Ordering in numerals across languages supports rapid information processing

Emmy Liu & Yang Xu (University of Toronto)
me.liu@mail.utoronto.ca

Introduction Recent work has shown that word ordering in sentence production supports information smoothing, also known as Uniform Information Density or UID [1, 2]. We investigate word ordering in compounds that have not previously been examined through the lens of information theory. Compounds are compositional structures that combine linguistic stems into one unit [3]. Compounding is a powerful mechanism for generating expressions, but the ordering of constituents can affect semantic processing (e.g., "twenty-one" vs "one-twenty"). We propose an alternative theory to UID that explains compound ordering from information gain [4]. This view predicts information front-loading as opposed to information smoothing such that the listener may gain information as rapidly as possible. We test this theory of rapid information gain (RIG) against UID in numeral expressions. Numeral systems provide a closed domain for easier analysis while relying on compounding to form new terms, e.g. "twenty-one" is composed of "twenty" and "one". In numerals with two constituents, the larger constituent is the base and the smaller is the atom. Greenberg (1978) and others have suggested that numeral systems beginning with the atom-base ordering typically switch to the base-atom ordering beyond 20 while the reverse never happens [5]. There is also a general preference for base-atom order across languages (Table 1). We test RIG and UID in explaining these phenomena, showing that RIG better accounts for the cross-linguistic data.

Methods We sampled numeral terms in 334 languages evenly from 53 language families [6]. To facilitate the information-theoretic analyses, we calculated numeral probabilities based on normalized term frequencies in 8 languages from the Google Ngrams corpora during 1900-2000 [7]. We decomposed a compound numeral into atom and base, ignoring connectives, e.g., "twenty-one" → ["twenty", "one"]. We reversed the attested order to form the alternate order (e.g., "one-twenty"). For each order, to calculate the information content of a compound, we used the formula $log_2\left(\frac{1}{U}\right) = log_2\left(\frac{1}{P(t)}\right) + log_2\left(\frac{1}{P(t|w_1)}\right) + \cdots + log_2\left(\frac{1}{P(t|w_1 \ldots w_n)}\right)$ where U is the surprisal of the compound utterance, t is the target, and the $w_i s$ are constituents. This reduces to $log_2\left(\frac{1}{U}\right) = log_2\left(\frac{1}{P(t)}\right) + log_2\left(\frac{1}{P(t|w_1)}\right) + log_2\left(\frac{1}{P(t|w_1 w_2)}\right)$ in this case [8]. For every numeral expression we computed d, or deviation from the UID ideal as $d = \frac{n}{2(n-1)} \sum_{i=1}^{n} \frac{I_{i-1} - I_i}{I_0} - \frac{1}{n}$ [9], where n is the total number of constituents and I_n is the surprisal of the target after n constituents have been communicated. We calculated cumulative surprisal as $c = \sum_{i=1}^{n} I_i$. We performed these calculations from 1-100 in each language, as well as in a universal "template" language with base-atom as the attested order and atom-base as the alternate.

Results and conclusion Overall, both UID and RIG identify the attested numeral orders as more efficient than the alternate orders. However, RIG accounts for the atom-base to base-atom order switch at 20, but UID does not (Figure 1). In the range 11-19, UID still shows strong support for the atom-base ordering, but RIG predicts that both orderings have approximately equal cumulative surprisals. For these numerical ranges, we conducted a permutation test with 100,000 trials and for each repetition, calculated the mean difference in cumulative surprisal between the two orders. This corroborated the null hypothesis for the range 11-19, but there is high statistical significance ($p < 0.004$) in rejecting the null in the range 21-29 (Figure 2), suggesting that the effect of information front-loading is significant in the range 21-29 and above. These results show that RIG offers a better account for constituent order in compounds than UID. The method outlined in this work can be extended to longer constituents, as well as to different semantic domains. Our work suggests that fine-grained ordering of lexical compounds facilitates rapid information processing and brings opportunities to characterize cross-linguistic universals in lexical design from a processing view-point. Future work should delineate when UID and RIG apply, and how the RIG principle might account for ordering beyond numerals.

