Complexity in complementation

Cuyckens, H.; Fonteyn, L.; Petré, P.; Kristiansen, G.; Franco, K.; Pascale, S. de; ... ; Zhang, W.

Hubert Cuyckens, Lauren Fonteyn and Peter Petré

Complexity in Complementation

Understanding Long-term Change in Verb Complementation in terms of Inter- and Intra-individual Variation

Abstract: This paper presents the contours of an innovative study probing the impact of interindividual differences in cognitive representations on long-term population-level language change. Focussing on the system of English verb complementation, the study seeks to offer a unified cognitive and socio-linguistic account of syntactic variation and change that has so far been missing. In particular, it investigates (i) idiolectal variation in the constraints on alternating complementation patterns; (ii) the interaction between the linguistic behaviour of individuals and the (changing) distribution of grammatical variants at the population-level across different time stages. The research is carried out by modelling the observed patterns of stability or diffusion in terms of the weakening and strengthening of grammatical and cognitive constraints within a socially homogeneous group of individuals.

Keywords: syntactic change, finite vs. nonfinite complement clauses, idiolectal variation, population-level change

1 Introduction

The present paper presents the contours of an innovative, cognitive-sociolinguistic study into grammatical variation and change, which is concerned with the interaction between long-term population-level change in syntactic variation, and idiolectal variation in such change. Specifically, this study contributes to a theory

Hubert Cuyckens, Dept. of Linguistics – KU Leuven, Blijde-Inkomststraat 21, Box 3308, 3000 Leuven, Belgium, e-mail: hubert.cuyckens@kuleuven.be
Lauren Fonteyn, Leiden University Centre for Linguistics, Arsenaalplein 1, 2311CT Leiden, The Netherlands, e-mail: l.fonteyn@hum.leidenuniv.nl
Peter Petré, University of Antwerp – Dept. of Linguistics, Prinsstraat 13, S.D.216, 2000 Antwerpen, Belgium, e-mail: peter.petre@uantwerpen.be

Note: Authorship is equally shared.

https://doi.org/10.1515/9783110733945-022

of language as a complex adaptive system (Beckner et al. 2009), in which language is viewed as a self-organizing network which shows properties at the macro-level that are not recurrent at the individual micro-level, but nevertheless emerge out of complex behaviour at that individual level. Attention is given to the changes in "variant grammatical production" (or mixed usage), whereby the language user actually uses the variants in alternation (under varying conditions), rather than consistently opting for one of the variants at all times. The domain of research is variation in clausal verb complementation – in particular, the alternation/competition between finite and nonfinite complement clauses (CCs) with selected matrix verbs (or Complement-Taking Predicates [CTPs]), as illustrated in (1) and (2):

(1) a. I *remember* a detective *coming in* and *speaking* to me. (Old Bailey Corpus, t-18811212)
    b. I *remember that* a detective came in and spoke to me. (adapted from (1a))

(2) a. They *believed that* the Bible was the word of God. (1821, CLMET)
    b. They *believed* the Bible *to be* the word of God. (adapted from (2a)) (= ACI construction)
    c. The Bible *was believed* to be the word of God. (adapted from (2a)) (= NCI construction)

The time frame envisaged is the (Late) Modern English period (1700–1920). During this period type (1a) was established as a competitor of (1b), while (2a, b) lost ground vis-à-vis (2c).

