"To take or not to take phraseology into account. The place of multi-word sequences in corpus data and experimental data"

Gilquin, Gaëtanelle

Abstract
The aim of this study is to determine the place of multi-word sequences in two types of linguistic data, namely corpus data and experimental data. The focus will be on two High frequency verbs, give and take, and their use in collocations and idioms. By comparing the phraseological usage of these two verbs in the corpus and experimental data, it will be possible to determine whether, as is commonly assumed, (recurrent) multi-word sequences are stored holistically in the mental lexicon.

Document type: Communication à un colloque (Conference Paper)

Référence bibliographique
To take or not to take phraseology into account

The place of multi-word sequences in corpus data and experimental data

Gaëtanelle Gilquin

Université catholique de Louvain, FNRS – Belgium

Keywords: high frequency verbs, psycholinguistics, elicitation, holistic storage, mental lexicon

1. Introduction

The aim of this study is to determine the place of multi-word sequences in two types of linguistic data, namely corpus data and experimental data. The focus will be on two high frequency verbs, *give* and *take*, and their use in collocations and idioms. By comparing the phraseological usage of these two verbs in the corpus and experimental data, it will be possible to determine whether, as is commonly assumed, (recurrent) multi-word sequences are stored holistically in the mental lexicon.

2. Data and methodology

The corpus data consist in 1,000 occurrences of each verb randomly selected from the Switchboard corpus (telephone conversations) and 1,000 occurrences from the Frown corpus (various written genres). Both corpora represent present-day American English.

For the experimental part, an online sentence production test was designed in which participants were asked to give the first sentence that came to their minds for each of the stimuli shown on the screen. The test comprised 20 stimuli, presented one at a time and in random order, and included two occurrences of the target verbs *give* and *take*. 40 native speakers of American English took part in the experiment, which resulted in a list of 80 sentences for each verb.

All the data thus collected were classified according to the sense they illustrated. The two categories that will interest us here are those of collocations (defined for our purposes as light verb constructions such as *give a smile* or *take a look*) and idioms (semantically non-transparent combinations such as *give rise to* or *take advantage of*).

---

3. Results

3.1. Holistic storage?

As can be expected from the nature of language in general (see e.g. Sinclair 1991 or Stubbs 1993) and of high frequency verbs in particular (cf. Lewis 1997:25), a large proportion of the occurrences of *give* and *take* in the corpus data consists in phraseological uses. In fact, for both verbs, the most frequent use is the delexical one (collocations), with 17% for *give* and 19% for *take*. Idioms represent a percentage of 5% and 11%, respectively.

These multi-word sequences, however, should arguably not make their way into the experimental data. If, as claimed by several psycholinguists (e.g. Wray 2002:9), such sequences are stored in the mind as whole units, they should not be elicited by one of their constituents. This, according to Nordquist (2004), explains why, when presented with the stimulus *I* in an experimental task, subjects do not access a sequence such as *I guess*, which has its own independent representation in the mental lexicon. In the present case, this would imply that the simple stimulus *take* does not elicit sequences such as *take a look* or *take into account*. Yet, the experimental data reveal that subjects produce quite a number of these. Most of the time, the proportions are smaller than in the corpus data, but in the case of idioms with *give*, the percentage is actually higher in the elicitation data than in the corpus data (16% vs. 5%).

These results seem to suggest that multi-word sequences are not, or at least not necessarily, stored holistically. In the next section, we will investigate a factor that may be invoked to account for the elicitation of such sequences, viz. frequency.

3.2. The role of frequency

It has been argued that the determining factor for holistic storage is frequency. Only if a particular sequence is very frequent in language does it become entrenched in the mental lexicon and does it acquire an existence of its own, independently of the items it consists of. Thus, Vogel Sosa & MacFarlane (2002) have demonstrated that subjects less readily identify the word *of* in highly frequent phrases (e.g. *kind of*) than in infrequent phrases (e.g. *cup of*), which suggests holistic storage in the former case, but not in the latter. It can therefore be hypothesised that, whereas the multi-word sequences that are not recurrent in language may be elicited by the stimuli *give* and *take*, those that are frequent will have lost all connection with the verbs and so will not be produced by the subjects.

The data, however, do not provide such a clear picture. Granted, some very frequent sequences such as *take place, take advantage of or give rise to* are never elicited, while others not represented in the corpus are (e.g. *take it or leave it, give it a go*). But some of the idioms and collocations produced by the respondents are among those frequently recurring in the corpus data. *Take care*, which is the most frequent collocation with *take* (130 occurrences in the corpus), and *give attention*, the second most frequent collocation with *give*, both appear in the experimental data. And *give somebody a break*, which is the idiom most often elicited (8 occurrences, i.e. 10% of all the elicited sentences for *give*), is in the top five list of idioms with *give* in the corpus.
4. Conclusion

Several studies have underlined the discrepancy between corpus data and experimental data (see e.g. Roland & Jurafsky 2002) and on the basis of the literature, we would have expected this to apply to phraseology as well. If it is true that multi-word sequences are stored holistically in the mental lexicon, they should not be associated with the elements that compose them – at least not when the sequence is sufficiently frequent in language to have become entrenched. This study, however, has shown that this is not so obvious and that there might still be some association between a multi-word sequence and its components, even in the case of a recurrent multi-word sequence. This conclusion is similar to that drawn by Schmitt et al. (2004), who demonstrate on the basis of dictation tasks that some, but not all recurrent clusters are stored in the mind as wholes. Not only does this challenge one of the most common claims found in the literature of phraseology, but it also calls for more research on the relation between corpus data and experimental data and, more generally, between language and cognition.

References