Table 1-Switch in numeral ordering conventions between 11-19 and 21-29 across languages

Number of languages	*No switch*	*Switched*
atom-base → base-atom	11	52
base-atom → atom-base	271	0

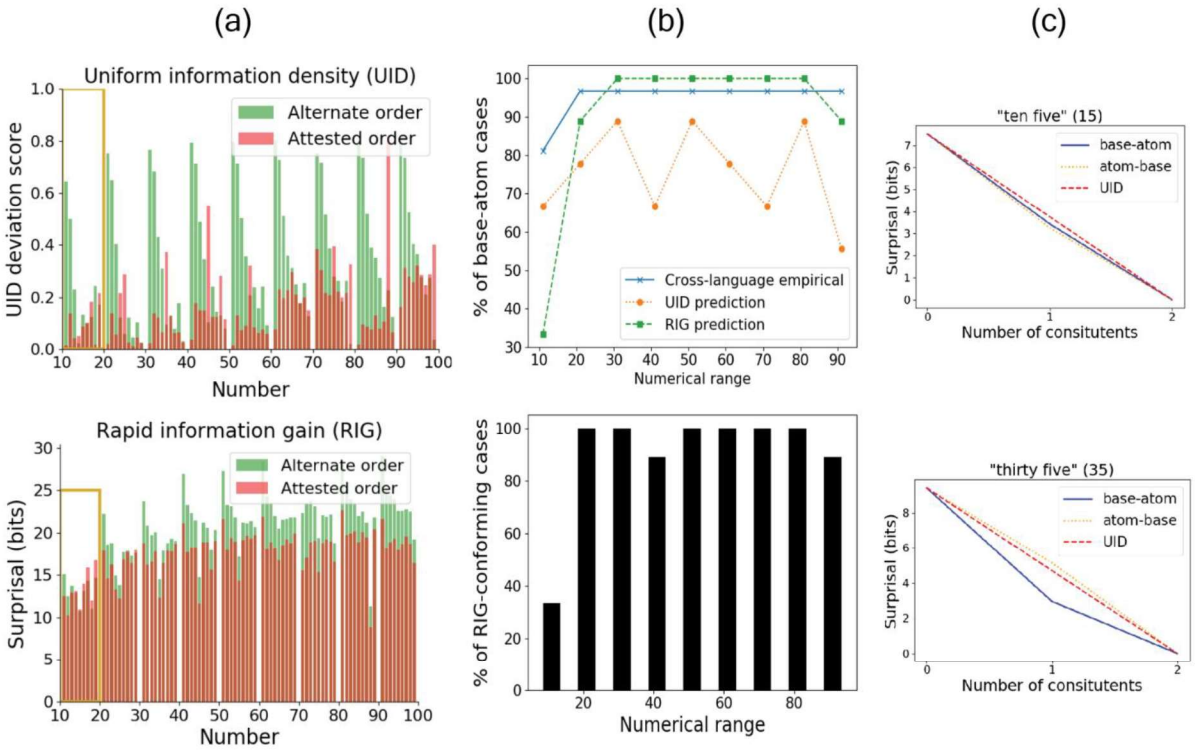

Fig. 1: Summary of main results that compare UID and RIG in accounting for cross-linguistic tendencies in numeral ordering. (a) The UID deviation and cumulative surprisal of numerals 1-100 in the template language (b) predicted base-atom percentages from the two theories against the empirical base-atom percentages across 334 languages (c) Information profiles illustrating increased departure from UID in the attested order as the number expressed increases

[1] F. T. Jaeger, "Redundancy and syntactic reduction in spontaneous speech," 2006.

[2] R. Levy and F. T. Jaeger, "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing Systems 19*, Vancouver, 2007.

[3] R. Lieber and P. Štekauer, The Oxford Handbook of Compounding, Oxford University Press, 2011.

[4] M. Oaksford and N. Chater, "Information gain explains relevance which explains the selection task," *Cognition,* pp. 97-108, 1995.

[5] J. H. Greenberg, "Generalizations about Numeral Systems," in *Universals of Human Language, Volume 3: Word Structure*, Stanford University Press, 1978, pp. 249-295.

[6] B. Comrie and E. Chan, "Numeral systems of the World's Languages," 30 Oct 2018. [Online]. Available: https://mpi-lingweb.shh.mpg.de/numeral/.

[7] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak and E. L. Aiden, "Quantitative analysis of culture using millions of digitized books," *Science,* pp. 176-182, 2011.

[8] R. Levy and F. T. Jaeger, "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing Systems*, 2007.

[9] L. Maurits, A. Perfors and D. Navarro, "Why are some word orders more common than others? A UID account," in *Neural Information Processing Systems*, Vancouver, 2010.