NOISE RESISTANCE AT MULTIPLE LEVELS OF LINGUISTIC STRUCTURE

Rachael Bailes[*1], Joel C. Wallenberg[2], and Christine Cuskley[1]

[*]Corresponding Author: rachael.bailes@ncl.ac.uk
[1]School of English Literature, Language, and Linguistics, Newcastle University, UK
[2]Department of Language and Linguistic Science, University of York, UK

Design for noise resistance is a fundamental requirement of successful communication systems both in technology and the natural world (Shannon, 1948; Brumm & Slabbekoorn, 2005). Redundancy, one way of resisting noise, has been observed at multiple levels of linguistic structure (Fenk-Oczlon, 2001; Aylett & Turk, 2004; Pluymaekers, Ernestus, & Baayen, 2005; Jaeger & Levy, 2007). The ultimate function of noise resistance features is to minimise the loss of *information*, which is quantified in bits for each signal unit in terms of its probability ($log_2 \frac{1}{p_i}$), from Shannon (1948)'s well-known formula for the average information of a set of signals.

We demonstrate that the *ordering* of high and low information units can yield more or less uniform distributions of information, and that more uniform distributions are functional for noise resistance. We also show that speakers are biased toward producing utterances with more uniform distributions (Cuskley, Bailes, & Wallenberg, 2021). Sentences of 10 words were extracted from the *The Penn-York Computer-annotated Corpus of a Large amount of English* (PY-CCLE; 2015), from which we derived four versions of each sentence: Optimised (where the information distribution was made maximally uniform by reordering units with a dedicated algorithm); Asymmetric (where the units were ordered by information content values such that the distribution was entirely asymmetric); Original (the original order of units in the natural sentence), and; Random (where the ordering of units was randomly shuffled). Each set of versions was subjected to two noise conditions: Single Unit Noise (where 3 single units were deleted from the 10 unit string in each version) and Clustered Noise (where 3 sequential units were deleted). The results show that (i) natural sentences were more uniform than random and asymmetric orders, (ii) more uniform distributions lost less information overall, and (iii) uniform distributions of information are 100% successful at preventing *catastrophic information loss* (defined as the loss of $\geq$50% of a distribution's information content). These findings indicate that linguistic units are dynamically ordered across the sentence so as to optimise resistance to noise.

We expand on the theoretical ramifications of an information-loss threshold and the effect of clustered noise as they apply to linguistic communication. Noise (anything that prevents the receiver from observing a complete signal, and only the intended signal) is particularly problematic for language when it comes in a large chunk, due to language's famous "duality of patterning" (Hockett, 1958). Linguistic structure is multi-layered, and the structure at each layer must be inferred from the information a previous layer provides.

An information theoretic approach suggests that each layer of structure serves as a *signal* for another layer. Each layer of structure is a message for which another layer is the signal, and so on. While noise might take a form that affects particular units in isolation (e.g., omitting or altering one particular word in a sentence; (Ryskin, Futrell, Kiran, & Gibson, 2018)), chunks of noise that have the potential to obscure multiple linguistic units are a larger danger from the perspective of catastrophic signalling failures. Even a small amount of noise that spans multiple units - e.g. obscuring the last few phonological segments of one word and the first few of the following word - may well result in a failure to reconstruct *both* words. Because linguistic structure at levels other than the strictly phonetic consists of categorical units, analogue types of noise have the potential to obscure multiple categorical linguistic symbols at once, and so result in knock-on effects disrupting multiple units at higher levels. This may also predict different distributions of information at different levels. Phonemic categories are the first to be signalled, and terminal fading of phonetic material presents a particular noise constraint on this level of signalling. Since information loss compounds, conservation at this first level is crucial. The risk of compounded information loss combined with phonetic processing constraints therefore predict the *frontloading* of phonemic information within a word (King & Wedel, 2020). That phonetic articulation is particularly amenable to redundancy through lengthening (Aylett & Turk, 2004) may serve to mitigate the asymmetry this introduces.

We thus explicate a model of how different levels of linguistic structure signal one another in information theoretic terms. This information-theoretic paradigm for linguistic processing specifies that (i) linguistic communication (i.e. 'message' reconstruction) is a multilayered process of sub-setting the space of possible outcomes based on signal information, and (ii) signals at lower linguistic levels subset the space of possibilities at higher linguistic levels in a serial 'daisy-chain' of inter-level communication, such that (iii) information loss at lower levels of linguistic structure **compounds** information loss at higher, propositional levels.

From this perspective, it follows that linguistic communication is constrained not only by the absolute amount of noise, but by the potential for analogue noise to cause catastrophic signalling failure between linguistic levels. Linguistic planning is biased toward more uniform distributions of information, suggesting adaptation for the prevention of such catastrophic communication failure.

References

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, *47*(1), 31–56.

Brumm, H., & Slabbekoorn, H. (2005). Acoustic communication in noise. *Advances in the Study of Behavior*, *35*, 151–209.

Cuskley, C., Bailes, R., & Wallenberg, J. (2021). Noise resistance in communication: Quantifying uniformity and optimality. *Cognition*, *214*, 104754.

Ecay, A. (2015). *The Penn-York Computer-annotated Corpus of a Large amount of English based on the TCP (PYCCLE-TCP).* Public release 1. https://github.com/uoy-linguistics/pyccle.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. *Typological Studies in Language*, *45*, 431–448.

Hockett, C. F. (1958). A course in modern linguistics.

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

King, A., & Wedel, A. (2020). Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing. *Open Mind*, *4*, 1–12.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America*, *118*(4), 2561–2569.

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, *181*, 141–150.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.