# Average surprisal of parts-of-speech

Hannah Kermes and Elke Teich (Universität des Saarlandes, Germany)

We present an approach to investigate the differences between lexical words and function words and the respective parts-of-speech from an information-theoretical point of view (cf. Shannon, 1949). We use average surprisal (AvS) to measure the amount of information transmitted by a linguistic unit. We expect to find function words to be more predictable (having a lower AvS) and lexical words to be less predictable (having a higher AvS). We also assume that function words' AvS is fairly constant over time and registers, while AvS of lexical words is more variable depending on time and register.

Our assumptions are based on well known differences between lexical words and function words with respect to frequency, word length, number (open vs. closed class) and information content, lexical words being the main carriers of meaning (Biber et al., 1999). Besides, Piantadosi, Tily, and Gibson (2011) show that average information content is a better predictor for word length than frequency. According to Quirk et al. (1985, 72) the choice is larger in typical contexts of lexical words than of function words and Linzen and Jaeger (2015) provide evidence that the number of choices in a particular context affects the predictions of people for upcoming syntactic construction.

As an example we look at the development of scientific English. We assume that due to specialization, scientific texts exhibit greater encoding density over time Halliday (1988); Halliday and Martin (2005), i.e. more compact, shorter linguistic forms are increasingly used, in order to maximize efficiency in communication. One feature of linguistics densification is the extensive use of lexical words (often approximated by lexical density). Thus, we expect to see differences in the AvS of lexical words in scientific writing over time and with respect to general language.

## Data and Methodology

To test our assumption about the constancy/variability in AvS of lexical words and function words over time and registers, we focus on the period of Late Modern English using two data sets the Royal Society Corpus (RSC, Kermes et al., 2016) and the Corpus of Late Modern English Texts, version 3.0 (CLMET, Diller et al., 2011).

The RSC is a historical corpus of written scientific English based on the first two centuries of the Philosophical Transactions of the Royal

Society of London (1665–1869) and comprises approx. 35 million tokens. With its long and continuous history the journal provides a very good basis for diachronic analysis of English scientific writing. The RSC is annotated for lemma and parts-of-speech using TreeTagger (Schmid, 1994, 1995). CLMET is a register-mix corpus with a similar size, time span (1710–1920) and comparable annotation (Penn Treebank tag set, Marcus et al., 1993).

As a measure of surprisal, we use a model of AvS, i.e. the average amount of information a word encodes in number of bits, calculated as

$$AvS(unit) = \frac{1}{|unit|} \sum_i -\log_2 p(unit|context_i)$$

i.e. the (negative log) probability of a given unit (e.g. a word) in context (e.g. its preceding words) for all its occurrences (cf. Genzel & Charniak, 2002). In general, surprisal Levy (2008) captures the intuition that the less probable a linguistic unit is in a given context, the more surprising or informative that unit will be and the more bits are needed to encode it (and vice versa). This allows us to investigate the differences between lexical words and function words with respect to information content and predictability synchronically and diachronically looking at the distribution of AvS for each part-of-speech group including the range/spread of AvS, (relation of) mean and median. The AvS values for each token are annotated in the corpora for an easy access.

We extract all words, excluding non-word items from each corpus with information about its parts-of-speech (UPenn tagset), AvS and time period. For a better abstraction we group the parts-of-speech into function words (*article, preposition, pronoun, modal, conjunction* and the auxiliaries *be* and *have*), lexical words (*noun, adjective, verb, adverb*), and *other*.

## AvS of parts-of-speech

Figure 1 displays the distribution of AvS values for parts-of-speech in the RSC. Function words are to the left of the diagram (*article, preposition, pronoun, modal, conjunction* and the auxiliaries *be* and *have*), lexical words to the right (*noun, adjective, verb, adverb*).

Generally, we can observe the following differences in the distribution of AvS for lexical words and function words. Lexical words have an almost equal mean and median, the distribution has a large spread/range and is mostly symmetric with a relatively flat curve. Function words have a lower mean than lexical words, the median is often lower than the mean with distinct peaks mostly to the lower end.

