Modern Journal of Social Sciences and Humanities
ISSN: 2795-4846 Vol. 2 (2022)
https://mjssh.academicjournal.io/index.php/mjssh

A USEFUL TOOL IN TEACHING VOCABULARY IS ANALYSIS OF CORPUS DATA

Davlatalieva Zarina Asqarali qizi
A teacher of Namangan Institute of Engineering and Technology

Abstract: is to identify the level of correctness, volume, depth and effectiveness of the knowledge acquired by students, obtain information about the nature of cognitive activity, the level of independence and activity of students in the educational process, as well as determine the effectiveness of methods, forms and methods.

Keywords: corpus, nature.

A useful tool in teaching vocabulary is analysis of corpus data. It provides valuable information for both students and teachers about how language is used in real-life situations. A corpus is a collection of authentic texts (written or spoken transcripts) that are stored in an electronic form. Its size can range from a few sentences to millions of words. Linguistic information is typically presented in the form of concordances. A concordance is a list of all the occurrences of a particular word or phrase in a corpus, presented within the context. Concordances are obtained using the software called a concordance. One of the first teachers who used a concordance was Tim Johns, who was the author of the Data Driven Learning (DDL) (Johns, 1991). DDL is an approach to language learning based on the assumption that the use of authentic language together with a concordance will enable the learners to observe the language as it is used in real-life situations. What is more, in DDL the learning process is based on the learner's discovery of rules and patterns of language use.

A variety of exercises and tasks are proposed by authors, such as Scrivener (1994), Carter (1998), De Carrico (2001), Nation (2001), Thornbury (2002). They include:

- matching pictures to lexical items;
- matching parts of lexical items to other parts, e.g. beginnings and endings;
- matching lexical items to others, e.g. collocations, synonyms, antonyms, lexical phrases;
- word-building – using prefixes and suffixes to build new lexical items from given words;
- classifying items into word families;
- using given lexical items to perform a specific task;
- filling in crosswords, grids or diagrams;
- filling in gaps in sentences;
- memory games

In addition to the limitations of corpus analysis we have already noted, Wray discusses two others. One is the big discrepancy in the estimates by different researchers of the proportion of the corpus they analyzed which could be considered to consist of formulaic sequences. Leaving aside any problems with the reliability of the individual analyses, there are clearly validity issues here related to differing theoretical and operational definitions of formulacity. Secondly, Moon among others has found that numerous formulaic expressions that are very familiar to native speakers do not occur at all even in the mega-corpora.

Structural analysis

A variety of formal criteria have been proposed to assist in the identification of formulaic sequences. The two mostly widely recognized ones are non-compositionality and fixedness, which are characteristics of some idioms and other formulaic expressions to a lesser degree. Non compositionality means that the sequence is not interpretable as a literal statement. It may contain individual words that never occur except as part of that expression. Fixedness refers to the degree to which the order of the words in the sequence can be changed, individual words can be replaced by others, items can be inserted, or items can be infected. The fact that these criteria turn out to be continua contributes to the difficulty in drawing the line between formulaic and non-formulaic expressions.

Phonological analysis

In the case of spoken language, certain phonological features have been investigated as possible indicators of formulaic sequences. These include speech rate, pausing, stress patterns and clarity of articulation. The investigation of phonological criteria is likely to involve elicitation of data by means of a structured research design rather than analysis of an existing corpus. Apart from the relatively limited size of spoken corpora, the transcription of the oral texts in a general corpus may not meet the specific requirements of a phonological analysis. In addition, there are certain variables that need to be controlled in the interests of internal validity, such as whether the talk is spontaneous or prepared, what the topic is and the nature of the speaking task to be performed. As with other kinds of research involving the elicitation of spoken language data, there is tension between the control and manipulation of key variables needed to obtain interpretable results and the desirability, in the interests of external validity, of recording speech which is as natural and unmonitored as possible.

Pragmatic/functional analysis

Another analytical criterion recognizes that formulaic sequences have important roles in the performance of speech acts and are commonly associated with particular speech events. This provides an alternative approach to identifying them when data-gathering focuses on the particular social setting in which they typically occur (see Kuiper, Chapter 3). It also gives another perspective on the lack of transparency that the more fixed formulaic sequences tend to exhibit. Idioms are said to lack semantic transparency because their meaning is not interpretable from knowledge of the individual lexical components. To this we can add pragmatic transparency, which refers to the need for knowledge of the social context in which particular formulaic expressions are used in order to be able to understand their role in the discourse.

Once a corpus has been compiled, it needs to be analyzed to be of any value. The computer revolution has also changed this aspect, with powerful new programs that can explore the corpus and isolate are aspects of language behavior than ever before. The three major kinds of information these programs can provide about language are how frequently various words occur, which words tend to co-occur, and how the structure of language. is organized. The last aspect has much to do with the recurring lexico grammatical theme running through this book, and will be expanded upon in the next chapter. Of the other two aspects, let us look at frequency.

Frequency

Probably the most basic thing that can be learned from studying the language contained a corpus is how frequently any particular word occurs. In fact, because counting frequency of occurrence is such a basic procedure, this was the major form of information that came out of the corpora. To derive this frequency information, the computer program simply counts the number of occurrences of a word (in any combination of its base, inflected, and derivative forms) in a corpus and shows the result on the screen in a matter of seconds. Figure 1 illustrates frequency lists from three different corpora. The first is from the corpus, and indicates the fifty most frequent words in the English language. The second list comes from the CANCODE corpus of spoken language, and represents the most frequent words in spoken English discourse. The third list shows which words are most frequent in the specialized genre of automotive repair manuals. Word counts, like these have provided some very useful insights into the way the vocabulary of English works. One of the most important is that the most frequent words cover an inordinate percentage of word occurrences in language.

However, because the most frequent content words are also the most likely to be polysemous, students must learn more than 2,000meaning senses if they are going to have control over this important vocabulary. In addition, the words make up the majority of tokens in any discourse, so if they are not known, language users will be unable to make accurate guesses about the meanings of the remaining less frequent words, many of which are likely to be unknown. A second insight is that the most frequent words in English tend to be grammatical words, also known as function words or functions (words that hold little or no meaning, and primarily contribute to the grammatical structure of language). This stems from the commonsense fact that such grammatical words are necessary to. the structure of English regardless of the topic. Articles, prepositions, pronouns, conjunctions, forms of the verb be, and so on, are equally necessary whether we are talking about cowboys, space exploration, botany, or music In contrast to grammatical words, however, content words are affected by the type of corpus. We can see that automotive-specific words such as valve and engine are extremely common in the AUTOHALL corpus but do not appear in the most frequent words of general English. The third insight is that spoken and written discourse differs considerably. A close look at the frequency lists above suggests the nature of some of these differences. In particular, a number of content words, such as know, well, got, think, and right, appear much "higher on the spoken list than on the written list.

References:

1. Ferry, B. (2008). Using corpora to augment teacher learning in environmental education. Proceedings Ascilite Melbourne, 295-298.
2. Foster, P. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing, M. Bygate, P. Skehan, and M. Swain (eds), 75– 93. Harlow: Longman.
3. Habbash, M. (2015). Learning English vocabulary using corpus linguistics: Saudi Arabian EFL instructors in focus. European Scientific Journal, 11(35), 446- 457.
4. Helm, F., Guth, S., & Farrah, M. (2012). Promoting dialogue or hegemonic practice: Power issues in tele-collaboration. Language Learning & Technology, 16(2), 103-127.
5. Huang, Y., Hwang, W., & Chang, G. (2010). Guest editorial-innovations in designing mobile learning applications. Educational Technology & Society, 13(3), 1-2