IVITRA Research in Linguistics and Literature

24

Computational Phraseology

EDITED BY
Gloria Corpas Pastor
and Jean-Pierre Colson

John Benjamins Publishing Company

Computational Phraseology

IVITRA Research in Linguistics and Literature

Studies, Editions and Translations

ISSN 2211-5412

This series aims to publish materials from the IVITRA Research Project. IVITRA carries out research on literary, linguistical and historical-cultural studies, and on history of literature and translation, specially those related to the Crown of Aragon in the Middle Ages and the Renaissance. The materials in the series will consist of research monographs and collections, text editions and translations, within these thematic frames: Romance Philology; Catalan Philology; Translation and Translatology; Crown of Aragon Classics Translated; Diachronic Linguistics; Corpus Linguistics; Pragmatics & Sociolinguistics; Literary and historical-cultural studies; and E-Learning and IST applications.

A complete list of titles in this series can be found on *benjamins.com/catalog/ivitra*

Volume 24

Computational Phraseology
Edited by Gloria Corpas Pastor and Jean-Pierre Colson

Computational Phraseology

Edited by

Gloria Corpas Pastor
University of Malaga

Jean-Pierre Colson
University of Louvain

John Benjamins Publishing Company
Amsterdam / Philadelphia

www.ivitra.ua.es
ISIC/2012/022

Universitat d'Alacant
Universidad de Alicante

DOI 10.1075/ivitra.24

Cataloging-in-Publication Data available from Library of Congress:

LCCN 2019057308 (PRINT) / 2019057309 (E-BOOK)

ISBN 978 90 272 0535 3 (HB)
ISBN 978 90 272 6139 7 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Foreword

Aline Villavicencio

In natural languages the availability of recurrent and prefabricated units facilitates communication by expressing complex concepts or ideas that are shared among a given linguistic community in a compact and precise way. These expressions are used both in informal scenarios (*making ends meet* as having enough money just for the essentials) and for technical and scientific communications (*long short-term memory* as a specific neural network architecture). They feature in child language from an early age, as word compounding can be an efficient strategy for increasing vocabulary coverage during language acquisition. However, they are also a common cause of confusion, especially for non-native speakers, as they can be conventionalised and semantically opaque (*clear up a mystery* vs. *illuminate up a mystery*), which may lead to them being mislearned (*to all intents and purposes* as *to all intensive purposes*). Moreover, languages differ in their inventory of expressions, and even if an expression in one language has an equivalent in another they may be realized in different ways which are lexically unrelated (*kick the bucket* as *die* is equivalent to *bater as botas* lit. *beat the boots* in Brazilian Portuguese, while its literal translation *chutar o balde* means *give up*).

Not surprisingly, multiword expressions have long been of interest to psychologists and linguists, as they are considered to be at the interface of the lexicon and grammar, and this may pose interesting questions for models of language acquisition and linguistic representation. For translation they are an indication of the fluency, adequacy and overall quality of the translated text and knowledge about a specific domain (e.g. *short term memory* in psychology vs. computer science). For natural language processing, the accurate handling of multiword expressions can lead to more natural and precise models of language, and has been found to bring improvements for many tasks and applications, like parsing, information retrieval, machine translation, and generation.

To date much progress has been made in profiling and in computationally modelling expressions. Since Firth's well known quote that the "collocations of a given word are statements of the habitual or customary places of that word."

https://doi.org/10.1075/ivitra.24.forvil

(Firth, 1957) various measures have been designed for detecting these expressions in corpora, such as Pointwise Mutual Information (Church & Hanks, 1990), and the development of tools that incorporate them like the Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) and the mwetoolkit (Ramisch, 2015). These, along with the creation of very large corpora, have facilitated enormously the discovery and cataloguing of expressions for particular domains and languages.

The book **Computational and Corpus-based Phraseology**, published by John Benjamins and edited by Gloria Corpas Pastor and Jean-Pierre Colson, provides a timely overview of recent advances in research on phraseology, covering a variety of methods and resources. With 16 chapters it presents a cross-section of types of phraseological units, and a variety of possible solutions for the challenges they present. Phraseological units in different languages including German, Czech, English, Russian and Italian provide scenarios that range from monolingual corpus studies of specific types, like verbal collocations, to the use of language independent methods for multiword expressions in general along with discussions related to multilinguality and translation.

What the book illustrates through this interesting overview of many languages and phraseological units is that there have been great advances towards understanding and modelling them. There are frameworks in place with customized measures and tools that can help extend these studies to phraseological units that have yet to be mapped in additional domains and languages. The editors are to be commended for the quality of this book, with chapters written by key players in the field. The result is a rich landscape of phraseology in different languages with creative computational solutions, discussed in a clear and didactic manner. Given the relevance of phraseological units for human and machine translation, as well as for language technology in general, this book will be of interest to a wide audience, from novice to expert researchers in the area.

The chapters

Profiling phraseology in different languages

Part of the book is dedicated to discussing specific types of phraseology in different languages. In the chapter, **Monocollocable words**, František Čermák analyses monocollocable – or cranberry – words, which are characterized by occurring in very few contexts (e.g. *ado* in *much ado about nothing*). Monocollocable words are collected for four languages (English, Italian, German and Czech) and their properties are compared with those of other types of multiword expressions. Dmitrij

Dobrovol'skij, focuses on the even more specific case of constructional phrasemes in German formed by *her* ('hither') and *hin* ('thither'), in the chapter, **German constructional phrasemes and their Russian counterparts**. These are semi-fixed expressions in which certain slots have to be filled. However, there may not be a systematically organised means for expressing the meanings of these expressions in other languages, and the discussion focuses on Russian.

These expressions may vary in terms of their degree of rigidity. Luigi Squillante analyses the syntactic transformations and lexical substitutions accepted in Italian MWEs in **Empirical variability of Italian multiword expressions as a useful feature for their categorisation**. Kathrin Steyer discusses language fixedness in German, from fixed lexical expressions to extended multiword expressions with facultative, but also recurrent contextual extensions of the core, in **Multiword patterns and networks**.

The realization of phraseological units may vary in different languages, and the book also provides a discussion of how these can be accommodated in translation studies and machine translation. Johanna Monti, Mihael Arcan and Federico Sangati, in the chapter **Translation asymmetries of multiword expressions in machine translation**, look at translation asymmetries between the source and target languages linked to MWEs and their impact for machine translation (MT), focusing on English and Italian using the TED-MWE corpus.

Measures for phraseology discovery

Part of the book is also dedicated to discussing measures for the automatic discovery of expressions. In **Computational phraseology and translation studies: from theoretical hypotheses to practical tools** Jean-Pierre Colson discusses some of the problems that phraseology causes for machine translation and proposes a new measure for identifying phraseological units, the Corpus Proximity Ratio. Alexander Wahl and Stefan Gries also address the extraction of formulaic language from corpora in **Computational extraction of formulaic sequences from corpora**, and test the proposed measure in scenarios that include the analysis of sequences that children may learn based on child-directed sentences. Michael Oakes discusses some widely used measures of lexical association or collocation strength, and how statistical significance can be calculated for measures that are derived from contingency tables, in the chapter **Statistical significance for measures of collocation strength**. Some of the work on discovery has also integrated syntactic information, and in **Verbal collocations and pronominalisation** Eric Wehrli, Violeta Seretan and Luka Nerima present a syntax-based collocation identification system enhanced

with an anaphora resolution module for cases in which the nominal element of a verbal collocation is pronominalized and involves anaphora resolution. The chapter by Anatol Stefanowitsch and Susanne Flach (***Too big to fail but big enough to pay for their mistakes***) examines the application of collostructional methods and measures the strength and direction of association between linguistic items in particular slots of the grammatical structure focusing on [*too* ADJ *to* V] and [ADJ *enough to* V] patterns. A detailed overview of the various stages involved in the discovery process, from defining search patterns to calculating association measures, is presented by Carlos Ramisch in the context of a general architecture for discovery, in the chapter **Computational phraseology discovery in corpora with the MWETOOLKIT**.

These expressions also vary in how their meanings can be derived from their component words. In **How context determines meaning**, Patrick Hanks argues that words in isolation have meaning potentials that can be activated in different contexts, and how meanings are associated with phraseological norms. Shiva Taslimipoor, Gloria Corpas Pastor and Omid Rohanian investigate the combination of knowledge-based and co-occurrence features for modelling semantic differences between words in both supervised and unsupervised scenarios in **Semantic discrimination based on knowledge and association**.

All we need is corpora

Given the importance of representative corpora for phraseological studies, the book also discusses the impact of different aspects of corpora. Peter Ďurčo discusses the use of comparable, monolingual and parallel corpora, in **Multiword expressions in comparable corpora**, analysing sketches of MWEs and their translation equivalents in various languages. The collocational behaviour of lexical items in general and specialised corpora is compared in terms of polysemy, choice and rank of collocates and semantic classes, in the chapter by Marie-Claude L'Homme and Daphnée Azoulay **Collecting collocations from general and specialised corpora: A comparative analysis**. The influence of corpus size and quality for finding translation equivalents for MWEs using comparable corpora is discussed by Ruslan Mitkov and Shiva Taslimipoor in the chapter **What matters more: The size of the corpora or their quality?**.

References

Church, K., & Hanks, P. (1990). Word Association Norms Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), 22–29.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford, UK: Oxford University Press.

Kilgarriff, A., Rychlỳ, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams, & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–115). Lorient: Université de Bretagne-Sud.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework* (Vol. XIV). Springer.

Introduction

Gloria Corpas Pastor and Jean-Pierre Colson
Universidad de Málaga / Université Catholique de Louvain

This volume is about *computational phraseology*, a fairly recent notion (Colson, 2003; Granger & Meunier, 2008; Heid, 2008). While *computational linguistics* as a whole has become one of the main research fields of linguistics, it is also worthy of note that recent years have seen many *computational* subdisciplines gaining some ground. At the time of writing, *computational sociolinguistics* yields 3,400 hits[1] on Google, *computational psycholinguistics* 32,400, computational *discourse analysis* 1,660 and *computational construction grammar* 3,690. *Computational phraseology* is also progressing, with almost 1,000 hits. There must be some logic to this more frequent use of the adjective *computational* in various fields.

Our point of view is that part of the explanation may be related to the *people* who carry out computational research in the various subdisciplines of linguistics. Their background can be very different: linguists, but also engineers, mathematicians, statisticians, computer scientists, programmers, etc. Bringing very diverse people together for research activities is a daunting challenge. Engineers, for instance, rely a lot on statistics and may not be aware of the terminology and cognitive aspects of linguistics; linguists, on the other hand, do not always realise that improving automated linguistic techniques also requires some mastery of computer science or statistics. The success of the *computational* subdisciplines of linguistics may precisely be due to the fact that they foster collaboration between linguists on the one hand, and engineers / statisticians / computer scientists on the other. These approaches work *bottom-up*: the research questions and the terminology come from raw linguistic data, and are then investigated within a collaborative project.

This is also the case for phraseology, the study of all fixed multiword expressions, from collocations and formulas to proverbs. However, *computational phraseology* is not yet very common, as mentioned above. There may be historical reasons for this. Phraseology as a discipline is mainly represented by Europhras,[2] the European

1. Google.com, last consulted on 23 May 2018. The search term used the quotation marks.

2. www.europhras.org

https://doi.org/10.1075/ivitra.24.00pas

association for phraseology, with a long research tradition coming from mainland Europe, and in particular from Russia and the German-speaking countries. Continental phraseology grew out of traditional linguistic approaches that focus on certain features of *phraseological units* or *phrasemes* (basically stability, idiomaticiy and gradability), and it has only recently started using automated techniques. On the other hand, the NLP and Computational Linguistics community have mainly focussed on the polylexical nature of these units, with a clear preference for terms like *multiword expression* (MWE), *multiword unit* or and *polylexical expression*. In this tradition, the focus has been on automatic identification, extraction and processing of MWEs, with little or no reference to other linguistic features, apart from idiomaticity (Monti et al., 2018). Two worlds apart…

On the other hand, phraseology also has close historical links to corpus linguistics: Sinclair (1991) lays stress on the importance of the *idiom principle*, according to which roughly 50 percent of any text consists of phraseology in the broad sense. More recent studies carried out within the framework of corpus linguistics have explored other aspects of phraseology in various corpora of L1 or L2 users (Granger and Meunier, 2008). Our purpose here is not to expand on the subtle differences between computational linguistics and corpus linguistics, but one of the reasons why *computational phraseology* is relatively infrequent a notion may be due to the competition with alternative terms referring to corpora, such as *phraseology and corpora* or *corpus-based phraseology*.

As a matter of fact, many researchers from computational and corpus linguistics were actually dealing with phraseology without using this term. This certainly holds true for more than 50 years research on collocations and their automatic extraction from corpora (Gries, 2013). While *collocation* is sometimes used by computational linguists in the general sense of fixed expression, there is now a broad consensus as to the position of *collocations* at the left-hand side of a continuum ranging from weakly idiomatic expressions to idioms and proverbs. In other words, studying collocations is hardly possible without taking into account the whole spectrum of weakly idiomatic / fixed and idiomatic / very fixed and highly idiomatic expressions. Besides, a broad array of phraseological studies deal with the complex interweaving between idiomaticity, language and culture. To give just one example, the Chinese 4-character idiomatic expressions known as *chéngyǔ* (成语) only make sense with reference to Chinese culture. They indeed correspond to an older period of the language, are often linked to literature and are fixed in the linguistic competence of native speakers of Chinese as sequences consisting (in principle) of just 4 characters.

It doesn't take a specialist in phraseology to realise that most idiomatic expressions in any language are thus part of a complex network of cultural and linguistic elements. For English, the high proportion of phrases of maritime origin comes to

mind, but more subtle links with culture or history can also be traced down. The popular phrase *be over the moon*, for instance, finds its origin in English nursery rhymes from the 18th century (Oxford Dictionary, online edition 2018), which underlines one particular literary genre and an aspect of British culture that had a major impact on English phraseology.

Against the background of history, society and culture, fixed associations in language can only be studied efficiently if a more general perspective is taken, which is precisely one of the goals of phraseology. Likewise, the notion of *computational phraseology* has the advantage of applying automated or corpus-based approaches to both linguistic and socio-cultural associations. This approach is also compatible with the recent developments of *construction grammar*, which sees language as a complex network of constructions, i.e. of pairings of form and meaning at different levels of abstraction and schematicity, and in relation with the culture of a specific language.

Let us take the example of the partly schematic *All-cleft* Construction, as in *All I had to do was find the correct answer*. This construction displays a complex network of schematic (i.e. interchangeable) and specific slots, and requires at least the pronoun *All* and two verbs as fixed slots: *All NP/PRO VP VP*. Corpus results show that a personal pronoun often follows the first slot (*All I had to go upon was…*), but other features display some regularity and might therefore be captured by statistical metrics: *All he achieved was, All he did the whole time was, All he had really expected was…*, etc. It is also clear that this construction can at any time head towards the phraseological end of the spectrum and yield a cliché or formula, as in *All you need is love*. The interplay with culture and society is clear in the last example, as well as the fuzzy border between partly fixed or partly schematic constructions and phraseology. Approaching such phenomena from the point of view of computational phraseology has the advantage of allowing for socio-cultural elements in the description, and of being compatible with a more general perspective on language.

In this book, the various aspects of the interplay between corpora, automated approaches and phraseology are illustrated.

In the first chapter, František Čermák discusses the results of a corpus-based, multilingual investigation into a special category of phrasemes that has received little attention, viz. *monocollocable words*, i.e. words that are so restricted in their combinations that they (almost) only occur in a limited numbers of phrasemes, e.g. the word *ado* in *much ado about nothing*. Interestingly, such a phenomenon turns out to be present in many languages, but extracting relevant examples poses many problems to corpus-based or computational linguistics. In this contribution, a first selection of relevant cases by means of a statistical method had to be completed with manual annotation.

Another challenge for computational phraseology is the improvement of machine translation for phrasemes or *multiword expressions* (MWEs). In the next chapter, Johanna Monti, Mihael Arcan and Federico Sangati present an original way of tackling this thorny issue, by investigating the translation asymmetries of MWEs between English and Italian in corpora. For instance, an idiom such as *spill the beans* may be translated in another language by just one word, or vice versa. This has a major impact on the quality of statistical machine translation (SMT), as the latter largely relies on translation corpora. The originality of this contribution is to rely on a MWE-annotated bilingual corpus, and to compare it with the results obtained by machine translation for MWEs, which opens up many possibilities for further research.

Another fascinating and recently investigated area of research for corpus-based phraseology is the interaction with constructions, as defined by construction grammar. In his contribution to this volume, Dmitrij Dobrovol'skij describes the German constructional idioms [vor sich her + V] and [vor sich hin + V] based on corpus examples. Constructions of this type are not only problematic for translation into another language, but they are particularly difficult to describe in dictionaries, in spite of their great importance in the language. This study shows what is at stake in the analysis and description of such borderline cases between phraseology and construction grammar, for which large linguistic corpora are of the essence. The author also pleads for a fruitful collaboration between phraseology and construction grammar, in order to shed light on the broad array of partly compositional constructions such as those under investigation.

In chapter four, Jean-Pierre Colson argues that a corpus-based and computational approach may shed some fresh light on the intertwining of phraseology, culture and translation. For instance, in spite of the largely similar and very frequent words of which they are made, the communicative phrasemes *That's life* and *This is the life* have a totally different meaning, which is to be situated against the backdrop of cultural elements, idiomaticity and recurrent patterns, and may create translation problems. Phraseology is a daunting challenge for human translators, as they have to decode very accurately all idiomatic meanings in the source text, and look for tentative equivalent phrases or constructions in the target text. Similarly, machine translation produces many cases of wrong translations because of phraseology. The author pleads for more theoretical research taking into account the diversity of languages, and also for practical tools, of which an example is presented, the *IdiomSearch* tool, based on the automatic extraction of phraseology by means of a clustering algorithm.

One of the key issues of computational phraseology is to find which algorithms are best suited for the automatic or semi-automatic extraction of phrasemes, with possible differences according to the phraseme category. In chapter five, Alexander

Wahl and Stefan Gries propose an algorithm for the automatic extraction of formulas, based on the progressive extension of bigrams according to the association strength. This methodology can also be used to help predict word sequences that young children will learn through language input. This shows that there is clearly a link between research on formulaic language and computational phraseology. While there is still room for improvement in the accuracy of the results obtained, this approach offers a comprehensive statistical discussion of the issues at stake in the extended bigram approach, which is one of the promising avenues of research in the thorny issue of automatic extraction of phraseology.

Carlos Ramisch's contribution, in the next chapter, provides an overview of the complex interplay between phraseology and computational linguistics: for instance, natural language processing (NLP) uses existing phraseological resources, but on the other hand also contributes to the creation of new ones. As in the preceding two contributions, the focus of this chapter is on the extraction of phraseology, but from the point of view of the practical tools, as opposed to sometimes technical algorithms that are hard to reduplicate. The author describes the *mwetoolkit*, a combination of programs that may be combined with corpora in order to extract phrasemes by statistical scores. As the author points out, the very complete set of tools provided by the *mwetoolkit* might still be improved by means of a graphical interface and an implementation as a web application, but it already offers a concrete example of how the manipulation of comprehensive tools plays a crucial role in computational phraseology.

As pointed out by several contributions in this volume, huge linguistic corpora are necessary for gaining useful information in computational phraseology. In chapter seven, Peter Ďurčo therefore starts from a freely available collection of impressive size, the *Araneum corpora*. He goes on to discuss the respective advantages and drawbacks of comparable corpora, as opposed to monolingual and parallel corpora, for the analysis of phraseology. Crucially, the study shows that the recourse to corpora of unrelated texts is very useful for computational phraseology, provided that the corpora are compiled with the same methodology.

Along the phraseological cline, many weakly idiomatic combinations are traditionally regarded as *collocations*. In chapter eight, Marie-Claude L'Homme and Daphnée Azoulay shed some new light on one hitherto unexplored aspect of the behaviour of collocations: the difference between collocational patterns in general vs. specialised corpora. For the purpose of this study, they used collocates associated with 15 lexical items, extracted from a specialised corpus on the environment and a general corpus. Interestingly, some significant collocational behaviour can be found in both corpora, which may have consequences for future research and practical applications in terminology and lexicography.

The recourse to corpora in computational phraseology also poses the question of the choice between quantity and quality: is it preferable to use huge corpora of average quality, or to choose smaller ones with a higher degree of reliability? In chapter nine, Ruslan Mitkov and Shiva Taslimipoor are confronted with this problem in their search for translation equivalents of verb-noun collocations across comparable corpora (English and Spanish). This type of study is of central importance to the research on machine translation, as there is a debate between the *big data* approach and a more finely tuned selection of corpora. Their results actually show that both aspects should be taken into account.

The extraction of collocations from corpora has been investigated with several methodologies and statistical scores, but the question of the significance of the scores is a complex one, which is examined by Michael Oakes in the next chapter. By weighing the respective advantages and drawbacks of the currently used statistical scores, he underlines the relationship between the frequency, distribution and statistical scores. Giving a measure of the collocational strength is another issue, because the raw statistical results do not necessarily imply a gradation in the strength of the association.

In addition to the complex statistical framework and the variety of possible scores for the automatic extraction of collocations, a recurrent theme in computational phraseology is the possible relationship between syntactic structure and collocation extraction. In chapter eleven, Eric Wehrli, Violeta Seretan and Luka Nerima show that parsing is, on the one hand, beneficial to collocation extraction and that the latter can also be useful for improving parsing. In other words, the identification of collocations and syntactic parsing are claimed to be interrelated processes, which sheds more light on the interface between syntax and phraseology. Another original feature of this contribution is the inclusion of anaphora resolution in the extraction of collocations.

Phraseology as a whole is characterised by a high degree of frozenness, but there is also some kind of variation possible. The many systematic or contextual variants of phrasemes or multiword expressions have been thoroughly investigated in the literature, but the question remains what is the best description of those variations and to which level of description they are linked, such as grammar or statistical association. In chapter twelve, Luigi Squillante analyses the variation of multiword expressions in the case of Italian, and reaches the conclusion that the recourse to linguistic corpora and to grammatical principles offers a better methodology for describing this phenomenon.

In recent years, there have been many contacts between phraseology and another major theoretical approach to language in which idioms play an important role, namely construction grammar. As constructions are defined as (partly

arbitrary) pairings of form and meaning, at various levels of abstraction, they also include phrasemes, but construction grammar offers fresh insights into the complex interplay between syntax and idiomaticity. *Collostructions* (a portmanteau word, from *collocations* and *constructions*) are particularly interesting at the crossroads of phraseology and construction grammar. In chapter thirteen, Anatol Stefanowitsch and Susanne Flach expand on collostructional analysis by starting from the patterns [too ADJ to V] and [ADJ enough to V].

Another way of looking at the interplay between constructions and phraseology, is to consider that what is at stake is a complex series of frozen lexical building blocks and of syntactic patterns. Using tools and corpora developed at the Institute for the German language in Mannheim, Kathrin Steyer shows in chapter fourteen that linguistic creativity in multiword expressions is actually rooted in a number of syntactic patterns. She also demonstrates that a corpus-based or corpus-driven approach to computational phraseology, even though it relies on huge collections of linguistic data, must always be refined at the light of an appropriate selection of the relevant results.

