Enhancing the linguistic discovery potential of historical corpora: A twin-track approach using ARCHER

Introduction

ARCHER (A Representative Corpus of Historical English Registers; Biber et al., 1994) is a major resource for studying register variation and diachronic change in British and American English, from the 17th to the 20th centuries. It is by no means the only corpus in the field: temporally, it overlaps in part with multi-register corpora such as COHA,[1] CLMET3.0,[2] Penn Parsed Corpus of Modern British English,[3] A Corpus of Late Modern British and American English Prose (COLMOBAENG),[4] in addition to register-specific corpora, notably TIME (magazine corpus),[5] the Hansard Corpus,[6] and the Old Bailey Corpus.[7] However, ARCHER fills a unique position in that it comprises a wide and balanced range of historical registers (advertising, drama, fiction, sermons, journals, legal, medicine, news, early prose, science, letters, diaries), across two national varieties of English (American, British) and a substantial time depth (four centuries of modern English). In line with current practice, the corpus is also enriched with extensive structural markup, using XML. In this paper we describe how we implemented a series of enhancements to ARCHER through multiple layers of annotation, which vastly increase its potential for linguistic exploration. We provide two case studies, based on alternative annotation schemes and search interfaces, that demonstrate the benefits of using these annotations in comparison to the corpus in its raw, unannotated state.

Methods

A recurrent issue with historical texts for linguistic analysis and Natural Language Processing (NLP) tools is that, until around 1850, spelling conventions in English were not well established. Our first step in enhancing the corpus, therefore, was to normalise spellings, to permit easier retrieval of variant linguistic forms from earlier periods. For this purpose, we used the VARiant Detector software (VARD; Baron and Rayson, 2008) which applies a combined set of methods (Levenshtein edit distance, phonetic matching, known variants, frequency information, spelling rules and adapting weights based on training corpora) in order to detect potential historical spelling variants and match them to Present-Day English (PDE) equivalents. For ARCHER, the VARD tool was manually trained on smaller corpus samples and then run automatically over the remaining text (Schneider et al., 2017). VARD retains the original spellings in the corpus and inserts PDE equivalents alongside them, and both versions will be available to corpus users. VARD achieves above 94% precision and a relative error reduction rate of above 60%, thus increasing the percentage of words that are consistent with PDE spelling from 97.1% to 99.1% in the ARCHER corpus. VARD also outperforms letter-based statistical machine translation approaches (Schneider et al., 2017).

Normalisation reduces the error rate of the subsequent POS-tagging step by about 50% (Rayson et al., 2007). Schneider et al. (2016) show that tagging accuracy on normalised texts from ARCHER is between 88% (for the 17th century) and 92-95% (for PDE). The normalised-spelling version of ARCHER provides the input for two parallel annotation pipelines, one using annotations developed at Zurich, the other annotations developed at Lancaster. Each scheme has its own strengths, and by offering both, corpus users have more scope and flexibility in conducting linguistic enquiries.

The Zurich pipeline (Lehmann & Schneider, 2012) has two main stages: automatic POS-tagging by TreeTagger (Schmid, 1994) and dependency parsing (Schneider, 2008; Schneider, 2012). The web-based interface allows both regular

expression queries (on the raw or the tagged text), and syntactic dependency queries. The PDE spelling results are shown by default. The original spelling, as well as the full XML and the syntactic analysis, are shown by clicking on individual hits. The system includes many tabulation functions, such as summation over types (lemmas, POS tags, etc.) and metadata (such as registers, periods).