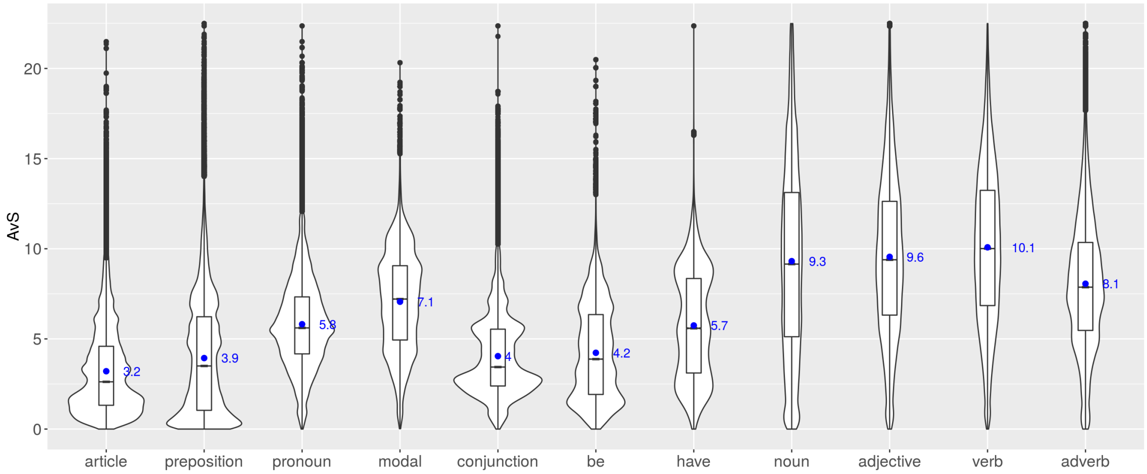

*Figure 1*. AvS of parts-of-speech in the RSC

Function words behave more diverse than lexical words. *Articles, prepositions* and *conjunctions* are positively skewed with the major peak shifted to the lower end of the distribution and the median being lower than the mean. The pronounced peak of *prepositions* to the far lower end is related to the preposition *of* in complex noun phrases. *Pronouns, modals* and the *auxiliaries* show characteristics of both function words and lexical words. *Pronouns* have a mostly symmetric distribution, mean and median being almost equal with a pronounced peak around the median. *Modals* and *auxiliaries* have flatter curves than the other function words, with less distinct peaks. The mean of *modals* is high in comparison to the other function words.

We compare these distributions to the AvS of parts-of-speech in CLMET (Figure 2) to see whether our findings are specific for scientific language or whether they have a more general character.

In general, we can observe a similar picture. There are differences between function words and lexical words with both groups exhibiting more or less the same general properties. A closer look reveals differences between CLMET and the RSC for specific parts-of-speech:

- the distribution of *nouns* is less symmetric in CLMET with a distinct peak to the higher end.
- the curves of *modals* and *auxiliaries* is more pointed with at least two peaks.
- the positive skew of *articles* and *prepositions* is less pronounced

The tendencies can be related to linguistic complex structures (such as complex noun phrases) being more common and thus more predictable in the RSC than in CLMET.

