System and use of wh-relativizers in 200 years of English scientific writing

Marie-Pauline Krielke, Stefan Fischer,
Stefania Degaetano-Ortlieb, Elke Teich
(Saarland University)

1. Introduction

We investigate the diachronic development of wh-relativizers in English scientific writing in the late modern period, characterized by an initially richly populated paradigm in the late 17th/early 18th century and a reduction to only a few options by the mid 19th century. To explain this reduction, we take the perspective of rational communication, according to which language users, while striving for successful communication, seek to reduce their effort. Previous work has shown that production effort is directly linked to the number of options at a given choice point (Milin et al. 2009, Linzen and Jaeger 2016). This effort is appropriately indexed by *entropy*. The more options with equal/similar probability, the higher the entropy, i.e. the higher the production effort. Similarly, processing effort is correlated with predictability in context – *surprisal* (Levy 2008). Highly predictable, conventionalized patterns are easier to produce and comprehend than less predictable ones. Assuming that language users strive for ease in communication, diachronically they are likely to (a) develop a preference for *which* options to use and discard others to reduce entropy, and (b) converge on *how* to use those options to reduce surprisal. We test this for the changing use of wh-relativizers in scientific text in the late modern period.

Many scholars have investigated variation in relativizer choice in standard spoken and written varieties (e.g. Guy and Bayley 1995; Biber et al. 1999; Lehmann 2001; Hinrichs et al. 2015), in vernacular speech (e.g. Romaine 1982, Tottie and Harvie 2000; Tagliamonte 2002; Tagliamonte et al. 2005; Levey 2006), and from synchronic and diachronic perspectives (e.g. Romaine 1980; Ball 1996; Hundt et al. 2012; Nevalainen 2012, Nevalainen and Raumolin-Brunberg 2002). While stylistic variability of the different options in written present day English is well known (see Biber et al. 1999; Leech et al. 2009), we know little about the diachronic development of relativizers according to register, e.g. in scientific writing. Also, most research only considers most common relativizers (e.g. *which*, *that*, *zero*) still in use in present day English. Here, we study a more comprehensive set of relativizers across scientific and "general language" (mix of registers) from a diachronic perspective.

Possible paradigmatic change is analyzed by diachronic word embeddings (cf. Fankhauser and Kupietz 2017), allowing us to select items affected by change. Then we assess the change (reduction/expansion) of a paradigm estimating its entropy over time. To check whether changes are specific to scientific language, we compare with uses in general language. Finally, we inspect possible changes in the predictability of selected wh-relativizers involved in paradigmatic change estimating their surprisal over

time, looking for traces of conventionalization (cf. Degaetano-Ortlieb and Teich 2016, 2018).

2. Data and Methods

For scientific writing we use the Royal Society Corpus (RSC v4.0; Kermes et al. 2016), consisting of the *Proceedings and Transactions of the Royal Society of London* spanning 1665-1869 with approx. 32 million tokens, including metadata (e.g. author, publication year) and linguistic annotation (e.g. tokens, lemmas, parts of speech). Tagging accuracy is 95.1% on normalized word forms (based on TreeTagger (Schmid 1994) and VARD (Baron and Rayson 2008)).

For general English we use the Corpus of Late Modern English Texts (CLMET v3.1; Diller et al. 2010), spanning 1710-1920 with approx. 40 million tokens from several genres (e.g. narrative, drama), processed with the same tools (TreeTagger, VARD).

To investigate diachronic change in the wh-relativizer paradigm, we use word embeddings. Calculation is based on the RSC for each decade, allowing explorative diachronic comparison (cf. Fankhauser and Kupietz 2017). From this, clusters of wh-words emerge (Figure 1, blue-green shades encode decreasing, yellow-red shades increasing relative frequency over time). After discarding words that are not relativizers (e.g. *hence*), we consider the following set: *wherewith*, *whereupon*, *whereby*, *whence*, *whereas*, *wherein*, *whereof*, *wherefore*, *wherever*, *whereon*, *whither*, *which*, *whose.*

