MWEs cross-linguistically: benefits and challenges from a historical perspective

Natalia Filatkina, University of Trier, Germany, Department of Germanic Linguistics
filatkina@uni-trier.de

Since its establishment in the 1940s and the development into an international branch of linguistics in the 1970s, **phraseological studies** have proven that phrasemes are a universal phenomenon typical for all modern languages but strongly depending on the communicative and cultural conventions of a given language. Until very recently, phraseological research has mostly addressed modern languages spoken in Europe. Nowadays, research on phraseology continues to gain new dimensions, and this is true even of such traditional subfields as **contrastive studies**. For example, research is now being conducted across many genetically unrelated and geographically divided languages far beyond the European borders. Dialectal materials and data from languages with a strong oral tradition and a lesser degree of standardization is increasingly being included.

A significant shift towards the investigation of the structure of phrasemes, their potential for variation and modifications occurred in the 1990s, partly driven by the advent of **corpus and computer linguistics**. Within the framework of Machine Translation and Natural Language Processing, the scholarly interest in MWEs can be traced back even to the 1960s. Despite the long existing tradition, MWEs are still considered to be "a pain in the neck" or a "tough nut"[1] from a technical and theoretical point of view. More linguistic knowledge is required in order to support the corpus compilation and the development of annotation tools and standards, or as Rayson/Piao/Sharoff/Evert/Moirón (2010 44: 2) put it:

> "[…] it has become increasingly obvious that in order to develop more efficient algorithms, we need deeper understanding of the structural and semantic properties of MWE's, such as morpho-syntactic patterns, semantic compositionality, semantic behaviour in different contexts, cross-lingual transformation of MWE properties etc."[2]

What both research directions – (computational) phraseology and computer linguistics – have been lacking so far is **a historical dimension** that sheds a new light on theoretical questions as well. Bennett/Durrell/Scheible/Whitt (2013: 8) give an answer to the question why a historical dimension should be included in the current research on MWEs:

> "It is in the nature of historical corpora that they involve methodological problems which can differ substantially from those presented by the compilation of corpora of living languages, and the tools used for analyzing a modern language may be quite unsuitable for the historical stages of the same language."[3]

The extensive research on historical German texts carried out in the HiFoS Group at the University of Trier (Germany) (www.hifos.uni-trier.de, PI: Natalia Filatkina) was one of the early attempts to address the challenges mentioned above, to develop a strong inclusive theory of historical formulaic language and to apply computer linguistic approaches to the study of variation and dynamics of historical formulaic patterns. With regard to the subject of the planned volume, the following findings of the HiFoS Group can be relevant:[4]

- Historical formulaic patterns show a high degree of variation and allow for the conclusion that a pattern becomes formulaic through **a complex processes of variation and change** that take place in different linguistic domains, in various domains at the same time and in close interaction of all the domains: in structure, semantics, pragmatics, ways of syntactic contextualisation, distribution

[1] Sag, I.A. et al. (2001): *"Multiword expressions: A pain in the neck for NLP"*, in: *LinGO* 2001-2003. For detailed analysis with regard to German cf. Moulin, C./Gurevych, I./Filatkina, N./de Castilho, R. E. (in print): *"Analyzing formulaic patterns in historical corpora."* In: Gippert, J./Gehrke, R. (eds.): *Proceedings of the LOEWE Conference "Historical Corpora 2012";* Filatkina, N. (2009): *"Historische formelhafte Sprache als „harte Nuss" der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt"*, in: *Linguistik online* 39/3, 75–95.

[2] Rayson, Paul/Piao, Scott/Sharoff, Serge/Evert, Stefan/Moirón, Begoña Villada (2010): Multiword expressions: hard going or plain sailing? In: *Language resources and evaluation*, 44/1, 1–5.

[3] Bennett, P./Durrell, M./Scheible, S./Whitt, R. J. (2013) (eds.): *New methods in historical corpora.* Tübingen.