Patrick Hanks addresses in chapter fifteen, the central issue of meaning and phraseology. If we claim that phraseology plays such an important role in language, there must be a way of connecting it to a semantic theory and, from a practical point of view, of explaining the meaning of words (partly) by phraseology. As lexical semantics turns out to be inconceivable without recourse to diverse contexts and preferred patterns, P. Hanks claims that the use of large electronic corpora will make it possible to map recurrent patterns of phraseology onto prototypical or stereotypical beliefs about meanings. In other words, phraseology may very well serve as a crucial link between semantics and language use.

In the last chapter, Shiva Taslimipoor, Gloria Corpas Pastor and Omid Rohanian present the results of a new methodology designed for establishing discriminative semantic differences. This volume therefore concludes with a central issue in present-day and future work on computational phraseology, namely the complex links between phraseological associations and semantic ones. Indeed, Taslimipoor et al. reach signifiant results for semantic discrimination by having recourse to a number of techniques used in corpus linguistics (association scores, frequency on huge linguistic corpora) but also vector models and a knowledge-based ontology. Crucially, they show that, to some extent, phraseological association (as in the case of collocations) also contributes to the semantic network of words.

References

Colson, J.-P. (2003). Corpus Linguistics and Phraseological Statistics: a Few Hypotheses and Examples. In H. Burger, A. H. Buhofer, & G. Gréciano (Eds.), *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zu Methodologie und Kulturspezifik der Phraseologie* (pp. 47–59). Baltmannsweiler: Schneider Verlag.

Granger, S., & Meunier, F. (Eds). (2008). *Phraseology. An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139

Gries, S. Th. (2013). 50-something years of work on collocations. What is or should be next… *International Journal of Corpus Linguistics*, 18(1), 137–165. https://doi.org/10.1075/ijcl.18.1.09gri

Heid, U. (2008). Computational phraseology: An overview. In S. Granger, & F. Meunier (Eds.), *Phraseology. An interdisciplinary Perspective* (pp. 337–360). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139.28hei

Monti, J., Seretan V., Corpas Pastor, G., & Mitkov, R. (2018). Multiword units in machine translation and translation technology. In J. Monti, V. Secretan, G. Corpas Pastor, & R. Mitkov (Eds.), *Multiword Units in Machine Translation and Translation Technology* (pp. 1–37). Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Monocollocable words

A type of language combinatory periphery

František Čermák
Charles University

How often do people, even native speakers, wonder, on hearing a familiar proverb, such as Much Ado about Nothing, what ado in this proverb really means? Most will know the proverb but their knowledge of ado is often restricted to a particular lexical neighbourhood without realising that it is in fact strongly and prohibitively limited to it in this way. It is not common to give much thought to words in combinations and modes of their combination and realise that some, such as auspices, aback, standstill, ado, may not depend on how the speaker would like to use them and what they choose to say but on what the language dictates to users, that is the way how they must be used. This does not mean that there is much liberty in the use of other words either but these limitations are not immediately obvious as in this case: here, words are in their usage severely restricted to one or few more combinations only. These monocollocable words (as they are termed here), to be found, probably, in all languages, are an obstacle in understanding a foreign language, while, on the other hand, textbooks and dictionaries never really give the user much warning that there is a difficulty related to them if these should be used correctly.

Keywords: collocation, combination, corpus, distribution, monocollocable, periphery

0. Opening

On the basis of large 100-million word-corpora (such as the British National Corpus) comprehensive lists have been identified systematically for the first time ever in four different languages, namely English, Italian, German and Czech. The algorithm used (Herfindahl-Hirschmann Index) produced rather crude extensive and comprehensive lists that had to be manually verified against human knowledge and contained many different collocations based on frequency. Their selection, with hundreds of these words in each of the four languages, was based on the manual filtering, and has been presented in both alphabetical and frequency lists for each

https://doi.org/10.1075/ivitra.24.01cer

language and, next to their corpus frequencies, those combinations where they are customarily used (followed in some cases by other, less restricted combinations or collocations). The alphabetical dictionary (list) that has been prepared presents them as a ready-made product, lumped together with words that belong to a larger lexeme, i.e. in fixed collocations where they occur (thus, *aback* appears here in *(to) be taken aback*). Since they usually exhibit a peculiar feature to combine with a closed and very small class of words that they combine with (1 up to +/−7), this (closed) class membership is signalled here by explicitly listing their combinatory companions in [], such as *[be, stay] at/in the forefront of*.

Going through these findings, it is readily apparent that we are dealing here with more than special idiom components, as there are numerous multi-word terms involved that behave in the same way.

Identification of these words which have been known under various labels (such as cranberry words in English) but never really assembled in a systematic way, shows that these are, mostly,

- forms, not lemmas,
- to be identified semantically nor formally from the traditional point of view,
- definable by their restricted collocability only,
- never to be identified by their relation to homonyms,
- limited to a single sense of a word only,
- parts of phrasemes/idioms and multi-word terms only.

1. By way of introduction

Let us pin down the basic terms that should be dealt with here first.

1. Webster's Dictionary of English suggests that the ***combination*** "according to" was first used in the 14th century, though it is not clear when it started to be felt as
2. a ***multiword preposition***, i.e. as a ***stable form***, a ***unit***. Tracing this would be a good example to see how reluctant the old lexicographers have been in recognising a multiword lexeme of this kind, i.e. a combination of several forms having the function of a single one. Not straying far from *according to*,
3. its cognate *accordance* (both derivate from *accord*) suggests more intricacies here pointing to a vast realm of ***lexical combinations***. In the Longman Dictionary of the English Language (5th ed.), the word *accordance* is given as a separate lemma without any definition, but followed by a combination *in accordance with something* (marked as formal) with a somewhat circuitous explanation "according to a rule, system, etc.", i.e. referring back broadly *to according to* and its three definitions preceding it.

4. Turning from the form to ***usage***, however, the BNC records less than 1 per cent of the use of *accordance* as an independent noun, while circa 99% belong to the combination *in accordance with* only). This seems to almost nullify its independent contemporary existence in everyday language. Taking for granted that the above examples have a common root (*accord*) and may be viewed as familiar to a degree, let us look at a word form that might be, if unknown, quite unfamiliar.
5. *Aloof* is used and hence known thanks to ***combinations with very few***, i.e. four or five verbs only, *namely remain/stay aloof*, *keep/hold aloof* (from somebody), *stand aloof* (from something) and the user, even if knowing these combinations, may still be unaware of what *aloof* (coming from Dutch *a-luff* or *aloef*, "aside from the windward side of a ship") might independently mean.
6. But there are monocollocable words (MWs) that are restricted to a ***single combination only***, such as *ado* in the Shakespearean proverb *Much ado about nothing*, better known in a short form as *without more/further ado*.
7. In order to expand the view and include inflectional languages at least, it must be noted that a typical feminine gender noun in Czech, having usually 14 specific case endings, may end up here are having most of these lost (i.e. 13), retaining only a single one (Locative case plural), as in *nechat někoho na* ***holičkách*** ("leave sb in lurch"). The presupposed Nominative case form (i.e. the lemma, as well as the rest) does not exist for this word. Hence, MWs may be recruited from inflectional forms of a lemma that does not exist.

The above observations point to some of the salient problems and aspects of multiword units, though not all. The above examples of MWs hopefully show what is involved here primarily:

A. not all senses (meanings) of a form are used independently (independent meaning having a sufficient freedom of combinability, and collocability with other, different forms, words)
B. not all lexical forms are used independently being exclusively bound to others, i.e. parts of combinations
C. some of the bound lexical forms are relatively frequent, others less so.

In the following, features hinted at and briefly illustrated above so far using English, mostly, and which are believed to be relevant for many, if not all, languages, will be taken up in more detail. Specific forms in the examples, such as *accordance, aloof*, etc. have been mentioned here due to their prominent combinatorial features. *Monocollocable words* (MW), however, represent a little-studied subset of these, cf. *aloof* (*hold, keep, remain, stay*, with the BNC adding *be, stand)*.

Recently, they have been systematically inspected in large corpora in four languages (see more in Čermák, 2014, and Čermák et al., 2016).

2. Substance and definition of monocollocable words

The fact that all words combine with others is so much taken for granted that it is generally not stressed. In fact it is an expression of the very substance of the language system and it must be viewed as a specific *combinatorial law* of words that has no exception (not even in interjections). The fact that most words combine with others, only due to semantic rules (which are much more fundamental than those of grammar) is hardly reflected. This is also because the number of collocates of any word is so high that human memory is not capable of capturing it and enumerating its collocates with ease. Yet, the whole field of word combinations is numerically a continuum, one end starting with such words whose collocate numbers are unusually very small, some of these being limited in their language behaviour to a very few or even a single other word (i.e. a collocate). Here, it is best to observe how the linguistic life of such words in such numerically-limited combinations looks: not only their unclear semantic character becomes obvious, and often not even that. It is just not possible to even give it a word class membership (e.g. *to and **fro***) since both meaning and other functions of a form are due to a repeated and sufficient use of it in many contexts, which is not the case here, with MWs. In fact, the use of such an MW represents its sole context and mode of use, without which the form does not exist.

Empirically, it has been found that MWs are identifiable by this kind of combinatorial behaviour, namely collocations with 1 up to +/–7 collocates. This severe limitation representing a profound and highly unusual type of lexical behaviour is in fact a specimen of a serious anomaly in language behaviour. Since behaviour of such MWs, often being part of idioms or multiword terms (see more in Section 4), precludes any growth in the number of their collocates (which would cease to be enumerable and acquire, eventually, a meaning and further functions), they may be viewed as members of a closed paradigm, to use the term of L. Hjelmslev.

3. Are there monocollocable words on the language periphery only?

The above examples and features (in 1, 1–7) seem to clash with each other, namely a severe combinatorial restriction, which has to be viewed as a kind of language anomaly (restriction), and a rather high frequency of some of MWs (though not all), such as in *get **rid** of, be **born**, be taken **aback***, etc.

The former feature is so prominent and largely unusual in language, that it may be viewed as peripheral, i.e. part of the combinatorial periphery of the language, given the generally-shared but mistaken notion, that words combine rather freely. The latter covers forms in combinations (restricted to them only) that, as language nominations, are not peripheral at all and are, in fact, quite frequent. Hence, to

avoid any misunderstanding, out of many interpretations of what might be viewed as language periphery, only the one adhering to combinatorial restriction or anomaly in lexical combination (see also Čermák 2015) is used here.

It is not difficult to observe this kind of phenomenon, which we prefer to call *monocollocable words* (collocability amounting to lexical combinatory capacity), elsewhere, in fact in many languages. Since this is a phenomenon of language (combinatory) periphery, it is probably quite general if not universal, though it has been given a special name only in some languages and national linguistics until now (cf. *monokolokabilní slovo* in Czech such as *vyjít/dát* ***najevo*** "come to light/make obvious", *unikales Morphem* or *unikales Wörter* in German such as *in* ***Bezug*** *nehmen* "take in consideration", *restmorphem* in Swedish, such as ***körs****bär* "cherry", or unique constituents, such as Russian *biť* ***bakluši*** "hang around, lie about"), etc. Its occurrence seems to be strongly tied to the typology of the given language, appearing as morphemes (root morphemes) in isolating languages such as Chinese or English on the one hand, up to, along a scale, morphologically bound forms in inflectional languages, such as Czech. To avoid terminological and theoretical confusion, let us state here in general that by word, both independent and bound words (morphemes) are to be understood here. For example, it is estimated that Chinese compound words may have no less than 10% of monocollable words, such as ***hú****dié* "butterfly" where the first constituent *hú*, corresponding to a special Chinese character, is limited to this combination only.

4. Distribution of monocollocable words

Under different labels, MWs words have been touched upon unsystematically and in scarce cases in various languages before, but this did not yield any consistent and overall picture of the field. On the basis of these rather accidental observations (in some notable cases more detailed, cf. Dutch below) it might be worthwhile to pay it special attention. Due to the scalar character of language, or rather, its lexical combinations (ranging from a few to many combinations for various lexical items) a tangible assumption is that it is a more general phenomenon to be found in any language.

Recently, an attempt has been made, using four large corpora (of some 100 million words in each case), to systematically register MWs in four different languages. Let us inspect the field by first using corpus-gleaned examples published in Čermák et al. (2016) as the results of their research. To make basic usage easier, most prominent collocations in which the MWs occur are given as their usual text neighbourhood or context; more will be introduced later. The four languages were chosen with respect to their typological character, including a Slavic language, a

Romance language, a Germanic language and English. Typologically, all of them are mixed, though Czech is typically inflectional, followed by Italian, while German is known for its agglutination and English is analytical. Let us have a look at a few examples.

Czech
(vyjít) ***vstříc****, (dát)* ***najevo****, (vydat)* ***všanc****, (udělat/být na)* ***úkor****, (otevřít/nechat/okna aj.)* ***dokořán****, (ani)* ***nedutal****, (nechat/zůstat aj. na)* ***holičkách****, (dát na)* ***rozmyšlenou****,* ***tratoliště*** *(krve)*

Italian
(fare) ***capolino****, (parte)* ***integrante****, (capro)* ***espiatorio****, (a)* ***malapena****, (all')* ***impazzata****, tacco/tacchi a* ***spillo***

German
ausfindig *(machen), (zum)* ***Vorschein*** *(kommen), (im)* ***Laufe*** *(der Jahre/Zeit), (zur/in)* ***Gänze****, (auf)* ***Anhieb****, (auf dem)* ***Prüfstand*** *(stellen)*

English
(at the) ***forefront*** *(of),* ***vantage*** *(point), (be on)* ***tenterhooks****, (take)* ***umbrage****, (make)* ***headway****, (to and)* ***fro***

Examples from other languages are not so easy to find but they do exist, whether registered by chance or systematically. Such may be seen in Dutch where an attempt has been made to partially extract a middle-sized dictionary (Visser, 1964) following a formal pattern, namely that of *ge-Consonant* offering some interesting results, such as ***geaard*** *(zijn),* ***gefeliciteerd*** *(hartelijk),* ***geboortig*** *(uit),* ***geboren*** *(zijn, worden*, also attributively before noun), ***gebrand*** *(zijn, worden*, also attributively before noun), ***gedachtig*** *(zijn, aan),* ***gedisponeerd*** *(zijn, over, van),* ***gedrongen*** *(worden, zijn*, also attributively before noun), ***geef*** *(te*, to be distinguished from verb forms), ***gefundeerd*** *(zijn, worden),* ***gehecht*** *(zijn, aan)*, etc.

However, such an approach to Dutch or any other language is both laborious and unsystematic. A similar approach may be applied partially to *a-words* in English dictionaries yielding such MWs as ***afloat, afoot, afoul, aground, alight, aloof, amiss, anew, apace, aplomb, aright, arrears, astray, astride, asunder***, etc., but it is no more comprehensive than the Dutch case.

Let us just add a few more examples from French: *(ętre)* ***désireux*** *(de),* ***hocher*** *(la tête),* ***parce*** *(que),* ***tandis*** *(que)*;

Finnish: *tulla* ***julki***, ***siristellä*** *silmiää*; and Russian, with its once-famous and profusely discussed example, *быть* ***баклуши***.

The obvious presupposition that MWs are to be found anywhere due to the basic combinatory character of any language has to be verified, but these and other findings do support this view.

5. Language combinations and language periphery

These word forms, or rather their limited combinations, are often identical with idioms or, sometimes, with multiword terms, this severely limited combination being their basic identification feature. The scale from which they are recruited in various languages is rather broad: some, such as *to and **fro*** or *under the **auspices** of*, are rather frequent, some less so, but most are inscrutable and "unusable" without knowledge of those few combinations where they belong. It is obvious that dictionaries usually err in listing them just as any other entries (which they are not, due to their restricted class membership in closed classes).

From the synchronic point of view, the substance of monocollocable words consists of their striking combinatorial restriction, i.e. a type of collocation anomaly. This anomaly is best understood against the background of the vast majority of lexemes of any language which seem to be (almost) limitless as far as the possibilities of combination with other words are concerned. Hence this fringe phenomenon of such a limitation, i.e. a combinatorial periphery of language, must appear to be anomalous (more in Čermák 1982, 2007, 2014). To use Louis Hjelmslev's terms, most words, lexemes, belong to open collocation paradigms (classes), being without apparent numerical limits as to the number of their members, and their membership often grows gradually. MWs belong to small and closed (collocation) paradigms (classes), i.e. those with a limited membership, their number of collocates ranging from one to a few, usually plus/minus seven.

Let us have a look at more examples of the different size of the closed collocation paradigm for *aback [taken, take]*, *needless [to say], standstill [bring, come, be at], vantage [point; find, have; view etc.], breach [contract, duty, peace, condition, warranty, confidence, rules]*, the last case bordering on an open type of paradigm with other potential collocates.

Taking a closer look, it is evident that there is no form (word) in the system that combines with all others, even in their usual function. Thus, there is no adjective to be able to qualify all nouns, or a verb combining with all nouns, etc. Although combinatorially the collocational capacity of words is vastly different, being on a scale from very many to many and, finally, few, none of them, with the exception of closed paradigm members, give the impression that the user can enumerate them.

Let us have a look at examples of this severe limitation of combinatorial capacity. Using BNC information, it is clear that *lack* always requires the preposition *of* (in 100% of cases) to follow it and introduces another noun (if used as a noun), e.g. *lack of time* (as against verbal to *lack time*). Next to this 100% usage (though homonymous with that of the verb), the form *aback* is used in 97% of cases (of 303 occurrences), i.e. collocations with *taken* (*she was slightly taken aback by his response)*, while the remaining few cases consist of *took* (9x) and *takes* (1x), all of

them, however, being forms of the same single lexeme *take*. A similar case is that of *forefront* whose 494 occurrences stick to the use of prepositions *at/in the (forefront of)*, rarely alternating with *to* and *from*, all being of the prepositional type of use amounting to 98%, while only about one percent (seven cases) goes/stands without a preposition.

These cases represent and illustrate the role of MWs in idioms (*lack of, be taken aback, at/in the forefront of*), but these words, or rather forms, are also to bound (next to proper names, such as *Gordian knot*) in multiword terms. For instance, a 100% use of the adjective *stainless* is to be found with the noun *steel* (only in *stainless steel*). A little less prominent is the case *rancid* whose dominant use is found in collocations with *oil/fat/butter/margarine*, these being rather semantic variants of the same thing, while collocations such as *rancid meat/smell/air* seem to be, invariably, a kind of semantic extensions of the same thing and meaning only.

It is to be acknowledged that all of the examples above are free forms, i.e. words which seem to be different from what L. Bloomfield speaks about where *cran-* is a part of the compound form *cranberry*. This distribution of monocollocable forms in compounds is frequent, too, and, to extend the view, it is frequent also in derivatives, in derived words, such as *conceive, perceive, receive* (where *-ceive* is not to be found elsewhere in English, being anomalous and monocollocable, too). However, for practical reasons, these types of compounds and derivatives containing monocollocable morphemes, being very large, is left aside, while only combinations of words are considered.

Thus, on the basis of combination restrictions, there are two levels of forms to be recognised that seem to have a monocollocable character, namely

A. ***monocollocable bound morphemes*** (roots, such as *cranberry*)
B. ***monocollocable words*** (such as *(be taken) aback* = MW).

Above, the point has been made that monocollocable words (B type, MW) form a prominent part of the language periphery if viewed as an extreme of combinatorial capacities where all of their collocates can easily be enumerated, and are forms of only a few or even a single combination (idiom, term or proper name). This does not mean, however, that from the point of view of their function their nature is peripheral, too. Thus, while some of these forms may only be found in a single collocation (*be taken aback*), this collocation, usually an idiom, may as a whole be repeated and found rather often; consider for example the frequency of *taken aback* in BNC being 281. Sometimes the difference in frequency of components of a combination, such as *to and fro*, may be staggering – compare the frequency of *to* (2 599 205) and *fro* (283, which amounts to the frequency of the collocation). It is evident that the frequency of the monocollocable word is influenced by the (high or low) use of its combination.

It seems that there is precious little that one can, from a sole combinatory neighbour of a form (i.e. a collocate) or a few of its combinations, derive and assume by way of generalisation, a point that Leonard Bloomfield has already made. Thus, this lack of any firm ground for analogy and inference makes it impossible for the word class membership and meaning of many MWs to be determined, of course. It is only safe to state that their sole place (role) is in the combination(s) where they occur. Obviously, next to this impossibility to determine parts of speech membership, what one is dealing with here, are, roughly, autosemantic words only, while on the other hand, it is dubious that any such monocollocable element, due to its different character, is to be found among grammar words (or current affixes). As to the function of their respective combinations, these being mostly idioms and terms, it seems that monocollocable words seem to be somewhat complementary, as most idioms having these monocollocable words appear, syntactically, to have the character of verbs or adverbs, e.g. *be taken aback, to and fro*, while most terms seem to be nouns functionally, e.g. *stainless steel*.

An estimate of the average number of collocates (plus/minus seven) given above is what it says, which happens to be a pragmatic estimate that seems to be valid in most cases. Although the size of the collocation paradigm may vary, it is nearly impossible to observe where it ends – and one deals with the potentiality of combination against the observed reality. Sometimes, there is more than one such paradigm, see the case of *vantage* [*point; find, have; view* etc.], where at least three such classes are to be identified (separated by a semicolon), while there is only one class that is closed (*vantage point*), properly speaking. Hence, there seems to exist a kind of polysemy or, rather, a combinatory homonymy of different types of collocational paradigm.

As to their form, MWs seem to prefer short words, though often, as in the case of some terms with special Latin collocates (*anorexia nervosa, myocardial infarction*), it may not be so. Often MWs function, however, only thanks to a combination with a grammar word only, see *lack of, on the verge of*, etc.

Synchronically, the special and historically interesting etymology of some MWs (such as in *vantage*, being originally a very old variant of *advantage*) does not seem to influence their usage or their membership of the MW group.

Having briefly inspected the MWs of the four languages studied so far (in Čermák et al. 2016), it is not surprising to see that there is precious little, apart from the combinatorial limitation, that they would have in common; all languages seem to be rather different and unique. Their combinatorial uniqueness is due to a specific development of each word in these languages.

Numerically, combinatorial types differ very much, spanning (a) many thousands of collocates (such as *great* in English. not really enumerable) on the one extreme and (b) very few on the other. It is the latter extreme, where the number

of collocates is very small (+/−7, such as *aback*), these belonging to combinations, that are stable and fixed, usually.

Let us examine some Czech examples introducing various subtypes of MWs. To make obvious the unique character of MWs here, their closed paradigms (classes) are marked by [ ]:

(1) **očitý** [*svědek*]/***očité*** *svědectví*, [vydat] **všanc**,

(2) [*ani*] ***nedutal***,

(3) [*nechat/zůstat*] ***holičkách***,

(4) [*otevřít, být, nechat, zůstat; dveře*] ***dokořán***,

(5) ***širé*** [*moře, pole, nebe, svět, pláň, lány*]

MW types shown here should be understood as including words, i.e. forms that are inflected in some languages (2, 3).

