From Knobhead to Sex Goddess: Swearwords in English Subtitles, Their Functions and Representation as Linguistic Linked Data

Anna Bączkowska,
English Language and Theoretical Linguistics Division, Institute of English and American Studies, University of Gdańsk, Poland, anna.baczkowska@ug.edu.pl

Dagmar Gromann,
Centre for Translation Studies, University of Vienna, Austria, dagmar.gromann@univie.ac.at

Purpose: Swearwords represent an important social vehicle for human communication that, beyond mere insults, are conventionally used for expressing solidarity, bonding, and banter, among other functions. Representing an empirically validated typology of such functions and annotated swearwords in subtitles as Linguistic Linked Open Data (LLOD) provides a rich linguistic research platform for foul language. As a first case study, we classify swearwords in the English subtitles of Bridget Jones’s Diary into twelve unique functions and represent them with OntoLex-Lemon (Cimiano et al., 2016).

Design/methodology/approach: Extending existing lists of swearwords (e.g., as proposed by Dewaele, 2016) with, e.g., Wiktionary lists. This approach automatically detects swearwords in English subtitles by means of a profanity checker. Two linguistic experts annotated each extracted example with functions from the proposed typology, including pain, flirting/teasing, banter, and giving advice. The annotated primary corpus data and their metadata and functions are represented as LLOD resources with the OntoLex-Lemon model and OntoLex-Frac module for corpus data, extolling its usefulness with some sample exploration of the LLOD swearwords.

Findings: The initial calculation of the kappa value provided moderate inter-annotator agreement of 0.429, upon which we refined the definitions of the functions and discussed examples to generate a gold standard. The function typology thereby created is easily reusable as an LLOD resource, for which we provide sample uses on the proposed dataset.

Research limitations/implications: This initial case study is limited to subtitles of a single, yet rich, movie, which naturally limits the variety of swearwords explored in this first approach. Nevertheless, it provides an excellent first testbed for porting foul language and its diverse functions to the LLOD.

Practical implications: Representing functionally classified swearwords as LLOD allows, for instance, for exploring relations between swearwords, frequencies of
functions, and the attitudes of movie characters with simple queries.

**Originality/Value:** To the best of our knowledge, this is the first proposal to represent swearwords and their functions as LLOD, which provides an interesting use case for existing modeling approaches for the LLOD community, and an incentive for linguists investigating offensive language to benefit from the interoperability and ease of reuse of LLOD resources.

**Keywords:** swearwords, functions, subtitles, Linguistic Linked Open Data

**Research type:** Research paper

**Bibliography**