The Lancaster pipeline contains five stages. First, morphosyntactic annotation is applied using CLAWS, a hybrid probabilistic and rule-based word class tagger (Garside & Smith, 1997), with output in the C7 tagset. We then use the Template Tagger, which is a template-based patching tool (Fligelstone et al., 1997) to introduce corrections to the C7 tags (thereby gaining 1-2% in accuracy), in addition to output in a more refined tagset known as C8 (see Leech et al., 2009). C8 tags make some linguistically useful distinctions not otherwise available in C7, including auxiliary vs. lexical uses of verbs BE, DO and HAVE, and relativizer vs. interrogative uses of pronouns *which*, *who, whose* and *whom*. Two further layers of annotation are added by the UCREL Semantic Analysis System tagger (USAS; Rayson et al., 2004). Lemmas (dictionary head words) are added using manually-coded rules developed by Beale (1987). Subsequently, coarse-grained semantic tags are assigned using a large knowledge base (two dictionaries) and six disambiguation methods. Finally, the combined set of annotations is formatted according to the formal requirements of CQPweb (see Hardie 2012), a web-based interface which offers access to the corpus via its powerful Corpus Query Processor.[8]

Results

For the purpose of demonstration, we offer two very brief case studies, one with the Zurich and one with the Lancaster pipeline. Each study explores the usability of the respective annotation scheme and query interface for pilot studies.

On the Zurich system, the pilot study reports the development of pre- and postmodification of NPs. Figure 1 shows a cross-tabulation of premodification (as a proportion of pre- and postmodification) by period and register. Particularly striking, and highly significant statistically, are the increases in medicine and science, where the figures double (cf. Leech et al. 2009) while other registers remain at similar levels, or increase less. The results substantially tally with Biber & Clark's (2002) claims that specialised registers have seen a shift away from postmodification towards premodification: in other words, from a more verbose to a more compact type of NP modification.

Crosstabulation of Query 'r1=ncmod1 ' according to categories 'text_type' and 'halfct'. Relative Frequency.

Tabulate: halfct Go! Crosstabulate: X text_type / Y halfct show Count of Instances Go! New Query Go!

halfct/text_type	Advertising	Diary	Drama	Earlyprose	Fiction	Journal	Legal	Letters	Medicine	News_periodicals	Science	Sermons	total
1600	0	0	798.362	682.073	0	619.247	688.44	0	0	0	0	511.159	704.908
1650	0	1091.711	653.3	0	584.671	921.804	641.037	861.373	814.875	826.772	807.665	645.299	772.326
1700	0	986.94	610.867	0	594.076	825.851	736.923	808.359	880.831	936.873	846.816	632.364	761.206
1750	1346.771	1218.352	727.827	0	657.36	1024.805	719.129	904.622	767.812	948.387	834.003	712.333	867.303
1800	1414.069	1174.472	681.645	0	683.784	920.495	686.159	825.23	907.069	1007.928	998.684	710.867	876.477
1850	1697.918	1089.102	716.64	0	692.302	1037.226	830.421	878.543	976.517	1197.283	1062.773	784.388	988.321
1900	1993.144	1242.746	675.952	0	752.109	1070.389	912.874	885.47	1101.002	1338.303	1318.611	784.481	1071.324
1950	2362.935	1000.847	707.244	0	775.401	1123.451	1200.251	900.945	1922.213	1633.556	1701.195	727.62	1258.962
total	1820.202	1128.602	698.149	682.073	694.212	1004.321	819.262	868.676	1093.543	1157.881	1131.632	707.522	963.634

Figure 1: Cross-tabulation of NP premodification by period and register

The case study on the progressive (e.g. *I was saying, she will be arriving*) explores the extent to which CQPweb queries using automated annotations echo previous reports – based on close manual analysis – of dramatic increases of the construction in the modern period (cf. Hundt, 2004; Smitterberg, 2005; Leech et al. 2009). Our results confirm such findings: in all registers, in both national varieties, there is an overall frequency increase between the earliest and latest sampling periods. The progressive is more prevalent and more consistently growing in registers at the more

'oral'/popular pole of the stylistic spectrum (e.g. drama, journals, diaries, letters,; cf. Biber & Finegan, 1997) than at the literate/specialised pole (science, medicine, law); cf. Figure 1.