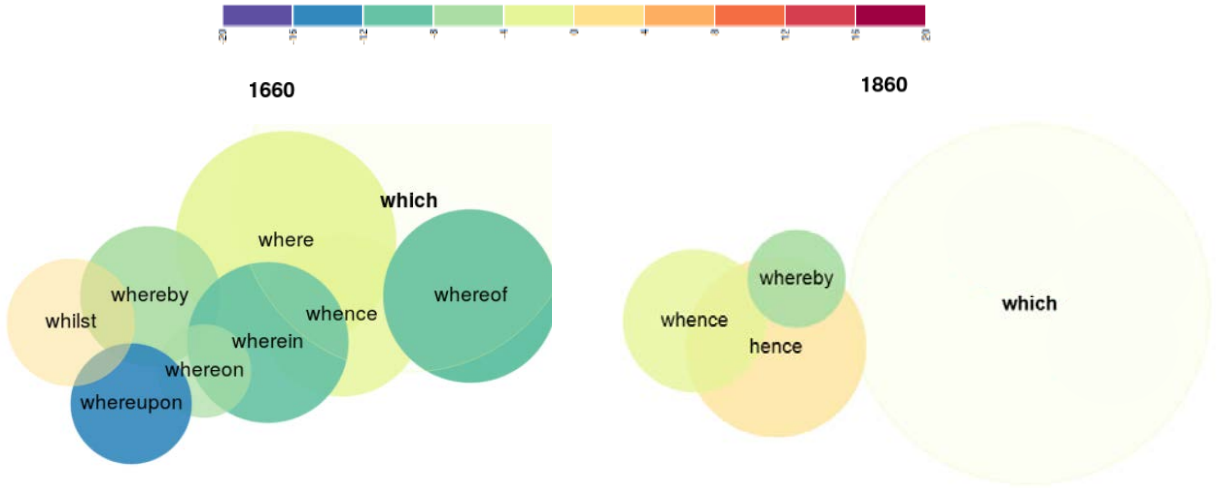

Figure 1: Word embeddings of the RSC in 1660s (1665-1669) vs. 1860s (1861-1869)

To assess a possible paradigm reduction, for each decade, we calculate the entropy (1) of wh-relativizers in scientific texts (RSC) and general language (CLMET) (Section 3.1).

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (1)$$

To investigate syntagmatic contexts of change, we calculate the surprisal (2) of selected wh-relativizers and their contexts (Section 3.2) based on conditional probabilities from a 4-gram language model (three preceding words), considering also the average surprisal per 50-year periods.

$$S = -log_2\, p(w_n|w_{n-3}, w_{n-2}, w_{n-1}) \quad (2)$$

3. Analysis

3.1 Wh-relativizers in scientific and general language

Inspecting the diachronic word embeddings of the RSC across decades, we observe a rather diversified cluster of wh-words in the 1660s and a fairly sparse cluster in 1860s (see Figure 1). Figure 2 shows a relative decline in frequency of the whole wh-cluster in both RSC and CLMET, with wh-relativizers being more frequent in scientific than in general language.

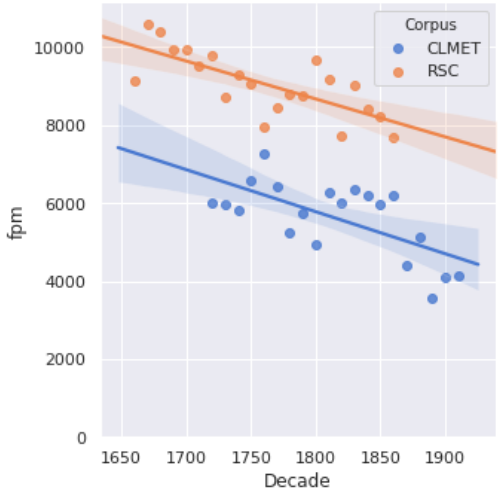

Figure 2: Frequency per million (fpm) of wh-words shown by word embeddings in the RSC and the CLMET

Inspecting frequencies of the wh-words identified as wh-relativizers (cf. Section 2), *which* is most frequent, proportionally increasing over time in scientific writing, while the others gradually fade out (Figure 3). Except for *which*, in general language wh-relativizers are relatively infrequent throughout.

