Automatic normalization of historical English for neologism detection

Mika Hämäläinen | University of Helsinki
Tanja Säily | University of Helsinki
Eetu Mäkelä | University of Helsinki

Previous studies of historical neologisms in English have mostly been limited to lexicographic resources, such as the *Oxford English Dictionary* (OED; Nevalainen 1999), which exhibits a bias towards well-known texts (Hoffmann 2004; Brewer 2007). Corpus-based studies, on the other hand, have tended to focus on individual affixes or present-day data (Palmer 2015; Renouf 2007).

We base our study on the *Corpora of Early English Correspondence* (CEEC, 1400–1800), which consist of letters written by authors from various social backgrounds. A pilot study of neologisms in the CEEC (Säily et al. in press) identifies a need for normalization strategies beyond the existing tools: there is a great deal of spelling variation in the corpus, and our quantitative approach of automatically mapping each word in the corpus to the OED and contemporary published texts requires even the most infrequent words to be normalized.

Several character-based machine translation models have been proposed to solve the normalization problem (Pettersson et al. 2013; Samardzic et al. 2015). Previous research conducted on the CEEC shows that a character-based neural machine translation (NMT) approach is the single most accurate method (Hämäläinen et al. 2018). The idea of such a translation technique is that instead of translating full sentences consisting of words, the system learns to translate from character to character within individual words. Essentially, the NMT model will learn to map the letters of a word following a historical spelling to the letters in the modern spelling variant.

Our findings suggest that adding more annotations, e.g. century, social metadata or pronunciation, does not improve accuracy. However, using a dictionary to filter the top 10 normalizations produced by the NMT model together with a lemmatizer does improve the results.

We provide insight into the accuracy of different NMT normalization models and an initial qualitative analysis of the newly found neologisms with the more complete normalization of the corpus. Our results will benefit future efforts to normalize historical corpora, and they will also provide new evidence of the extent to which different social groups have engaged in neologizing in the history of English.

Brewer, C. 2007. *Treasure-house of the language: The living OED*. New Haven: Yale University Press. | **CEEC**. *Corpora of Early English Correspondence*. Compiled by T. Nevalainen et al. at the Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/. | **Hämäläinen**, M., T. Säily, J. Rueter, J. Tiedemann & E. Mäkelä. 2018. Normalizing early English letters to present-day English spelling. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (ACL Anthology W18-45), 87–96. Stroudsburg, PA: ACL. | **Hoffmann**, S. 2004. Using the OED quotations database as a corpus – a linguistic appraisal. *ICAME Journal* 28. 17–30. | **Nevalainen**, T. 1999. Early Modern English lexis and semantics. In R. Lass (ed.), *The Cambridge history of the English language, III: 1476–1776*, 332–458. Cambridge: CUP. | **OED**. *Oxford English Dictionary*. OED Online. OUP. http://www.oed.com. |**Palmer**, C. C. 2015. Measuring productivity diachronically: Nominal suffixes in English letters, 1400–1600. *English Language and Linguistics* 19(1). 107–129. | **Pettersson**, E., B. Megyesi & J. Tiedemann. 2013. An SMT approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA*, 54–69. Linköping: Linköping University Electronic Press. | **Renouf**, A.. 2007. Tracing lexical productivity and creativity in the British media: "The chavs and the chav-nots." In J. Munat (ed.), *Lexical creativity, texts and contexts*, 61–89. Amsterdam: Benjamins. |**Säily**, T., E. Mäkelä & M. Hämäläinen. In press. Explorations into the social contexts of neologism use in early English correspondence. *Pragmatics & Cognition* 25(1). | **Samardzic**, T., Y. Scherrer & E. Glaser, 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference*. https://archive-ouverte.unige.ch/unige:82397.