A diachronic perspective on efficiency in language use: *that*-complement clause in academic writing across 300 years

Stefania Degaetano-Ortlieb, Pauline Krielke,
Franziska Scheurer and Elke Teich
Saarland University

Efficiency in language use and the role of predictability in context have attracted many researchers from different fields (Zipf 1949; Landau 1969; Fidelholtz 1975, Jurafsky et al. 1998; Bybee and Scheibman 1999; Genzel and Charniak 2002; Aylett and Turk 2004; Hawkins 2004; Piantadosi et al. 2009, Jaeger 2010). The analysis of reduction processes, where linguistic units are reduced/omitted has enhanced our knowledge on efficiency in communication. Possible factors affecting retention or omission of an optional element include discourse context (cf. Thompson and Mulac 1991), the amount of information a unit transmits given its context (known as *surprisal*, cf. Jaeger 2010) or the complexity of the syntagmatic environment (Rohdenburg 1998). So far, the role change in language use plays has been less considered.

Taking a diachronic perspective, we hypothesize that as conventionalized uses establish, omission will be favored. We analyze VERB+*that*-clauses in scientific texts, assuming that due to specialization and institutionalization (Ure 1982) specifically scientific language undergoes processes of conventionalization (Halliday 1988). These clauses combine with few semantic domains (mental, e.g. *think, know*, and communication, e.g. *say, suggest*; Biber 1999), allowing us to analyze a confined set of options. Previous work has shown a major influence of the matrix verb for retention/omission (cf. Elsness 1984, Underhill 1988, Thompson and Mulac 1991, Roland et al. 2005), e.g. high frequency verbs (e.g. *say*, *know*) favor omission. Retention is favored in environments in which a *that*-clause becomes less predictable based on its context, e.g. in cases of an intervening noun phrase between verb and *that*-clause (cf. Thompson and Mulac 1991, Biber 1999, Jaeger 2010).

For our analysis, we use the Royal Society Corpus (RSC) (Kermes et al. 2016) built from the Transactions and Proceedings of the Royal Society of London. It consists of around 300 million tokens spanning from 1665-1996 (v5.0) comprising metadata (e.g. author, publication year), linguistic information (e.g. tokens, lemmas), and each word's surprisal as a measure of predictability in context (cf. Jaeger 2010). We calculate surprisal as the negative log-probability of a word given its three previous words (see equation 1). For diachronic comparison, calculation is based on decades.

$$\text{Surprisal}(w_i) = -\log_2 p(w_i | w_{i-1}\, w_{i-2}\, w_{i-3}) \quad (1)$$

Low surprisal indicates highly predictable words, high surprisal less predictable words. The more predictable words are in a context, the more their use is conventionalized in that context. Considering VERB+*that*-clauses, we inspect surprisal of the matrix verb and the complementizer. A considerable decrease in surprisal over time of the verb and complementizer will indicate development of a conventionalized use (cf. Degaetano-Ortlieb and Teich 2016, 2018).

Preliminary results show a tendency towards a conventionalized use of the pattern, surprisal of the verb (Figure 1) and especially the complementizer (Figure 2) decreasing steadily over time. We consider zero alternatives (*that*-omission) by inspecting particular complementation patterns (e.g. VERB+[*that* or zero]+noun-phrase) and comparing whether as a pattern becomes more conventionalized (decrease in surprisal), *that*-omission increases. Moreover, we analyze in more detail the syntagmatic contexts of *that* and zero alternatives observing which additional contextual factors contribute to a reduced variant as the pattern increases predictability over time.

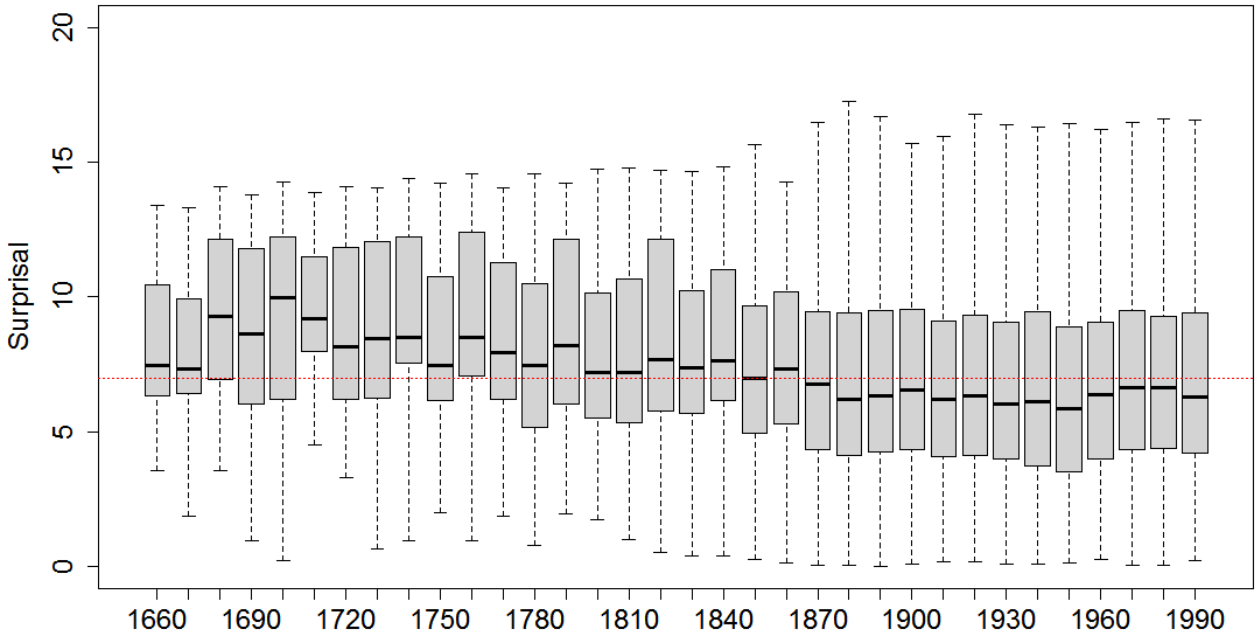

Figure 1: Surprisal of the matrix verbs in the VERB+*that*-clause pattern across decades in the RSC

