As a first test of this hypothesis, we used information-theoretic methods [3] to test robustness of communication with regard to the grammatical feature of case. This approach follows recent work in semantic typology [4,5]. We trained graph based parsing models [6-7] with and without grammatical case for a sample of 13 languages in the UD treebank V2.4 [8] and, using 5-fold cross-validation, compared the expected information loss between a speaker and a hearer attempting to reconstruct the semantic dependencies. If grammatical case aided reconstructive robustness of semantic dependencies, we expect that a language will have greater information loss (higher cross-entropy) when lesioned (i.e. stripped of case) than when intact. The data in our sample are consistent with this prediction (Table 1)—i.e., lesioned languages result in greater loss of information about semantic dependencies (binomial test: $p < 0.05$). As a preliminary check that information loss does not blindly increase due to the addition of any possible feature, we augmented English with a concreteness feature take from [9] and find the effect in the opposite direction (ΔCE = -0.16). Ongoing work is expanding this study to more grammatical features/languages.

To further test our hypothesis, we constructed a simple model of communication to probe whether the word order strategies and grammatical features of natural languages achieve a near-optimal trade-off between communicative robustness and algorithmic efficiency. Noting the isomorphism between Abstract Meaning Representation graphs and dependency parses, we constructed a generative model of "languages"—i.e., functions that encode a meaning graph into a linear string. For our analysis, we focused only on transmitting information about dependencies between verbs, arguments and adjuncts, and their semantic roles—i.e., where there is considerable variation in morphology across extant languages [10]. As a result, we only included a subset of all possible grammatical features (number, gender–assigned as in Spanish, and case) and semantic roles (agent, patient, location, beneficiary, instrument, duration and manner). We find that natural languages lie towards the optimal trade-off between algorithmic complexity and communicative robustness (Figure 1). These languages, however, are not strictly optimal with respect to our framework; however, there are clear avenues for further investigation. In this regard, our model serves as a first step in a rational analysis of grammatical features to generate new hypotheses about language complexity. In addition to the quantitive results in Figure 1, we find that our framework accounts for certain qualitative properties of natural language, including dependency length minimization [11] along with other notions of *psycholinguistic efficiency*.

Table 1: Our cross-entropy (CE) distortion measure for each language in our sample.

	CE Case	CE No-Case	△ CE
Basque	13.48	14.14	0.65
Bulgarian	6.23	6.24	0.01
Chinese	20.27	20.96	0.69
Danish	16.93	17.07	0.13
Erzya	10.69	11.02	0.33
Estonian	9.20	10.60	1.40
Hebrew	19.43	19.54	0.10
Hindi	7.88	7.96	0.08
Hungarian	27.17	28.45	1.28
Persian	20.48	20.54	0.06
Serbian	14.98	15.05	0.07
Turkish	13.32	13.67	0.35
Urdu	16.45	16.33	-0.12

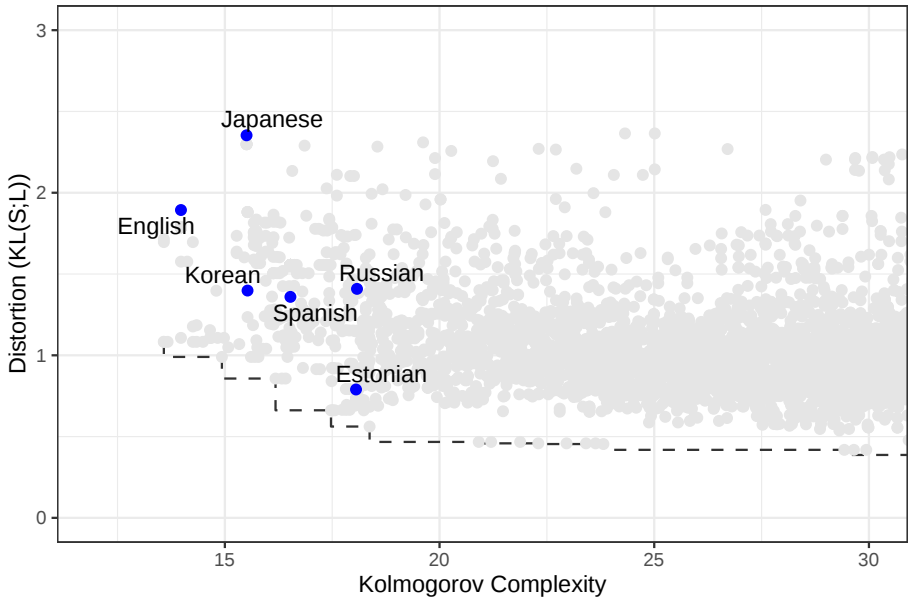

Figure 1: Using MCMC methods, we simulated possible languages (gray dots) that could adopt different word orderings and make use of different grammatical features, which could be expressed implicitly or derivationally in words, via agreement/governance inflections or lexicalized. The dotted line reflects the observed optimal trade-off frontier. Solid lines reflect natural languages.

References

[1] Dryer, M. S., & Haspelmath, M. (2013). WALS [2] Kibort, A., & Corbett, G. G. (2008). doi: http://dx.doi.org/10.15126/SMG.18/1.16 [3] Tishby, N., Pereira, F. C., & Bialek, W. (2000). arXiv preprint physics/0004057 [4] Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). PNAS, 115(31) [5] Steinert-Threlkeld, S. (in press). 22nd Amsterdam Colloquium [6] McDonald, R., & Pereira, F. (2006). EACL [7] Le, P., & Zuidema, W. (2015). arXiv preprint arXiv:1504.04666 [8] Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., et al., (2016). LREC [9] Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Behavior research methods, 46(3) [10] Nichols, J. (1986). Language [11] Futrell, R., Mahowald, K., & Gibson, E. (2015). PNAS, 112(33)