Phraseology in Dictionaries and Corpora. Introductory Remarks

Phraseology *in* the dictionary, phraseology *in* the corpus – this sounds less complex than it is: upon closer examination, we are faced with a multi-faceted matter. Not only can we understand 'phraseology' as a scientific discipline (i. e., in terms of phraseological research), but we can also think of it as an object of study (i. e., the treasure of phraseological units occurring in clearly defined linguistic material or even in a language as a whole). After all, phraseology is not contained per se in a dictionary or a corpus, and thus "simply placed", at our disposal for other purposes – starting from general interests for private use, through instructional and educational purposes, right up to phraseological research. Rather, the dictionary and corpus (or, more correctly: different kinds of dictionaries and corpora) are different in this and other respects and, additionally, stand in multiple complex interrelations. But, first and foremost, phraseological material is not "simply given" – not in a corpus or in a dictionary.

In order to provide a general framework for the individual contributions to this volume, and with regard to the variety of paths and ways suggested and pursued here, it seems reasonable to discuss the problems mentioned, albeit in a most general manner, notwithstanding the fact that some, or even most, of the following remarks may appear to be more or less self-evident to a phraseologically oriented audience.

Assuming that we are concerned with a corpus of text, 'corpus' is usually understood as a more or less systematic compilation of linguistic material, serving to make empirical observations, i. e., to collect specifically linguistic data, which may refer either to an individual phraseological unit (in this case resulting in what might be termed a single case study) or to a particular set of phraseological units. In any case, the objective is a statement about the presence or occurrence of particular elements, or a generalizing statement, the validity of which may be intended to go beyond the corpus under study.

In the simplest case, we are merely concerned with symptomatic description: some occurs in the material under study, or it does not occur; this is, after all, a simple binary, dichotomous classification. The subsequent information about frequency of occurrence leads to the categories of quantity or degree, providing information on how often, or to what degree, a particular phenome-

non occurs. This may, but need not, go beyond the symptomatic state, not least depending on whether the frequencies are relativized in one way or another, i. e., related to some totality, or with regard to some comparable amount of data. In this case, we are, however, still concerned with descriptive procedures, and in principle there is still no need to reflect on the quality of the given corpus: it is as it is, and all statements concern no more and no less than the available data analyzed. In case one intends to make far-reaching conclusions (or, to be more cautious, assumptions or hypotheses), which go beyond the data observed, one usually regards the corpus as a "sample" for a (assumed) totality, or even for (a given) language as a whole. From a theoretical point of view, this last assumption is difficult to sustain because "language as a whole" does not exist, regarding language to be an inherently heterogeneous, dynamic and continuously evolving system: 'language' is neither the sum of all texts ever produced, nor is it the sum of all texts ever to be produced – 'language' is tangible only as an abstract construct, based on linguistic observations and generalizations.

What can be done instead – and this would be a theoretically substantiated procedure – is to create models on the basis of observed data, the validity of which can subsequently be applied to other (including larger) data sets of linguistic material, in the form of hypotheses and testing. However, this employs inferential processes, which is true even for the (statistical) comparison of two samples, insofar as what is tested here is the assumption that both samples originate from one and the same population. In any case, the qualitative and quantitative formulation of hypotheses, empirical testing, and the final interpretation represent a sine qua non condition.

In principle, these remarks concern dictionaries as well as corpora. Given that a text (not necessarily to be specified here in detail) is a constitutive minimal unit of a given corpus, then the principally limited set of combined texts T_1, T_2, … T_n represents the corpus. The texts can be genuinely oral, written or transliterated, and they can be available in electronic (digital) form – which is the standard today – or not. In any case, corpora must be compiled, and depending on the nature of the corpus, various material will be represented to varying extent – which, as has been said above, constitutes a problem for generalizing ambitions of relevance, rather than for descriptive procedures.