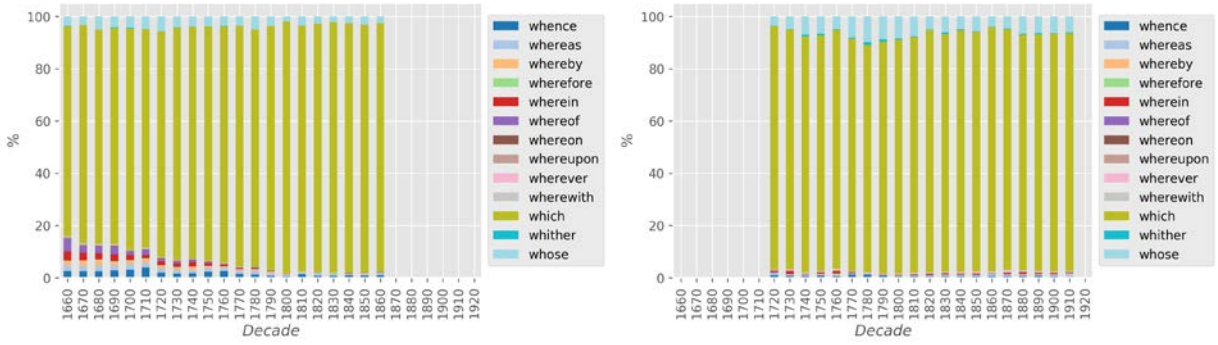

Figure 3: Distribution of wh-relativizers over time (RSC left, CLMET right)

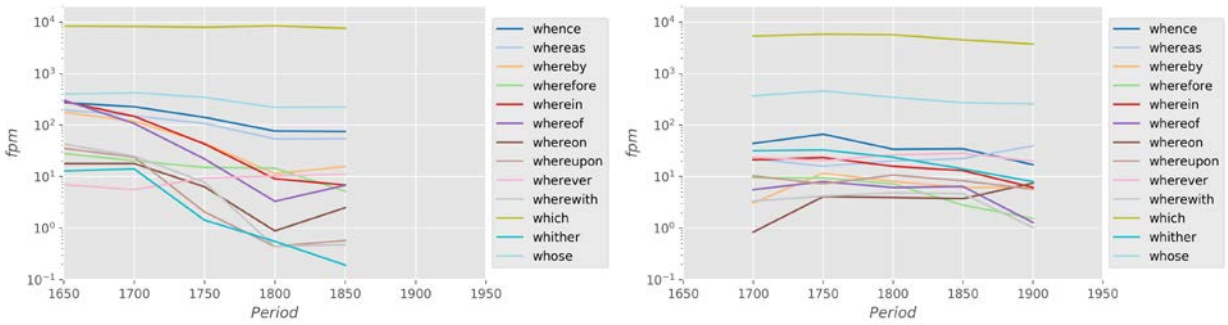

Figure 4: Log scaled frequency per million of wh-relativizers (RSC left, CLMET right)

Except for *which* and *wherever* (slight increase), in the RSC all other options go down in frequency (Figure 4), indicating a reduction of options over time, while frequencies per million in CLMET stay relatively stable.

To assess whether we encounter a true paradigm reduction, we compare the entropy of the wh-relativizers across decades. Figure 5 shows a significant reduction in entropy for scientific writing diachronically, while for general language entropy is fairly stable, i.e. the reduction of the wh-paradigm and the changing use of relativizers is clearly specific to scientific writing.

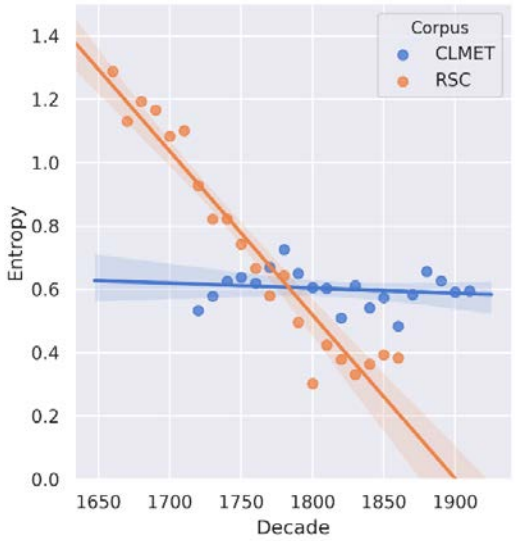

Figure 5: Entropy of the wh-paradigm in the RSC and the CLMET

3.2 Syntagmatic environments of wh-relativizers in scientific writing

Further, we investigate whether the syntagmatic environments of wh-relativizers also change, possibly resulting in conventionalized usage in scientific writing. We analyze the average surprisal of wh-relativizers over time (here: 50-year periods), inspecting whether particular options become favored and if so, in which contexts.