These types (all of them being parts of idioms) illustrate a frequent case such as a verb and noun forms *nedutal* and *holičkách* (2, 3) that are used as the inflected forms occurring only here, and nowhere else. Their "corresponding" lemmas (**dutat* ("utter a sound") and **holička* ("something bare, hairless") are dubious at best, or they do not exist at all), this being a favorite object of contest for many traditional linguists not being able to reconcile with the idea that there is no lemma, even though a form does exist.

The remaining types show a real "mono"-case (1) form which is limited to a single collocate ("eye-witness", "eye-testimony") or case where a strictly limited number of collocates combines with the MW *dokořán* (No 4 based on five collocates, meaning "open broad, ajar"). Finally (5) six collocates are usually found only to combine with the MW *širý* which happens to be a form of the common adjective *široký* ("broad"), used without any restriction elsewhere in the language.

These must be complemented by such MWs that do not happen to be semantically special as to enter idioms. Instead, they form a part multi-word terms, very common in scientific use, such as ***rovným*** *dílem,* ***tekoucí*** *voda,* ***zhoubný*** *nádor,* ***polehčující*** *okolnost, sluníčko* ***sedmitečné***, etc.

Oddly enough, these two rather numerous fields where MWs are to be found, find their small but highly interesting complement from everyday speech, namely ***zout*** *(si) boty* (put/pull down one's shoes); there is no other way how to express this common meaning and notion. Hence it is suggested to call this small group *everyday common terms*, as it does represent a transition between terms and idioms.

Obviously, MWs recruit from various type of lexical fields, and also from other languages as old borrowings or are remnants of old inflection types. There is little systematicity to be found here, however interesting this often seems.

6. Identification of MWS in corpus

Only occasional mentions of MWs of forms for various languages have been made, all of their listing having been fortuitous, hand-picked by chance observation, and there was no method available to identify them. Recently it has been suggested that data obtained by an algorithm based on the Herfindahl-Hirschman index (Cvrček 2013) might serve to get a systematic coverage of a whole corpus at last. In contrast to the expanded approach used for Czech (Čermák 2014), this Herfindahl-Hirschman index has been used as the sole basis and method on the English, Italian, German and Czech corpora (Čermák et al. 2016); the HHI index is just one from several indexes that have been considered, especially Simpson's diversity index. Unfortunately, this index gave very large lists of candidates of MWs out of which, a time-consuming manual filtering was necessary, being controlled against the corpora, before a list was achieved.

For practical purposes all of the lists, based on this index, being rather technical in their nature, have been converted and ordered by frequency. It has been decided, because of too many irregular items, to start at the threshold of frequency of 200 and higher.

However, two notes of warning have to be made here. Although this is the first attempt ever at a comprehensive and exhaustive coverage of a language, those MWs below the arbitrarily chosen threshold have not been selected (such as, for example, *look askance*). Second, a 100-million-word corpus (in each language), though being definitely quite large, should not be substituted for the whole language. Thus, some MWs have not found their way here because the corpora used do not have them (e.g. *blithering idiot*).

Having filtered off all collocationally irrelevant and infrequent collocates, the result has been annotated by adding the MW's membership of either an idiom (introduced by @) or a term (introduced by #) or both. Likewise all MWs pertaining to a closed class (paradigm) of its collocates, have been annotated as []. The final results have been organised into two dictionaries, an alphabetical one (with full idiom lemmas added), and one based on frequency.

This list has been alphabetised and organised on two modifications. First, all remaining irrelevant forms have been deleted in such a way that, next to MWs, only those collocates have been retained that belong to the same stable combination (idiom or term) since this alphabetical dictionary presents only multiword units in fixed forms. Accordingly, those words forming an integral part of such idioms or multiword terms that were not found in the corpus, have been added to make the multiword lemmata complete.

The corpora used are all about the same size, having 100 million tokens. These are comparable with the British National Corpus, used for English, Coris/Codis used

for Italian (thanks to its authors), the Czech National Corpus, SYN2010, and an ad hoc corpus of German of the same size whose composition was, however, different (72% of which was comprised of Wikipedia texts and almost 20% of fiction texts).

Ultimately of a morphological nature, this limitation manifests itself in a high degree of homonymy and (morphological) conversion. In practical terms this means that if, for example, only one of the existing homonyms displays a tendency towards monocollocability and, at the same time, it proves to be of a relatively low frequency, the HHI of the homonym is relatively low. Since in the case of Czech a manual cross-control based on a simple programme listing only textual neighbours (for a list of words, i.e. MWs) has been used, much better and comprehensive results have been achieved, but that, too, is quite laborious. So, in the final analysis, the HHI approach proved itself to be helpful, but obviously, better tools for identification should be sought (see more in Čermák, 2014 and Čermák 2016).

It is just difficult to guess how many MWs are to be found in language; the BNC for English (see also the notes above) with its 100 million words and imposed restrictions is far too small. The Czech data, added by those from a Dictionary of Czech Idiomatics and Phraseology suggest that there might be a few thousand MWs available. To give at least a brief and limited view of MWs, see the list below.

A view of MWs from the letter A *(in Čermák et al. 2016)*

[ ] member of a closed paradigm, @ part of an idiom, # part of a multi-word term, x number of occurrences in BNC

aback [ ] @ (303 x): *take aback*

accordance [ ] @ (8850 x): *in accordance with sth*

adrift [ ] @ (218 x): *be adrift; cast adrift; set adrift; come adrift; leave adrift; go adrift*

afield [ ] @ (350 x): *further afield; far afield; farther afield*

afloat [ ] @ (241 x): *keep afloat; stay afloat; be afloat; get afloat; stay afloat; leave afloat*

ago [ ] (19324 x): *year(s) /month(s) / week(s) /days ago; long/(Adj/Pron): time ago*

almighty [ ] @ (381 x): *almighty God/Christ/Father/Lord/Jesus*

aloft [ ] @ (202 x): *hold aloft; be aloft; leave aloft; hoist aloft; bear aloft; carry aloft; go aloft; up and aloft*;

keep aloft; perch aloft; stay aloft; high aloft

aloof [ ] @ (226 x): *be aloof; remain aloof; hold aloof; keep aloof; stand aloof; seem aloof*

alright @ (8329 x): *be alright; do alright; get alright; go alright*

anorexia # (317 x): *anorexia nervosa; anorexia disorders; anorexia patients; anorexia treatment; anorexia*

symptoms; anorexia sufferer(s) ; anorexia problem

attorney # (685 x): *attorney general; district attorney; appointed attorney; new attorney; state attorney*

auspices [ ] @ (364 x): *under/through/outside the auspices of sth*
aware @ (10478 x): *be/become/make (well/more/fully) aware*

7. Outlook and applications

It is obvious that the whole field of MWs, so far underexplored, may serve several goals, e.g. to delimit those MWs that belong to idiomatics and dictionaries as well as language textbooks where they deserve a specific treatment. This points, among many things, to their membership in closed paradigms which should be learned as a whole and which are largely unpredictable. A similar goal, though not much pursued, is identical for multiword terms, where even their identification and listing may be useful.

No doubt, more data and more languages should be studied along these lines, as this combinatorial feature seems to be universal with many practical applications.

References

Čermák, F. (2014). *Periferie jazyka: slovník monokolokabilních slov* [Language Periphery. A Dictionary of Monocollocable Words]. Prague: NLN

Čermák, F. (2015). Phraseology and Idiomatics: Substance and Vagaries of Views. In V. Jesenšek, & D. Dobrovolski (Eds.), *EUROPHRAS 2104 Proceedings* (pp. 41–62). Maribor: Zora.

Čermák, F., Čermák, J., Obstová, Z., & Vachková, M. (2016). *Language Periphery: Monocollocable Words in English, Italian, German and Czech*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.74

Čermák, F. (2009). *Slovník české frazeologie a idiomatiky*. Vol. 1 Přirovnání (Similes, with J. Hronek and J. Machač et al., 2nd revised and enlarged edition); 2 Výrazy neslovesné (Non-Verb Expressions, with J. Hronek and J. Machač, 2nd revised and enlarged edition); 3 Výrazy slovesné (Verb Expressions, with J. Hronek and J. Machač, 2nd revised and enlarged edition); 4 Výrazy větné (Sentence Expressions, 1st ed. with J. Holub, R. Blatná and M. Kopřivová). Prague: LEDA.

Čermák F. (2016). Monokolokabilní slova v češtině: jejich hlavní aspekty. [Monocollocable Words in Czech. Their Main Aspects]. *KGA*, 13, 3–12.

Čermák, F. (1982). *Idiomatika a frazeologie cestiny* [Idiomatics and Phraseology of Czech]. Prague: Univerzita Karlova.

Čermák, F. (2007). *Frazeologie a idiomatika ceská a obecná* [Idiomaticity and Phraseology of Check]

Čermák, F. (2007). *Frazeologie a idiomatika česká a obecná* [Czech and General Phraseology]. Prague: Nakladatelství Karolinum.

Cvrček, V. (2013). *Kvantitativní analýza kontextu*. [Quantitative Analysis of Context] Prague: Nakladatelství Lidové noviny.

Visser, F.T. (1964–1973). *An Historical Syntax of the English Language*. Leiden: Brill Archive.

Translation asymmetries of multiword expressions in machine translation

An analysis of the TED-MWE corpus

Johanna Monti[1], Mihael Arcan[2] and Federico Sangati[1]
[1]Università degli Studi di Napoli "L'Orientale" / [2]Insight Centre for Data Analytics

Machine Translation (MT) is now extensively used both as a tool to overcome language barriers on the internet and as a professional tool to translate technical documentation. The technology has rapidly evolved in recent years thanks to the availability of large amounts of data in digital format and in particular parallel corpora, which are used to train Statistical Machine Translation (SMT) tools. The quality of MT has considerably improved but the translation of multiword expressions (MWEs) still represents a big and open challenge, both from a theoretical and a practical point of view (Monti, 2013). We define MWEs as any group of two or more words or terms in a language lexicon that generally conveys a single meaning, such as the Italian expressions *anima gemella* (soul mate), *carta di credito* (credit card), *acqua e sapone* (water and soap), *piovere a catinelle* (rain cats and dogs). The persistence of mistranslation of MWEs in MT outputs originates from their lexical, syntactic, semantic, pragmatic but also translational idiomaticity. Therefore, there is a need to invest in further research in order to achieve significant improvements MT and translation technologies. In particular, it is important to develop resources, mainly MWE-annotated corpora, which can be used for both MT training and evaluation purposes (Monti and Todirascu, 2016).

This work focuses on the translation asymmetries between English and Italian MWEs, and how they affect the SMT performance. By translation asymmetries we mean the differences which may occur between an MWE in a source language and its equivalent in the target language, like in many-to-many word translations (En. *to be in a position to* → It. *essere in grado di*), many-to-one (En. *to set free* → It. *liberare*) and finally one-to-many correspondences (En. *overcooked* → It. *cotto troppo*). This chapter describes the evaluation of mistranslations caused by translation asymmetries concerning multiword expressions detected in the TED-MWE corpus (http://tiny.cc/TED_MWE), which contains 1,500 sentences and 31,000 EN tokens. This corpus is a subset of the

https://doi.org/10.1075/ivitra.24.02mon

TED spoken corpus (Monti et al., 2015) annotated with all the MWEs detected during the evaluation process. The corpus contains the following information: (i) the English source text, (ii) the Italian human translations (from the parallel corpus), and (iii) the Italian SMT output. All the annotators were Italian native speakers with a good knowledge of the English language and with a background in linguistics and computational linguistics. They were asked to identify all MWEs in the source text together with their translations in approximately 300 random sentences each and to evaluate the automatic translation correctness. The identified MWEs and the evaluation of both the human and the machine translation are also recorded in the corpus. This chapter will discuss (i) the related work concerning the impact of anisomorphism (the absence of an exact correspondence between words in two different languages) and the consequent translation asymmetries on MWEs translation quality in MT, (ii) the corpus, (iii) the annotation guidelines, (iv) the methodology adopted during the annotation process (Monti et al., 2015), (v) the results of the annotation and finally (vi) the evaluation of translation asymmetries in the corpus and ideas for future work.

Keywords: machine translation, translation asymmetries, multiword expressions, TED-MWE corpus

1. Introduction

Multiword expressions, i.e. groups of two or more words that convey a single, usually non-compositional meaning, such as *credit card*, *get off*, *European Union*, *pay attention*, still represent a true bottleneck in Natural Language Processing (NLP), Machine Translation (MT) and Translation Technology (TT), despite the remarkable advances achieved in these fields in recent years. MWEs are very frequent and productive linguistic phenomena both in everyday language and in language for special purposes. In addition, they are the result of human creativity, which is not ruled by algorithmic processes, but by very complex processes, which are not fully representable in a machine code since they are driven by flexibility and intuition. MWEs represent, therefore, a very frequent source of mistranslations in MT because of intrinsic ambiguities, structural complexity, lexical asymmetries between languages and, finally, cultural differences (Monti, 2014).

Processing and translating MWEs is a crucial task in many NLP applications such as multilingual terminology extraction, machine translation (MT), cross-lingual information retrieval (CLIR) and cross-language information extraction (CLIE) among others. In particular, CLIR and CLIE success in retrieving relevant information relies on the quality of MT (Fu et al., 2009) and therefore inaccurate or incorrect translations may cause serious problems.

Even the dominant paradigm, SMT, and also the more recent neural machine translation (NMT) technology face several difficulties in translating these types of constructions, since they tend to translate on a word-by-word basis and are not able to reconstruct the intended meaning, as it can be easily verified using the available online MT systems. For instance, if we translate the English sentence, "*Every kid in the world is the apple of their parents' eye*." into Italian with Google Translate, which is now based on the neural approach, the result (https://translate.google.it/?hl=it as of June 2018) is the following: "*Ogni bambino al mondo è la mela dell'occhio dei loro genitori*". Here, the meaning of the idiomatic MWE *to be the apple of someone's eye(s)* is non-compositional and corresponds to the Italian idiomatic MWE *essere la luce degli occhi di qualcuno*, but the translation system is not able to translate it correctly.

MT has enormously improved in the last decades, but processing and translating MWE still represents one of the most important challenges. The traditional word-based alignment approach, following IBM Models (Brown et al., 1993), shows many shortcomings related to MWE processing, especially due to its inability to handle many-to-many correspondences. Since alignment is performed only between single words, i.e. one word in the source language only corresponds to one word in the target language, these models are not able to handle MWEs properly. Figure 1 presents a typical MWE misalignment in a word-based SMT system, namely Giza++ (Och and Ney, 2003).

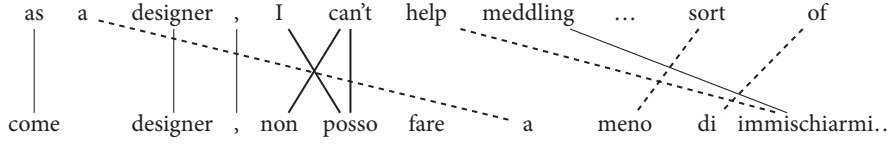


Figure 1. Example of a GIZA++ misalignment between the English MWE *I can't help* and its Italian MWE translation *non posso fare a meno di* (lit. not can do to less than). Dotted lines are indicating incorrect alignments, and tick lines (both continuous and dotted) are those adjacent to MWE tokens in the source or target sentence

The phrase-based (PB) alignment approach (Koehn et al., 2003) is better at dealing with MWEs as it considers many-to-many word alignments. However, many combinations of words or n-grams have no linguistic significance (*the war*), while others are linguistically meaningful (*cold war*). In the widely used PB-SMT systems, phrases are sequences of contiguous words, which are not linguistically motivated and do not implicitly capture all useful MWE information, although they are able to translate contiguous MWEs and sometimes also discontiguous ones. The correct translation of MWEs occurs on a statistical basis if the constituents of MWEs are aligned as parts of consecutive phrases (n-grams) in the training set.

Furthermore, MWEs are not generally treated as a special case when correspondences between source and target language do not consist of consecutive many-to-many source-target correspondences. MWE processing and translation within SMT started being addressed only very recently and different solutions have been proposed so far, but they are basically considered either as a problem of automatically learning and integrating translations into an SMT system or as a problem of word alignment. The most used methodologies are identification of possible monolingual MWEs. This phase can be accomplished using different approaches, by means of (i) morpho-syntactic patterns (Okita and Way, 2010; Dagan and Church, 1994), (ii) statistical methods (Vintar and Fišer, 2008) and finally (iii) hybrid approaches (Wu and Chang, 2004; Seretan and Wehrli, 2007; Boulaknadel et al., 2008; Daille, 2001).

Furthermore research was performed on the extraction of the equivalent translations of the identified monolingual MWEs according to the different alignment methodologies. Current approaches to MWE processing integrate phrase-based models with linguistic knowledge, such as hand-crafted dictionaries and grammars or data-driven ones, in order to identify and process MWEs as single units.

Finally, the new neural approach to MT in which a large neural network is trained by deep learning techniques is still in its pioneering stage and little has been reported about the improvements it can bring to MWE processing and translation.

One of the main problems in translating MWEs is represented by their translation idiomaticity, i.e. it is not usually possible to translate MWEs literally. In addition to that, their internal structure may greatly vary from one language to another one. This property, which goes under the name of non-literal translatability, means that an MWE cannot be translated from one language to another on a word-for-word basis (Sag et al., 2002; Barreiro, 2008; Monti, 2012), and is characteristic of the majority of MWEs, in particular those with limited or no variation of distribution of their internal constituents. This is the case for idioms (e.g. *it's raining cats and dogs!* → It. **sta piovendo cani e gatti*), but also of many collocations (e.g. *heavy rain* → It. **pioggia pesante*), fixed expressions (e.g. *by and large* → It. **da e largo*), proverbs (e.g. *there's no such thing as a free lunch* → It. **non esiste una cosa come un pranzo gratuito*) and phrasal verbs (e.g. *bring somebody down* → It. **portare qualcuno giù*) amongst others.

The anisomorphism between languages leads to translation asymmetries, i.e. the differences which may occur between a MWE in its source language and its translation, like in many-to-many translations (En. *to be in a position to* → It. *essere in grado di*) but also in many-to-one (En. *to set free* → It. *liberare*) and one-to-many (En. *overcooked* → It. *cotto troppo*) correspondences.

MWEs are sometimes discontinuous, i.e. it is possible to insert an element between the constituents of a multiword. As an example, it is possible to insert an NP into the verbal MWE *take into account* as in *take something into account*.

Translation asymmetries are one of the main sources of mistranslations in MT and one of the possible solutions to this problem is to develop large linguistic resources, mainly MWE-annotated corpora, which can be used both for MT training and evaluation purposes (Monti and Todirascu, 2016).

This chapter presents the results of the English-Italian TED-MWE project and in particular: (i) the related work, (ii) the MWE-TED corpus, the annotation guidelines and methodology, (iii) the results of the experiment and finally (iv) the evaluation of the translation asymmetries and the mistranslation in the TED-MWE corpus.

2. Related work

Studies on translation asymmetries and their impact on MT quality are underrepresented in recent NLP studies. The definition of Translation asymmetries in MT can be dated back to Pause's paper on *Interlingual strategies in translation* (1997), but the concept of a source language structure translated with a different structure in the target language was already discussed in Dorr (1994), who classified human and MT *divergences* in six different types. Dorr identifies a specific class for MWE translation, namely *Conflational or Inflational Divergence*. A conflational divergence is when two or more words in the source language are translated by one word in the target language. The inflational divergence, instead, arises when one word in the source language is translated by two or more words in the target language.

This classification has been used in Lin et al. (2005), Mahesh et al. (2005) and lately by Kauffmann (2013), who devotes a few words to the problem of MWEs in conflational divergences. According to Kauffman, large-scale monolingual lexicons of multi-word expressions (and collocations) and bilingual lexicons that record their translations represent a possible solution to the processing of such divergences in MT. MWE multilingual lexicons as well as parallel corpora annotated with MWEs represent invaluable linguistic resources for MWE processing and translation, but recent surveys (Constant et al., 2017 and Losnegaard et al., 2016) have highlighted that these types of resources are still lacking and this fact may hinder research both on the translation of MWEs across languages and NLP involving two or more languages.

Translation asymmetries represent an important clue as to the occurrence of MWEs in parallel corpora and are at the heart of a few studies which aim to detect MWEs using unsupervised or semi-unsupervised methods. Melamed (1997) develops a method for the discovery of MWE on the basis of their translational entropy in parallel corpora. A statistically-driven alignment-based approach to MWE identification in technical corpora, including parallel corpora, is shown in Caseli et al.

(2009); they examine how a second language can provide relevant clues for this tasks and extract sequences of length 2 or more in the source language that are aligned with sequences of length 1 or more in the target (m:n alignments). Bouamor et al. (2012) address non-compositional contiguous MWE sequences and present a method combining linguistic and statistical information to extract and align MWEs in a French-English parallel corpus. The extracted bilingual MWEs are integrated into MOSES to show that MT quality can be improved by the use of such units. In recent years, different approaches have been adopted with reference to MWE identification from the translational asymmetries (misalignments) in parallel corpora, such as Lambert (2005), who use an asymmetry-based approach and focus on alignment sets in which source-to-target links proposed by Giza++ are different from target-to-source alignments, or Tsvetkov and Wintner (2010), who focus on misalignments to develop an unsupervised algorithm for identifying MWEs in (small) bilingual corpora, using automatic word alignment extraction of MWEs of various types, lengths along with their translations. Other works are based on extraction of bilingual MWEs, such as Thurmair and Aleksić (2012), who extract terms and lexicon entries directly from SMT translation models, or Arcan et al. (2017), who propose a framework for extracting bilingual terms from a post-edited corpus and using them to enhance the performance of an SMT system embedded in a collaborative CAT environment. Moirón and Tiedemann (2006) focus on Dutch expressions and their English, Spanish, and German translations in the Europarl corpus (Koehn, 2005). MWE candidates are ranked by the variability of their constituents' translations. To extract the candidates, they use syntactic properties (based on full parsing of the Dutch text) and statistical association measures. Sangati and van Cranenburgh (2015) focus on identification and extraction of MWEs from a large set of recurring syntactic fragments from a given treebank. They use these fragments to identify MWEs as a parsing task (in a supervised manner) and compare various association measures in re-ranking the expressions underlying these fragments in an unsupervised fashion.

3. The TED-MWE corpus

Annotated parallel corpora are a very important resource for MT, but to present there are only very few small-sized corpora, containing, aligned sentences representative of a specific type of MWE and for a limited number of language pairs, which are also very difficult to reuse in research settings different from the original ones (Monti and Todirascu, 2016). To our knowledge, none of the corpus resources developed so far encode multiword expressions of all different types in a parallel corpus. Therefore we developed the TED-MWE corpus, which is based on the

web inventory named WIT3 (https://wit3.fbk.eu), a collection of transcribed and translated talks (Cettolo et al., 2012). The core of WIT3 is the TED Talks corpus that basically redistributes the original content published by the TED Conference website. Since 2007, the TED Conference posted all video recordings of its talks together with subtitles in English: almost all talks have been translated by volunteers into more than 80 languages and the translated talks range in number from several hundred (e.g. such as for the Dutch, German, Hebrew, Romanian languages) to just one (e.g. for Hausa, Hupa, Bislama, Ingush, Maltese). The WIT3 corpus re-purposes the original content in a way that is more convenient for MT researchers. For our experiments, we used the 2014-released WIT3 TED data set for the English-Italian language pair, which contains the training data of 190,000 parallel sentences, needed to build an SMT system. In addition, we used the 2014 TED development set (1,000 sentences) and the 2010/2011/2012 test sets (1,500 sentences each).