By focussing on variation and change in English verb complementation, this study is designed to provide new insights regarding the history of English syntax. First, the variation between finite and nonfinite complement clauses allows us to empirically verify Denison's claim (1998: 256) that "a long-term trend in English has been the growth of nonfinite complement clauses at the expense of finite clauses" (see also Rohdenburg 2014). Second, the domain of variation in verb complementation lends itself well to the study of (individual) cognitive motivations beyond (population-level) social factors, as syntax has been identified as a domain with many mixed-users and less social sensitivity to variation (e.g., Labov 2001; Nevalainen, Raumolin-Brunberg, and Mannila 2011). Finally, the study of changing syntactic variation within the domain of clausal verb complementation has led to profound insights from several research angles, including detailed cognitive-functional accounts of verb complementation patterns that shed light on their internal functionality and diffusion (e.g., Rohdenburg 2006;

Cuyckens, D'hoedt and Szmrecsanyi 2014)[1] in the language at large, as well as (some) sociolinguistic accounts (e.g., Nevalainen, Raumolin-Brunberg, and Mannila 2011). However, at present, there are virtually no accounts that consider the emergence and diffusion of verb complementation patterns (or even grammatical variation more generally) by unifying the sociolinguistic focus on population dynamics with the functionalist focus on the semantic or structural factors that constrain and motivate the existence of grammatical variation. We suggest that a better grasp of the cognitive representations of individual language users is a requirement to attain such a unification, especially as recent work in (historical) linguistics has shown that, when there is diachronically unstable variation, there may be non-trivial differences in how individuals believe the variation is conditioned. As such, these studies (e.g., Schmid and Mantlik 2015; Anthonissen 2019; Fonteyn and Nini 2020; Petré and Anthonissen 2020) have shown that it cannot simply be assumed that population-level generalizations over the factors conditioning variation in different time stages are representative of the constraints employed in the cognitive models of each individual in the population.

2 Objectives of the Study

The study is designed to offer a unified cognitive and sociolinguistic account of syntactic variation and change that has so far been missing. Focussing on the system of English verb complementation, it investigates (i) idiolectal variation in the constraints on alternating complementation patterns; (ii) the interaction between the linguistic behaviour of individuals and the (changing) distribution of grammatical variants at the population-level across different time stages. This will be done by describing and measuring the extent of inter-individual variation in verb complementation, and modelling the observed patterns of stability or diffusion in terms of the weakening and strengthening of grammatical and cognitive constraints within a socially homogeneous group of individuals. As an example, consider the complementation patterns following the private factual verbs *believe* and *suppose*, and the public factual verbs *acknowledge* and *declare*, and produced by Daniel Defoe (°1660), William Wake (°1657), and Benjamin Hoadly (°1676). A brief examination of Figure 1 reveals that there appears to be inter-

1 Cuyckens, D'hoedt & Szmrecsanyi (2014), for instance, show that Rohdenburg's (1996) Complexity Principle "needs to qualified, in that not all types of structural complexity have a discouraging effect on nonfinite complementation"; their research also suggests an increase of nonfinite vs. finite complement clauses with selected matrix verbs.

individual variation, with Defoe preferring the NCI construction (*Scotland was acknowledged to be independent*) over the ACI, which is preferred by Wake (*They acknowledged His Majesty to have all that Authority*). Hoadly, in contrast, uses both ACI and NCI relatively commonly at the cost of *that*-complements. It is also noticeable that variation between authors appears to be more consistent than the variation between semantically different clusters, which is harder to interpret.