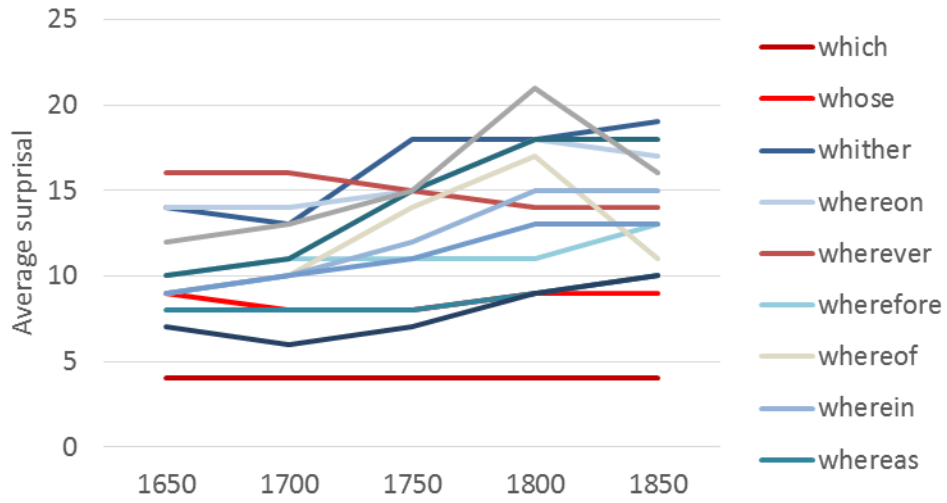

Figure 6: Average surprisal of wh-relativizers measured on periods of 50 years

Increasing surprisal may indicate a linguistic option on its way out, decreasing or stable surprisal indicates that an option is becoming a preferred choice. Figure 6 shows an increase in average surprisal for all but *which*, *whose*, and *wherever*. The wh-relativizers with the highest rates of decline in frequency and highest increase in surprisal are notably pronominal adverbs (i.e. wh-word and a preposition as in *whereof*). The survival of *which* and the decline of the pronominal adverbs suggest a potential replacement by more analytic structures such as preposition+*which*, e.g. *by which*, *of which*, typically introducing relative clauses with adverbial gaps (Biber et al. 1999: 624). Considering the most frequent context preceding *which* (based on part-of-speech 3-grams), Figure 7 shows a remarkable increase of the [DT-NN-IN] 3-gram (noun phrase with preposition) in frequency over time. Accordingly, surprisal for *which* after [DT-NN-IN] decreases (Figure 8), becoming thus more predictable and indicating a conventionalized use.

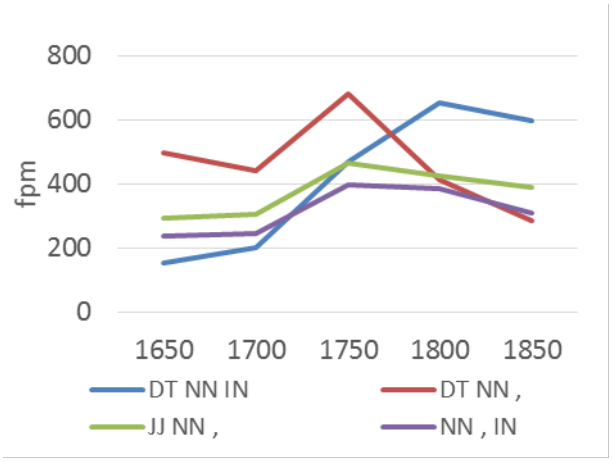

Figure 7: Most frequent PoS 3-grams preceding *which*

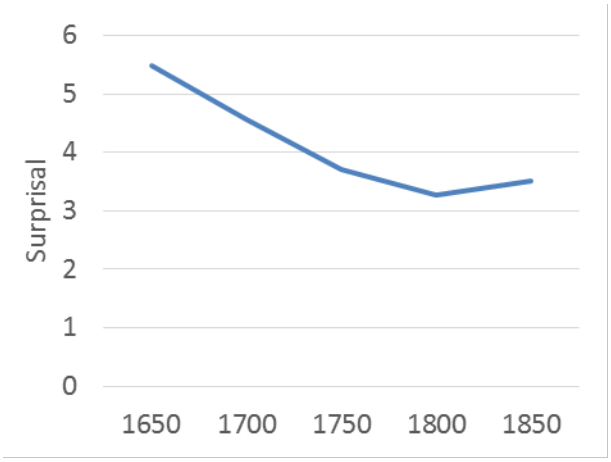

Figure 8: Average surprisal of *which* after [DT-NN-IN]

While preposition+*which* (e.g. *of which*) is not a new phenomenon, diachronically it increases in frequency by around 30% (Figure 9). Slightly decreasing surprisal values for *which* after a preposition indicate higher predictability of the construction over time (Figure 10).

