STRUCTURAL FEATURES AND DEFINING COLLOCATIONS IN ENGLISH

https://doi.org/10.5281/zenodo.7792111

Xamroqulova Shaxzoda
Samarkand state institute of foreign languages, master student

Abstract: *Since the definition, application, and evaluation of statistical association measures that take into account the linguistic properties of terms and collocations are the main objectives of this thesis, it is first necessary to substantiate in detail the distinctive characteristics of these linguistic expressions that have been proposed in the scientific literature. One can assume that there is a great amount of study material on the topic of collocations and words, so any attempt to present an overview will inevitably need to focus in on the key points, especially in the context of a computational approach like this one.*

Key words: *collocation, definition, research, perspective, lexicon.*

Introduction

Linguists had known for a while that words in natural language are neither arbitrarily put together into phrases and sentences nor are they merely restricted by the laws of syntax. Curiously, a significant portion of modern mainstream linguistics, which has been primarily concerned with examining language from a theoretical perspective, has taken this fundamental fact about collocations and, at the same time, their rather diverse and apparently idiosyncratic behavior, out of focus. The Chomskyan tradition of generative linguistics, in particular, safely demotes all lexical and syntactic peculiarities into the lexicon

If research or application task is to extract collocations from unrestricted text, computational linguists - who work in the fields of computational linguistics and natural language processing (NLP) - typically acknowledge that the two fields of linguistic research have produced a wide range of distinct definitions.

As a result, computational implementations generally do not take into account insights about the linguistic features of collocations. Of course, this is one of the holes that this thesis seeks to address. We shall compile and evaluate the linguistic characteristics of collocations taken from several linguistic research strands for this goal. On the one hand, we'll concentrate on four distinctive linguistic characteristics of collocations that can be codified algorithmically from a computer standpoint. On the other hand, we will make use of linguistic traits that will enable us to distinguish between collocations and non-collocations in linguistic terms and to define several linguistic subtypes of collocations.

Main part

Defining Collocations from the Lexicographic Perspective

Lexicologists and lexicographers are the linguists who normally have a keen interest in studying collocations and their linguistic characteristics. Of course, this is a result of lexicographers having to consider how to convey information about collocations in a linguistic dictionary or lexicon.

Collocation research in German linguistics is based on a totally different conceptualization, one that is formed from a phraseological-semantic point of view. In particular, Hausmann (1985) and Burger (2003) classify collocations according to the semantic specificity of their elements rather than concentrating on the prescriptive accuracy of collocational language use. As such, content words, including as verbs, adjectives, and nouns, are crucial collocational elements. As contrast to having equal status, the members of a collocation have a directed connection with one another. While the collocate is dominated by the base, the collocational base is described as the dominant constituent. In particular, the base is the semantically independent portion, but it requires the collocate to fully convey its meaning. The following German preposition-noun-verb (PNV) collocations serve as examples of this (and their English translations).

a. “zur Verf¨ugung stellen” (to make available)

b. “in Erw¨agung ziehen” (to take into consideration)

The degree of fixity between the various components of a collocational statement is another key feature in Hausmann's idea of collocations. On the one hand, there are fixed word combinations, most of which are idioms, to which the definitions of base and collocate above rarely apply.

Defining Collocations from the Frequentist Perspective The concept of collocation in its original sense is inextricably linked to the British contextualist linguistic movement and its pioneer, John R. Firth. However, as was already mentioned above, Firth should also be given credit for helping to establish the frequentist or empiricist tradition of British (corpus) linguistics, whose leading figures are Michael A. K. Halliday and John Sinclair. Firth's work laid the foundation for this tradition. Their study, and that of Firth, was centered on the idea that the empirical, even statistical, aspect of language use in text corpora could act as a framework to define and explain natural language. The empirically driven and statistical methodologies used in modern computational linguistics have many roots in this linguistic heritage, in fact. In particular, the idea of co-occurrence, which runs like a thread through the corpus linguistics heritage, has developed into a defining quality in practically all computer linguistics applications to collocation extraction.

Defining Collocations from a Computational Linguistics Perspective

The research on automatic algorithms to extract collocations from machine-readable natural language text was influenced in different ways by the numerous methods previously outlined for defining and identifying collocations from a linguistic perspective. Early approaches, like Berry-Rogghe, closely followed the theoretical guidelines on collocations that linguistic research had developed. As a result, these early approaches attempted to examine linguistic theories or even aimed to advance them. The requirements of computability and applicability are more important in more recent approaches, which still describe and emphasize the foundational work done by linguists. As a result, the concept of collocation is defined and used in a much broader and practical sense than in linguistics.

Collocation extraction from text corpora receives a full chapter in Manning & Schutze's well-known and lauded textbook on statistical NLP. This prominence unquestionably reflects the significance of collocation recognition as a processing stage in a natural language processing pipeline, one that ideally comes before the semantic module. Collocations, on the

other hand, have proven to be an ideal linguistic construction to apply and adapt common statistical machinery and measures to issues in natural language processing, thanks to their frequentist properties as framed by the British contextualists, as a result of the empirical turnaround of the 1990s in computational linguistics.

Conclusion

The core of word and collocation extraction still relies on a high-quality lexical association measure, despite all the exciting directions for future study, it should be emphasized once more. On the one hand, there are technological domains and topic areas that either have no terminological resources at all or only a dearth of them. This means that not every subject area benefits from the Umls resource to the same extent as the biomedical industry. Collocational lexicons have also proven to be infamously underspecified. Contrarily, the nature of natural language, i.e. its inventiveness and production, ensures that new collocations and new terms are continually being coined, even when resources like the Umls are available for a certain domain.

REFERENCES:

1. Benson, Morton, Evelyn Benson & Robert Ilson (1986b). Lexicographic Description of English. Studies in Language Companion Series, No 14. Amsterdam: John Benjamins.

2. Chomsky, Noam (1995). The Minimalist Program. Cambridge, Mass.: MIT Press. 3. Jurafsky, Daniel & James A. Martin (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, NJ: Prentice Hall. 4. Miller, George A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11):39–41.