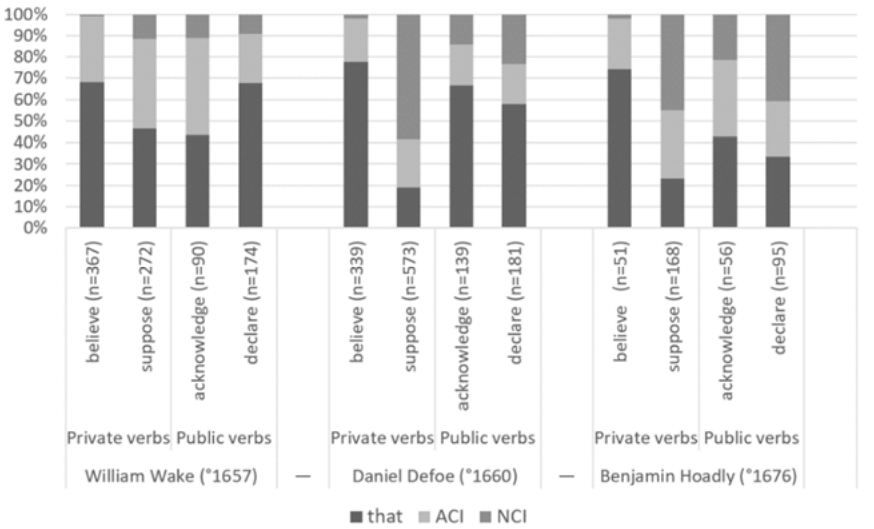

Fig. 1: Distribution of complementation patterns in three authors

In light of such observations, the question arises whether such inter-individual frequency fluctuations can be explained stylistically (e.g., differences in the text genres the authors generally produce and come into contact with), or by the fact that these individuals have constructed a different understanding of how these verb complementation patterns are conditioned. Furthermore, the question arises how the distribution and conditioning of patterns observed in the linguistic output of specific individuals relates to the (changing) distribution and grammatical conditioning of the complementation patterns in the language at large.

To improve our understanding of inter-individual syntactic variation and the micro- and macro-level dynamics of syntactic change, we set out to test the four hypotheses below. While Hypotheses 1–3 are not strictly about change, our main claim is that understanding how individual variation is cognitively constrained

is a prerequisite to understanding language change, as formulated for instance in Hypothesis 4.

- *Hypothesis 1: Constraints on variation in complementation differ between individuals in non-trivial ways*. This hypothesis is in line with recent findings on individual differences in constraints on syntactic structures (Fonteyn and Nini 2020; Standing and Petré 2021).
- *Hypothesis 2: Individuals' constraints are not random but make sense cognitively,* i.e., intra-individual variation cognitively optimizes certain functional divisions of labour between variants. This hypothesis is in line with the expectations of usage-based cognitive-functional linguistics. Cognitive variables include widely attested functional factors, as well as the impact of analogy between semantically similar matrix verbs. The hypothesis does not preclude probabilistic assignment of constraints. The input individuals receive from their social network is diverse, and the factors involved partly overlap, meaning that a particular choice of complement in a specific utterance will be determined by a range of motivations with variable weights. Still, in line with the available evidence (Fonteyn and Nini 2020; Standing and Petré 2021), it may be assumed that individuals will have distinct profiles preferring certain divisions of labour more than others. Also, even while complementation is not expected to be socially indexed, subconscious alignment with one's social network may still be considerable.
- *Hypothesis 3: Constraints cluster around a limited number of types of generalization inter-individually*. In addition to a presumably limited typology of cognitive learning styles, functional constraints, which may have initially been pragmatic and wide-ranging, are narrowed down over time by mutually reinforcing processes of social conventionalization and cognitive entrenchment (Schmid and Mantlik 2015: 620; Schmid 2020), which will limit the range of inter-individual variation.
- *Hypothesis 4: In the long run, simpler, more general constraints will impact the population-level more than more local constraints*. Some strategies of constraining variation will have a bigger impact on the population-level behaviour than others. This hypothesis relates to the findings of Dąbrowska (2008, 2020), which suggest the emergence of broad generalizations that are reflected only in a minority of individuals. The current project will extend this claim by testing it diachronically.

3 Methodology and Design

The study is broken down into five steps, each comprising different tasks contributing to answering the research questions and testing the four hypotheses associated with them.

Step 1: Corpora and selection of individuals. To establish long-term effects of individuals on the population-level system of complementation, we employ data from a representative number of individuals covering a period beyond what is in the purview of connected generations. Obviously, combining longitudinal study (multiple periods), multiple individuals and quantitative methods very quickly leads to an explosion of data. To maintain feasibility, and maximize comparability, we have therefore chosen to restrict the data set to a selection of 40 individuals who meet the following criteria:

- Their active careers roughly cover the period *1700 until 1920*;
- They are clustered around *4 generations* (broadly defined; 10 individuals/generation);
- For each individual there are at least *500,000 words* of digitized text available;
- They are *prolific male authors* who spent the majority of their careers/lives in *London*;
- The number of individuals with and without a university degree is roughly equal.

