Neologisms in early English letters: How to find them and what they can reveal

Tanja Säily, Eetu Mäkelä & Mika Hämäläinen (University of Helsinki)

We apply a big-data approach to discovering neologisms in a sociohistorical corpus. Our contribution to historical lexicology is threefold: (1) rather than focusing on individual affixes as much of previous corpus research has done, we analyse a wider range of neologisms; (2) to enable such a large-scale investigation, we develop a semi-automated pipeline; and (3) while building upon existing lexicographical research and resources, we aim to cover a wider social spectrum, as previous work is admittedly biased towards published texts by well-known authors (cf. Brewer 2007). Indeed, our main interest is in the social embedding of neologisms – in identifying the groups of people who are most likely to engage in lexical innovation, the ways in which the innovations propagate in the speech community, and the purposes for which the neologisms are created and established.

Our material consists of the *Corpora of Early English Correspondence* (CEEC; c.1400–1800), the letters in which are sampled for social representativeness and augmented with metadata such as the gender and social status of the writers and recipients. To discover neologisms, we automatically map each word in the CEEC to the *Oxford English Dictionary*, the *Middle English Dictionary* and databases of contemporary published texts, including *Early English Books Online*. If the first attestation date in our corpus is earlier than in the other resources, we are dealing with a neologism candidate. These candidates are then filtered and categorized in the FiCa interface (Säily et al. in press), with additional information on e.g. semantics and etymology automatically retrieved from lexicographical resources, after which the sociolinguistic analysis and interpretation may begin.

We present a case study of 17th-century neologisms identified through our pipeline. In addition to analysing their social embedding, we will discuss details of the pipeline under development, including our methods of spelling normalization required to map the words across the resources (Hämäläinen et al. 2018). Pilot results from the 18th century indicate that while the social rank of professionals, including authors, expectedly uses a great deal of innovative vocabulary, so do the lowest social ranks, and there are highly creative individuals who stand out from the rest. Besides advancing the field of historical lexicology, our discoveries benefit historical lexicography, as we will submit our antedatings to the OED and to the Middle English Compendium, and the source code for our pipeline will be freely available.

References

Brewer, Charlotte. 2007. *Treasure-house of the language: The living OED*. New Haven: Yale University Press.

CEEC = *Corpora of Early English Correspondence*. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg et al. at the Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/

Hämäläinen, Mika, Tanja Säily, Jack Rueter, Jörg Tiedemann & Eetu Mäkelä. 2018. Normalizing early English letters to Present-day English spelling. Beatrice Alex et al. (eds.), *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 87–96. Stroudsburg, PA: ACL. https://aclweb.org/anthology/W18-4510

Säily, Tanja, Eetu Mäkelä & Mika Hämäläinen. In press. Explorations into the social contexts of neologism use in early English correspondence. *Pragmatics & Cognition*.