THE COURSE OF GRAMMATICAL CHANGE IN SCIENTIFIC WRITING
Interdependency between convention and productivity

STEFANIA DEGAETANO-ORTLIEB; KATRIN MENZEL; ELKE TEICH
SAARLAND UNIVERSITY

We present an empirical approach to analyze the course of usage change in scientific writing. A great amount of linguistic research has dealt with grammatical changes, showing their gradual course of change, which nearly always progresses stepwise (see e.g. Bybee et al. 1994, Hopper and Traugott 2003, Lee 2011, De Smet and Van de Velde 2013). Less well understood is under which conditions these changes occur. According to De Smet (2016), specific expressions increase in frequency in one grammatical context, adopting a more conventionalized use, which in turn makes them available in closely related grammatical contexts.

Our approach is based on information theory (cf. Shannon 1948), where the amount of information conveyed by a linguistic unit is measured based on the probability of a unit given a context, termed *surprisal*. If surprisal is low, the unit has high probability of occurring in a given context, if surprisal is high, the probability of that unit to appear in a given context is low. A greater number of low surprisal units indicates conventionalized use, while a greater number of high surprisal units indicates productive use.

To observe the course of usage change in scientific writing, we use two scientific corpora: the Royal Society Corpus (RSC; Kermes et al. 2016) built from the Proceedings and Transactions of the Royal Society of London spanning from 1665 to 1869, and the Scientific Text Corpus (SciTex; Degaetano et al. 2013) consisting of research articles from the 1970/80s and 2000s. Based on the linguistic unit under investigation, surprisal is calculated for each unit and annotated into both corpora.

In our analyses, we investigate different linguistic phenomena undergoing usage change. Consider, e.g., at the morphological level the combining form *–lysis*, which increases in frequency over the 350 years investigated and denotes decompositions or dissolution. Examples of *–lysis* in different word classes are *analysis*, *electrolysis* (noun), *analytical*, *paralytic* (adjective), *analys/ze*, *hydrolys/ze* (verb), and *analytically*, *electrolytically* (adverb). Surprisal is calculated based on the different forms of *–lysis* given the preceding part of the word and two previous words (i.e. their preceding context). Preliminary results (see Figure 1) show that nouns become quite predictable over time with an increase of low surprisal values (dark grey bars, reaching almost 80%,), i.e. they adopt a quite conventionalized use. Moreover, since noun forms of *–lysis* become more predictable, *–lysis* spreads out to other word classes: adjectives, verbs, and adverbs. Forms in these word classes (see e.g. verb in Figure 1) increase in predictability (i.e. low surprisal) only in later time periods. As specific forms become more predictable in one of these word classes (e.g. different verb forms of *analyze*), other forms are coined (*co-analyzed*, *preanalyzed*, *re-analyzed*). Thus, while units become more predictable, i.e. more conventionalized, these units are used in other contexts, resulting in an increase in productivity.