Data are retrieved from the following sources.

1. The *Early Modern Multiloquent Authors* Corpus (EMMA, Petré et al. 2019). From this corpus we include ten authors, born between 1657–1676, in our sample; each of them meets the above criteria. Together they represent a first generation (broadly defined).
2. ECCO-TCP and Evans-TCP: These are two public domain databases of texts manually transcribed from scans by the Text Creation Partnership. Here again, ten authors, born between 1685–1737, are selected. They serve as a second generation.
3. The Hansard Corpus. Twenty prolific members of parliament active between 1803–1919 who meet the criteria will be represented by their speeches and interventions. They serve as the third and fourth generation.

All texts will be labelled for text form, prototypical text type, and genre (following the procedure described in Petré et al. 2019) to ensure that any comparative claims across generations can be controlled for genre variation.

Step 2: Selection of verbs and data retrieval. From the data set resulting from Step 1, all tokens for a selection of verbs are retrieved, which vary between the complementation patterns given in (1) and (2) (these patterns constitute the dependent variable in the statistical models described below). A set of six verbs is selected from a list of Complement-Taking Predicates (matrix verbs) showing variation between *that*- and *-ing*-clauses (see Quirk et al. 1985; Declerck 1991). The six selected verbs were also in use, with overall stable semantics, in the period under study. Another set of six verbs is selected from those matrix verbs varying between the *that*-clause, the ACI, and the NCI constructions. Each set is divided into two smaller subsets, with each subset containing verbs that are semantically closely related, but unrelated to verbs from the other subset. The verbs in semantic cluster 1 are "private factual verbs", "express[ing] intellectual states" (Quirk et al. 1985: 1181); those in semantic cluster 2 are "public factual verbs", i.e., "speech act verbs introducing indirect statements" (Quirk et al. 1985: 1180). The general setup is summarized in Table 1.

Tab. 1: Case study design

	Semantic cluster 1	**Semantic cluster 2**
Variation set 1 (*that*/*-ing*)	*remember, recall, forget*	*admit, mention, deny*
Variation set 2 (*that*/ACI/NCI)	*believe, suppose, suspect*	*acknowledge, declare, report*

This design allows us to test if closely related verbs are more likely to be constrained in more similar ways than unrelated verbs. Based on the data of two authors, we roughly estimate the total data set to contain about 26,000 data points (capping high-frequency verbs to 200 random instances per author).

Step 3: Analysis of variables. Every instance retrieved is manually coded for seven to eight functional variables (testing Hypothesis 2 & Hypothesis 3): semantic (*Step 3.1*), structural (*Step 3.2*), and discourse-related (*Step 3.3*). These are the independent variables (fixed effects) in our model.

Step 3.1: Semantic variables

– *Meaning of the complement clause* (CC). This variable has two values: the complement clause may denote a state, as in *She had very much regretted*

being from home, or an event/action, as in *This will make him regret he ever married you*.

– *Animacy of the subject of the CC*.

Step 3.2: Structural variables. The choice between finite and nonfinite complementation entails a choice between more or less explicit grammatical options. The finite complement is more explicit in that its subject is always explicitly encoded and the verbal phrase has full tense and modal potential. In this regard, it has been claimed by Rohdenburg (1995: 151) that "the more explicit [option] will tend to be favoured in cognitively more complex environments". Cognitively complex environments are here operationalized in terms of the following structural complexity factors.

