6.9   Computational aspects of modification in fixed expressions

Villada Moirón M.B., Bouma G.

Fixed expressions are linguistic expressions that show irregularities concerning their syntactic behavior and their meaning. Native speakers have no difficulty in producing and interpreting fixed expressions. They know what words within fixed expressions can be modified and under what circumstances their figurative meaning disappears. Language learners as well as computers have trouble learning and efficiently using fixed expressions.

Example (1) shows the Dutch fixed expression *de knuppel in het hoenderhok gooien* ("cause some commotion"). When *knuppel* and *hoenderhok* co-occur with the verb *gooien* these words denote a different meaning from the one they have outside this expression. To test whether this expression is fixed, one cannot replace the words by synonyms, because the original meaning is lost (2,3). Some fixed expressions may admit modifiers (4) but not all (5).

1   De boze rekeninghouder gooit uiteindelijk de knuppel in het hoenderhok.
2   De boze rekeninghouder gooit uiteindelijk de stok in het hoenderkok.
3   De boze rekeninghouder gooit uiteindelijk de knuppel in het kippenhok.
4   Ethica Dupuis gooit in het debat over de gezondheidszorg verschillende knuppels in het hoenderhok (taken from De Volkskrant op CD-ROM, 1997).
5   *Jan heeft Piet een leuke loer gedraaid.

Due to their idiosyncratic behavior, fixed expressions are not as productive as regular phrases and need to be treated separately in a computational grammar. Two problematic aspects are: how to predict when the insertion of modification preserves the original meaning and how to represent these expressions in a grammar (see figure 15 for an example).

Figure 15 ▸
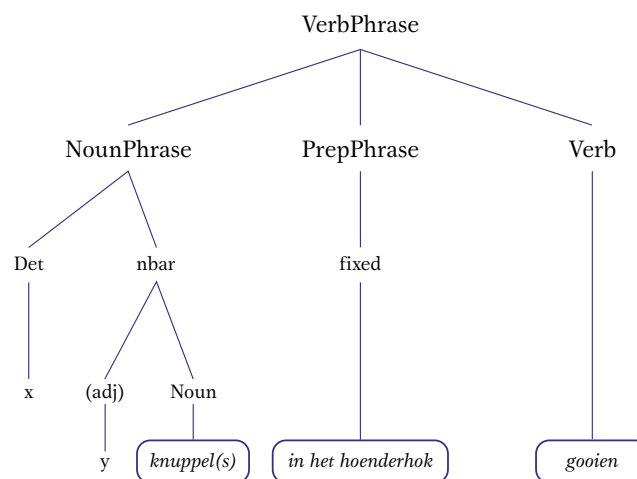
The goal of this project is to understand the grammar of fixed expressions in particular their potential for modification. This investigation attempts to infer which restrictions affect modification within fixed expressions by observing their behavior in real language use. Large collections of text (corpora) provide evidence about the linguistic performance of humans when uttering fixed expressions. We pursue a corpus-based approach to identify Dutch fixed expressions and explore allowable modification. To accomplish this, we develop statistical models that extract fixed expressions from corpora. Such models are equipped with statistical tests that measure the strength of the association between words inside candidate expressions. The statistical models provide us with a list of fixed expressions. Then, we semi-automatically determine the behavior of the extracted expressions in corpora aiming to propose a description of modification phenomena within fixed expressions.