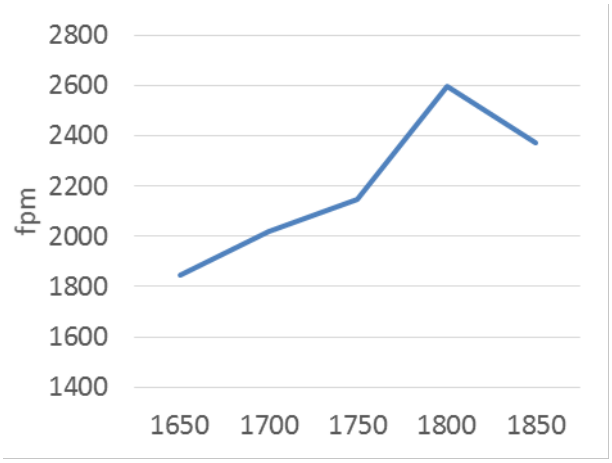

Figure 9: Prepositions preceding *which*

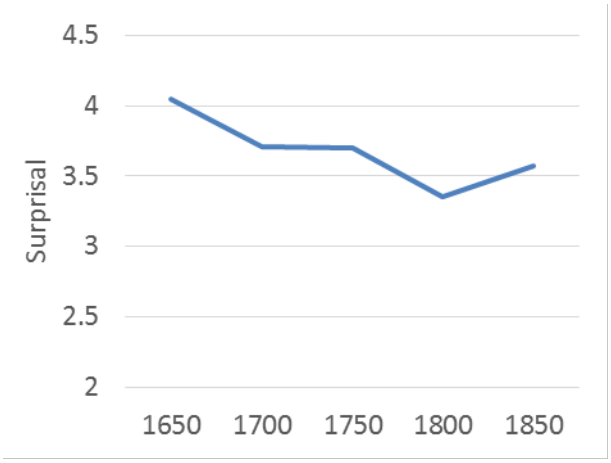

Figure 10: Average surprisal of *which* after preposition

period	freq pM	freq raw	3-gram	type
1650	24.00	62	, out of	quantification
	13.16	43	the first of	quantification
	11.23	34	, some of	quantification
1700	21.96	75	, some of	quantification
	17.86	61	, one of	quantification
	12.01	41	, out of	quantification
1750	22.86	145	the manner in	manner
	13.72	87	, one of	quantification
	13.56	86	, some of	quantification
1800	16.46	150	the mode in	manner
	15.80	144	, one of	quantification
	13.61	124	by means of	manner
1850	25.24	265	the manner in	manner
	14.48	152	, each of	quantification
	13.91	146	, one of	quantification

Table 1: Top three lexical 3-grams preceding *which*

Considering the top three lexical 3-grams with a preposition preceding *which* (Table 1), we see expressions of quantification and manner, the latter increasing over time (Figure 11). Thus, increase in prepositions preceding *which* is not attributable to an increasingly analytic way of combining prepositions and wh-relativizers (*e.g. by which*) but rather to a gradual increase in the use of relative clauses with adverbial gaps describing *manner*. This development is possibly driven by the absence of a relative adverb for manner adverbials in English and the increasing need for a condensed way to express manner in

scientific writing. In fact, Biber et al. (1999: 629) show for contemporary English that in scientific writing, with a preference for preposition+*which*, a manner adverbial gap is most commonly marked by *in which*. The steep increase for manner+*which* (Figure 11) is a significant hint towards a diachronic development filling a semantic gap in scientific discourse.

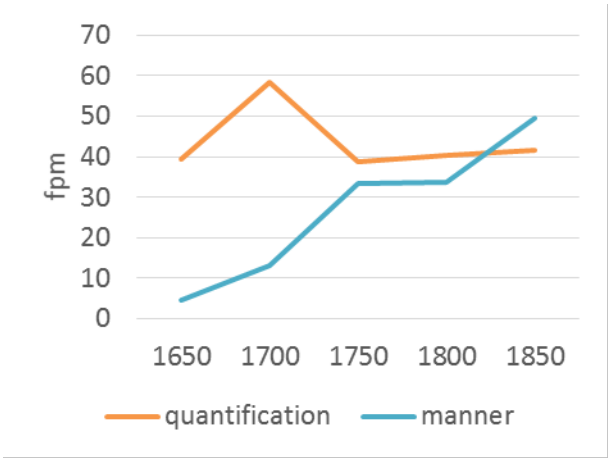

Figure 11: Expressions of quantification and manner with preposition preceding *which*