If such a corpus is large enough, phraseological units of different kinds will occur; these are then, in a certain sense, "given", but still not detected. In order to identify these successfully (i. e., to identify and extract from the corpus), search queries are needed, the quality of which depends on the kind of corpus given. In the case of an electronic corpus (which is the standard case today), the possible search and retrieval strategies essentially depend on the

pre-processing procedures, which, in turn, pre-suppose human resources of one kind or another. If the corpus is not, or not yet, specifically pre-processed, or annotated, (e. g., grammatically tagged, syntactically parsed), such search strategies can only rely primarily on object-language based user knowledge: knowing a given phraseological unit one can search for it, or for its individual components, by way of string searches, i. e., simple chains of characters as parts of the requested unit. It holds true here that you can only search for what you know. The same relates to searching for concordances, that is, the presence and occurrence of linguistic units in their immediate contexts. However, the possibility of an automatic, computer-based search for phraseological units and their extraction from corpora also exists: in this case, one relies on the phraseological criterion of frozenness, according to which a phraseological unit is composed of more than one lexical entity, stereotypically fused, or merged, into one whole. Works along this line usually count the individual components' frequency of occurrence and then calculate a measure of association, or correlation, most of which are not void of statistical problems, which need not be discussed in detail here. Lexical co-occurrences, which are usually only phraseological "candidates", can be detected this way and at the next step they should be separated from fixed lexical combinations, general collocations, etc., and be verified again (and, eventually, classified), necessarily relying on human resources. The identification of specifically phraseological units can be undertaken based on either introspection – a procedure rather unreliable due to the subjectivity involved – or in the form of interviews and surveys with informants – a procedure which, depending on the method chosen, may cause problems in its own right, but at least the results and decisions come from a broader base and a wider range. As a matter of fact, comparisons with specific databases (if these do exist), or with relevant dictionaries, are also of concern here, depending of course on these sources' quality. If the corpus is annotated – which in turn presupposes the prior employment of human resources – and if phraseological units are specifically marked by way of meta-linguistic tags (which is, however, to this day a general desideratum in phraseological research), then more promising search strategies (including meta-lingual) are at our disposal, given the existence of adequate interfaces, mediating between user and data. This, however, asks for the prior detection, identification, and annotation of phraseological material in earlier phases of corpus pre-pressing, and achieving this state is still a long way away for contemporary phraseological research.

As compared to a corpus, linguistic material in a dictionary has not simply been gathered within it (in the sense outlined above), but it must be specifically collected, compiled and processed, before it finds its way into the dictionary, in

its ultimate form. This circumstance is a trivial but crucial difference between the status of a corpus and a dictionary, in both temporal and qualitative aspects.

Generally speaking, one can say that a dictionary is intended to cover a given language's lexical treasury, or a clearly defined part of it. In this framework, the dictionary's material can in principle be based on a single text as well as on a text corpus, and the dictionary derived from these sources can be of both a natural kind (i. e., contain not only phraseological units) and a specific phraseological dictionary. Moreover, a dictionary is usually characterized by lemmatization of its entries (unless we are concerned with a specific word form dictionary), accompanied and complemented by further (meta-lingual) information, starting from orthography and pronunciation, through grammatical (part of speech, gender, etc.), to explanatory information about origin, meaning, usage, translations, synonyms, equivalences, etcetera. Depending on the kind of information given (and depending on whether this information is given in one and the same or one or more other language/s), we are concerned with different kinds of dictionaries.

Strictly speaking, phraseological material can thus be included in such a dictionary in a narrow sense of the term only under the assumption of a phraseme's word equivalence – otherwise, we would be concerned with a dictionary in the broader sense of this term; these dictionaries contain *inter alia* phraseological units, asking for specific search and query strategies, which likewise holds true for specific phraseological dictionaries. Search strategies thus depend, among other features, on the dictionary's quality: If the dictionary is not given in an electronic form, one can only search manually, and the success rate will largely depend on the arrangement of the dictionary entries; but even in the case of electronically available material, specific search and query strategies are needed, and usually it is necessary to know what exactly one is looking for, the success often depending on the kind of dictionary one is using (mono- or multilingual, dictionaries for special purposes such as for language learning, specific phraseological dictionaries, etc.).

Most of these remarks likewise hold (albeit in a somewhat different way) for specific data base systems, which for specific purposes may take the function of (phraseological) dictionaries, which usually consist of two parts, the database itself, and the database management system – they, too, ask for human resources, including specific interfaces capable of mediating between users and the data base structure.

