Mika Hämäläinen[1], Tanja Säily[1], Daniela Landert[2]

Data-Driven Neologism Mining in a TV Corpus

Keywords *neologisms, TV corpus, subs, data mining*

Contribution short paper

Affiliation 1: University of Helsinki, Finland;
2: University of Basel, Switzerland

Abstract New words emerge in language all the time, and sometimes they become a part of the language for a long time, while sometimes they disappear from use as soon as they appeared. Following our previous methods in historical data (Säily et al., 2021), we focus on neologisms in a more contemporary setting. Our aim is to study the emergence and use of neologisms in the TV Corpus, which contains 325 million words of subs (Davies, 2021).

Due to the massive size of the corpus, studying neologism candidates by hand would be a time-consuming task. Therefore, we apply a filtering approach to create an initial list of neologism candidates. First, we extract the publication year of the TV show episode or movie where each lemma in the corpus first appears. This gives us a list of the earliest attestation of every lemma in the corpus. We found that comparing this list to the earliest attestations in the Oxford English Dictionary (OED Online, n.d.), and considering the words that appear in our corpus the same time or before their recorded earliest attestation in the OED potential neologism candidates does not yield enough results, unlike in our previous studies with historical data (Säily et al., 2021). For this reason, we use a large corpus called Corpus of Historical American English (COHA) (Davies, 2012) to do this filtering. We thus compare the earliest occurrences of words in the TV Corpus to the earliest occurrences in COHA producing a list of words that appeared earlier in the TV Corpus than in COHA. This list of candidates will then be gone through manually by carefully studying each occurrence of a potential neologism.

The use of novel vocabulary in television series has been studied by e.g. Bednarek (2018: Chapter 9). We aim to scale up this research by using a significantly larger corpus and automating comparisons with dictionary and corpus data. By comparing the TV Corpus with COHA and the OED and by utilizing the metadata associated with them, we are able to analyse the diachronic development of neologism use in English-language television series as well as register variation in their frequency, types, functions and semantics. As an example, science fiction series often seem to use words related to technological innovations (e.g. biodome in our data), and in some series neology may act as a characterization device (Reichelt, forthcoming).

Bibliography

Bednarek, M. (2018). Language and Television Series: A Linguistic Approach to TV Dialogue. Cambridge University Press.

Davies, M. (2021). The TV and Movies corpora: Design, construction, and use. International Journal of Corpus Linguistics.

Davies, M. (2012). The 400 million word corpus of historical American English (1810–2009). In English Historical Linguistics 2010: Selected Papers from the ICEHL 16.

OED Online (n.d.). Oxford University Press. http://www.oed.com/
Reichelt, S. (forthcoming). Innovation on screen. Marked affixation as characterization cue in Buffy the Vampire Slayer.

Säily, T., Mäkelä, E., & Hämäläinen, M. (2021). From plenipotentiary to puddingless: Users and uses of new words in early English letters. In Multilingual Facilitation.