– *Intervening material,* measured in terms of the number of words**,** between (i) the Complement-Taking Predicate (CTP) and the *that*- or zero-complementizer, (ii) between the CTP and the gerund, or (iii) between the CTP and the infinitive marker *to*. Not considered intervening material are subjects from the CC that were raised to object position (e.g., *Bible* in (2) *They considered the Bible to be* ...) and precede the *to*-infinitive. For the same reason, negative markers preceding the gerund or the *to*-complementiser are also not considered.
– *Complexity of the CC*. The CC may show the following three types of predicate structure: (i) *[V]* = (intransitive) verbs without argument/modifier; (ii) *[V (arg|mod)]* = verbs with one argument/modifier; (iii) *[V (arg|mod)+]* = verbs with at least argument + modifier, at least two arguments, or at least two modifiers.
– *Negation in the complement clause*.
– *Active or passive voice of the complement clause*.

Step 3.3: Discourse factors

– *Denotation*. Two values are distinguished: (i) *different*: subject of the CTP and CC-subject denote different entities, as in *Do you remember a gang coming there in Oct. 1746*?; (ii) *same*: CTP- and CC-subjects denote the same entities; in this condition, the CC-subject may be controlled by the matrix subject (*Debbenham denied ever knowing Mr. Athill*), or not (*He denied his ever giving an order to the prisoner*).
– *Information status of CC-subject*. This factor is about whether the subject of the complement clause has been previously mentioned in the discourse. We adopt the method developed in Götze et al. (2007), who distinguish between

given-active (mentioned within the preceding 5 clauses), *given-inactive* (mentioned earlier still), *accessible*, and *new* subjects (not yet mentioned). Accessible items are defined on the basis of cultural conventions (*God*) and conventional meronymic relations (e.g., the use of *flowers* in the sentence after a discussion of gardens is deemed accessible).

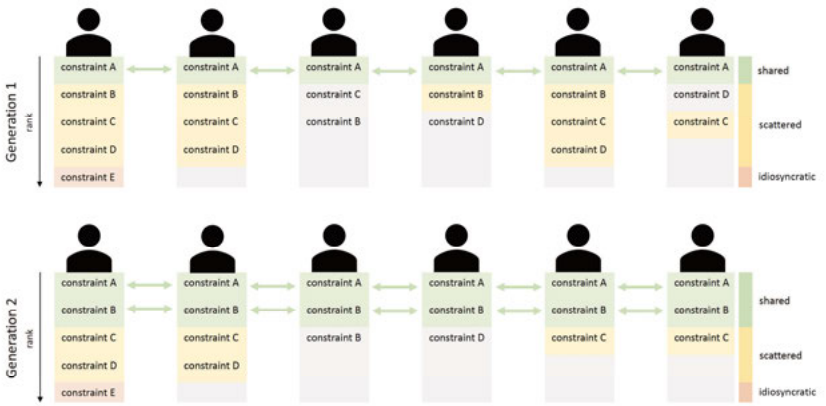

Fig. 2: Fictional example of constraint ranking comparisons

Step 4: Statistical analysis of individual profiles. In order to chart and quantitatively compare the differences in constraint models between individuals (Hypothesis 1), the project's methodological toolkit is extended with multifactorial classification models, which include the Conditional Inference Tree and Random Forest algorithms (e.g., Tagliamonte and Baayen 2012; Szmrecsanyi et al. 2016). An important advantage of these statistical methods is that they are robust even with smaller data sets (Fonteyn and Nini 2020), and that they are able to cope with complete and quasi-separation as well as multicollinearity (cf. Tomaschek, Hendrix, and Baayen 2018). The latter is particularly relevant for cases where multiple constraints may determine the choice of a variant, and reinforce each other: if, for instance, lack of co-reference and the overall complexity of the clause are both valid motivations for opting for a finite complement clause, then it is expected that the two will often overlap and may be collinear to a certain extent. Such high degrees of overlap between the predictor constraints will cause problems in an analysis that relies on traditional multifactorial regression models, as such models ultimately suggest that one of the two variables is entirely superfluous. Such a result is, however, misleading, as it is at odds with the nature

of language, in which redundancy and reinforcement strategies, such as those offering protection against information loss (Van de Velde 2014: 142), are pervasive.