The TED-MWE corpus is the result of the annotation of the English-Italian WIT3 TED data set with MWEs of different types. Besides the WIT3 English-Italian parallel corpus, the TED-MWE corpus also contains the Italian output for the English source sentences obtained using the statistical translation toolkit Moses (Koehn et al., 2007), where the word alignments were built with the GIZA++ toolkit (Och and Ney, 2003). The IRSTLM toolkit (Federico et al., 2008) was used to build the 5-gram language model. The parameters within the SMT system were optimised on the development data set using MERT (Clark et al., 2011; Bertoldi et al., 2013).

The TED-MWE corpus is available for download at: http://tiny.cc/TED_MWE. In the next sections we describe the guidelines used for the annotation, the methodology adopted for the annotation process and the results of the annotation process.

4. The annotation guidelines

The judgement of whether an expression should qualify as an MWE relies on the annotation guidelines, which are based on (i) the PARSEME MWE template and (ii) the testing of MWE properties.

The PARSEME MWE Template (Savary et al., 2015) was designed to provide information and examples for MWEs in different languages along comparable dimensions of classification. These dimensions are: syntactic structures (e.g. nominal, verbal, adjectival, prepositional and clausal MWEs), the fixedness/flexibility of MWE parts (such as passivisation or modification), the different levels of idiomaticity (lexical, syntactic, semantic, pragmatic, statistical) and finally the rhetoric relations within an MWE.

In addition to the template, annotators were provided with a set of tests (Monti, 2013) to be used to assess whether a certain group of words can be considered as a MWE on the basis of the following properties:

- **Non-substitutability**: one element of the MWE cannot be replaced without a change of meaning or without obtaining a non-sense (*in deep water / in hot water; gas chamber* – *gas room);
- **Non-expandability**: insertion of additional elements is not possible (*get a head start* – *get a quick head start);
- **Non-reducibility**: the elements in the MWE cannot be reduced and pronominalisation of one of the constituents is also not possible (*take advantage* – *what did you take? advantage; *Did you take it?);
- **Non-literal translatability**: the meaning cannot be translated literally. The difficulty of a literal translation across cultural and linguistic boundaries is mainly a property of MWEs with limited or no variation of distribution, such as idioms (e.g. *it's raining cats and dogs* → It. *sta piovendo cani e gatti), but also of many collocations (e.g. *heavy rain* → It. *pioggia pesante), fixed expressions (e.g. *by and large* → It. *da e largo), proverbs (e.g. *there's no such thing as a free lunch* → It. *non esiste una cosa come un pranzo gratuito), phrasal verbs (e.g. *bring somebody down* → It. *portare qualcuno giù);
- **Invariability**: Invariability can affect both the morphological and the syntactic level, whereby the inflectional variations of the constituents of the MWEs are not always possible. Invariability affects the head elements as well as its modifiers (*fish out of water*– *fishes out of water; *dead on arrival*– *dead on arrivals; *in high places*– *in high place), syntactical variations inside an MWE may also not be acceptable (*credit card*– *card of credit);
- **Non-displaceability**: displacement and a different order of constituents are not possible (*wild card*– *is wild this card?; *back and forth* - *forth and back);
- **Institutionalisation of use**: certain word units, even those that are semantically and distributionally "free", are used in a conventional manner. The Italian expression *in tempo reale* (a loan translation of the English expression 'in real time') is an example of this feature since its antonym *in tempo irreale (*in unreal time) seems to be unmotivated and not used at all.

In order to consider a certain word unit as an MWE it is sufficient that it shows at least one of the above-mentioned properties. Nevertheless, during the annotation process, the property which turned out to characterise the majority of MWEs was the non-literal translatability.

5. The annotation methodology

The annotation was organised in three distinct phases: individual annotation, inter-annotation check and validation.

Individual annotation

During the first phase, thirteen annotators with linguistic background in Italian and English were asked to annotate 1,529 sentences in the TED-MWE corpus. The sentences were organised in a spreadsheet (see Figure 2) containing the following information: (i) the English source text, (ii) the Italian manual translations (from the parallel corpus) and finally (iii) the Italian SMT output.

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	MWE				
				SOURCE TEXT	MANUAL TEXT	MANUAL CHECK (Y/N)	AUTO TEXT	AUTO CHECK (Y/N)
369	people sort of think i went away between "titanic" and "avatar" and was buffing my nails someplace, sitting at the beach.	la gente pensa quasi che me ne sia andato tra "titanic" e "avatar" e che mi stessi girando i pollici seduto su qualche spiaggia.	persone come pensare partii tra "titanic" e "avatar" e fu buffing mie unghie da qualche parte, seduto in spiaggia.	buffing my nails	girando i pollici	Y	buffing mie unghie	N

Figure 2. Annotation Phase 1 – Individual annotation

The annotators were asked to identify all MWEs in the source text together with their translations in approximately 300 random sentences each and to evaluate the correctness of the automatically translated MWE. If the manual or the SMT generated translations were wrong, the annotators were asked to specify the correct translations. The annotation work was organised in such a way that each sentence was annotated by at least two annotators. The annotation took into account all MWE types detected in the source text with no restrictions to a particular type of MWE and in particular, both continuous and discontinuous MWE types were recorded in the dataset. The MWEs identified during the annotation process were recorded as sequences of tokens with no further information about their internal syntactic structure or semantic features.

Inter-annotation validation

In the second phase, each annotator was confronted with the anonymised annotations of the other annotators on his/her annotation subset, in order to decide about his/her choices, i.e. to confirm or change the annotations for each source text/manual/SMT set.

Evaluation

Finally, we have randomly selected about half of the annotated sentences (801) and asked the annotators to integrate and resolve the possible annotation conflicts (see Figure 3).

SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	ANN #	SOURCE TEXT	MANUAL TEXT	MWE MANUAL CHECK (Y/N)	AUTO TEXT	AUTO CHECK (Y/N)
26	"don," i said, "just to get the facts straight, you guys are famous for farming so far out to sea, you don't pollute."	"don", gli ho detto "tanto per capire bene, voi siete famosi per fare allevamento così lontano, in mare aperto, che non inquinate."	"non", ho detto, "per ottenere i fatti dritto, siete famosa per coltivare così lontano in mare, non inquinante."						
				3	to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				9	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				13	get… stright	capire bene	Y	per ottenere… dritto	N
				FINAL	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N

Figure 3. Annotation Phase 3 – Validation

6. The results of the annotation process

Based on the annotation process, out of 1,529 annotated sentences, 541 (35.9%) showed a good inter-annotation agreement, i.e. at least two annotators completely agreed on the annotations. In total we have collected 2,484 English MWEs types out of which 2,391 (96%) are contiguous and 93 (4%) are discontinuous. At least two annotators agreed for the 27% (671) of the MWEs and in 45% of them (1,115) at least two annotators showed an agreement (at least one word in common).

As a final step we have randomly selected about half of the annotated sentences (800) and asked the annotators to integrate and resolve the possible annotation conflicts. This resulted in a total of 799 English MWE types (931 tokens), of which 729 (91%) are contiguous and the 9% (70) are discontinuous.

Most MWEs have length of 2 (515) and 3 (261), but there are MWEs up to the length of 8. In 52% of the cases (471) the annotators have evaluated the automatic translation to be incorrect. Out of the 729 continuous MWEs, 253 occur only once in the whole English corpus and are therefore excluded from the final data set used for the experiments, which contains the remaining 476 English MWEs.

7. Translation asymmetries and mistranslations in the TED-MWE corpus

The fact that translation asymmetries (or divergences) between a source language and a target language may cause mistranslations is a well-known problem, but to the best of our knowledge it has not been systematically studied in the context of Machine Translation, in particular, with reference to the English-Italian language pair.

In our study, we have analysed all the asymmetries (about 250 different ones) which occur in the TED-MWE corpus and evaluated their impact on the quality of the MT output. In order to do so, a further annotation step was required: all the MWEs found in the corpus together with the correct Italian manual translation and their incorrect MT generated were annotated with POS information, as shown in Figure 4:

	A	B	C	D	E	F	G	H	I	J	K	L
499	no, i actually fell out of love with this fish because, i swear to god, after that conversation, the fish tasted like chicken.	no, in effetti mi disamorai di questo pesce perché, ve lo giuro dopo quella conversazione, il pesce sapeva di pollo.	no, ho cadevano d' amore con questo pesce perché, lo giuro su dio, dopo quella conversazione, i pesci assaggiato di pollo.	fell out of love	mi disamorai	Y		cadevano d' amore	N	V Part Prep N	Pron V	V Prep N
500	i was imagining a "march of the penguins" thing, so i looked at miguel.	immaginavo una cosa tipo la marcia dei pinguini così guardai miguel.	immaginai "marcia dei pinguini" cosa, così ho analizzato miguel.	looked at	guardai	Y		ho analizzato	N	V Prep	V	V
501	with pivot, you can drill into a decade.	con pivot si può osservare un decennio.	con perno, puoi bucare in un decennio.	drill into	osservare	N	instillare	bucare	N	V Prep	V	V
502	was coming into our living rooms with his amazing specials that showed us animals and places and a wondrous world that we could never really have previously	entrava nei nostri salotti con i suoi fantastici documentari che ci mostravano animali e luoghi e un mondo meraviglioso che prima non avremmo neanche potuto	stava arrivando nei nostri salotti con i suoi fantastici specialità che ci ha mostrato animali e luoghi e un mondo meraviglioso che potremmo mai immaginato prima	was coming into	entrava	Y		stava arrivando	N	V Prep	V	V

Figure 4: Annotation with POS information

A sample of about 500 MWEs incorrectly translated into Italian are taken into account. The mistranslations occur mainly with nominal and verbal MWEs. Discontinuous MWEs are mainly verbal ones and account for about 10% of translation errors. Examples of wrong translation correspondences for discontinuous MWEs are:

– [Verb … Adjective] as in *Not even the truth will **set** them **free** → Nemmeno la verità li **renderà libero*** (instead of *Neanche la verità riesce a **liberarli***)
– [Verb … Noun] as in *Is there any chance that politicians, that the country generally, would take a finding like that seriously and **run** public **policy** based on it? → C'è una possibilità è che i politici, che il paese generalmente, vorrebbe una scoperta simile seriamente e **correre politica** pubblica basato su?* (instead of *Esiste la possibilità che i politici, e la nazione in generale, possano prendere una scoperta come quella seriamente e **portare avanti** una **politica** pubblica basata su di essa?*)
– [Verb … Particle] as in *I'll **get** my sleeve **back**. → **Prenderò** mia manica* (instead of ***Tiro su** la manica*.)

The Table below shows the most mistranslated MWEs, in absolute terms.

Table 1. Translation errors per source MWE

Source MWE	#
Noun Noun	98
Verb Particle	86
Adjective Noun	54
Verb Preposition	36
Verb Noun	21
Verb Adverb	12
Verb … Noun	12

On the other hand, if we take into account the correspondences between source and target MWEs the picture changes. There are 262 different types of source-target MWE correspondences in the selected corpus and the most frequent mistranslations concern the following translation asymmetries:

Table 2. Translation errors per source-target MWE correspondences

Source MWE	Target MWE	#
Verb Particle	Verb	54
Noun Noun	Noun Preposition Noun	50
Adjective Noun	Noun Adjective	25
Noun Noun	Noun Adjective	25
Noun Noun	Noun	17
Verb Preposition	Verb	14

The one-to-many correspondences produce incorrect translations only in very few cases and concern the following structures [Verb → Verb Noun], [Adjective → Adjective Adverb], [Noun → Noun Adjective], and [Verb → Verb Preposition Noun].

Mistranslation due to many-to-one correspondences are numerous (153): the majority (140) are due to 2:1 correspondences and include 35 different types of correspondences among which the ones which produce the highest number of translation errors are:

– [Verb Particle → Verb] correspondence (54 translation errors), such as in *We **put out** a lot of carbon dioxide every year → Abbiamo **messo fuori** un sacco di anidride carbonica ogni anno*. (instead of *noi **emettiamo** molta co2 ogni anno)*
– [Noun Noun → Noun] correspondence (17 translation errors), such as in *I decided I was going to become a **scuba diver** at the age of 15. → Ho deciso che sarei diventato un **tuffatore bombole** all'età di 15 anni*. (instead of *Ho deciso che sarei diventato un **sommozzatore** all'età di 15 anni*.)
– [Verb Preposition → Verb] correspondence (14 translation errors), such as […] *You are not going to **get to** the correct answer. → […] Non vanno a **raggiungere** la risposta giusta* (instead of *Non potrete **ottenere** la risposta corretta*.)

The occurrence of mistranslation in many-to-many correspondences is shown in Figure 5. These types of correspondences represent the widest group with 378 translation errors in the corpus.

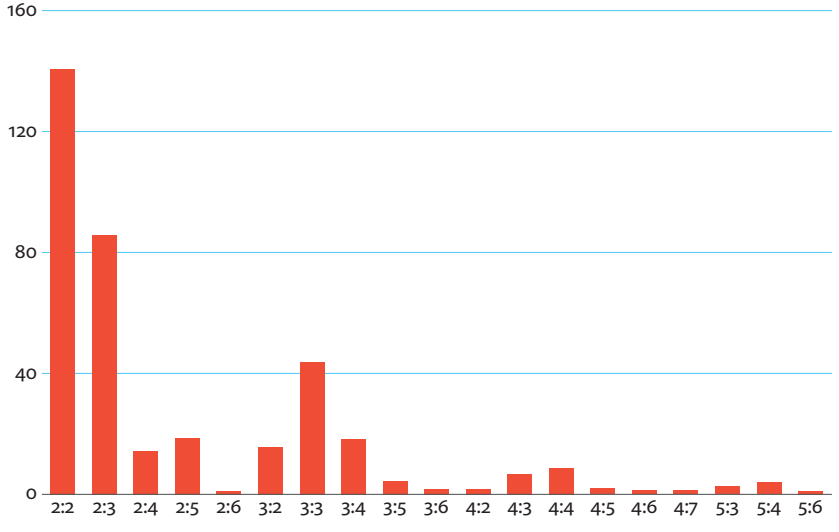


Figure 5. Translation errors per many-to-many correspondences

The main sources of errors are represented by:

- 2:2 correspondences with 141 translation errors, among which the [Adjective Noun → Noun Adjective] correspondence is the most problematic case (25 translation errors) as in *It struck me how much this dive, these deep dives, was like a **space mission**. → Mi colpì quanto questa immersione, queste immersioni profonde, era come uno **spazio missione***. (instead of *Sono rimasto fulminato da quelle immersioni profonde, era come una **missione spaziale***.)
- 2:3 correspondences with 86 translation errors, among which the [Noun Noun → Noun Preposition Noun] correspondence causes 50 errors, such as in *It's a **fish farm** in the southwestern corner of Spain. → È un **pesce fattoria** in un angolo sudovest della Spagna* (instead of *È un **allevamento di pesci** nell'angolo sudoccidentale della Spagna*)
- 3:3 correspondences with 44 translation errors: this translation asymmetry shows a high grade of variability with 40 different correspondences. An example is the [Noun Noun Noun → Noun Noun Adjective] asymmetry, as in *5.93 million years ago was when our earliest **primate human ancestors** stood up. → 5.93 milioni di anni fa era quando i nostri primi **primate antenati umani** si alzò*. (instead of *5.93 milioni di anni fa fu il periodo i nostri **antenati primati umani** si alzarono in piedi*).
- 3:4 correspondences with 19 translation errors, among which the [Adjective-Noun Noun → Noun Preposition Noun Adjective] one represent the most troublesome class as in *I hope that you will agree with me that gamers are a human resource that we can use to do **real-world work** → Spero che sarete d'accordo con me che giocatori sono una risorsa umana che possiamo usare per fare **funzionare reale*** (instead of *Spero che siate d'accordo con me che i giocatori abituali sono una risorsa umana che possiamo utilizzare per fare del **lavoro nel mondo reale***)
- 2:5 correspondences with 19 translation errors, among which there are nine different correspondences. An example is the [Verb Particle → Verb Preposition Determiner Adjective Noun] correspondence as in *If you're getting queasy, **look away** → Se vi steste queasy, **guarda***. (instead of *Se vi sentite male **guardate da un'altra parte***.)
- 3:2 correspondences with 16 translation errors. An example is the [Verb Particle Noun → Verb … Noun] correspondence, as in *He **set up** a **camera** in front of gamers. → Così **ha creato** una **telecamera** davanti ai giocatori mentre erano giocare*. (instead of ***Ha messo** una **telecamera** di fronte ai giocatori*).
- 2:4 correspondences with 15 translation errors. An example is [Noun Noun → Noun Preposition Noun Adjective], as in *The average young person today in a country with a strong **gamer culture** will have spent 10,000 hours playing online → A media oggi giovani in un paese con un forte **giocatore culture** avranno speso 10.000 ore davanti giochi online dall'età di 21 anni*. (instead of *Il tipico giovane medio oggi giorno in un paese con una forte **cultura di giocatore abituale**, avrà passato 10.000 ore giocando online, all'età di 21 anni*.)

7. Conclusions and future work

In this chapter, we have dealt with the concept of translation asymmetries of multi-word expressions in Machine Translation with reference to the English-Italian language pair. The study is based on the analysis of the TED-MWE corpus, containing MWE-annotated sentences of an English-Italian parallel corpus, complemented and compared with an Italian MT output also annotated with MWEs. The MT output has been further analysed in terms of translation divergences, looking at the correspondence patterns between the two languages under examination.

The rationale for taking on translation asymmetries is to observe the cases where structures of both source and target language are divergent, and where these divergences are the cause of mistranslations. This analysis might prove to be useful in relation to better MWE processing and translation, since it conducts a thorough analysis of the patterns which may create problems to an accurate and fluent MT output.

Future work will concern a more fine-grained analysis of the types of errors that occur for different translation asymmetries, which cause a one of the largest translation error class. This analysis will help researchers to understand whether specific translation asymmetries are related to specific error typologies.

References

Mihael, A., Turchi, M., Tonelli, S., & Buitelaar, P. (2017). Leveraging bilingual terminology to improve machine translation in a CAT environment. In *Natural Language Engineering*, 23(5), 763–788. https://doi.org/10.1017/S1351324917000195

Barreiro, A. (2008). *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation* (PhD Thesis, Universidade do Porto).

Bertoldi, N., Haddow, B., & Fouet, J.B. (2009). Improved minimum error rate training in MOSES. *Prague Bull. Math. Linguistics*, 91(1), 7–16.

Bertoldi, N., Cettolo, M., & Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of Machine Translation Summit* XIV. Nice, France.

Bouamor, D., Semmar, N., & Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey

Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008). A multi-word term extraction program for arabic language. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., & Mercer R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2), 263–311.

Caseli, H., Villavicencio, A., Machado, A., & Finatto, M. J. (2009). Statistically-driven alignment- based multiword expression identification for technical domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications* (pp. 1–8). Singapore: Association for Computational Linguistics.

Cettolo, M., Girardi, C., & Marcello, F. (2012). Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)* (pp. 261–268). Trento, Italy.

Clark, J., Dyer, C., Lavie, A., & Smith, N. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*-Volume 2 (pp. 176–181). Association for Computational Linguistics.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Rasmich, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: a survey. *Computational Linguistics*, 43(4), 837–892. https://doi.org/10.1162/COLI_a_00302

Dagan, I., & Church, K. (1994). Termight: Identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing* (pp. 34–40). Association for Computational Linguistics. https://doi.org/10.3115/974358.974367

Daille, B. (2001). Extraction de collocation à partir de textes. In *TALN 2001 (Traitement automatique des langues naturelles*).

Dorr, B. J. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4), 597–663.

Marcello, F., Bertoldi, N., & Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association* (pp. 1618–1621). Brisbane, Australia.

Fu, B., Brennan, R., & O'Sullivan, D. (2009). Cross-lingual ontology mapping–an investigation of the impact of machine translation. In A. Gómez-Pérez, Y. Ding, Y. Yong (Eds.), *Asian Semantic Web Conference* (pp. 1–15). Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-10871-6_1

Kauffmann, A., & Azar, J. (2013). *Structural Asymmetries in Machine Translation: The case of English-Japanese*. (PhD Thesis, University of Geneva).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 177–180). Prague, Czech Republic: ACL.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*-Volume 1 (pp. 48–54). Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit* (pp. 79–86). Phuket, Thailand: AAMT.

Lambert, P., & Banchs, R. (2005). Data inferred multi-word expressions for statistical machine translation. In *Proceedings of Machine Translation Summit* X (pp. 396–403).

Lin, S.-C., Wang, J.-C., & Wang, J.-F. (2005). Translation Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation. In *Proceedings of 17th Conference on Computational Linguistics and speech Processing (ROCLING 2005)*. Tainan, Taiwan.

Losnegaard, G. S., Sangati, F., Parra Escartín, C., Savary, A., Bargmann, S., & Monti, J. (2016). PARSEME Survey on MWE Resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia.

Melamed, I. D. (1997). Automatic discovery of noncompositional compounds in parallel data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 97–108).

Moirón, B. V., & Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word expressions in a multilingual context* (pp. 33–40).

Monti, J., & Todirascu, A. (2016). Multiword units translation evaluation in machine translation: another pain in the neck? In G. Corpas Pastor, J. Monti, R. Mitkov, & V. Seretan (Eds.), *Workshop proceedings for Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2015)* (pp. 25–30). Geneva: Editions Tradulex.

Monti, J., Sangati, F., & Arcan, M. (2015). TED-MWE: a bilingual parallel corpus with MWE annotation. In C. Bosco, S. Tonelli, & F. M. Zanzotto (Eds.), *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)* (pp. 193–197). Torino: Accademia University Press srl/Centro Altreitalie. https://doi.org/10.4000/books.aaccademia.1514

Monti, J. (2013). *Processing in Machine Translation. Developing and using language resources for multi-word unit processing in Machine Translation* (PhD Thesis in Linguistica Computazionale, Università degli Studi di Salerno, a.a. 2011–2012).

Monti, J. (2014). An English-Italian MWE dictionary. In *Clic-it Proceedings* 2014. Pisa University Press srl.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29 (1), 19–51.

Okita, T., & Way, A. (2010). Statistical Machine Translation with Terminology. In *Proceedings of the First Symposium on Patent Information Processing (SPIP)* (pp. 1–8).

Pause, P. E. (1997). Interlingual strategies in translation. In C. Hauenschil, & S. Heizmann (Eds.), *Machine Translation and Translation Theory* (pp. 175-190).

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science*, vol 2276. Berlin/ Heidelberg: Springer. https://doi.org/10.1007/3-540-45715-1_1

Sangati, F., & Cranenburgh A. V. (2015). Multiword Expression Identification with Recurring Tree Fragments and Association Measures. In *Proceedings of Annual conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 10–18). Denver, CO: Association for Computational Linguistics.

Savary, A., Sailer, M., Parmentier, Y., Rosnes, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegaard, G. S., Parra Escartín, C., Waszczuk, J., Constant, M., Osenova, P., & Sangati, F. (2015). PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*. Poznań, Poland.

Seretan, V., & Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)* (pp. 401–410). Toulouse, France.