In conclusion, a variety of differences can be observed between phraseology in dictionaries and in corpora. Dictionaries, which are more than simple word lists (or word form lists), represent the result of linguistic processing, based on

the analysis of linguistic material and its lexicographic (or, eventually, phraseographic) treatment. As compared to this, text corpora as specific compilations of linguistic material may be pre-processed in different ways and to different degrees. Notwithstanding, these differences as to the phraseological material's status in dictionaries and in corpora, manifold and possibly complex relations between these two must be taken into account: on the one hand, work with corpora can be a presupposition for the compilation of dictionaries, the collection, identification, verification, or quantification of phraseological units; on the other hand, the manual or automatic search for phraseological units in corpora may be oriented towards and based on specific dictionary material.

Given these multiple and multi-layered distinctions and manifold relations and interrelations, it seems reasonable to organize the contributions to this volume in a simple and straightforward alphabetical way, rather than in thematically defined sections. All contributions are preceded by short abstracts, but the following short summaries may also be helpful for the reader and serve as a first orientation. Presentations based on these contributions were held at the EUROPHRAS conference (a traditional biannual conference organized by the European Society of Phraseology – EUROPHRAS) hosted by the University of Maribor between the 27th and the 31th of August I 2012.

In TORBEN ARBOE's (Aarhus) contribution *Phraseology – Central Parts of Culture Treated in a Dictionary*, we find a report about a Jutlandic dialectal dictionary, which contains collocations and idiomatized set phrases; by way of an illustration, the author discusses selected examples from the domain of domestic animals.

ELENA BERTHEMET's (Brest) *Colidioms. A Contribution to Cross-Cultural Research* presents an overview of a project aiming at the design of a multilingual phraseological database for phraseological systems of different languages, including semantic as well as syntactic information.

The starting point of JEAN-PIERRE COLSON's (Brussels/Louvain-la-Neuve) study *Corpus-Driven Phraseology Assessment: an Experiment* is the observation that the use of phraseology (in a broad understanding of this term, including multi-word expressions) by non-native speakers is characterized by the underuse of some structures and simultaneous overuse of others.

COSIMO DE GIOVANNI (Cagliari) aims at a revision of the relation between synonymy and collocation, on the one hand, and between corpus and bilingual dictionaries, on the other. In his study *La synonymie collocationnelle. Entre corpus et dictionnaire bilingue*, relevant examples are analyzed in order to demonstrate the difference between corpus evidence and lexicographic treatment.

MARCEL DRÄGER, RENÉ FRAUCHIGER, MARLÈNE LINSMAYER und ALESSANDRA WIDMER (Basel), in their article *Kollokationenlexikographie. Ein Bericht aus der Praxis* show how quantitative-statistical co-occurrence analyses must necessarily be complemented by subsequent qualitative editing procedures in compiling collocation dictionaries.

In her study *Étude du figement dans les Curiositez françoises (1640) d'Antoine Oudin*, CLAIRE DUCARME (Liège) raises the problem of frozen structures of old languages, or of past conditions of languages, when it is not possible to study the frozenness of phraseological items using modern speakers' competence or intuitive/introspective methods; analyzing material from a French 17th century dictionary, she focuses on both internal (implicitly or explicitly given by the lexicographer) and external (obtained through the consultation of other lexicographic and literary sources) indications, aiming to identify frozen structures and to determine their status.

PETER GRZYBEK's (Graz) und DARINKA VERDONIK's (Maribor) contribution *General Extenders: From Interaction to Model* is based on the assumption that general extenders represent a separate category in the linguistic and phraseological system of a given language; their study attempts to show that the frequency of occurrence of general extenders is organized in a regular and law-like manner, as the result of a diversification process, and presents a relevant theoretical model.

AI INOUE's (Yokosuka) *Study of a new phraseological unit – 'be on against' as an example* concentrates on a recent phenomenon to be observed in present-day English that has not yet found entrance into relevant dictionaries, when two prepositions (termed 'complex prepositions') are put together with a single meaning, becoming established as new phraseological units.

