Abstract

This thesis presents a corpus-based investigation that focuses on adjectival similes. As similes are often overshadowed by metaphors, the existing research on similes using empirical evidence is relatively scarce. The presented work provides a complex account of adjectival similes in corpora using data from multiple heterogeneous sources. The methodology for mining adjectival similes from corpora is thoroughly described, along with the associated pitfalls. As every corpus requires a different approach, the CQL query designs are presented individually for each researched corpus. The corpus-based mining yielded a list of 309 unique adjectival similes. The obtained data are subjected to meticulous scrutiny in the form of both formal and content analyses. Following a statistical overview of the data collected from each corpus, the 60 most frequent similes are presented in detail, including their meanings and functions, typical targets, related similes, and frequencies. These similes are then searched in respected online dictionaries, and the findings suggest that dictionaries struggle to keep up with corpus evidence. The formal analysis shows a strong preference for monosyllabic constituting elements. Regarding their content, adjectival similes can be divided into five general ground-centred groups and three source-centred ones, suggesting that general patterns exist despite sometimes being very broad. Lastly, the English simile sample is compared to a Czech one, revealing a significant lexical overlap. Adjectival similes in their prototypical form are relatively infrequent compared to other idiomatic multi-word units. However, they are firmly established in the English lexicon, and the most frequent examples belong to the 'similes minimum' – similes any proficient speaker of the language would know. While this work represents a complex presentation of adjectival similes based on empirical evidence, further research is necessary in the area of their transformations (namely compound adjectival similes), whose frequencies are often much higher.

Keywords: corpora, CQL queries, dictionaries, frequency, idiomaticity, phraseology, simile