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.


The English comparative alternation revisited:
A fresh look at theory and data

Stefan Hartmann | Universität Bamberg
Lukas Sönning | Universität Bamberg

English offers two strategies for forming the comparative degree of adjectives: inflection (suffixation by -er) and periphrasis (with the degree adverb more). Past research has documented a considerable number of adjectives that allow both strategies (e.g. commoner, more common) and identified a wide range of probabilistic constraints underlying this alternation (e.g. Hilpert 2008; Mondonf 2009; Cheung & Zhang 2016). The choice between the inflected and the periphrastic form appears to be sensitive to properties of the adjective, such as phonological, morphological, and lexical features, and contextual constraints, e.g. position in the phrase/clause and syntactic complexity. Empirical