INTRODUCTION TO THE CORPUS ANALYSIS OF PHRASEOLOGY

Rahimova Mahzuna Husan kizi

National University of Uzbekistan,

Faculty of Foreign Philology,

Linguistics (German)

Annotation: This article argues that the use of digitized dictionary texts for corpus analysis offers numerous advantages and opens up expanded perspectives in phraseology. The systematic use of such resources can provide deeper insights into the development and structure of phrasemes, which is invaluable for linguistic research.

Keywords: Phraseology, corpus analysis, phrasemes, idioms and proverb lexicon, corpora.

A successful corpus analysis in the field of (historical) phraseology requires input that is independent of expert intuition. The goal of this analysis is to search large corpora in a targeted manner to find meaningful evidence for phrasemes[3]. These evidences are crucial for a) establishing a chronological dating and b) analyzing, verifying, and describing the semantic structure of the phraseme in context.

The procedure is similar to that of single-word lexicography, with the addition of a search for variants in phraseography. A central problem is the lack of specific input necessary to know which phrasemes and their variants existed historically. An example illustrates this problem: The phrase "jemandem die Leviten lesen" (to read someone's riot act) is considered common today, but no lexical variants are known.

The analysis of the DWDS core corpus shows that "Leviten" appears as a collocator only 139 times, without identifying any further variants.[5] In addition, the historical German Dictionary of Proverbs provides at least some lexical variants, such as "Epistel" or "Vers." However, it remains unclear how common these variants actually were, making cross-checking with a text corpus essential.

Corpus analysis versus looking up dictionaries. Looking up the Duden dictionary of idiomatic expressions doesn't yield many new insights, as it lacks further references. The German Dictionary of Proverbs couldn't help much either, as the high hit rate is primarily due to the explanations of the common variant.

However, the full-text search in the digitized German Dictionary of Proverbs makes it possible to discover both lexical variants and phraseological synonyms. For example, alternatives such as "Ich hab' ihm den Kümmel gerieben" (I rubbed him the caraway) can be found.

Digital corpus analysis. This approach is far more than simple reference; it is corpus research. Digital access to full texts transforms printed reference works into electronic corpora, opening up new possibilities for phraseological research.

Corpus analysis plays a central role in phraseological research. This particularly concerns the search for phrasemes, which is often treated as a special case. According to Rothkegel, it is about determining under which conditions a word chain is considered a lexical-semantic unit and how this can be determined. The most commonly used method for identifying multi-word units is collocation analysis, which analyzes large numbers of neighboring lexemes in a short time and thus provides valuable clues to phraseologized word combinations.

Despite these advances, Heid still sees a need for development, especially with longer and more varied phrases. Many idiomatic combinations are only recognized automatically, which is effective, but does not capture all facets of the language. The interpretation of whether a phraseme is a phrase is often left to experts. These must not only evaluate the semantic features, but also determine the variants and boundaries of the phrasemes, which means additional research effort.

A particular challenge is the historical dimension of phrasemes, which arises from their cultural embedding and motivational history.[2] This leads to a competence problem, as experts cannot easily assess how phrasemes functioned at different historical language stages. The availability of contemporary sources decreases with time as we look back through language history, further complicating analysis.

The automatic extraction of phraseological material, especially for more complex phrases, proves unrealistic. Collocation analysis is highly dependent on sufficient data; for earlier centuries, such as the 19th century, data are often sparse. Statistical conclusions are therefore often unreliable, as relevant evidence for many variants is missing or scarce.

Overall, it appears that the focus on idiomatic phrasemes and a diachronic perspective significantly worsens the cost-benefit ratio of corpus analysis.

Search for linguistically related noun forms. In (historical) phraseology, real evidence is of central importance, as it provides information about the variants and morphosyntactic variations in which phrasemes existed[5]. Verifying the actual occurrence of these phrasemes in large historical corpora is therefore essential. However, a lack of evidence does not negate the fundamental existence of a phraseme.

Evidential dictionaries play a crucial role here, as they list realia for individual lemmas and phrasemes and thus confirm the actual occurrence of the lemma. While the "Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart" (Grammatical-Critical Dictionary of High German Dialect) offers relatively few evidences due to its lexicographical approach and often relies on self-formulated example sentences, the "Deutsches Wörterbuch" (German Dictionary) (DWB) provides an extensive collection of evidence texts. This collection is particularly valuable because it can be searched using corpus analysis methods and can contain extensive semantic assignments[1].

However, the marking of phraseological entries in the German Dictionary of German Proverbs (DWB) is less consistent than in Adelung's work. The evidence comes predominantly from literary works and provides an indication of the fundamental existence of an expression. To prove the phraseological status—namely, the stability or commonality of an expression—multiple evidence is required, ideally including a list of existing nominal forms.

An example illustrating this process is the idiom "das Herz abfressen," (to eat one's heart away), which is related to the still-common phrase "jemandem das Herz brechen" (to break someone's heart). Evidence shows that this phrase was used in historical contexts to express emotional states.[4] A comparison with the original sources could expand the textual context and thus provide additional insights into the semantics.

The digital version of the German Dictionary of German Proverbs (DWB) not only provides access to literary sources but also opens up access to extensive historical collections of proverbs.[5] These collections are often only accessible with considerable effort, but references in the DWB make phraseographic work considerably easier. A simple analysis of the DWB based on the names of authors of historical proverb collections allows relevant phrases to be found quickly.

Overall, the DWB provides a valuable basis for the analysis and research of linguistic expressions, their use and their development over the centuries.

Literature

1.Adelung, Johann Christoph (ed.) (1793–1801/1970): Grammatisch-kritisches Wörterbuch der hochdeutschen Mundart. Mit beständiger Vergleichung der übrigen Mundarten, besonders aber der Oberdeutschen. Nachdruck hrsg. v. Helmut Henne. Hildesheim/New York.

2.Brückner, Dominik/Knoop, Ulrich (2003): "Das Klassikerwörterbuch. Begründung und Erläuterung eines digitalen Wörterbuchprojekts zum differenten Wortschatz in der klassischen Literatur". Zeitschrift für germanistische Linguistik 31: 62–86.

3.Burger, Harald (1989): "Phraseologismen im allgemeinen einsprachigen Wörterbuch". In: Hausmann, Franz Josef et al. (ed.): Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. Bd. 1. Berlin/New York, de Gruyter: 593–599.

4.Čermák, František (2006): Statistical Methods for Searching Idioms in Text Corpora". In: Burger, Harald/Häcki Buhofer, Annelies (ed.): Phraseology in Motion. Methoden und Kritik. Bd. 1. Baltmannsweiler, Schneider Verlag Hohengehren: 33–42.

5.Dräger, Marcel (2008): "Kurz angebunden. Historisch-lexikographische Betrachtungen einer Redewendung". Erscheint in: Földes, Csaba (ed.) (im Druck): Phraseologie disziplinär und interdisziplinär. Tübingen, Gunter Narr.