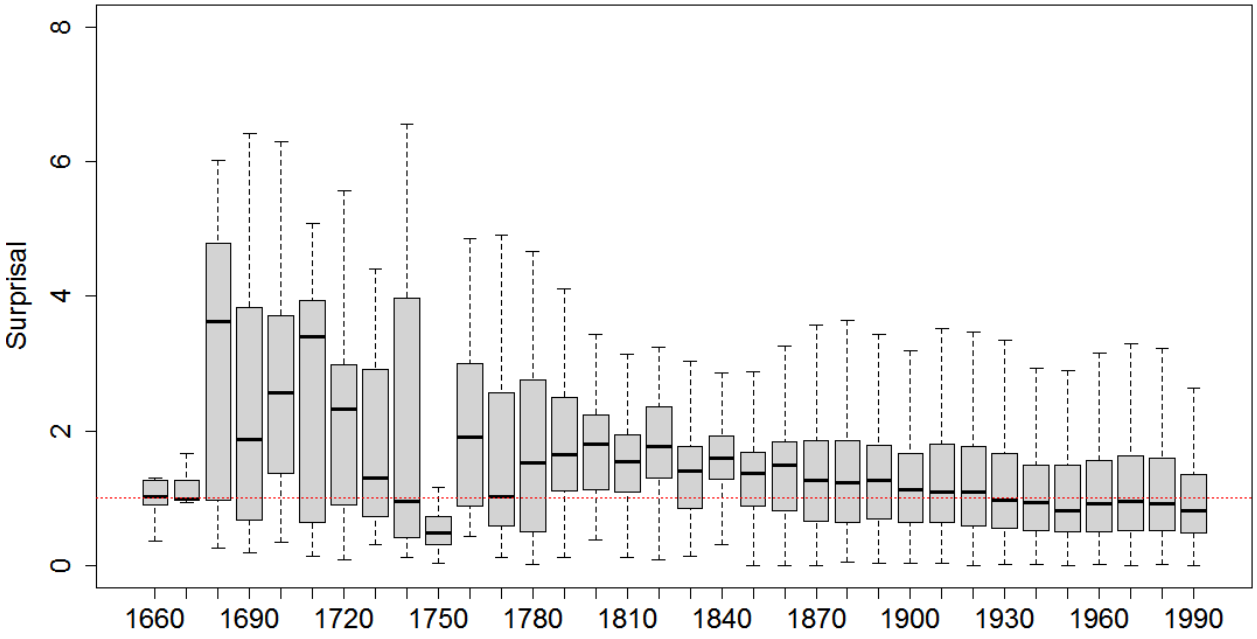

Figure 2: Surprisal of the complementizer in the VERB+*that*-clause pattern across decades in the RSC

References

Aylett, M. P., & 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.

Biber, D. (1999). A register perspective on grammar and discourse: Variability in the form and use of English complement clauses. *Discourse Studies,* 1(2): 131–150.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics*, *37*(4): 575–596.

Degaetano-Ortlieb, S. & Teich, E. (2016). Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Berlin, Germany, pages, 165-173. ACL.

Degaetano-Ortlieb, S. & Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING*, pages 22–33, Santa Fe, NM, USA. ACL.

Elsness, J. (1984). That or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies*, 65: 519–533.

Fidelholz, J. (1975). Word frequency and vowel reduction in English. In *CLS-75*, pp. 200–213. University of Chicago.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the Association of Computational Linguistics*, pp. 199–206, Philadelphia, PA.

Halliday, M.A.K. (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.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Jaeger, T.F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61: 23-62.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymond, W. D. (1998). Reduction of English function words in Switchboard. In *ICSLP-98*, Vol. 7, pp. 3111–3114, Sydney.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. & Teich, E. (2016). The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the 10th LREC*, Portoroz, Slovenia.

Landau, M. (1969). Redundancy, rationality, and the problem of duplication and overlap. *Public Administration Review*, 346–358.

Piantadosi, S. T., Tily, H., & Gibson, E. (2009). The communicative lexicon hypothesis. In *The 31st annual meeting of the Cognitive Science Society* (CogSci09), pp. 2582–2587.

Rohdenburg, G. (1998). Clausal complementation and cognitive complexity in English. In F. W. Neumann & S. Schulting (Eds.), *Anglistentag*, Erfurt, pp. 101–112. Trier: Wissenschaftlicher Verlag.

Roland, D., Elman, J. L., & Ferreira, V. S. (2005). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*, 1–28.

Thompson, S. A., & Mulac, A. (1991). The discourse conditions for the use of complementizer that in conversational English. *Journal of Pragmatics*, 15: 237–251.

Underhill, R. (1988). *The discourse conditions for that-deletion*. San Diego State University.

Ure, J. (1982). Introduction: Approaches to the study of register range. *International Journal of the Sociology of Language*, 35: 5–23.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.