Challenges of parsing a historical corpus of Scientific English

Tom S Juzek[1], Stefan Fischer[1], Pauline Krielke[1], Stefania Degaetano-Ortlieb[1], and Elke Teich[1]

[1]Saarland University

In this contribution, we outline our experiences with syntactically parsing a diachronic historical corpus. We report on how errors like OCR inaccuracies, end-of-sentence inaccuracies, etc. propagate bottom-up and how we approach such errors by building on existing machine learning approaches for error correction. The Royal Society Corpus (*RSC*; Kermes et al. 2016) is a collection of scientific text from 1665 to 1869 and contains ca. 10 000 documents and 30 million tokens. Using the *RSC*, we wish to describe and model how syntactic complexity changes as Scientific English of the late modern period develops. Our focus is on how common measures of syntactic complexity, e.g. length in tokens, embedding depth, and number of dependants, relate to estimates of information content. Our hypothesis is that Scientific English develops towards the use of shorter sentences with fewer clausal embeddings and increasingly complex noun phrases over time, in order to accommodate an expansion on the lexical level.

To test this hypothesis, high-quality annotations are needed on a syntactic level. We use the Stanford Parser with off-the-shelf settings (Klein and Manning 2003) to parse the *RSC*. This is done after normalising orthography, for which we use VARD (Baron and Rayson 2008). For high-accuracy part-of-speech tagging, i.e. above 95% accuracy for contemporary English, we use the parser's default POS tagger (Manning 2011). Such a set-up would typically yield a parsing accuracy rate of about 82% to 87% for contemporary English (cf. Rehbein and Ruppenhofer 2018). However, it has become clear early on that sequences from the *RSC* are parsed at a considerably lower accuracy rate. We sampled and evaluated 100 parses and observe that sequences from 1665 are parsed at an accuracy rate of 40.7%. Sequences from 1850 only slightly improve to 45.3% (these numbers are in parts lowered by very strict evaluation criteria).

A deeper analysis reveals that a good deal of errors are true parsing errors. However, a considerable number of parses come out wrong, because previous errors propagate bottom-up. That is, OCR errors, spelling variants, inaccurate end-of-sentence detection, and inaccurate POS tagging impair correct parses. See Figure 1 for details. It is likely that further error categories will surface as the sample is extended.

Correcting these errors will need its own procedures and tools. For the earlier sequences from 1665, genuine parsing errors are the main issue. To improve their parses, we will have to enrich the parser, by training it with input from that period. We are currently exploring semi-automatic, integrative approaches with human supervision for this training. Concretely, we are considering an approach put forward by Rehbein and Ruppenhofer (2017, 2018), which combines information-theoretic measures, like entropy, for error detection with semi-supervised machine learning approaches, notably active learning, for error correction. For sequences from 1850, errors propagating from pre-parsing steps are the main cause of inaccurate parses. Here, we are exploring neural-networks for a better end-of-sentence detection and a noisy channel model for OCR corrections (cf. [withheld]).

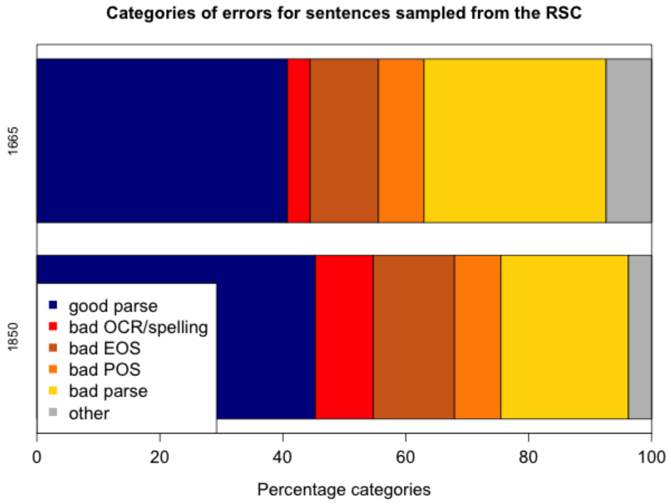

Figure 1: An overview over the error categories for the parse trees sampled from the Royal Society Corpus: Evaluations from 1665 vs evaluations from 1850.

References

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

Kermes, Hannah et al. (2016). “The Royal Society Corpus: From Uncharted Data to Corpus”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia: European Language Resources Association (ELRA).

Klein, Dan and Christopher D. Manning (2003). “Accurate Unlexicalized Parsing”. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. ACL ’03. Sapporo, Japan: Association for Computational Linguistics, pp. 423–430.

Manning, Christopher D. (2011). "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander F. Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 171–189.

Rehbein, Ines and Josef Ruppenhofer (2017). "Detecting annotation noise in automatically labelled data". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1160–1170.

— (2018). "Sprucing up the trees – Error detection in treebanks". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 107–118.