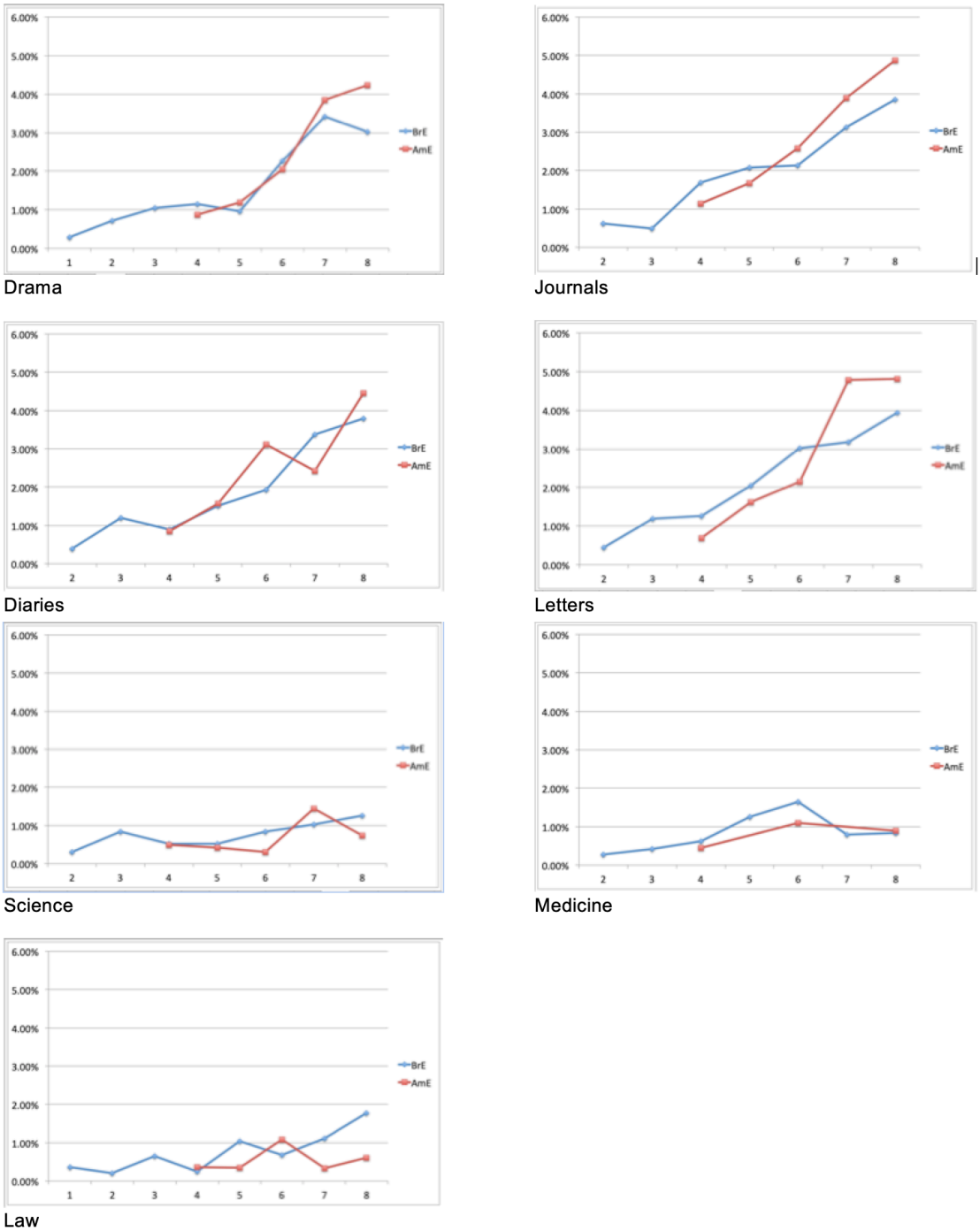

Figure 2: Progressives in ARCHER as a proportion of finite verb phrases: four popular and three specialised registers

In addition to consolidation of the construction within the English aspectual system (Denison 1998), this expansion may well be related to the 'situational immediacy' (Smitterberg 2005) of the progressive, conveying greater vividness and involvement with one's interlocutor. We find that many uses in the popular registers are in present tense, with first person subject and a verb conveying affect or cognition, cf. (1):

1) A. You terrify me.
   B. I*'m seeing* what makes you unhappy
   (1938crot_d7a, Drama, American English, 1900-49)

Using the USAS semantic tags, it is possible to explore the classes of lexical verb that feature most prominently in the progressive across periods and registers.