These algorithms, then, will be used to determine which language-internal factors an individual employs to condition or constrain the grammatical variation attested in their linguistic output (as obtained in Steps 2–4). Subsequently, these individualized statistical models are subjected to an elaborate model comparison, involving a systematic comparison of the variable importance or "constraint" rankings between individuals, and at different time stages (Hypothesis 1 & Hypothesis 2). These model and ranking comparisons enable us to trace the degree of conventionalization of linguistic constraints across time (Hypothesis 4), distinguishing between (i) idiosyncratic, (ii) sparse and scattered, or (iii) consistently shared across the individuals in the sample.

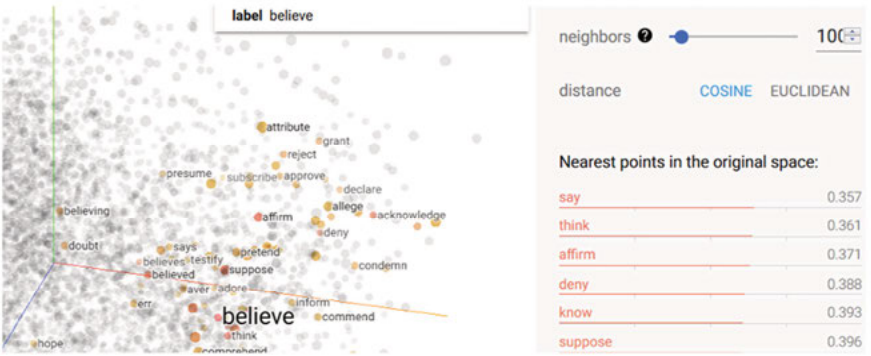

Fig. 3: Near-synonyms of *believe* in 1680–1700 (Using Word2Vec [Mikolov et al. 2013] on an 800m word data set)

Step 5: Charting local semantic clustering versus global generalizations (Hypothesis 3)**.** As stipulated in *Step 2*, manual multi-factorial analysis will be restricted to 12 verbs to maintain feasibility. To compensate this relatively narrow sampling of verbs, the findings will be extrapolated by using deep learning techniques (cf. Figure 3). Specifically, distributional models will be used to identify near-synonyms of the verbs in our selection. With the help of the algorithm *CIRCE* (Verheyen et al. 2019), we can semi-automatically identify the complementation patterns of these near-synonyms, and broadly assess (i) to what extent the hypothesis holds that more closely related verbs have more similar complementation patterns, and (ii) for which individuals.

4 Conclusion

In sum, the main objective of this study is to establish the impact of interindividual differences in cognitive representations on long-term population-level language change. The specific domain of research is variation in the English system of clausal verb complementation. While a certain amount of variation can probably be accounted for by a desire of varying itself, it has been shown that the choice between complement variants is influenced by various functional (i.e., semantic, structural and discourse) factors such as animacy (human or abstract) or complement clause length. At the same time, existing studies have experienced difficulties with robustly accounting for the variation by means of population-level (social) variables only. When social clues are insufficient to determine usage, cognitive mechanisms may come into play that are different between individuals and therefore cannot easily be averaged over. The project outlined here seeks to advance our insight in the functionality of abstract grammatical variation of this kind by putting individual-level analysis centre-stage.

References

Anthonissen, Lynn. 2019. Constructional change across the lifespan: The nominative and infinitive in early modern writers. In Kristin Bech & Ruth Möhlig-Falke (eds.), *Grammar – discourse – context: Grammar and usage in language variation and change*, 125–156. Berlin: De Gruyter Mouton.