Sinha, R. M. K., & Thakur, A. (2005). Translation divergence in English-Hindi MT. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005)* (pp. 245–254). Budapest, Hungary.

Thurmair, G., & Aleksić, V. (2012). Creating term and lexicon entries from phrase tables. In *Proceedings of the 16th Conference of the European Association for Machine Translation* (*EAMT 2012*).

Tsvetkov, Y., & Wintner, S. (2010). Extraction of multiword expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 1256–1264). Association for Computational Linguistics.

Vintar, S., & Fišer, D. (2008). Harvesting multi-word expressions from parallel corpora. In *LREC. European Language Resources Association*.

Wu, C.C. & Chang, J. S. (2004). Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses. *Computational Linguistics and Chinese Language Processing*, 9(1), 1–20.

German constructional phrasemes and their Russian counterparts

A corpus-based study

Dmitrij Dobrovol'skij
Russian Language Institute and Institute of Linguistics,
Russian Academy of Sciences / Stockholm University

In this article I examine a group of semi-fixed German expressions that are irregular with regard to the relationship between form and meaning, namely constructional phrasemes with the deictic elements *her* 'hither' and *hin* 'thither' [*vor sich her* + V] and [*vor sich hin* + V]. These constructions pose considerable difficulties not only for the description of their semantics, but also for translation into other languages. Languages such as Russian, English and French do not have exact equivalents of the German deictic elements *hin* and *her*. In cases where the German deictic elements *her* and *hin* are constituents of relatively fixed and irregular constructions, their meaning fits even less well their standard definition. Using corpus examples, I propose a means of describing these constructional phrasemes in a German-Russian dictionary.

Keywords: constructional phraseme, German, Russian, corpora, deictic elements, lexicography, phraseology, construction grammar

1. Introduction

This paper presents some findings of my research on constructional phrasemes. *Constructional phrasemes* are brought into consideration as one of the most significant classes of multiword expressions in (Dobrovol'skij, 2011). This class is related to the *formal* or *lexically open idioms* which are defined as "syntactic patterns dedicated to semantic and pragmatic purposes not knowable from their form alone" (Fillmore et al., 1988, p. 505); cf. also the class of *schematic idioms* addressed in (Croft and Cruse, 2004, p. 248). I describe constructional phrasemes as syntactically autonomous expressions with a fixed composition in which certain

https://doi.org/10.1075/ivitra.24.03dob

slots have to be filled; for more detail see (Dobrovol'skij, 2011; 2015). Russian constructional phrasemes are treated in (Baranov and Dobrovol'skij, 2013, pp. 86–90). This category should be considered a subclass of constructions in the sense of Goldberg (1995).

In what follows I focus on German constructional phrasemes with deictic elements and their Russian equivalents.

2. German deictic elements *hin* and *her*: Semantics and combinatorial potential

The German system of linguistic means for expressing spatial deixis exhibits a variety of non-trivial characteristics. The fact that there is an enormous number of adverbials (often as components of verbal word-formation or co-occurrences with varying degrees of fixedness and idiomaticity) with the deictic elements *hin* 'thither' and *her* 'hither' indicates that this kind of deixis has practically been grammaticalised in German. In Russian (as in English, French and many other languages) there are no systematically organised means for expressing these meanings. An additional problem connected with deictic words of this type is that their description in German dictionaries and grammars is incomplete and in some senses even misleading.

The semantics of these deictic elements has traditionally been described as indicating "toward the speaker" for *her* and "away from the speaker" for *hin*. Such characterisations, however, distort somewhat the real state of things. Thus someone who knocks on the door to be admitted to a room is more likely to ask *Darf ich herein?*, literally, "May I come in hither?" – that is, toward a place where the speaker him- or herself is located – than *Darf ich hinein?*, literally, "May I come in thither?", or toward a place where the speaker has not yet arrived. Thus the traditional rule is clearly being violated. In other words, it is not only the speaker that can be the subject of deixis; in certain situations "deictic authority" can be delegated to another person (for more detail, see Dobrovol'skij and Padučeva, 2008). In cases where the German deictic elements *hin* and *her* are used within relatively fixed and irregular constructions, their meaning conforms even less well to the standard definitions.

The present study will examine the two constructional phrasemes *vor sich hin* (literally "before/in front of oneself thither," which means approximately 'to/for oneself, quietly, not intensely') and *vor sich her* (literally, "before oneself hither", the basic meaning of which is 'from before / in front of oneself in the same direction'). The direction indicated by *hin* and *her* in these constructional phrasemes would seem to be opposite the expected one; that is, these deictic elements are used precisely the other way round. See contexts (1) and (2):

(1) *Er verflucht den unglücklichen Zufall in den ersten Jahren laut, später, als er alt wird,* ***brummt*** *er nur noch* ***vor sich hin****.* (RNC: F. Kafka. Der Prozess)
'Over the first few years he curses his unhappy condition out loud, but later, as he becomes old, he just *grumbles to himself*.' (trans. David Wyllie)

(2) *Frauen, Männer und Kinder* ***schieben*** *Rollwagen mit Grills, Herdplatten, Friteusen oder ganze Garküchen* ***vor sich her****, reihen sich nebeneinander auf und beginnen zu kochen.* (DWDS: Zeit-Corpus 2008)
'Women, men and children *push in front of them* carts with grills, cooktops, deep fryers or entire food stalls, line up next to each other and begin to cook.'

The speaker here is obviously irrelevant, since both examples are narrative contexts. It would seem that the closest contender for the role of subject of deixis should be the subject of the action. In which case, based on the meaning of the constructions *vor sich hin* and *vor sich her*, one might expect that *hin* and *her* would exchange places. Thus the people pushing carts in front of them (context 2) are clearly executing a movement away from rather than toward themselves. And in accordance with the standard rule, motion away from oneself, i.e. 'thither', (in contrast to motion 'hither') should be marked by means of *hin* rather than *her*. Why then does *schieben* 'push, move' normally collocate with *vor sich her* rather than *vor sich hin*? And conversely, the man grumbling to himself (context 1) is performing an action directed not out from but into himself, as it were. In that case it would seem more natural to mark the direction of the action with *her* rather than *hin*, but in fact replacing *hin* with *her* is impossible here.

This "deictic paradox" demands commentary, and Dobrovol'skij and Padučeva (2008) propose an explanation of the phenomenon. Other interpretations are, of course, possible, including the assumption that the deictic elements in these constructions are completely demotivated, i.e. have dissolved, as it were, in the meaning of the whole. I shall not dwell on this problem here, for the present study has a different goal. Employing the data of German-Russian parallel corpora, I want to identify the means of translating these constructions into Russian and refine their combinatorial profile and meaning structure.

The German-Russian and Russian-German parallel corpora of the Russian National Corpus (RNC) are still relatively small, consisting of some 10 million running words. Corpora of that size do not enable us to conduct a statistical analysis that would produce representative findings. Nevertheless, even relatively limited parallel corpora can yield useful and significant information about the meaning and functioning of linguistic units, especially if these units have not yet been subjected to thorough semantic and/or contrastive analysis. This is fully applicable to constructions with a varying degree of fixedness and idiomaticity, especially constructional phrasemes – that is, phrasemes with at least one lexically fixed

constituent and with empty slots that can be filled by various words.[1] Depending on how a slot is filled, the meaning of the constructional pattern underlying such phrasemes changes, as do the means for translating the phraseme into another language.

The use of parallel corpora allows us not only to identify appropriate means of translation, but also to refine our notions of the semantics and co-occurrence of the German constructional phrasemes under investigation. From this perspective, the fact that we cannot apply statistical methods does not prevent us from achieving our objectives, since quantitative analysis is not among the tasks addressed in the present study.

My investigation makes a certain contribution to Construction Grammar and the theory of phraseology, since the more concrete constructions in different languages that can be described through the use of authentic corpus data, the more successful will be the development of a typology of constructions. In other words, it will become clearer what types of constructions exist in different languages and how language-specific or universal they are. The class of constructional phrasemes has long remained on the periphery in the description of phraseology. Yet corpus data shows that this class is no less important to communication than idioms. Finally, the present study is of immediate value to bilingual lexicography.

Data was drawn especially from the corpora of parallel texts in the Russian National Corpus (RNC). Monolingual German corpora were also used: *Das Deutsche Referenzkorpus* (DeReKo) of the Institute of German Language in Mannheim and the *Corpora des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts* (DWDS) of the Berlin-Brandenburg Academy of Sciences.

The study consists of two parts. The first (Section 3) examines the constructional phraseme *vor sich her* within a group of related constructions, and the second (Section 4) treats *vor sich hin*.

3. Construction *vor sich her*: Underlying pattern, semantics and Russian counterparts

This construction is a particular instance of the pattern [*vor* N_{dat} *her*], the meaning of which is not described in dictionaries and cannot be derived from the semantics of its constituents: *vor* N_{dat} 'before / in front of sb' + *her* 'hither' (in the direction of

1. Cf. the notion of constructional idiom in (Booij, 2013, p. 258): "A constructional idiom is a (syntactic or morphological) schema in which at least one position is lexically fixed, and at least one position is variable".

the subject of deixis.[2] Often the N_{dat} slot is filled by the reflexive pronoun *sich*. See context (3). Because *sich* is followed by the dative, in the first and second persons the construction has the form *vor mir/dir/uns/euch her*.

(3) *Das Tier, auf dessen Rücken ich saß, schwamm nicht, sondern lief mit unglaublicher Geschwindigkeit auf dem Grunde des Meeres weg und* ***trieb*** *Millionen von Fischen* ***vor sich her****.*

(RNC: G.A. Bürger. Die Abenteuer des Freiherrn von Münchhausen)

Животное, на спине которого я сидел, не плыло, а с неимоверной быстротой бежало по морскому дну, ***гоня перед собой*** *массу всяких рыб.*

(пер. на рус.: В. Вальдман)

'The animal I rode did not swim; he galloped with incredible swiftness along the bottom, *driving before him* millions of fish.'

The data indicates that the meaning of *vor sich her* and, consequently, its translation into Russian, vary depending on the semantic class of the verb with which the expression co-occurs. Thus for my purposes it is more convenient to speak about the construction [*vor sich her* + V].

The adverbial form *vor sich her*, which is the nucleus of the constructional phraseme [*vor sich her* + V], has a basically spatial meaning. Here we are concerned with a realisation of the pattern [$Prep_{spatial}$ N_{dat} *her*]. Referring to (Marcq, 1988), Krause and Baerentzen (2010, pp. 21, 43, 46) observe that constructions such as [*hinter* N_{dat} *her*], [*vor* N_{dat} *her*], [*neben* N_{dat} *her*] and [*zwischen* N_{dat} *her*] have meanings of their own and express the idea of joint forward motion in one direction and at the same speed. It is important to note that this is the meaning of the construction and not the meaning of the phrasal verb *her*-V combined with *vor/hinter/zwischen/neben sich* 'before <in front of> / behind / between / alongside <next to / beside> oneself'.

> Es ist also wenig nützlich, Partikelverben anzunehmen wie *herrennen, herlaufen, herstiefeln, herstolpern, herstöckeln, hertorkeln, herkriechen* in den Fällen, wo man es mit den Kombinationen *vor* + Dativ, *hinter* + Dativ und *neben* + Dativ + *her* zu tun hat, die alle eines gemeinsam haben, nämlich auszudrücken, dass zwei Teilnehmer sich in der gleichen Geschwindigkeit fortbewegen. Das Verb präzisiert nur, auf welche Art und Weise das geschieht, aber es ist zum Ausdruck dieser Relation nicht notwendig. Wörterbücher[…] und Grammatiken tendieren zur Einordnung als Verbalpartikel, müssen sich dann aber den Vorwurf gefallen lassen, dass sie sehr unvollständig sind – und auch unlogisch, denn semantisch betrachtet gehört *her* zu *hinter* / *vor* / *neben* + Dativ und nicht zum Verb.
>
> (Krause and Baerentzen, 2010, p. 21)

2. In the canonical communicative situation this is the speaker; in hypotaxis it is the subject of the matrix clause; in a narrative, the narrator or a character from whose perspective events are viewed.

'It is hardly useful to assume the presence of the verbs *herrennen, herlaufen, herstiefeln, herstolpern, herstöckeln, hertorkeln, herkriechen* in all cases in which we are concerned with combinations of *vor* + dative, *hinter* + dative and *neben* + dative. Common to all these instances is the meaning that two participants of the situation are moving at the same speed. The verb merely specifies how this movement occurs, but has nothing to do with the idea itself of this relationship. Dictionaries and grammars prefer to interpret these constructions on the basis of the phrasal verb, but they are forced in that case to concede that the description is incomplete and illogical, since semantically *her* belongs to *hinter/vor/neben* + dative, and not to the verb.'

3.1 The constructional phraseme [*vor sich her* + v] and its underlying pattern

This constructional phraseme must be analysed among similar realisations of the pattern [$Prep_{spatial}$ N_{dat} *her*] that underlies it. Let us examine in more detail three of the possible such realisations: [*vor* N_{dat} *her* + V], [*hinter* N_{dat} *her* + V] and [*neben* N_{dat} *her* + V]. They were selected because there was a sufficient number of their instances in the RNC. I will first dwell on the expressions [*hinter sich her* + V] and [*neben sich her* + V] (see Examples (4)–(6)) and then go on to [*vor sich her*+ V]. Since all three expressions are based on the pattern [$Prep_{spatial}$ N_{dat} *her*], meaning joint spatial movement, it is natural to expect that they will co-occur with verbs of motion indicating both independent forward movement (4)–(5) and the causation of motion (6)–(7). This is basically the case.

(4) *Willy **geht neben mir her**. Was für ein Leben!*
(RNC: E.M. Remarque. Der schwarze Obelisk)
*Вилли **шагает рядом со мной**. – Что за жизнь!*
(пер. на рус.: В. Станевич)
'Willy *is walking beside me*. What a life!' (trans. Denver Lindley)

(5) *Er **saust** schon die ganze Zeit wie blödsinnig **hinter mir her!***
(RNC: E. Kästner. Emil und die Detektive)
*Он давно уже **гонится за мной**, словно бешеный.*
(пер. на рус.: Л. Лунгина)
'He has been *chasing me* like mad for a long time!'
(trans. Margaret Goldsmith)

(6) *Er **schleppte** einen Kranz **hinter sich her**.* (RNC: F. Dürrenmatt. Justiz)
***За собой** Эшисбургер **тащил** венок.* (пер. на рус.: С. Фридлянд)
'He *was dragging* a wreath *behind him*.'

(7) *[…] so daß ereignisreiche Jahre viel langsamer vergehen als jene armen, leeren, leichten, die der Wind* ***vor sich her bläst****, und die verfliegen.*
(RNC: Th. Mann. Der Zauberberg)
[…] и годы, богатые событиями, проходят гораздо медленнее, чем пустые, бедные, убогие; их как бы ***несет*** *ветер, и они летят.*
(пер. на рус.: В. Станевич)
'[…] years rich in events pass much more slowly than do paltry, bare, feather-weight years that are *blown before the wind* and are gone.'
(trans. John E. Woods)

It is obvious from an analysis of these examples that there is no single way of translating *vor sich her, hinter sich her* and *neben sich her*. This fact alone confronts the lexicographer with the need to search for non-trivial means for registering these units in a bilingual dictionary. The first question to arise is which meaning of the word *her* we should attach to the given expressions.

The problem is that independently of the verbs that govern them, adverbial constructions of the type *vor sich her, hinter sich her* and *neben sich her* are scarcely translatable into Russian. The only reasonable solution here lies in describing their semantics and translating the instances of the V-construction.

Of the contexts in the RNC, the most frequent realisation of the construction [*hinter sich her* + V] was *hinter sich ziehen* 'pull behind oneself' while there was one occurrence each of the verbs *schleppen* 'drag', *schleifen* 'pull, drag', *zerren* 'haul', *locken* 'entice, lure' and *sausen* 'rush'. Only *sausen* (see Example 5 above) is a verb of motion, while all the others denote causation of motion. Russian translations of these forms can be tentatively divided into two groups: those which do not indicate any direction of movement (8)–(10) and those which do (11)–(13).

(8) *Er wartete nicht ab, was Momo dazu sagen würde, sondern* ***zog*** *sie an der Hand* ***hinter sich her*** *zum Auto.* (RNC: M. Ende. Momo)
Не дожидаясь ответа, он ***втащил*** *Момо в машину.*
(пер. на рус.: Ю.И. Коринец)
'Without even waiting for an answer, he *seized* Momo's hand *and towed her to* the car.' (trans. J. Maxwell Brownjohn)

(9) *[…] und* ***locken*** *den Rasenden heimtückisch* ***hinter sich her*** *in dunkle Gänge.*
(RNC: G. Meyrink. Der Golem)
[…] коварно ***заманивают*** *безумного в темные коридоры.*
(пер. на рус.: Д. Выгодский)
'[…] when he is beside himself with fury, slyly *lure* him *into* dark corridors.'
(trans. Mike Mitchell)

(10) *[…] und* ***zog*** *Marie am Arm* ***hinter mir her****, hinaus.*
(RNC: H. Böll. Ansichten eines Clowns)
[…] ***схватил*** *Марию за руку и* ***потащил*** *ее из дома.*
(пер. на рус.: Р. Райт-Ковалева)
'[…] and *pulled* Marie *after me* by the arm, out.' (trans. Leila Vennewitz)

All of the translations in (8)–(10) are adequate in the sense that they convey the basic meaning of the original texts, but they cannot be used in a bilingual dictionary as Russian equivalents of the given German expression. Thus *он втащил Момо в машину* 'he towed her to the car' in (8) does not describe the movement of the Agent, who was himself getting into the car at the moment this action was performed, whereas on the basis of the Russian translation one might think that he was already in the car. In (9) *коварно заманивают безумного в темные коридоры* 'slyly lure him into dark corridors' can be construed as meaning that the enticers are located somewhere outside, but *hinter sich her locken* clearly indicates that they are themselves in the dark corridors and are luring someone to follow them. It is clear from the context (10) that the Agent was himself leaving the house at the moment he pulled Marie by the arm out of the building (*схватил Марию за руку и потащил ее из дома*), but the expression *потащил ее из дома* 'pulled her out of the building' corresponds to the German *zog sie hinaus* 'pulled her out'. Thus the idea contained in the adverbial *hinter mir her* remains unexpressed.

(11) *Durch die halboffene Tür schlüpfte eine Frauengestalt und* ***zog*** *ein Kind* ***hinter sich her****.* (RNC: E. Kästner. Pünktchen und Anton)
В приотворенную дверь протиснулась какая-то женщина. ***За собой*** *она* ***вела*** *ребенка.* (пер. на рус.: Е. Вильмонт)
'A woman slipped through the half-open door, *leading* a child *behind her*.'

(12) *Johann Buddenbrook kam zornig herbei, den Kerzenlöscher* ***hinter sich her schleifend****.* (RNC: Th. Mann. Buddenbrooks)
Иоганн Будденброк направился к сыну, сердито ***волоча за собой*** *гасильник.*
(пер. на рус.: Н. Ман)
'Johann Buddenbrook walked over angrily, *dragging* the candel-snuffer *behind him*.' (trans. John E. Woods)

(13) *„Aber gewiss", rief die Frau und* ***zog*** *ihn eiligst* ***hinter sich her****.*
(RNC: F. Kafka. Der Prozess)
– Ну конечно же, – воскликнула она и торопливо ***потянула его к столу****.*
(пер. на рус.: Р. Райт-Ковалева)
'"Yes, certainly," the woman cried, and *pulled* K. *along behind her* as she rushed to them [the books on the table].' (trans. David Wyllie)

In contrast to Examples (8)–(10), the Russian translations of *hinter sich her* in contexts (11) and (12) can be regarded as candidates for inclusion in a dictionary. The forms *вести за собой* 'lead after oneself' and *волочить за собой* 'drag after oneself' fairly accurately convey the notion of parallel movement contained in the German expressions *hinter sich her ziehen* and *hinter sich her schleifen*. The translation of (13) is interesting primarily because it shows the potential variety of ways to get across meanings that Russian has no standard means for conveying. The direction of the caused movement is conveyed here through *к столу* 'toward the table', which does not occur in the German original.

Deserving special attention are contexts in which the construction [*hinter sich her* + V] is used not to denote motion in the strict sense, but in the derived meaning of occurring after some event; see (14).

(14) *Der Golem* ***zieht*** *eine unheimliche Gefolgschaft* ***hinter sich her***.
(RNC: G. Meyrink. Der Golem)
За *Големом всегда такая страшная* ***свита*** *событий*.
(пер. на рус.: Д. Выгодский)
'The Golem *brings* some macabre things *in its wake*.' (trans. Mike Mitchell)

It is entirely possible to regard this sort of use as an idiomatisation of the original expression. When *hinter sich her ziehen* is employed not to denote the causation of motion,[3] but in a metaphorical sense based on a transfer from the spatial to the temporal sphere, the expression can be viewed as a potential idiom. This example demonstrates one possible way in which weakly idiomatic expressions (here, constructional phrasemes) can become idioms. How they are translated differs from the means used in Russian to convey the original meaning of the expression. In addition to the literary solution proposed in (14), other possible Russian equivalents include the verb phrase *вести за собой* 'lead to, entail', the collocation *иметь своим следствием* 'have as a consequence', etc.

3.2 Semantic derivation, construction polysemy and lexicographic description

Even more interesting in this regard is the central object of my investigation: the constructional phraseme [*vor sich her* + V]. It serves as the basis of the full-fledged idiom *etw. vor sich her schieben* – literally, "push sb ahead of oneself" – meaning 'to postpone or defer indefinitely some action or decision because one is unable to mobilise one's inner resources'. The Duden phraseological dictionary (Duden 11) does not feature this idiom, in spite of the fact that it is fairly common in both oral

3. The standard Russian translation in that case is *тащить за собой* 'drag after oneself'.

and written speech and is established in the language system as an independent lexical unit with a metaphorical meaning. See (15):

(15) *Der Bund **schiebt** einen seit Jahrzehnten angehäuften Schuldenberg von mehr als 900 Milliarden Euro **vor sich her**.* (DWDS: Zeit-Corpus 2008)
'The federal government *is deferring* action on a debt accumulated over decades of more than 900 billion Euros.'

The idiom *etw. vor sich her schieben* arose from a semantic reinterpretation of the construction [*vor sich her* + $V_{CausMotion}$], that is, a realisation of the constructional phraseme [*vor sich her* + V] in which the verbal slot is filled by a verb denoting causation of motion. On the whole, the group of constructions described here can be represented in the form of a chain of successive stages of lexico-semantic specification: [$Prep_{spatial}$ N_{dat} *her*] → [*vor sich her*] → [*vor sich her* + V] → [*vor sich her* + $V_{CausMotion}$] → [*vor sich her schieben*] → [IDIOM [*vor sich her schieben*]].

Analysis of examples drawn from the parallel corpus shows that the constructions of this group exhibit a language-specific configuration of semantic features. Consequently, they do not have simple Russian equivalents. As the corpus data indicates, in translation this meaning is often not conveyed at all, or it is distributed among other elements of the context or is clear from the situation. In a number of cases the expression (*прямо*) *перед собой* '(directly) in front of oneself' corresponds to the adverbial *vor sich her*. Thus Russian translations convey the purely spatial component of the semantics of this construction (the locus of the object vis-a-vis the subject), while the notion of 'parallel motion' remains unexpressed.