In our presentation we will expand on the findings, and discuss comparative results from the normalised, annotated flavours of ARCHER against the 'raw', unannotated original corpus.

References

Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, 22 May 2008.

Beale, A.D. (1987). Towards a Distributional Lexicon. In: R. Garside, G. Leech and G. Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman, pp. 149 - 162.

Biber, D. and E. Finegan. 1997. 'Diachronic relations among speech-based and written registers in English.' In: T. Nevalainen and L. Kahlas-Tarkka (eds.). *To Explain the Present. Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Mémoires de la Société Néophilologique de Helsinki, pp. 253-75.

Biber, D., Finegan, E., Atkinson, D. (1994). ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers. In: Fries, U., Tottie, G. & Schneider, P. (eds.) *Creating and Using English Language Corpora*. Amsterdam: Rodopi, pp.1-14.

Biber, D. & Clark, V. (2002). Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb? In Teresa Fanego, Maria Jose Lopez-Couso, and Javier Perez-Guerra (eds.), *English historical syntax and morphology*. Amsterdam: John Benjamins, pp. 43-66.

Denison, D. (1998). 'Syntax.' In: Suzanne Romaine (ed). *The Cambridge History of the English Language*. Vol. IV: *1776-1997*. Cambridge: Cambridge University Press, pp. 92-329.

Fligelstone, S., Pacey, M., & Rayson, P. (1997). How to generalise the task of annotation. In. R. Garside, G. Leech, and A. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 122-136.

Garside, R., & Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 102-121.

Hardie, A. (2012) CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), pp. 380-409.

Hundt, M. (2004). Animacy, agentivity, and the spread of the progressive in modern English. *English Language and Linguistics* 8(1), pp. 47-69.

Leech, G., Hundt, M., Mair, C. & Smith, N.. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Lehmann, H. M. & Schneider, G. (2012). BNC Dependency Bank 1.0. *Studies in Variation, Contacts and Change in English*, Volume 12: Aspects of corpus linguistics: compilation, annotation, analysis. Helsinki: VARIENG.

Rayson, P., Archer, D., Piao, S.L., & McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.

Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the bard: Evaluating the accuracy of a modern POS tagger on early Modern English corpora. In *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.

Schneider, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, University of Zurich.

Schneider, G. (2012). Adapting a parser to historical English. Studies in Variation, Contacts and Change in English, Volume 10*: Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Helsinki: VARIENG.

Schneider, G., Hundt, M., & Oppliger, R. (2016). Part-Of-Speech in Historical Corpora: Tagger Evaluation and Ensemble Systems on ARCHER. *Proceedings of KONVENS,* Bochum, Germany.

Schneider, G., Pettersson, E., & Percillier, M. (2017). Comparing Rule-based and SMT-based Spelling Normalisation for English Historical Texts. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pp. 40-46.

Smitterberg, Erik. (2005). *The Progressive in 19th-Century English: A Process of Integration*. Amsterdam and New York: Rodopi.

[1] https://corpus.byu.edu/coha/
[2] https://perswww.kuleuven.be/~u0044428/
[3] https://www.ling.upenn.edu/histcorpora/PPCMBE2-RELEASE-1/index.html
[4] http://www.helsinki.fi/varieng/CoRD/corpora/COLMOBAENG/index.html
[5] http://corpus.byu.edu/time
[6] https://www.clarin.ac.uk/hansard-corpus
[7] http://www1.uni-giessen.de/oldbaileycorpus/

[8] http://cwb.sourceforge.net/