[4] For publications cf. the website of the Research Group: www.hifos.uni-trier.de/Publikationen.htm. The work has been supported by the Alexander von Humboldt Foundation in the framework of Sofja Kovalevskaja Award 2006 for Dr. Natalia Filatkina.

in texts, stylistic connotations, frequency of use, degree of familiarity and so on. The idiom *Perlen vor die Säue werfen*, for example, occurs in German texts from the 9th to 16th century 33 times demonstrating each time a different structure and syntactic contextualisation as well as semantic change from a very narrow sense in religious contexts only to a broader one. The idiom changes with regard to its pragmatic function from didactic to commentarial and stylistic connotation from a noble expression of the Biblical origin to a colloquial one. Furthermore, the restriction to religious texts becomes obsolete from the 15th century onwards. However, the diachronic study of variation of formulaic patterns is often completely neglected, even in publications claiming the status of reference works on language change. This fact stands in striking contrast to language change studies in the field of phonology, morphology, single word lexicon or other linguistic domains that date back to the establishment of historical linguistics as a scientific discipline in the 19th century. Common criteria known from existing language change theories do not apply to formulaic patterns in the same way as, for example, to sound, grammar or even lexical change. In order for language historians to carry out extensive research into variation models and their dynamics, this fact must be taken into consideration while answering the question about the depth of corpus annotation. It also sheds a new light on the process of (semi-)automatic identification of formulaic patterns.

- For historical times, the decision about the formulaic character of a certain unit often cannot be made on the basis of one particular language as the pattern might occur there only once. **The cross-linguistic approach advances to a necessary method of historical analysis**, determining even the decision making at the core level of definitions.[5] An example here is the German version of the widespread proverb *Big fish eat little fish*: Though its existence in the modern highly stable syntactical, morphological and lexical form in English and French can be traced back to the 13th century, the today's degree of fixedness is not reached in German until the 17th century. With regard to its Biblical origin, it is a striking fact that the occurrence in written tradition starts earlier even in Yiddish than it does in German.
- **The definition/classification criteria of formulaic patterns** do not entirely match the criteria established for phrasemes on the basis of modern languages (polylexicality, syntactic stability, idiomaticity). Polylexicality confronts the lack of orthographic norms or the problem of word/sentence boundaries and idiomaticity – the difficulties of hermeneutic interpretation of meaning caused by culture and time distances between present day and historical data. One of the more widely accepted criterion for a formulaic pattern both in scholarly research on phraseology and in computer linguistics is its repetitious occurrence. It would seem a truism that this phenomenon can and indeed must be documented in order to employ the criterion. Thus, it cannot be put at the centre of linguistic analysis of the historical data. This is why HiFoS used the term *formulaic pattern*, which allows for including in historical analysis even units that a) are highly flexible in their grammatical structure, lexical constituents, and meaning, but have a stable underlying syntactical pattern; b) consist of only one word or are much longer than a sentence; c) are central to the text because of a specific pragmatic function, and d) might occur in texts only once. Corpus and computational studies in the field of formulaic patterns should take these circumstances as a starting point in order to facilitate linguistic research and to take it to a new level.
- Diachronically, **essential shifts in the usage of single types of formulaic patterns** can be observed. Less standardized and codified languages like older German, Yiddish and Luxembourgish, contemporary dialects and colloquial languages treat formulaic patterns in different ways. Here, recent research has been able to show clear distinctions especially with regard to frequency and functions of formulaic patterns. For formulaic patterns of the older German, the HiFoS Group had to develop **a new classification**.

[5] Therefore, HiFoS has established an international research network on historical formulaic patterns that currently unites cooperation projects from UK, Austria, France, Luxembourg, Switzerland and Germany and such languages as Classical Latin and Greek, Middle Hebrew, Classical and Modern Arabic, Yiddish, Luxembourgish, Old Syrian, mediaeval Romance languages and Yudeo-Spanish.