It turns out that it is rather difficult to find translation equivalents suitable for inclusion in a dictionary. I have in addition drawn upon materials from the Russian-German corpus of parallel texts. The use of one or another linguistic unit in the text of a translation is of great diagnostic value for identifying the particular semantics and pragmatics of the unit. This has to do with the fact that when a translator intuitively senses the need to use some expression even if there is no equivalent in the original text, s/he has identified the meanings on which the expression is focused. Contexts such as (16) demonstrate that the German construction [*vor sich her* + V_{Motion}] indicates parallel motion at the same speed.[4]

(16) – *Говорю вам: **впереди меня шла**, шатаясь, тут же на бульваре.*
(RNC: Ф.М. Достоевский. Преступление и наказание)
„*Ich sage Ihnen ja: sie **ging** schwankend **vor mir her**, hier auf diesem Boulevard.*"
(Übers.: A. Eliasberg)
'"I tell you she *was walking in front of me*, staggering just here, in the boulevard."'
(trans. Constance Garnett)

4. As well as [*vor sich her* + $V_{CausMotion}$].

This meaning is not expressed explicitly in Russian, but it can be inferred from general knowledge of the situation.

Translations of the idiom *vor sich her schieben* in the parallel corpus are especially problematic. See (17):

(17) *[…] und dachte an Marie, an die Christen, an die Katholiken und* ***schob*** *die Zukunft* ***vor mir her****.* (RNC: H. Böll. Ansichten eines Clowns)
[…] начал думать о Марии, о протестантах, о католиках и о будущем. (пер. на рус.: Р. Райт-Ковалева)
'[…] and thought of Marie, of Christians, of Catholics, and *contemplated* the future.' (trans. Leila Vennewitz)

The problem is that a translator – particularly a first-class translator, as in the case of (17) – can have individual priorities when translating certain text fragments. It is for that reason that some lexical units – here the idiom under consideration – sometimes remain untranslated. The German original says that the hero thought about Marie and other things but tried not to think about the future, but the translation does not convey this contrast.

As a possible dictionary equivalent of the German idiom I can propose the expressions *откладывать что-л.* 'postpone, defer sth' or *отгонять от себя мысли о чем-л.* 'banish / drive away thoughts about sth').

As for presenting constructional phrasemes of the type [$Prep_{spatial}$ *sich her* + V] in the dictionary, the most economical and user-friendly way of describing it is to indicate the meaning of the pattern underlying it [$Prep_{spatial}$ N_{dat} *her*] and enumerate all the prepositions that can appear in the position $Prep_{spatial}$. For example, *the expressions* ***hinter*** *jmdm.* ***her*** 'behind sb', ***vor*** *jmdm.* ***her*** 'before / in front of sb', ***neben*** *jmdm.* ***her*** 'next to sb' *and* ***zwischen*** *jmdm.* ***her*** 'between sb' *denote movement by the participants of the situation in the same direction and at the same speed*. It is important to include in the same zone of the entry examples illustrating the basic ways to use these constructions and translate them into Russian.[5]

4. The construction *vor sich hin*: Semantics, co-occurrence types and Russian counterparts

Let us begin with what is known about this construction from lexicographic sources. It is presented twice in (Duden GWDS): at *vor* (a) and at *hin* (b):

5. Cf. Steyer (2015, p. 282): "MWEs [multiword expressions] usually have multiple entries in the mental lexicon, on the one hand, as more or less specified lexical units (lexemes) and on the other hand, as (proto)typical realisations of a more abstract MW pattern".

a. **vor sich hin** (*ganz für sich u. in gleichmäßiger Fortdauer*): vor sich hin schimpfen, reden, weinen;
b. **vor sich hin** (*ohne die Umwelt zu beachten, für sich*) murmeln, reden, gehen (the meaning of *hin* in this co-occurrence is incorrectly described as "drückt die Erstreckung aus," i.e. 'indicates length or distance').

From this description, it can be inferred that the construction contains the semantic features 'duration' and 'introverted nature of the action' (that is, it is directed "inward" rather than "outward"). On the one hand, these semantic features are presented quite adequately. On the other, however, we can see from even random examples of the construction used with different verbs that this description is not exhaustive; see contexts (18) and (19):

(18) *Die Schlafkrankheit hat ihren Namen dadurch erhalten, dass die Kranken* ***vor sich hin dämmern***. (http://www.netzeitung.de/genundmensch/237937.html)
'Sleeping sickness gets its name from the fact that the afflicted *are semi-conscious*.'

(19) *Er* ***nickte vor sich hin*** *und sagte abschließend: „Das ist wichtig.*"
(RNC: M. Ende. Momo)
'He *nodded to himself*. "That's important, too."' (trans. J. Maxwell Brownjohn)

Context (18) has to do with patients who are in a not fully conscious state – it is hardly a question of deliberately ignoring the surrounding world or an action directed inward, but a state that the subject does not fully control. As for 'duration', this semantic component is not pertinent to (19).[6]

To deal with the semantics of *vor sich hin* we must attempt to systematise the verbs that can co-occur with this construction and identify the relevant semantic features of each of these constellations. In other words, at least at the first stage of the investigation what must be analysed is not the adverbial construction *vor sich hin*, which to all appearances has such rich and flexible semantics that it is not amenable to any one interpretation, but the verbal construction [*vor sich hin* + V],[7] which due to the presence of a verb of one or another type has been concretised along a certain semantic axis. Consider co-occurrences such as *leise vor sich hin fluchen* 'curse quietly to oneself', *vor sich hin starren* 'stare dead ahead of oneself'

6. All that is indicated in Example (19) is the direction of motion toward oneself. Consider the translation of this context in the RNC: *Он кивнул сам себе и закончил: – Вот что важно*. Note that his translation does not entirely accurately convey the semantics of the predicate *vor sich hin nicken*. *Er nickte vor sich hin* is understood to mean repeated action. In other words, in cases where the feature 'duration' is ruled out by the context, it is generally 'iteration' that is brought to the fore.

7. The fact that a construction can be part of another construction is generally well known and has been repeatedly discussed in the literature. See, for example, (Taylor, 2006).

and *vor sich hin welken* 'wither away', which differ considerably with respect to their semantic potential.

4.1 Types of verbs co-occurring with *vor sich hin*

Let us now turn to the types of verbs that co-occur with the construction *vor sich hin*. Analysis of examples drawn from the corpus shows that verbs of at least the following types are possible here.

i. Verbs that have a certain communicative potential ($V_{CommPotential}$): *sprechen* 'speak', *murmeln* 'murmur, mutter', *nuscheln* 'grumble', *lallen* 'babble', *brabbeln* 'babble', *fluchen* 'curse', *schimpfen* 'scold', *singen* 'sing', *summen* 'hum', *kichern* 'giggle', *grinsen* 'grin', *gähnen* 'yawn', *lachen* 'laugh', *weinen* 'cry', *heulen* 'howl'.

Included here are both "classic" verbs of speaking and verbs denoting physical activity that conveys something about inner states. Normally these actions are registered by persons in the vicinity of the subject, so that the action itself possesses communicative potential. When we yawn, laugh or hum we communicate some information about our own inner state to the other participants of the situation. In such cases the basic function of *vor sich hin* is to "delete" the notion of outward direction. Combined with these verbs, *vor sich hin* weakens or even annuls their communicative potential and transforms them to mean quasi-autistic reactions. Yet another typical semantic feature is 'weak intensity'. The presence of these features can easily be detected in Russian equivalents. See Examples (20) and (21):

(20) *[…] und* ***brabbelte****, wenn die Erinnerung daran zu kräftig in ihm aufstieg, beschwörend „Holz, Holz"* ***vor sich hin****.* (RNC: P. Süskind. Das Parfum)
[…] и, когда воспоминание с новой силой всплывало в нем, ***бормотал про себя****, словно заклиная: «Дрова, дрова».* (пер. на рус.: Э.Венгерова)
'[…] whenever the memory of it rose up too powerfully within him he would *mutter* imploringly, over and over, "wood, wood."' (trans. John E. Woods)

(21) *Gigi* ***pfiff*** *leise ein melancholisches Lied* ***vor sich hin****.* (RNC: M. Ende. Momo)
Джиги тихо ***насвистывал себе под нос*** *грустную песенку.*
(пер. на рус.: Ю.И. Коринец)
'Guido *was whistling* a soft and melancholy tune.'
(trans. J. Maxwell Brownjohn)

The feature 'introversion' (= "for/to oneself, not for/to others") is conveyed by the Russian expressions *про себя* 'silently' and *себе под нос* 'under his/her breath', and weak intensity is indicated by means of the morphological derivation [*на- + (ы)ва-*]; cf. *свистеть* 'whistle' vs. *насвистывать* 'whistle (a tune with weak intensity)', *петь* 'sing' vs. *напевать* 'hum'. See also (22), an untypical example in

which, however, the same principles apply; the Russian translation uses similar word-formation devices: *хрюкать* vs. *похрюкивать* 'grunt'.

(22) *Doch es gibt Betriebe, die in einer eigenen Küche Wurst herstellen. Dran glauben musste dafür das letzte Schwein, das in seinem Stall in der Ladenburger Straße einsam* ***vor sich hin grunzte****.*
(Nach DeReKo: Mannheimer Morgen, 19.08.2011)
Однако есть и предприятия, изготавливающие колбасу на собственной кухне. Ради этого пришлось отдать концы последней свинье, которая ***тихонько похрюкивала себе*** *в хлеву на Ладенбургерштрассе.* (MDRI)
'There are, however, establishments that make sausage in their own kitchens. That is why the last hog *quietly grunting to itself* in its stall on Ladenburger Strasse had to die.'

The semantic realisation of [*vor sich hin* + V] in (22) is close to the version of this construction [*vor sich hin* + $V_{Process/StateAnim}$] discussed below under point (ii). Indirectly this is indicated by the use of the particle *себе* ('self') in the Russian translation equivalent.

ii. Verbs denoting inactive, uncontrolled states of living organisms, usually humans, as well as corresponding indolent, generally uncontrolled processes ($V_{Process/StateAnim}$): *vegetieren* 'vegetate', *welken* 'wither', *dösen* 'doze, drowse, daydream', *kränkeln* 'ail, be unwell', *siechen* 'waste away', *dämmern* 'doze, drowse', *schlummern* 'slumber, doze', *leiden* 'suffer', *gammeln* 'loaf around'.[8]

Here the semantic contribution of *vor sich hin* is especially to intensify the notions of duration, inactivity, and weak control. Thus *dösen* ('to doze') in itself denotes a passive state. *Vor sich hin* emphasises that the subject is in a state that excludes active involvement and control (of both the situation and one's own state) and also stresses the duration of the state. For these reasons the co-occurrence of *vor sich hin* with verbs of type (ii) often produces expressions with a negative connotation. If by itself the verb *dösen* can be translated, for example, as *клевать носом* 'nod off', *vor sich hin dösen* suggests *пребывать в состоянии сонной апатии* 'be in a state of drowsy apathy', as in (23).[9]

---

8. The features [Anim] and [Inanim] (see below) should be understood not in the sense of the grammatical categories of animate and inanimate, but only in the sense of the opposition between a living organism (humans, animals, plants) and inert matter (substances, artefacts, institutions).

9. There are, of course, contexts in which it would be appropriate to translate *vor sich hin dösen* as *клевать носом* 'nod off'. The important point is that such "neutral" interpretations of the state are rather rare.

(23) a. *Julias Exmann entspricht einem klassischen Typ, der sich dem männlichen Ideal in Passivität verweigert und schon in der russischen Literatur* ***vor sich hin döste***. (DWDS: Zeit-Corpus 2008)
'Julia's ex-husband is the type of guy who in his passivity rejects the manly ideal and *was daydreaming* already in Russian literature.'

b. *An den ungeputzten Rotznasen der Kinder kann man erkennen, dass hier die Armut zu Hause ist, und aus dem Anblick der* ***vor sich hin dösenden*** *Erwachsenen spricht Perspektivlosigkeit*. (DWDS: Zeit-Corpus 2008)
'It is evident from the unwiped runny noses of the children that poverty dwells here, and the appearance of the *drowsy* adults speaks of hopelessness.'

Generally speaking, states denoted by type (ii) verbs are regarded as "bad" (even without *vor sich hin*), and processes are viewed as leading to a worsening of the state of the subject. Such, for example, is the verb *welken* 'wither, fade' about plants and figuratively about people. Consider also "bad" states such as *vegetieren* 'vegetate', *kränkeln* 'ail, be unwell' and *siechen* 'waste away'. There are cases, however, where the verb in itself does not mean 'sickly' or 'fading'. Thus in the semantics of the verb *leben* 'live' there is nothing inactive or uncontrollable. It is absolutely neutral in this regard and can only conditionally be included in group (ii). An interpretation of the predicate in this sense is possible precisely because it co-occurs with *vor sich hin*. Consider context (24), where *vor sich hin leben* is construed as 'living passively, without anything special happening, shut up in one's own little world'. The semantic features that this construction introduces into the utterance, therefore, can be (at least approximately) characterised as 'duration', 'passivity', 'externally uncontrollable' and 'negative evaluation'.

(24) *Monsieur Ravel ist hier ein schrulliger Typ mit Einstecktuch und Lackschuh-Manie, ein Neuerer der Herrenmode, der nebenbei auch komponierte und ansonsten eigenbrötlerisch* ***vor sich hin lebte***. (DWDS: Zeit-Corpus 2008)
'Monsieur Ravel is an eccentric guy with a fancy handkerchief and a penchant for leather shoes, an innovator of men's fashion who in addition also composed music and *lived a solitary life*.'

iii. Certain mental verbs and verbs denoting inner states (V_{Mental}); *denken* 'think', *überlegen* 'consider', *sinnieren* 'ruminate', *träumen* 'dream', *brüten* 'brood'.

(25) *Auch Herbert von Karajan* ***träumte*** *beim Auswendigdirigieren meistens nur so* ***vor sich hin***. (DeReKo: A00/JAN.00921)
'Herbert von Karajan also *dreamed*, mostly just *to himself*, while conducting from memory.'

It is not so easy to find a good answer to the question of just what *vor sich hin* contributes semantically to an utterance. Obviously it cannot be the idea of introversion

or direction "into oneself", as in co-occurrences with verbs of the type $V_{CommPotential}$, since verbs meaning 'think', 'dream', 'reflect', etc. already denote mental activity and inner states that by definition cannot be directed outward, and (perhaps with the exception of 'duration') the semantic features identified in co-occurrences with verbs of the type $V_{Process/StateAnim}$ are not relevant to V_{Mental}. This is understandable. The scale of activity-passivity or degree of controllability is difficult to apply to inner states. Nor is negative evaluation an inherent feature of the co-occurrences of group (iii); see context (25). Evidently, what *vor sich hin* brings to an utterance is the idea of ignoring the surrounding world, complete immersion in oneself, as well as a relatively prolonged state. These happen to be precisely the semantic features which – excessively extrapolated – are identified in (Duden GWDS) for all uses of the construction *vor sich hin*. Examples of this are *ganz für sich u. in gleichmäßiger Fortdauer* 'entirely for/to oneself and of uniform duration' and *ohne die Umwelt zu beachten, für sich* 'taking no note of one's surroundings, to oneself)'.

The notion of rejecting any sort of contact with the outside world is clearly observable in Russian translations of co-occurrences of type (iii). See (26):

(26) *Dann setze ich mich in meinen Stuhl und* ***brüte vor mich hin****.*
(RNC: E.M. Remarque. Der schwarze Obelisk)
Потом усаживаюсь в свое кресло и ***погружаюсь в мрачную задумчивость****.*
(пер. на рус.: В. Станевич)
'Then I sit down in my chair and *brood*.' (trans. Denver Lindley)

iv. Verbs relating to inanimate entities and denoting slow, mostly uncontrollable processes and corresponding states ($V_{Process/StateInanim}$): *tröpfeln* 'dribble', *köcheln* 'simmer', *brennen* 'burn', *glimmen* 'glow', *dümpeln* 'roll', *gären* 'ferment', *rosten* 'rust', *modern* 'decay', *plätschern* 'splash', *kriseln* 'go through a crisis'.

(27) *Wenn die Bakterien von keiner Zahnbürste gestört* ***vor sich hin gären*** *können und gleichzeitig genügend Zucker als Rohmaterial bekommen, dann werden große Mengen Milchsäure gebildet, die zuerst den Zahnschmelz angreift.*
(DeReKo: P00/APR.15369)
'If the bacteria are not disturbed by a toothbrush they can *ferment* and if at the same time they get enough sugar as raw materials, large quantities of lactic acid are formed that first attacks the tooth enamel.'

Type (iv) can be regarded as a metaphorical expansion of (ii). When it co-occurs with $V_{Process/StateInanim}$ verbs, therefore, *vor sich hin* exhibits quite a few semantic features that are typical of its co-occurrences with $V_{Process/StateAnim}$. Thus the corresponding processes are thought to lead to a worsening of the subject's condition; in other words, this verb phrase (as in type (ii)) often has the feature 'negative evaluation'. And this feature may already be present in the semantics of the verb itself

(e.g. *modern, dämmern, gammeln*) or it can be added to the construction by *vor sich hin* (as in context 27). The other characteristic features are 'duration', 'externally uncontrollable' and 'weak intensity.' The close semantic connection between [*vor sich hin* + $V_{Process/StateAnim}$] and [*vor sich hin* + $V_{Process/StateInanim}$] is also evident in the fact that the same verbs can be used in both constructions: for example, *kränkeln* 'ail, be unwell', *siechen* 'waste away', *dümpeln* 'stagnate', *dämmern* 'doze' and *vergammeln* 'waste, go bad'. A relationship of metaphorical inheritance is established between the uses of these verbs with an animate and inanimate subject, so that the basic non-trivial semantic features are repeated: for example, *siechen* meaning 'ail, waste away' (about humans) and in the meaning 'languish, grow feeble', as in Example (28):

(28) *Von Kommunikation zwischen den Funktionsträgern des ehedem badischen Musterklubs nicht die Spur. Das zeugt davon, wie marode der Traditionsverein inzwischen* ***vor sich hin siecht****.* (DeReKo: M00/APR.12681)
'Of the communication between the functionaries of the former Baden perfect club there remains not a trace. This shows how seriously the traditional league *has languished* in the meantime.'

With respect to cross-linguistic equivalence, the construction [*vor sich hin* + $V_{Process/StateInanim}$] exhibits a specific feature. Its most typical Russian equivalent is [$V_{Process/StateInanim}$ + *себе*], e.g. *кипит себе* ≈ 'boils', *коптит себе* ≈ 'smokes, fumes', *дымит себе* ≈ 'smokes', *ржавеет себе* ≈ 'rusts' and *чахнет себе* ≈ 'withers'. The particle *себе* ('self', roughly meaning 'to itself') expresses duration, weak intensity and uncontrollability. That is, it indicates that it is a question of prolonged, slow processes that transpire without any outside interference and not in conformity with the intent of an active Agent. Russian translation equivalents can also use *себе*, albeit to a lesser degree, for the construction [*vor sich hin* + $V_{Process/StateAnim}$].

There are another three types of co-occurrences with *vor sich hin*:

v. with verbs denoting certain kinds of activities that can be interpreted as indicating 'weak intensity' or 'introversion' (that is, 'to/for oneself rather than others') and 'uniform duration' ($V_{Activity}$): for example, *arbeiten* 'work', *sortieren* 'sort', *dilletieren* 'dabble', *essen* 'eat', *regieren* 'rule' and *suchen* 'seek';
vi. with verbs of motion (V_{Motion}): *gehen* 'go', *tanzen* 'dance', *taumeln* 'lurch', *hüpfen* 'hop', *fahren* 'drive, ride', *tappen* 'toddle', etc.;
vii. with verbs denoting directed visual perception (V_{Visual}): *blicken* 'glance, gaze', *starren* 'gaze, stare', *schauen* 'look, behold', *sehen* 'see', *glotzen* 'gawk' and *stieren* 'stare, gape'.

The constructions [*vor sich hin* + $V_{Activity}$] and [*vor sich hin* + V_{Motion}] occur in the corpora significantly less frequently than those discussed above.[10] From a semantic point of view they have their own configurations of features. As for (vii), it can be illustrated with examples such as (29).

(29) *„Messen Sie dem Lachen nicht zuviel Bedeutung zu", sagte das Mädchen zu K., der, wieder traurig geworden,* ***vor sich hin starrte*** *und keine Erklärung zu brauchen schien.* (RNC: F. Kafka. Der Prozess)
– И пожалуйста, не придавайте слишком много значения нашему смеху, – обратилась она к К., видя, что тот опять помрачнел и ***уставился перед собой****, не интересуясь никакими объяснениями.*
(пер. на рус.: Р. Райт-Ковалева)
'"Don't worry too much about him laughing," said the girl to K., who had become unhappy once more and *stared quietly in front of himself* as if needing no further explanation.' (trans. David Wyllie)

In co-occurrences of the type *vor sich hin starren* we are also witnessing a kind of transitional case, since [*vor sich hin* + V_{Visual}] permits various interpretations. On the one hand, it can be regarded as realising [*vor sich hin* + V], while on the other it is a homonymous construction [*vor sich* + *hin*-V]. For example, *vor sich hin starren* can also be construed as *vor sich hinstarren* (in the corpora it is found written both as one and two words), that is, as a combination of a phrasal verb *hin*-V with the locative adverbial *vor sich*. It is no coincidence that in such a case *vor sich hin* is often translated into Russian by means of the place adverbial *перед собой* 'before / in front of oneself' (e.g. context 29), whereas in examples of type (i) *vor sich hin* often corresponds to the Russian adverbial of manner *про себя* (to/for oneself). Nevertheless, it seems appropriate to consider *vor sich hin* with verbs denoting directed visual perception as a variant of the construction [*vor sich hin* + V], since the notions of 'introversion' and 'uniform duration' are more salient than the locative component.[11]

4.2 *Vor sich hin*: Semantic features

Thus I have established that the German adverbial construction *vor sich hin* can be embedded into seven constructional patterns with empty V-slots:

---

10. This is also related to the fact that the number of verbs that can potentially belong to group (v) is extremely limited. Only certain verbs denoting activity can participate in the construction [*vor sich hin* + $V_{Activity}$]. Group (vi) is also limited.

11. There is, however, yet another possibility; cf. (Dobrovol'skij, 2010a; 2010b.)

1. [*vor sich hin* + $V_{CommPotential}$]
2. [*vor sich hin* + $V_{Process/StateAnim}$]
3. [*vor sich hin* + V_{Mental}]
4. [*vor sich hin* + $V_{Process/StateInanim}$]
5. [*vor sich hin* + $V_{Activity}$]
6. [*vor sich hin* + V_{Motion}]
7. [*vor sich hin* + V_{Visual}]

Constructions 1–4 and 7 have been examined in some detail. Let us now attempt to determine how the meaning of *vor sich hin* is structured. As we have seen, the traditional lexicographic description is limited to the features 'oriented toward oneself' and 'uniform duration'. These features are neither sufficient nor necessary. Analysis of the corpus data shows that in addition to these features there are a number of others that are not accounted for in (Duden GWDS), e.g. 'weak intensity', 'externally uncontrollable', 'passivity' and 'negative evaluation'.