EMMERICH KELIH (Vienna), in his *Paarformeln und Binomiale im Slowenischen: Ein korpusbasierter Ansatz*, studies reversible binomials; subsequent to a short synopsis about reasons for the order of a binomial's components, he studies aspects of phraseological variability and argues in favor of taking into account frequency as an important factor related to a binomial's linguistic form.

From Dictionary to Corpus is the title of MARIE KOPŘIVOVÁ's und MILENA HNÁTKOVÁ's (Prague) article; the authors raise the question of how to identify idiomatic expressions in large corpora, when intelligent quest and search strategies are needed; they discuss and demonstrate the possibility of using special phraseological dictionary material for automatic search strategies apt to yield information of frequency and distribution of phraseological material in a corpus' texts.

NATAŠA KRALJ (Maribor), in her Study *Digitalisierung der Phraseologie und der Benutzer-Aspekt*, criticizes that there are no sufficient user-oriented studies on the usability of electronically based phraseological material, particularly as far as foreign learners, or foreign language learning material, is concerned.

CLAUDIA LÜCKERT, geb. AURICH (Münster), in her contribution *Prosodic Aspects of Proverb Change in English: Panini's Principle*, shows that not only set phrases, but also proverbs, tend to show specific patterns of semantic and phonological sequences, and that these tendencies are important from a diachronic perspective; in detail, she studies the influence of Panini's Principle (also known as Behaghel's Law), in proverbs, showing that this principle is at work, but that further phonological principles may need to be taken into account at the same time.

JASMINA MARKIČ's (Ljubljana) article *Acerca de la (in)traducibilidad de las unidades fraseologicas en la interpretacion de conferencias* deals with phraseological units and conference interpreting, and analyses examples of collocations, locutions, paremias and formulas which appear in Spanish speeches and are translated into Slovene.

MATEJ METERC (Ljubljana) presents an on-going project on the familiarity of Slovene paremiological units (*Online Questionnaire Providing Information on most Well-known and Well-understood Proverbs in Slovene Language*); the questionnaire, consisting of 918 units from two lexicographic sources, is presented online as a full text presentation. Further follow-up studies are planned.

VESNA MIKOLIČ (Koper) studies idiomatic expressions, using selected examples to be included in a Slovenian-English dictionary of tourism terminology. As the author argues in her contribution *Večbesedni termini v turističnem terminološkem slovarju*, the distinction between collocations, non-phraseological multi-word terms and phrasemes turns out to be important, since only the latter are included as entries (including idiomatic or non-idiomatic expressions), whereas terminological collocations are stated in the dictionary entry at one of the entry words.

PIOTR PĘZIK's (Łodź) *Graph-based Analysis of Collocational Profiles* focuses on the study of distributional characteristics of phraseological units; he discusses selected aspects of generating and using automatic collocation dictionaries in phraseological studies, with particular emphasis on graph-based methods of exploring and visualizing the (phraseological) material under study.

According to LIEZL POTGIETER (Stellenbosch), bilingual dictionaries are an inadequate resource for professional translators when translating idioms; the author makes some suggestions for improving the treatment of idioms in bilingual dictionaries and making them more user-friendly for translators.

MAKOTO SUMIYOSHI (Osaka) focuses on the analysis of *Valency Patterns in Dictionaries*; studying valency patterns in monolingual learners' dictionaries, valency pattern dictionaries, and authentic data, the author shows that the comparison of descriptions in these dictionaries, phraseologists, in collaboration with lexicographers, can contribute to the clarification of language change, phraseology thus being able to play a role in language research, especially lexicography, that is more important than usually assumed.

CLAUDIA MARIA XATARA (São Paulo) presents a Brazilian-Portuguese online dictionary of idioms (*Un projet phraséographique: critères et choix*), which contains definitions, additional information, illustrative examples, indications of synonymy (if any), and equivalents in Portuguese of Portugal and in the three variants of French (France, Belgium and Canada).

Peter Grzybek, Vida Jesenšek

ZORA
97

Phraseologie im Wörterbuch und Korpus

Phraseology in Dictionaries and Corpora

Herausgeber / Editors
Vida Jesenšek
Peter Grzybek

Maribor, Bielsko-Biała, Budapest, Kansas, Praha
2014