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; Plantadosi 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. Elness 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.

\[
\text{Surprisal}(w_i) = -\log_2 p(w_i|w_{i-1} w_{i-2} w_{i-3})
\] (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


