

Formal marking is redundant with lexico-semantic cues to meaning in transitive clauses

Richard Futrell (UCI), Evgeniia Diachek, Nafisa Syed, Edward Gibson, Evelina Fedorenko (MIT)

Contact: rfutrell@uci.edu

A great deal of work in linguistic typology has focused on the structure of simple transitive clauses, consisting of a subject (S), a verb (V), and an object (O). The crucial axes of crosslinguistic variation for these clauses are (1) the dominant order of these elements, e.g. SOV, SVO, VSO, etc. (Greenberg, 1963); (2) whether the order of these elements is fixed or flexible; and (3) whether there is morphological marking (e.g. case marking) distinguishing the subject from the object (Dryer, 2002). A common assumption is that these features trade off in the efficient communication of meaning: for example, a language with flexibility in the order of subject and object must have morphological case marking to distinguish the two, since word order is not informative (Kiparsky, 1997; Koplenig et al., 2017). Here we challenge this simple view with experimental evidence from two typologically distinct languages, demonstrating that in the vast majority of cases in actual usage, the subject and the object are clearly distinguishable based on their lexical semantics alone, leaving formal marking (in the form of word order or case marking) nearly entirely redundant.

We evaluated the extent to which subject and object are distinguishable based on lexical semantics alone using a hybrid corpus/experimental approach. In **Experiment 1** we extracted 500 transitive sentences at random from parsed corpora of English web text (UD English v. 2.1: Nivre et al., 2017), filtering out cases where subject or object were pronouns. We reduced each sentence to a **triplet**: the head noun of the subject, the head noun of the object, and the head lexical verb, converted to past tense to remove number agreement marking. When a triplet is presented in a shuffled order, it contains neither word-order nor morphological cues to meaning. We presented shuffled triplets to native English speakers on Mechanical Turk and asked them to identify which noun was the subject. Across three replications, participants chose the correct subject with ~90% accuracy (original study, n=79: 88.9%; replication 1, n=78: 90.0%; replication 2, n=84: 90.0%; replication 3, n=82: 90.6%). Thus in transitive clauses as actually produced in English, word order provides unique information about meaning in only 10% of instances.

In **Experiment 2** we repeated the above experiment in Russian, a language with flexible word order and differential accusative case-marking. We sampled triplets from a parsed corpus of Russian newspapers (UD Russian-SynTagRus). We removed morphological information by converting all nouns to nominative case and all verbs to the infinitive. Native Russian speakers selected the correct subject at a rate of 87.5% (n=93), nearly the same as the English accuracy.

We found that both English and Russian have the same high degree of redundancy of formal marking with lexical cues to meaning in transitive clauses. Efficiency-based theories of linguistic typology must take into account the vast redundancy of formal marking in usage. Our results support efficiency theories based on communication in a noisy channel, which requires high redundancy (Gibson et al., 2013a,b), or theories based on production constraints rather than comprehension accuracy (MacDonald, 2013). Our work also joins psycholinguistic work emphasizing the importance of lexico-semantic cues to meaning, which often override local and global syntactic constraints (e.g., MacDonald et al., 1994) and take up most storage space in our knowledge of language (Mollica & Piantadosi, under review).

References

- Dryer, M. S. (2002). Case Distinctions, Rich Verb Agreement, and Word Order Type. *Theoretical Linguistics* 28: 151–7.
- Gibson, E., Bergen, L., & Piantadosi, S. T. (2013a). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*.
- Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013b). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science* 24(7): 1079–88.
- Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.
- Kiparsky, P. (1997). The rise of positional licensing. In Ans von Stechow and Nigel Vincent, editors, *Parameters of Morphosyntactic Change*, pages 460–94. Cambridge University Press.
- Koplenig, A., Meyer, P., Wolfer, S., & Mueller-Spitzer, C. (2017). The statistical trade-off between word order and word structure: Large-scale evidence for the principle of least effort. *PLoS ONE*, 12(3), e0173614.
- MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101(4): 676–703
- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology* 4: 226.
- Mollica, F. & Piantadosi, S. T. (under review). Humans store 1.5 megabytes during language acquisition: Information theoretic bounds. <https://mollicaf.github.io/Papers/mollica201Xbits.pdf>
- Nivre, J. et al. (2017). Universal Dependencies 2.1, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-2515>.