4. Summary

We found a diachronic change in the paradigm of wh-relativizers in scientific writing from many options to fewer options with a more restricted use. Using entropy, we compared the development of the paradigm with general language where no such change occurs. Inspecting surprisal (i.e. predictability in context) of the wh-relativizers in scientific writing, we found that *which* and *whose* have stable surprisal over time, while the other wh-relativizers become increasingly less predictable. Our prior assumption, based on the survival of *which* and the decline of other wh-relativizers, that wh-relativizers have moved from synthetic to analytic formation has not been confirmed. Instead, we have observed an upward trend in the use of preposition+*which* for adverbial gaps, particularly for the expression of manner. The reduction of terms in the wh-paradigm lowers entropy and the convergence on particular usages of the remaining options lowers surprisal, thus making communication in the scientific domain more efficient for both producers and comprehenders. In future work, we will inspect potentially correlating expansions of paradigms, e.g. in the case of wh-relativizers this might be a rise in adverbial types as a substitute for relative clauses as part of a more comprehensive transformation of grammatical usage in scientific English (cf. Halliday and Martin (1993)).

Bibliography

Ball, C. N. (1996). A diachronic study of relative markers in spoken and written English. *Language Variation and Change* 8(2). 227–258. doi:10.1017/S0954394500001150.

Baron, A. & Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham, UK.

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

Degaetano-Ortlieb, S. and 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. and 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

Diller, H.-J., DeSmet, H. & Tyrkkö, J. (2010). A European database of descriptors of English electronic texts. *The European English Messenger*, *19*(2), 29–35.

Fankhauser, P. & Kupietz, M. (2017). Visualizing language change in a corpus of contemporary German. In *Proceedings of the 9th International Corpus Linguistics Conference,* University of Birmingham. Birmingham, UK.

Guy, G. R. & Bayley, R. (1995). On the choice of relative pronouns in English. *American Speech* 70(2), 148–162. doi:10.2307/455813

Halliday, M.A.K. & Martin, J.R. (1993). *Writing Science: Literacy and Discursive Power.* London: The Falmer Press.

Hinrichs, L., Szmrecsanyi, B. & Bohmann, A. (2015). Which-hunting and the Standard English relative clause. *Language* 91(4), 806–836

Hundt, M., Denison, D. & Schneider, G. (2012). Relative complexity in scientific discourse. *English Language and Linguistics* 16(02), 209–240. doi:10.1017/S1360674312000032.

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 Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Leech, G. N., Hundt, M., Mair, C. & Smith, N. (2009). *Change in contemporary English: A grammatical study*. (Studies in English Language). Cambridge, UK and New York: Cambridge University Press.

Lehmann, H. M. (2001). Zero subject relative constructions in American and British English. *Language and Computers* 36(1), 163–177.

Levey, S. (2006). Visiting London relatives. *English World-Wide* 27(1), 45–70. doi:10.1075/eww.27.1.04lev.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177.

Linzen, T. & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6), 1382–1411.

Milin, P., Kuperman, V., Kostic, A. & Baayen, R. H. (2009). *Analogy in Grammar: Form and Acquisition*, Chapter Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation, 214–252. Oxford: Oxford University Press.

Nevalainen, T. (2012). Reconstructing syntactic continuity and change in Early Modern English regional dialects: The case of who. In David Denison, Ricardo Bermúdez-Otero, Chris McCully & Emma Moore (Eds.), *Analyzing Older English*, 159–184. Cambridge: Cambridge University Press.

Nevalainen, T. & Raumolin-Brunberg, H. (2002). The rise of relative who in early Modern English. In P. Poussa (Ed.), *Relativisation on the North Sea Littoral*, 109–121. Munich: Lincom Europa.

Romaine, S. (1980). The relative clause marker in Scots English: Diffusion, complexity, and style as dimensions of syntactic change. *Language in Society* 9(02), 221–247. doi: 10.1017/S004740450000806X.

Romaine, S. (1982). *Sociolinguistic Variation in Speech Communities*. London: Edward Arnold.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Tagliamonte, S. (2002). Variation and change in the British relative marker. In Patricia Poussa (ed.), *Relativisation on the North Sea Littoral*, 147–165. Munich: Lincom Europa.

Tagliamonte, S., Smith, J. & Lawrence, H. (2005). No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change* 17(1). 75–112. doi:10.1017/S0954394505050040.

Tottie, G. & Harvie, D. (2000). It's all relative: Relativization strategies in early African American Vernacular English. In Shana Poplack (Ed.), *The English History of African American English*, 198–230. Oxford: Blackwell.