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, CLMET3.0, PennParsed Corpus of Modern British English, A Corpus of Late Modern British and American English Prose (COLMOBAENG), in addition to register-specific corpora, notably TIME (magazine corpus), the Hansard Corpus, and the Old Bailey Corpus. 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 (Garsside & 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).

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.

<table>
<thead>
<tr>
<th>Table:</th>
<th>NPs</th>
<th>Pre</th>
<th>Post</th>
<th>Pre/Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register</td>
<td>Total</td>
<td>Total</td>
<td>Total</td>
<td>Total</td>
</tr>
<tr>
<td>Advertising</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Diary</td>
<td>1091</td>
<td>778</td>
<td>313</td>
<td>0.4</td>
</tr>
<tr>
<td>Drama</td>
<td>602</td>
<td>473</td>
<td>129</td>
<td>0.2</td>
</tr>
<tr>
<td>Enquiry</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fiction</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Journal</td>
<td>619</td>
<td>427</td>
<td>192</td>
<td>0.3</td>
</tr>
<tr>
<td>Legal</td>
<td>698</td>
<td>444</td>
<td>254</td>
<td>0.4</td>
</tr>
<tr>
<td>Lenses</td>
<td>814</td>
<td>577</td>
<td>237</td>
<td>0.3</td>
</tr>
<tr>
<td>Medicine</td>
<td>836</td>
<td>577</td>
<td>259</td>
<td>0.3</td>
</tr>
<tr>
<td>News:Periodicals</td>
<td>867</td>
<td>645</td>
<td>222</td>
<td>0.3</td>
</tr>
<tr>
<td>Science</td>
<td>111</td>
<td>119</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>788</td>
<td>596</td>
<td>192</td>
<td>0.3</td>
</tr>
</tbody>
</table>

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.

![Graphs showing progressive verb usage in different registers](image)

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


---

1. [https://corpus.byu.edu/coha/](https://corpus.byu.edu/coha/)
2. [https://perswww.kuleuven.be/~u0044428/](https://perswww.kuleuven.be/~u0044428/)
3. [https://www.ling.upenn.edu/histcorpora/PPCMBE2-RELEASE-1/index.html](https://www.ling.upenn.edu/histcorpora/PPCMBE2-RELEASE-1/index.html)
5. [http://corpus.byu.edu/time](http://corpus.byu.edu/time)
6. [https://www.clarin.ac.uk/hansard-corpus](https://www.clarin.ac.uk/hansard-corpus)
7. [http://www1.uni-giessen.de/oldbaileycorpus/](http://www1.uni-giessen.de/oldbaileycorpus/)
http://cwb.sourceforge.net/