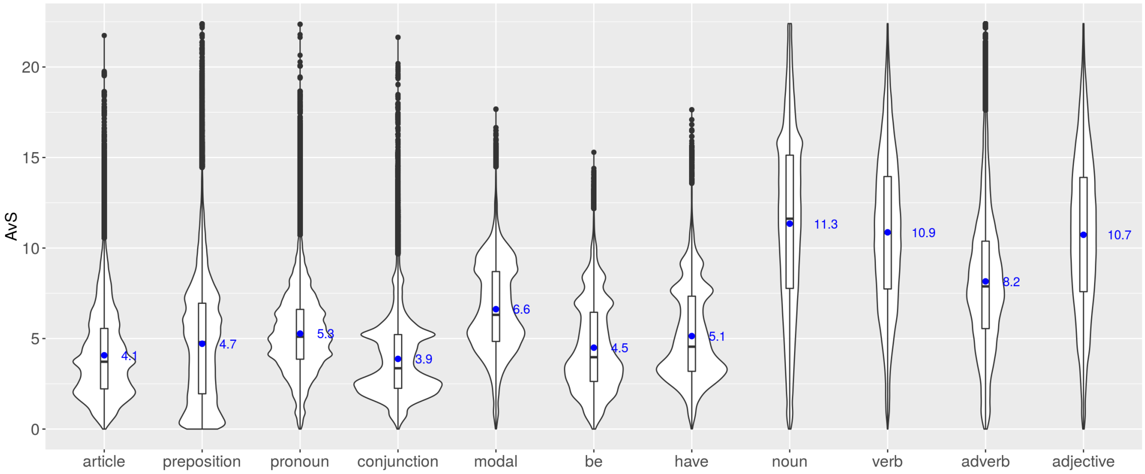

*Figure 2*. AvS of parts-of-speech in CLMET

## Diachronic development of AvS

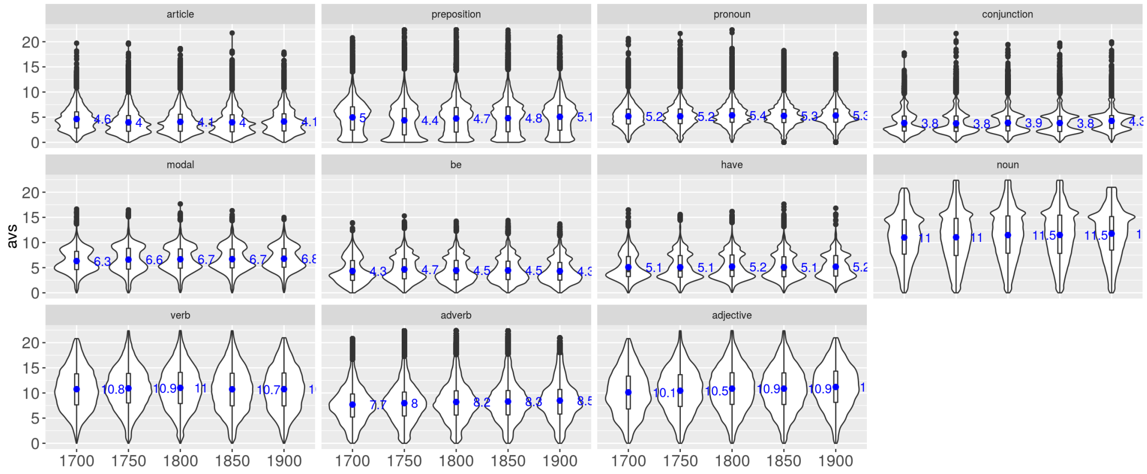

*Figure 3*. Diachronic development in CLMET

If we now look at the diachronic development of AvS of parts-of-speech, we can observe that the AvS remains relatively stable over time in CL-MET, mean, average and shape of the distributions hardly change (cf. Fig-ure 3). In the RSC, however, we can observe small changes for some of the parts-of-speech. For typical modifiers such as *adverbs, modals* as well as for *pronouns* AvS increases slightly. For *articles, prepositions* and *nouns* as well as for *verbs* and the auxiliaries *be* and *have* AvS decreases.

Overall, the range of the distribution increases for all parts-of-speech and there is a general trend to develop peaks to the lower end of the

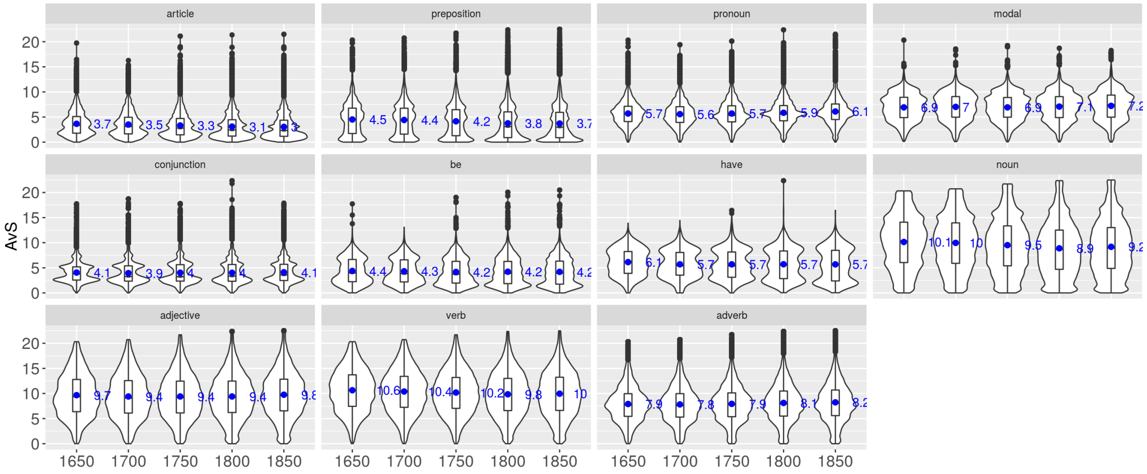

*Figure 4*. Diachronic development in the RSC

distribution. The differences that we observe in synchronic comparisons of CLMET and the RSC get stronger over time. In other words, scientific writing gets more distinct from “general language” over time with respect to the AvS of lexical and function words.

## References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Diller, H.-J., De Smet, H., & Tyrkkö, J. (2011). A European database of descriptors of English electronic texts. *The European English Messenger*, *19*, 21–35.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 199–206). Association for Computational Linguistics.

Halliday, M. (1988). On the Language of Physical Science. In M. Ghadessy (Ed.), *Registers of Written English: Situational Factors and Linguistic Features* (pp. 162–177). London: Pinter.

Halliday, M., & Martin, J. (2005). *Writing Science: Literacy and Discursive Power*. Taylor & Francis.

Kermes, H., Degaetano, S., Khamis, A., Knappen, J., & Teich, E. (2016). The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the LREC 2016.* Portoroz, Slovenia.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008*

*Conference on Empirical Methods in Natural Language Processing* (pp. 234–243). Honolulu.

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*. doi: 10.1111/cogs.12274

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011, March). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. doi: 10.1073/pnas.1012551108

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Longman.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing* (pp. 44–49). Manchester, UK.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop.*

Shannon, C. E. (1949). *The mathematical theory of communication*. Urbana/Chicago: University of Illinois Press.