Significantly, not one of these semantic components is a cross-cutting feature. The most stable among them is the core feature 'duration', but even this semantic component can be neutralised in certain contexts.[12] The broadly understood semantic feature 'introversion' is also among the most stable. What I am somewhat arbitrarily and not entirely accurately calling 'introversion' here has two rather autonomous aspects: first, actions with no external addressee (*бормотать* 'murmur', *мурлыкать себе под нос* 'hum tunelessly'), and second, processes and states that do not assume any external influence (*чахнуть* 'wither', *хиреть* 'languish, grow feeble'), and the first of these groups is heterogeneous. On the one hand, it includes predicates that are themselves inclined toward introversion, while on the other it has predicates that can be interpreted as such (for example, *экспериментировать* 'experiment', *работать* 'work' and other verbs from the construction [*vor sich hin* + $V_{Activity}$]).

Each of these types of verbal constructions in which *vor sich hin* participates has its own configuration of semantic features. Essentially this distribution can be described as ambiguity – a kind of chain polysemy. A more adequate model of meaning however, is something in the spirit of Wittgenstein's "family resemblance". On the one hand, such a model must register the most frequently occurring features in their prototypical configuration (evidently the combination of features 'duration', 'orientation toward oneself', 'weak intensity' and 'externally uncontrollable'). On the other, it must describe the modifications of this semantic structure depending

12. Especially together with verbs of types (v) and (vi), 'duration' is replaced by 'iteration' in some contexts.

on the context.[13] In each concrete occurrence of the construction [*vor sich hin* + V] certain aspects of meaning can be neutralised, "recede into the shadows", or even be completely eliminated, while others on the contrary are brought into focus or added to the basic configuration of features (e.g. the notion of "bad", which appears in (ii) and (iv) and is untypical of the other groups). To a significant degree, such an approach conforms to the principles of contemporary semantic theories; cf. (Padučeva, 2004). Such context-dependent modifications of meaning can be described with the help of semantic rules (in the sense of Apresjan, 2008).

5. Conclusion

The analysis has revealed several things, which I will now briefly summarise.

First of all, both phraseological theory and practical phraseography need to expand their subject domain. All languages have not entirely compositional constructions which, although they fall under the definition of phrasemes, remain practically unstudied within traditional phraseology. The more fixed co-occurrences with non-compositional semantics that we can investigate and describe lexicographically, the more valid will be our theoretical conceptions of the objectives and boundaries of phraseology. Such studies will also enable us to create an empirically based classification of constructions. In this respect collaboration between phraseology and Construction Grammar may prove to be very fruitful. Construction Grammar views all irregular formations not as exceptions to the rules, but as entirely normal ways to express assigned meanings. The appearance of such approaches to the study of language has been made possible thanks to, among other things, new tools of linguistic analysis, namely large text corpora. The methods of corpus linguistics enable us to register the combinatorial potential of each lexeme and regard frequent co-occurrences as units of linguistic description regardless of whether there is a semantic shift in the meaning of the constituents of such co-occurrences.

Second, the contrastive analysis of all units of language must consistently distinguish between translation equivalence and systematic equivalence. Not all correlates of an expression in the source language that can be found in texts in the target language meet the criteria of systematic equivalence. Studies based on data from parallel corpora provide information about translation equivalence.[14] To identify functional equivalence, i.e., get data for lexicography, additional information is

13. Cf. in this connection the notion of *coercion* often used in Construction Grammar.

14. However, deriving linguistic information from translation corpora may be useful, but with a necessary note of caution. Such issues have been discussed in translation studies.

needed, primarily about the conditions of use in each concrete context and the combinability of a particular unit.[15] Without such knowledge it is impossible to discover dictionary equivalences in a different language. This is especially true of phrasemes with open slots in their structure; that is, the kind of expressions analysed in the present article.

Third, the use of parallel corpora is an extremely effective tool in the study of constructional phrasemes. It must be taken into account, however, that currently the available corpora of parallel texts (including the German-Russian ones) are still quite modest in size. The contexts from the RNC containing the phrasemes *vor sich her* and *vor sich hin* do not reflect all relevant types of constructions, not to mention all possible translations of these phrasemes.

Funding

This paper is based on work supported by the RFFI under Grant 17-29-09154.

References

Apresjan, J. D. (2008). O proekte aktivnogo slovarja (AS) russkogo jazyka [On a Project of a Production Dictionary of Russian]. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2008"*, 7(14), 23–31.

Baranov, A. N., & Dobrovol'skij, D. O. (2013). *Osnovy frazeologii*. Moskva: Flinta, Nauka.

Booij, G. (2013). Morphology in CxG. In T. Hoffmann, & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 255–273). Oxford: Oxford University Press.

Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511180386

Dobrovol'skij, D. O. (2010a). Čto značit nemeckaja konstrukcija *vor sich hin*? In N. S. Babenko (Ed.), *Jazyk. Zakonomernosti razvitija i funkcionirovanija* (pp. 286–296). Moscow, Kaluga: Eidos.

Dobrovol'skij, D. O (2010b). Deiktische Konstruktionen des Deutschen aus lexikographischer Perspektive. In A. Dykstra, & T. H. Schoonheim (Eds.), *Proceedings of the XIVth Euralex International Congress* (pp. 219–220). Leeuwarden: Fryske Akademy.

Dobrovol'skij, D. O (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch, & A. Ziem (Eds.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze* (pp. 111–130). Tübingen: Stauffenburg.

Dobrovol'skij, D. O (2015). On the Systematic Variation of German Idioms: Converse Pairs as a Constructional Phenomeno. *Journal of Social Sciences*, 11 (3), 248–257. https://doi.org/10.3844/jssp.2015.248.257

15. "More extensive direct integration of the context should also narrow the current gap between lexical and textual equivalence." (Váradi and Kiss, 2001, p. 176)

Dobrovol'skij, D. O., & Padučeva, E. V. (2008). Dejksis v otsutstvie govorjaščego: o semantike nemeckix dejktičeskix elementov *hin i her*. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2008"*, 7(14), 140–146.

Duden 11 (2002). – *Duden – Redewendungen. Wörterbuch der deutschen Idiomatik*. 2., neu bearb. und aktualisierte Auflage. Duden Band 11. Mannheim etc.: Dudenverlag.

Duden-GWDS. (1999). - *Duden. Das große Wörterbuch der deutschen Sprache in zehn Bänden*. 3., völlig neu bearb. und erw. Aufl. Mannheim etc.: Dudenverlag.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and Idiomaticity in Grammatical Constructions. The Case of let alone. *Language*, 68(3), 501–553.

Goldberg, A., E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago/London: The University of Chicago Press.

Krause, M., & Baerentzen, P. (2010). *Spatiale Relationen – kontrastiv: Deutsch-Dänisch*. Tübingen: Julius Groos.

Marcq, P. (1988). *Spatiale und temporal Präpositionen im heutigen Deutsch und Französisch*. Stuttgart: Hans-Dieter Heinz Verlag.

MDRI – *Moderne Deutsch-Russische Idiomatik: Ein Korpus-Wörterbuch*, forthcoming. http://www.europhras.org/index.php?option=com_content&view=article&id=138%3Amoderne-deutschrussische-idiomatik-ein-korpus-woerterbuch&catid=36%3Aprojekte&Itemid=91&lang=de

Padučeva, E. V. (2004). *Dinamičeskie modeli v semantike leksiki*. Moscow: Jazyki slavjanskoj kul'tury.

Steyer, K. (2015). Patterns. Phraseology in a State of Flux. *International Journal of Lexicography*, 28(3), 279–298. https://doi.org/10.1093/ijl/ecv021

Taylor, J. R. (2006). Polysemy and the Lexicon. In G. Kristiansen et al. (Eds.), *Cognitive Linguistics: Current Approaches and Future Perspectives*. Berlin/New York: Mouton de Gruyter.

Váradi, T., & Kiss, G. (2001). Equivalence and Non-Equivalence in Parallel Corpora. *International Journal of Corpus Linguistics*, 6, 167–177. https://doi.org/10.1075/ijcl.6.3.13var

Corpora

DeReKo – Das Deutsche Referenzkorpus des IDS Mannheim im Portal COSMAS II (Corpus Search, Management and Analysis System): https://cosmas2.ids-mannheim.de/cosmas2-web

DWDS – Corpora des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts: http://www.dwds.de

RNC – Russian National Corpus = Корпус немецко-русских параллельных текстов Национального корпуса русского языка: http://www.ruscorpora.ru/search-para.html

Computational phraseology and translation studies

From theoretical hypotheses to practical tools

Jean-Pierre Colson
Université catholique de Louvain

The notion of phraseology is now used across a wide range of linguistic disciplines but it is conspicuously absent from most studies in the area of Translation Studies (e.g. Delisle, 2003; Baker and Saldanha, 2011). The paradox is that many practical difficulties encountered by translators and interpreters are directly related to phraseology in the broad sense (Colson, 2008, 2013), and this can also clearly be seen in the failure of machine translation systems to deal efficiently with the translation of phraseological units (PUs).

We argue that phraseology and translation studies have much to gain from cross fertilisation, because both disciplines are regularly criticised for their lack of coherent terminological description and for the insufficient number of reproducible experiments they involve.

Decoding phraseology in the source text is far from easy for translators and interpreters, all the more so as they are usually not native speakers of the source language. Finding a natural formulation in the target language and avoiding *translationese* requires an excellent mastery of the phraseology of the target language. Even experienced professionals sometimes fail to detect the fixed or semi-fixed character of a source text construction. We argue that algorithms derived from text mining and information retrieval techniques can be efficient and (computationally) cost-effective in order to build up unfiltered collections of recurrent fixed or semi-fixed phrases, from which translators could gain information about the number of PUs in the source text. Such an algorithm has been proposed in Colson (2016) and has been implemented in a web application enabling translators and language professionals to automatically retrieve most PUs from a source text. Other tools should be developed in order to bridge the gap between the findings of computational phraseology and the practice of translation and interpreting.

Keywords: phraseology, computational linguistics, translation, interpreting, text mining

https://doi.org/10.1075/ivitra.24.04col

1. Introduction

In the former Roman city of Timgad (North Africa), a Latin inscription was found on a stone in the ruins of the old forum: "*Venari, lavari, luderi, rideri, occ (hoc) est vivere,*" which could be literally translated "to hunt, to bathe, to play, to laugh: that is living".[1] It is fascinating to see how the crux of the phraseological matter was already present in this very old text. In the first place, it is an alternation of lexis, grammar and phraseology, because the sequence *hoc est vivere* clearly has a figurative meaning and was partly fixed in Latin. We can therefore see it as an illustration of the alternation of the *open choice principle* and the *idiom principle* (Sinclair, 1991). The link between phraseology and culture is obvious in this fragment: visions of life in the Roman culture have given rise to phraseological units of this type, and they have survived in many European languages today.

The translation of the apparently easy Latin phrase *Hoc est vivere* in this context is far from straightforward, because *that's life* would be used in English with a negative connotation: according to the OED (Oxford English Dictionary),[2] it is "an expression of resignation or acquiescence in things as they are". It is therefore synonymous with *such is life, life's like that*, for which the French phrase *c'est la vie* is also used in English. A better English translation of *Hoc est vivere* would therefore be *This is the life*, defined by the OED as "an expression of satisfaction".

It may therefore come as a surprise to learners of English as a foreign language that *this is the life* has a completely different meaning from *that's life*. It so happens that *This is the life* is also the title of a famous song and album by the Scottish singer Amy Macdonald,[3] and I have noticed that many people in France and francophone Belgium did not understand the true meaning of the title. Phraseology is not only an issue for language learners, but is also a major source of misunderstandings and wrong translations.

In this paper, I will try to show that a corpus-based and computational approach may shed some fresh light on the intertwining of phraseology, culture and translation.

---

1. http://chalcedon.edu/research/articles/inscription-at-timgad/, consulted on October 15th, 2016.

2. http://www.oed.com, consulted on October 26th, 2016.

3. Amy Macdonald, label Vertigo, 2007.

2. Phraseology and translation studies

Let us start from the traditional definition of phraseology (Burger et al., 2007) as the study of phraseological units (PUs) in the broadest sense, including partly fixed phrases (routines and formulae, collocations), and also very fixed ones (such as idioms and proverbs).

A thorough study of this phenomenon falls beyond the scope of the present contribution, but it is obvious that the notion of phraseology is now used across a wide range of linguistic disciplines: phraseology (proper, e.g. Burger et al., 2007), corpus linguistics (Sinclair, 1991, Granger and Meunier, 2008), discourse analysis (Schnur, 2014), pragmatics (Mellado Blanco, 2013), cognitive linguistics (Omaziü, 2005), etc.

Strangely enough, phraseology is, on the other hand, rarely mentioned by translation studies. Most publications mention the problem of expressions/idioms/collocations but they do not refer to phraseology as a theory or discipline. In Delisle (2003), for instance, expressions are just treated as a part of the lexicon. Phraseology is also conspicuously absent from a major reference work in the field, the Routledge Encyclopedia of Translation Studies (Baker and Saldanha, 2011), and the same holds true for collocations. The interest in phraseology (at least for collocations) in translation studies actually came from corpus linguistics (e.g. Teubert, 2002).

On the other hand, computational linguistics is now showing a growing interest in phraseology, particularly against the backdrop of automatic translation (e.g. Monti et al., 2013). At the same time, researchers in phraseology have taken a keen interest in translation theory and practice: Colson (2008, 2011, 2013), Corpas Pastor (2000, 2007, 2008, 2013), Leiva Rojo (2013).

Our point of view is that phraseology and translation studies have much to gain from cross-fertilisation, and that a computational and corpus-based approach would yield the best results by addressing problems posed by phraseology to both human and machine translation.

3. Problems posed by phraseology to human translation

Decoding phraseology in a source text is far from easy for translators and interpreters, all the more so as they are usually not native speakers of the source language. On the other hand, finding a natural formulation in the target language and avoiding translationese (Tirkkonen-Condit, 2002) requires an excellent mastery of the phraseology of the target language.

CBTS (corpus-based translation studies; Kruger et al., 2011) has not yet fully investigated the implications of phraseology for translation theory and practice.

Experiments with translation corpora may precisely shed some light on some crucial aspects of phraseology and of translation studies.

For instance, translation errors due to phraseology are present in many translation corpora, even in the official translations of international organisations, as in the following examples:

(1) Cost-cutting and cutting corners caused the biggest environmental disaster in history.
Réduire les coûts et arrondir les angles ont engendré le plus gros désastre écologique de l'histoire. (europarl.eu / linguee.com, 01/07/2015)

(2) Above all it is important to avoid cutting corners, which easily happens with the ongoing and multiple evaluations (…)
Il faudrait éviter avant tout de traiter les choses de manière superficielle, ce qui est souvent le cas avec les évaluations multiples (…)
(eur-lex.europa.eu / linguee.com, 06/04/2015)

(3) If we are going to move the goalposts, if that is the will of this General Conference, let us do that before the whistle.
Si nous voulons déplacer les poteaux de but, si telle est la volonté de la Conférence générale, alors faisons-le avant que la partie n'ait commencé.
(unesdoc.unesco.org / linguee.com, 06/04/2015)

In (1) and (2), the English phrase *to cut corners* has been wrongly translated into French. The dictionary[4] defines the figurative meaning of *cut a corner* or *cut corners* as "to pursue an economical or easy but hazardous course of action; to act in an unorthodox manner to save time; also, to act illegally". The French translation *arrondir les angles* is totally wrong in (1), because it means "atténuer les oppositions, les dissentiments"[5] (to smooth out oppositions or disagreements). The correct French translation of (1) and (2) should have been *rogner sur les coûts / lésiner sur les coûts*. In (3), a calque of the English phrase was used in French, instead of the correct translation *changer les règles du jeu (à la dernière minute)*.

If even translators working at international institutions are apt to fall into the trap of phraseology, this applies even more to less experienced translators and interpreters. There are numerous examples of errors of this kind in translations produced by students (Colson, 2010): phraseology is not decoded, and this results in a literal translation, a calque.

4. http://www.oed.com, consulted on October 26th, 2016.

5. Le Grand Robert de la langue française, http://gr.bvdep.com/, consulted on October 27th, 2016.

4. Problems posed by phraseology to machine translation

Phraseology has only recently been identified as one of the main sources of errors in automatic translation systems, including the most recent SMT systems[6] (Monti et al., 2013). According to a comparative study carried out by Barreiro et al. (2013) with two machine translation systems, *OpenLogos* and *Google Translate*, and four languages (English, French, Italian, Portuguese), the translation of phraseological units by SMT systems leaves a lot to be desired. In about 40 percent of the cases, Google Translate offered a wrong translation of phraseological units.

Indeed, it doesn't take much time to collect numerous examples of phraseological units that are wrongly translated by Google as literal grammatical constructions, even between international languages such as English and French, as shown by Examples (4), (5), (6), in which English phrases were wrongly translated into French by Google Translate.[7]

(4) A lot of things just come out of the woodwork.
Google Translate (GT): *Beaucoup de choses vient (sic) de sortir de la menuiserie*.
(Correct translation : *beaucoup de choses apparaissent comme par enchantement*).

(5) Let us give credit where credit is due.
GT: *Donnons crédit lorsque le crédit est dû*. (Correct translation : *il faut rendre à César ce qui appartient à César*).

(6) You would be up the creek without a paddle.
GT: *Vous seriez le ruisseau sans pagaie*. (Correct translation: *vous seriez dans le pétrin*).

It should be stressed that these are not isolated cases. Here are just a few examples of various English phrases that I have come across in my readings, and which are wrongly translated into French by Google Translate:

(7) come hell or high water, clutching at straws, the credits roll, cut the mustard, cutting corners, a dab hand, a dead giveaway, in deadly earnest, death by a thousand cuts, snatch defeat from the jaws of victory, do a double take, does what is says on the tin, don't mention the war, don't quote me on this, a cheap shot, the chickens are coming home to roost, chuffed to bits.

6. Statistical machine translation systems.

7. Checked on October 28th, 2016.

In spite of its relative frequency,[8] *chuffed to bits* (a popular phrase meaning "all excited", "thrilled") is translated into Spanish *chuffed en pedazos* and into French *chuffed aux bits* by Google Translate, and the same problem holds true for the other examples under (7).

5. Theoretical hypotheses

The major problems posed by phraseology to human and to machine translation, as briefly described under Sections 3 and 4, call for a theoretical explanation. It should be stressed, however, that the precise role of phraseology in language remains largely a theoretical conundrum.

According to John Sinclair, the idiom principle (1991) or phraseological tendency (1996) is the general rule:

> Most normal text is made up of the occurrence of frequent words, and the frequent senses of less frequent words. Hence, normal text is largely delexicalized, and appears to be found by exercise of the idiom principle, with occasional switching to the open-choice principle. (Sinclair, 1991, p. 113)

For Jackendoff (1995), there are about as many fixed expressions as there are single words in the dictionary, but others (such as Meľčuk, 1995) hold the view that fixed expressions far outnumber single words. Hanks (2010) points out that the meaning potential of a word consists of a puzzling mixture of terminology and phraseology. Erman and Warren (2000) claim that prefabs represent about 55% of the texts they have analysed. Crucially, they aslo stress that

> The identification of prefabs is difficult. There are two main reasons why this is so. One is that what is a prefab to some members of a language Community need not be a prefab to all members. Some prefabs will be known to practically all native and fluent speakers; others will be more limited in dispersion and entrenchment. (Erman and Warren, 2000, p. 33)

If the identification of phraseology by humans is so difficult, one possible solution would be to turn to automatic extraction. However, as stated by Gries (2013), after "50-something years of work on collocations", the results are still disappointing and "after many decades of 'more of the same', (…) it is time to explore new ways of studying collocations" (Gries, 2013, p. 159).

---

8. *Chuffed to bits* with quotation marks yields 229,000 occurrences on Google (October 27th, 2016) and 633 occurrences on the enTenTen13 corpus (20 billion tokens) of the Sketch Engine (http://sketchengine.co.uk).

I have therefore proposed a new metric (Colson, 2016) for the automatic extraction of phraseology: the Corpus Proximity Ratio (*cpr*):

$$cpr = \frac{n(x_{i_1} x_{i_2} x_{i_3} \dots x_{i_n})}{n(\max(||x_i - x_j||) \leq W)}$$
$$x_{i_1}, x_{i_2}, x_{i_3}, \dots x_{i_n}$$

Figure 1. The Corpus Proximity Ratio (CPR), J.-P. Colson (2016)

The *cpr-score* basically measures the ratio between the exact frequency of an n-gram in a corpus, and the frequency of the n-gram given a certain window between the grams; this window (*W*) must be set experimentally according to the corpus and the language. The *cpr-score* meets three criteria recommended by Gries (2013) for the improvement of automatic extraction of collocations: the measure is directional; the methodology uses recurrence across corpora; finally, it is extendable to multiword expressions, as it can be used for n-grams ranging from bigrams to 10-grams. From a psycholinguistic point of view, the score may be seen as a simulation by an algorithm of the Firthian principle: "You shall know a word by the company it keeps (Firth, 1957, p. 11). Indeed, the *cpr-score* is derived from metric clusters (Baeza-Yates and Ribeiro-Neto, 1999), in which proximity plays a key role, as opposed to parametric statistical scores based on the probability calculus.

As for many aspects of computational linguistics, testing algorithms and their feasibility for practical applications plays a key part in computational phraseology. Let us start from a concrete example, the English phrase *hit the road*. In order to compute the *cpr-score* of this n-gram in a corpus, computational phraseology faces the following problems. First of all, as mentioned above, the size of the corpus will be of the essence. On the basis of experiments, a corpus of at least 200 million tokens is recommended. The next step will be to read the whole corpus (the equivalent of about 500,000 A4 pages for a corpus of 200 million tokens), and check all offsets (positions in file) for 'hit', 'the', 'road'. For all of the offsets of 'hit'; the computer will check if there is an instance of 'the' within a given window (from 20 to 50 words) and, if so, if there is an instance of 'road' within the same window. All results will be stored, and the frequency will be compared with that of the exact phrase on the corpus. For most English corpora of that size, the *cpr-score* for *hit the road* will be very high: around 0.51 (significant at 0.065),[9] in spite of the very high frequency of the grams. Computational phraseology should in other words go hand in hand with research in computer science, because theoretical hypotheses can only be tested efficiently if very powerful computing techniques are available. In this case,

9. The significance of the cpr-score has been set experimentally.

the Lemur Toolkit, based on a query likelihood model, was used.[10] Computing the *cpr-score* of an n-gram on a corpus of 200 million tokens therefore only takes an average of 0.07 seconds on a run of the mill computer.

The *cpr-score* is currently being tested, and the preliminary results should therefore been taken with a note of caution. This applies in the first place to the way of extending the state of the art method to longer n-grams (from trigrams to 10-grams). Checking the precision and recall of the extracted phraseological units with a panel of native speakers (Gries, 2013) turns out to be nigh impossible for trigrams and fourgrams, let alone for 5 to 10-grams. It should be kept in mind that we are interested here in phraseology in the broadest sense, including formulaic language (Wray, 2009), and as already pointed out by Erman and Warren (2000), native speakers are not all aware of the subtle degrees of fixedness or idiomaticity in longer structures such as prefabs, routines and formulae or longer collocations. Preliminary tests with a panel of natives speakers indicate that the *cpr-score* reaches a precision score of about 90 percent, but recall is impossible to determine for longer n-grams, because no agreement can be reached on the exact number or the boundaries of longer phraseological units contained in a text or a corpus. The following sentences may serve to illustrate this problem.