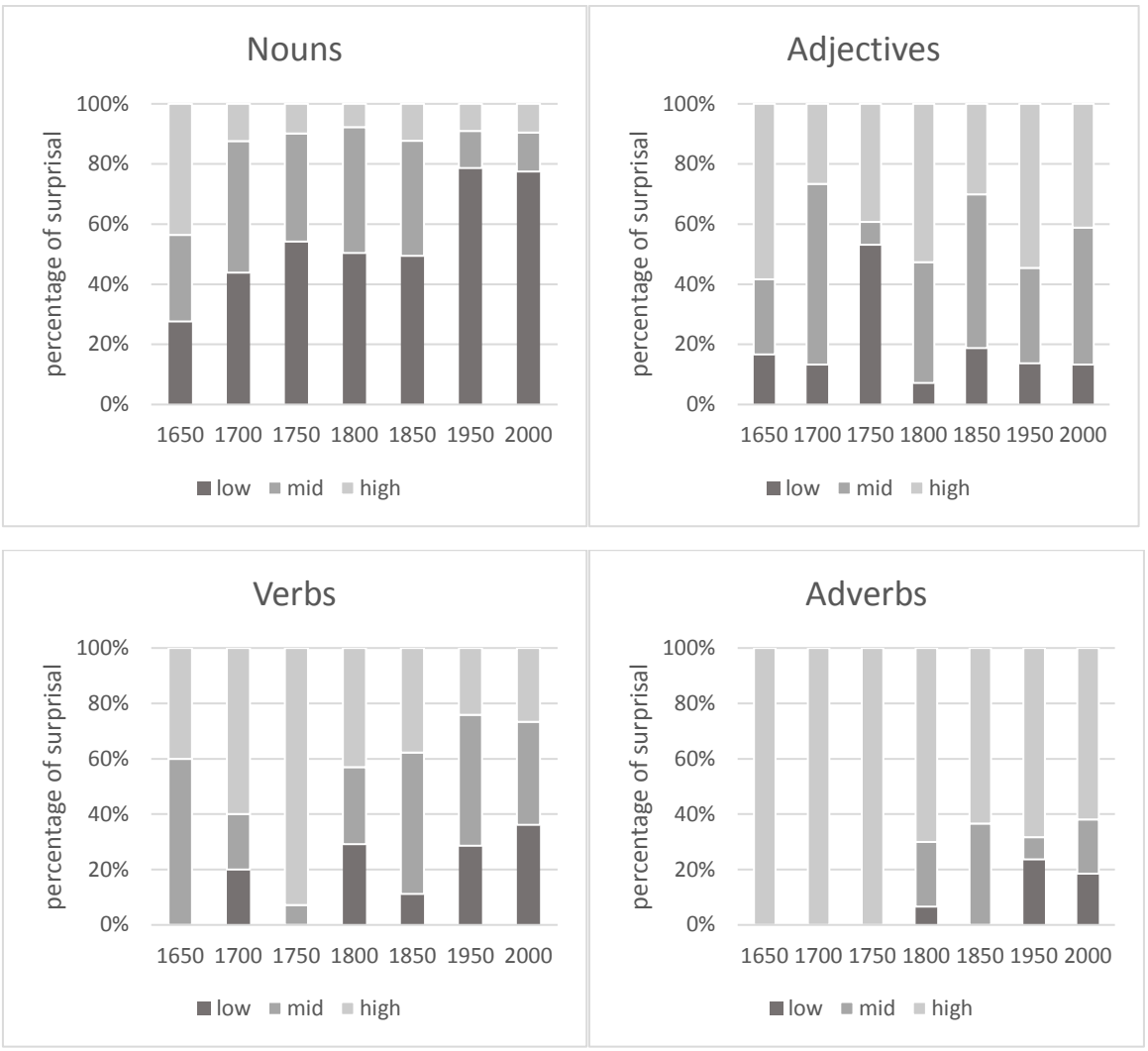

Figure 1: Percentage of surprisal in low, middle and high bins across time for different word classes of the combining form *–lysis*

References

Bybee J.L., Perkins R., and Pagliuca W. 1994, *The evolution of grammar: Tense, aspect and modality in the languages of the world*, The University of Chicago Press, Chicago.

Degaetano-Ortlieb S., Kermes H., Lapshinova-Koltunski E. and Teich E. 2013, *SciTex – A diachronic corpus for analyzing the development of scientific registers*, in Bennett P., Durrell M., Scheible S. and Whitt R. J. (eds.), *New methods in historical corpus linguistics. Corpus linguistics and interdisciplinary perspectives on language - CLIP, 3.* Narr, Tübingen, pp. 93-104.

De Smet H. and Van de Velde F. 2013, *Serving two masters: Form-function friction in syntactic amalgams*, in "Studies in Language" 37, pp. 534-565.

De Smet H. 2016, *How gradual change progresses: The interaction between convention and innovation*, in "Language Variation and Change" 28[1], pp. 83-102.

Hopper P.J. and Traugott E.C. 2003, *Grammaticalization*, Cambridge University Press, Cambridge.

Kermes H., Knappen J., Degaetano-Ortlieb S. and Teich E. 2016, *The Royal Society Corpus: From uncharted data to corpus*, in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2982-1931.

Lee J.W. 2011, *Much ado about a lot: A corpus study of 'much' as a negative polarity item*, in *Proceedings of the 20th International Conference on Historical Linguistics*. Osaka, Japan, July 26th. http://www.acsu.buffalo.edu/~jiwonlee/.

Shannon C.E. 1948, *A mathematical theory of communication*, in "Bell Syst Tech Journal" 27, pp. 379-423 (Part I) & 623-656 (Part II).

Bionote

Stefania Degaetano-Ortlieb is a PostDoc in the Collaborative Research Center *Information Density and Linguistic Encoding* at Saarland University. She has expertise in corpus linguistics, text mining, data analytics and probabilistic modelling for sociolinguistics, register/language variation and change. Recently, she obtained funding to work on linguistic profiles of social variables.

Katrin Menzel is a PostDoc and lecturer at the Department of Language Science and Technology at Saarland University. She wrote her PhD thesis on German-English contrasts in textual cohesion and now works in the Collaborative Research Center *Information Density and Linguistic Encoding* at Saarland University.

Elke Teich is a full professor of English Linguistics and Translation Studies at Saarland University, head of the Collaborative Research Center *Information Density and Linguistic Encoding* and PI in the CLARIN-D project. Teich's expertise ranges from descriptive grammar over register analysis to translatology and she has worked in machine translation, automatic text generation, corpus linguistics and digital humanities.

Author's address: s.degaetano@mx.uni-saarland.de; k.menzel@mx.uni-saarland.de; e.teich@mx.uni-saarland.de