Beckner Clay, Richard Blythe, Joan Bybee, Morten Christiansen, William Croft, Nick Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system. *Language Learning* 59. 1–26.

Cuyckens, Hubert, Frauke D'hoedt & Benedikt Szmrecsanyi. 2014. Variability in verb complementation in Late Modern English: Finite vs. non-finite patterns. In Marianne Hundt (ed.), *Late Modern English syntax*, 182–203. Cambridge: Cambridge University Press.

Dąbrowska, Ewa. 2008. The later development of an early-emerging system: The curious case of the Polish genitive. *Linguistics* 46. 629–650.

Dąbrowska, Ewa. 2020. Language as a phenomenon of the third kind. *Cognitive Linguistics* 31(2). 213–229.

Declerck, Renaat. 1991. *A comprehensive descriptive grammar of English*. Tokyo: Kaitakusha.

Denison, David. 1998. Syntax. In Suzanne Romaine (ed.), *The Cambridge history of the English language, Vol.* 4, 92–329. Cambridge: Cambridge University Press.

Fonteyn, Lauren & Andrea Nini. 2020. Individuality in syntactic variation: An investigation of the seventeenth-century gerund alternation. *Cognitive Linguistics* 31(2). 279–308.

Götze, Michael, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas & Ruben Stoel, 2007. Information structure. *Interdisciplinary Studies on Information Structure* 7. 147–187.

Labov, Willam. 2001. *Principles of linguistic change*, Vol. 2: *Social factors*. Oxford: Blackwell.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems,* Vol. 2, 3111–3119. Red Hook, NY: Curran Associates Inc.

Nevalainen, Terttu, H. Raumolin-Brunberg & H. Mannila. 2011. The diffusion of language change in real time. *Language Variation and Change* 23. 1–43.

Petré, Peter & Lynn Anthonissen. 2020. Individuality in complex systems: A constructionist approach. *Cognitive Linguistics* 31(2). 185–212.

Petré, Peter, Lynn Anthonissen, Sara Budts, Enrique Manjavacas, Emma-Louise Silva, William Standing & Odile A. O. Strik. 2019. Early Modern Multiloquent Authors (EMMA): Designing a large-scale corpus of individuals' languages. *ICAME Journal* 43. 83–122.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Rohdenburg, Günter. 1995. On the replacement of finite complement clauses by infinitives in English. *English Studies* 76. 367–388.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182.

Rohdenburg, Günter. 2006. The role of functional constraints in the Evolution of the English complementation system. In Christiane Dalton-Puffer, Dieter Kastovsky & Nikolaus Ritt (eds.), *Syntax, style, and grammatical norms: English from 1500-2000*, 143–166. Bern: Lang.

Rohdenburg, Günter. 2014. On the changing status of *that*-clauses. In Marianne Hundt (ed.), *Late Modern English syntax*, 155-181. Cambridge: Cambridge University Press.

Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford: Oxford University Press.

Schmid, Hans-Jörg & Annette Mantlik. 2015. Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles. *Anglia* 133. 583–623.

Standing, William & Peter Petré. 2021. Exploiting convention: Lifespan change and generational incrementation in the development of cleft constructions. In Isabelle Buchstaller & K. Beaman (eds.), *Language variation and language change across the lifespan*, 141–163. London: Routledge.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37. 109–137.

Tagliamonte, Sali & Harald Baayen. 2012. Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.

Tomaschek, Fabian, Peter Hendrix & Harald Baayen. 2018. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* 71. 249–267.

Van de Velde, Freek. 2014. Degeneracy. In Ronny Boogaart, Timothy Colleman & Gijsbert Rutten (eds.), *Extending the scope of construction grammar*, 141–179. Berlin: De Gruyter Mouton.

Verheyen, Lara, William Standing, Sara Budts & Peter Petré. 2019. Pattern matching or holistic retrieval: Finding bare clefts in unannotated corpus data. Presentation delivered at SLE52 in Leipzig, Germany.