(8) At the heart of the current turmoil is a decision by Saudi Arabia and other leading voices in the Opec oil cartel to get drawn into a turf war with the new generation of US shale producers.[11]

If asked to indicate PUs ranging from bigrams to 10-grams, native speakers may indeed hesitate about the number of constructions that would be relevant in this sentence: should they for instance underline *heart of*, *at the heart*, *at the heart of*, *at the heart of the*? Should they consider *Saudi Arabia* as a PU? And if so, is *Opec oil cartel* the complete phraseological unit or just *oil cartel*? Is *new generation* a collocation, or is it *the new generation of*?

In order to reach a workable compromise, the methodology implementing the *cpr-score* will check the smallest sequence (bigrams), adding one gram at a time, and check if the score increases or decreases, as exemplified for *at the heart of* in Table 1.

Table 1. Frequency of cpr-score of 3 n-grams on ukWaC corpus

n-grams	Hits in ukWaC	cpr-score
at the heart	1,851	0.60
at the heart of	1,815	0.72
at the heart of the	697	0.34

10. http://www.lemurproject.org/lemur/

11. The Guardian, January 7th, 2016.

Table 1 displays the frequency and *cpr-score* of *at the heart, at the heart of* and *at the heart of the* in a balanced British English corpus of 200 million tokens.[12] On the basis of the *cpr-score*, the three n-grams might be considered as PUs in the broad sense, but the extension to the following gram reveals that *at the heart of* yields the highest score, so this combination will be selected. Using the same methodology and the same corpus, the following PUs were extracted from Example (8) by means of the *cpr-score*:

Table 2. Frequency and cpr-score of PUs in the ukWaC corpus

Pus	Hits in ukWaC	cpr-score
at the heart of	1,815	0.72
current turmoil	3	0.60
decision by	89	0.20
Saudi Arabia	508	1.00
leading voices	5	0.63
oil cartel	4	1.00
get drawn into	23	0.68
turf war	10	0.83
new generation of	558	0.74

It is worthy of note that the *cpr-score* is far less dependent on frequency than other statistical scores such as *log-likelihood* or *t-score*. The PU *at the heart of* receives a high score of 0.72, combined with a high number of occurrences in the corpus (1,815), but *current turmoil, leading voices* and *turf war* also yield a high *cpr-score*, while the frequencies in the corpus are very low. In a corpus of about 200 million tokens, experiments with the *cpr-score* show that three occurrences are enough to detect a possible PU.

As shown by the PUs contained in Example (8), the *cpr-score* will in most cases be very high for idiomatic PUs. In Table 2, *turf war*[13] yields a score of 0.83, despite occurring just ten times in the corpus. It should be pointed out that *Saudi Arabia* logically receives the maximal *cpr-score* of 1.00, because named entities clearly belong to phraseology in the broad sense, which brings us to a theoretical remark on computational phraseology and cultural diversity.

Current research in computational and corpus linguistics does not always take into account the fundamental differences between Indo-European languages (to which belong three of the four most spoken languages of the world: English,

12. A randomly selected portion of 200 million tokens from the ukWaC corpus, University of Bologna, http://wacky.sslmit.unibo.it/doku.php?id=corpora.

13. *Turf war*, in its figurative meaning, is defined by the OED as: "colloq. (chiefly N. Amer.), a dispute over territory; freq. in extended use" (www.oed.com, consulted on November 1st, 2016).

Spanish, Hindi) and the other language families. Indeed, the most spoken language of the world (Mandarin Chinese) provides a partly different picture of lexis, grammar and phraseology. For one, it is mainly an isolating language (there is a very low morpheme-to-word ratio), and it does not mark words by a space, nor named entities by capital letters (as they do not exist). When applying computational phraseology to Chinese, it is therefore quite natural to treat for instance 上海 'Shanghai' as a phraseological unit, because the association between the two Chinese characters (hans) will be very high: a *cpr-score* of 0.88. But the point is that this really makes sense because *Shanghai* means something in Chinese: "on the sea". This will hold true for most cities in China. Beijing (北京), for instance means "capital of the north" or "northern capital". But is it so different in European languages? Yes and no: the etymology of London, Paris, Leiden, Brussels, Malaga and so on will reveal in most cases an original phraseological unit in the linguistic substrate. Both the Dutch city of *Leiden* and the French city of *Lyons*, for instance, derive from the Latin *Lugdunum*, a calque of Celtic *Lugdunos*, a hill or fort for the god Lug. In other words, not only should *New York* or *Port-la-Nouvelle* rightly be seen as PUs, but also many other European cities, if we take etymology into account, and this situation is naturally to be found in Chinese as well.

The word *apricot* may serve as another example of the fuzzy borderline between lexis and phraseology. The etymology of this word speaks volumes:

> originally < Portuguese albricoque or Spanish albaricoque, but subseq. assimilated to the cognate French abricot (t mute). Compare also Italian albercocca, albicocca, Old Spanish albarcoque, < Spanish Arabic al-borcoq(ue (P. de Alcala) for Arabic al-burqūq, -barqūq, i.e. al the + barqūq, <Greek πραικόκιον(Dioscorides, c100;later Greek πρεκόκκια and βερικόκκια plural), probably < Latin praecoquum, variant of praecox, plural praecocia, 'early-ripe, ripe in summer,' an epithet and, in later writers, appellation of this fruit, originally called prūnum or mālum Armeniacum.
> (OED online, consulted on November 1st, 2016)

In short, the English word *apricot*, consisting today of one morpheme, was originally a phraseological unit, and meant "early-ripe, ripe in summer", which appears in the two separate Greek morphemes *πραι* / *κόκιον* or in the two Latin morphemes *prae* / *cox*, which also gave *précoce* in French and *precocious* in English. This example shows *why* the interplay between lexis, grammar and phraseology may be more complex than in Chinese, namely because European languages have undergone so many influences; they are, in other words, very *hybrid*. When testing linguistic hypotheses in general, and computational phraseology in particular, one should therefore consider more isolating and less hybrid languages such as Chinese as very useful.

6. Towards new practical tools

While many theoretical aspects of phraseology need to be further investigated, there is also a need for practical tools for language learners, translators and interpreters. The awareness of the existence of chunks, prefabs, or phraseological units in the broad sense, was already considered as a priority by Michael Lewis (1993, 1997). It often comes as a surprise to learners that so many aspects of the language they are learning are actually more than just words in a syntactic structure. Further research in computational phraseology should pay attention to the interface between theoretical hypotheses and situations in which language users should be able to receive some feedback about the presence of phraseology in the texts they are confronted with.

As a tentative step in that direction, the *IdiomSearch* Project[14] tries to implement the *cpr-score* presented above in a user-friendly web application, as shown in Figure 2.

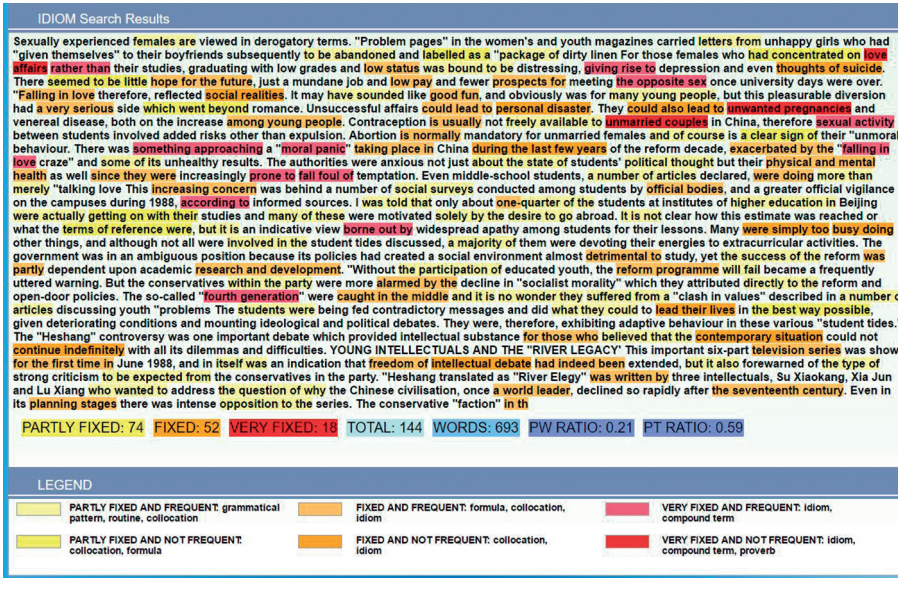


Figure 2. Screenshot of the IdiomSearch web application

Although this experimental tool still needs to be tested on a wider scale, it provides results that are compatible with Sinclair's hypothesis (1991) that about 50 percent of any text will reflect the *idiom principle*. We therefore propose the *PT ratio*: the

14. J.-P. Colson & Université catholique de Louvain, 2016, http://idiomsearch.LSTI.ucl.ac.be

number of phraseological units (in the broad sense) per text. It is simply the proportion of tokens that are part of a phraseological unit. Thus, in Figure 2, no less than 59 percent of all tokens are actually included in PUs.

Results obtained with the *IdiomSearch* tool suggest that it may be useful for language learners or translators to be aware of a whole host of weakly idiomatic PUs, which might go largely unnoticed if not highlighted by tools deriving from computational phraseology. Thus, in the text presented in Figure 2, the algorithm was able to extract PUs like *social realities, which went beyond, could lead to, personal disaster, something approaching, borne out by, fall foul of, caught in the middle, lead their lives, continue indefinitely, the best way possible*, etc. The top of the phraseological iceberg may consist of idioms and very fixed phrases, but most PUs are weakly idiomatic and partly fixed: collocations, routines and formulae or clichés. Those structures are however very useful for language learners, translators and interpreters.

The time factor is of crucial importance in the development of new practical tools derived from research in computational phraseology. Traditional tools for manipulating corpora, such as concordancers or the Sketch Engine[15] require a number of actions before the translator can find the information he or she needs: selecting a corpus, checking the word(s) in context, looking for a broader context, selecting possible collocates on the basis of a statistical score, etc.

On the other hand, the most efficient tool in terms of speed is undoubtedly the Web, but search engines are actually displaying a complex mixture of real world, linguistic and commercial information. Thus, a translator looking for common PUs with "comfort" will receive the following n-grams from Google while typing the word with a space at the end:[16] "comfort inn, comfort hotel, comfort inn laval, comfort inn levis, comfort zone", of which only the last could be seen as a PU. Using the already mentioned *cpr-score* on a corpus would on the other hand yield an whole host of interesting chunks and PUs around "comfort", of which a sample is given in (9):

(9) a great comfort, added comfort, all from the comfort of, can take comfort, can take some comfort from, comfort and convenience, comfort and cuisine, comfort and durability, comfort and enjoyment, comfort and flexibility, comfort and joy, comfort and luxury, comfort and performance, comfort and protection, comfort and quality, comfort and relaxation, comfort and reliability, comfort and safety, comfort and security, comfort and style, comfort and support, comfort and warmth, comfort and well-being, comfort blanket, comfort cooling,

15. http://www.sketchengine.co.uk

16. http://www.google.com, checked on November 9th, 2016

comfort eating, comfort factor, comfort food, comfort level, comfort of knowing, comfort zone, complete comfort, convenience and comfort, degree of comfort, ease and comfort, every modern comfort, extra comfort, find comfort in, for extra comfort, for maximum comfort, for your comfort, from the comfort of your own home, give comfort to, great comfort, health and comfort, high standard of comfort, in comfort and style, in the comfort and privacy, in the comfort of their own home, in the comfort of their own homes, in the comfort of your own home, increased comfort, joy and comfort, level of comfort, little comfort, maximum comfort, out of your comfort zone, passenger comfort, peace and comfort, performance and comfort, reasonable comfort, relative comfort, ride comfort, safety and comfort, security and comfort, source of comfort, space and comfort, standard of comfort, strength and comfort, style and comfort, such a comfort, support and comfort, take comfort from, take some comfort from, take some comfort, thermal comfort, too close for comfort, took comfort, total comfort, ultra-comfort gloss, warmth and comfort, wearing comfort, words of comfort

As illustrated by the results in (9), computational phraseology provides very powerful ways of extracting linguistic associations (e.g. *take some comfort from, comfort zone, too close for comfort*), but also associations that are simply related to society in the broad sense (*health and comfort, safety and comfort, space and comfort*). The use of *and* as a seed word is particularly useful in this latter case.

It should be pointed out that both linguistic and extra-linguistic associations of this type can be very useful for translators and interpereters. The importance of linguistic associations, especially in the case of phraseological units, has been illustrated in the preceding sections. Other types of associations can be useful in the case of cultural differences, but also of recent terms.

In the case of literary translation, it may be worth checking the linguistic and extra-linguistic associations of a given word before attempting to find equivalent associations in the target language. This will be particularly obvious for languages displaying major cultural differences, but it may even be applied to such close languages as English and French. The results in (10) and (11) were automatically obtained from comparable corpora[17] on the basis of the search term "tea" and the corresponding French word "thé", and selected by the highest *cpr-score*:

17. The corpora were compiled by using the WebBootCat tool provided by the Sketch Engine, following the method based on seed words described in Baroni et al., 2009. The corpora used for this experiment were comparable web corpora of 200 million tokens each.

(10) *Coffee and tea, Coffee or tea, Cup of tea, Fairtrade tea and coffee, Tea and cakes, Tea and coffee, Tea and coffee making, Tea and coffee making facilities, Tea and coffee provided, Tea and coffee will be provided, Tea and coffee will be served, a cream tea, a cup of tea, a cup of tea or coffee, a nice cup of tea, a tea strainer, afternoon tea, afternoon tea and cakes, be your cup of tea, black tea, Boston tea party, chamomile tea, clean tea towel, coffee and afternoon tea, coffee and tea, coffee and tea making, coffee and tea making facilities, coffee or tea, complimentary tea, cup of tea, drinking tea, enjoy a cup of tea, everyone's cup of tea, followed by tea, for a cup of tea, for afternoon tea, green tea, hairdryer and tea, have a cup of tea, having a cup of tea, herb tea, herbal tea, home for tea, hot cup of tea, hybrid tea, iced tea, in time for tea, including tea and coffee, leaf tea, like making tea, loose tea, lunch and afternoon tea, lunch or afternoon tea, make a cup of tea, making a cup of tea, mint tea, morning coffee and afternoon tea, mug of tea, mugs of tea, my cup of tea, nice cup of tea, not my cup of tea, over a cup of tea, peppermint tea, pot of tea, producers of tea, require towels and tea towels, serve tea, sipping tea, supply of tea, sweet tea, tea afterwards, tea and a biscuit, tea and a chat, tea and a piece of, tea and a sandwich, tea and biscuits, tea and cake, tea and cakes, tea and cocoa, tea and coffee, tea and coffee making, tea and coffee making facilities, tea and coffee provided, tea and coffee will be available, tea and scones, tea and soft drinks, tea and toast, tea bag, tea bags, tea bar, tea boy, tea break, tea ceremony, tea chest, tea cosy, tea dances, tea drinking, tea estates, tea industry, tea interval, tea ladies, tea leaves, tea lights, tea maker, tea making facilities, tea merchant, tea or coffee, tea party, tea plant, tea plantation, tea room, tea rose, tea service, tea shop, tea spoons, tea tent, tea together, tea tonight, tea towel, tea trade, tea tray, tea tree, tea tree oil, tea trolley, tea urn, tea will be provided, television and tea, television and tea and coffee making facilities, time for tea, to get some tea, to have tea with, towels and tea, traded tea, varieties of tea, welcome cup of tea, will be followed by tea*

(11) *Le thé à la menthe, cuillère à thé de, de thé noir, de thé vert, du thé vert, le thé vert, lui offrait du thé, thé bouillant, thé de compost, thé vert, thé à la menthe*

The discrepancy between the whole range of results obtained for English, and the very meagre results for French, speaks volumes. Major cultural differences can likewise be confirmed by means of computational phraseology, as it also extracts recurrent combinations belonging to the realm of terminology (e.g. *chamomile tea*, *leaf tea*, *mint tea*, *tea plant*) or culture and society (e.g. *enjoy a cup of tea*, *lunch and afternoon tea*, *nice cup of tea*, *tea and biscuits*, *tea will be provided*). Future research might improve the tools derived from computational phraseology, in order to make a clearer distinction between linguistic and cultural associations.

7. Conclusion

In spite of the diversity of its approaches and of the sometimes fuzzy terminology it uses for classifying its different categories, phraseology has now become a major branch of research in linguistics. There is, however, still a long way to go before it is widely accepted by translation studies. Computational phraseology can be particularly useful, both for human and machine translation, and also for a better understanding of linguistic aspects of meaning, which are deeply rooted in culture, but are also characterised by recurrent patterns, thereby falling within the scope of a statistical approach.

New tools derived from computational phraseology may directly contribute to improving the efficiency of computer-aided translation. From a theoretical point of view, they may in the long run even challenge the status of machine translation, of which one of the present shortcomings precisely lies in the imperfect rendering of phraseological units.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press, Addison Wesley.

Baker, M., & Saldanha, G. (Eds). (2011). *Routledge Encyclopedia of Translation Studies*. New York, NY: Routledge.

Baroni, M., Bernardini, S., Ferrares, A. & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*, 43, 209–26. https://doi.org/10.1007/s10579-009-9081-4.

Barreiro, A., Monti, J., Batista, F., & Orliac, B. (2013). When multiword go bad in machine translation. In J. Monti, R. Mitkov, G. Corpas Pastor, & V. Secretan (Eds.), *Workshop Proceedings: Multi-word units in machine translation and translation technologies, Nice 14th Machine Translation Summit* (pp. 26–33).

Burger, H., Dobrovol'skij D., Kühn, P., & Norrick, N. R. (Eds.) (2007). *Phraseologie / Phraseology. Ein internationals Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research*. Berlin/New York: de Gruyter.

Colson, J.-P. (2008). Cross-linguistic phraseological studies: An overview. In S. Granger, & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 191–206). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139.19col

Colson, J.-P. (2010). Fraseologie en (ver)taalcompetentie. In C. Van de Poel, & W. Segers (Eds.), *Tolk- en vertaalcompetentie. Onderwijs- entoetsvormen* (pp. 111–124). Leuven: Acco.

Colson, J.-P. (2011). La traduction spécialisée basée sur les corpus : une expériencedans le domaine informatique. In I. Sfar, & Mejri, S. (Eds.), *La traduction de textes spécialisés : retour sur des lieux communs* (pp. 115–123). Synergies Tunisie 2.

Colson, J.-P. (2013). Pratique traduisante et idiomaticité :l'importance des structures semi-figées. In P. Mogorrón Huerta, D. Gallego Hernández, & M. Tolosa Igualada (Eds.), *Fraseología, Opacidad y Traducción* (pp. 207–218). Frankfurt am Main: Peter Lang.

Colson, J.-P. (2016). Set phrases around globalization: an experiment in corpus-based computational phraseology. In F. Alonso Almeida, I. Ortega Barrera, E. Quintana Toledo, & M. E. Sánchez Cuervo (Eds.), *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics* (pp. 141–152). Newcastle: Cambridge Scholars Publishing.

Corpas Pastor, G. (2000). Acerca de la (in)traducibilidad de la fraseología. In G. Corpas Pastor (Ed.), *Las lenguas de Europa: Estudios de fraseología, fraseografía y traducción* (pp. 483–522). Granada: Comares.

Corpas Pastor, G. (2007). Europäismen– von Natur aus phraseologische Äquivalente? Von blauem Blut und sangreazul. In M. Emsel, & J. Cuartero Ota (Eds.), *Brücken: Übersetzen und interkulturelle Kommunikationen. Festschrift für Gerd Wotjakzum 65. Geburtstag*(pp. 65–77). Frankfurt: Peter Lang.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma* (Studien zur romanische Sprachwissenschaft und interkulturellen Kommunikation, 49). Frankfurt: Peter Lang.

Corpas Pastor, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In I. Olza, & E. Manero (Eds.), *Fraseopragmática* (pp. 335–373). Berlin: Frank & Timme.

Delisle, J. (2003). *La traduction raisonnée*. Ottawa: Presses de l'Université d'Ottawa.

Erman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, 20(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *In Philological Society* (Ed.), *Studies in Linguistic Analysis* (pp. 1–32). Oxford: Basil Blackwell.

Granger, S. & Meunier, F. (Eds). (2008). *Phraseology. An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139

Gries, S. Th. (2013). 50-something years of work on collocations. What is or should be next… *International Journal of Corpus Linguistics*, 18(1), 137–165. https://doi.org/10.1075/ijcl.18.1.09gri

Hanks, P. (2010). Lexicography, Terminology, and Phraseology. In *Proceedings of Euralex* 2010. Leeuwarden.

Jackendoff, R. S. (1995). The boundaries of the lexicon. In M. Everaert, E. van der Linden, A. Schenk, & R. Schroeder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133–165). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kruger, A., Wallmach, K., & Munday, J. (Eds.) (2011). *Corpus-Based Translation Studies. Research and Applications*. London: Bloomsbury.

Leiva Rojo, J. (2013). La traducción de unidades fraseológicas (alemán-español/español-alemán) como parámetro para la evaluación y revisión de traducciones. In C. Mellado Blanco, P. Buján, N. M. Iglesias, M. C. Losada, & A. Mansilla (Eds.), *La fraseología del alemán y el español: lexicografía y traducción* (ELS, Etudes Linguistiques / Linguistische Studien, Band 11)(pp. 31–42). München: Peniope.

Lewis, M. (1993). *The Lexical Approach*. Hove: Language Teaching Publications.

Lewis, M. (Ed.) (1997). *Implementing the Lexical Approach*. Hove: Language Teaching Publications.

Meľčuk, I. (1995). Phrasemes in language and phraseology in linguistics. In M. Everaert, E. van der Linden, A. Schenk, & R. Schroeder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 167–232). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mellado Blanco, C. (2013). La gramaticalización de las restricciones y preferencias de uso de las unidades fraseológicas del español y alemán desde un enfoque cognitivo-pragmático. In I. Olza, & E. Manero (Eds.), *Fraseopragmática*(pp. 303–331). Berlin: Frank & Timme.

Monti, J., Mitkov, R., Corpas Pastor G., & Seretan, V. (Eds.) (2013). *Workshop Proceedings: Multiword units in machine translation and translation technologies, Nice 14th Machine Translation Summit*.

Omazié, M. (2005). Cognitive linguistic theories in phraseology. *Jezikoslovlje*, 6(1), 37–56.

Schnur, E. (2014). Phraseological Signaling of Discourse Organization in Academic Lectures: A Comparison of Lexical Bundles in Authentic Lectures and EAP Listening Materials. *Yearbook of Phraseology*, 5(1), 95–121. https://doi.org/10.1515/phras-2014-0005

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Teubert, W. (2002). The role of parallel corpora in translation and multilingual lexicography. In B. Altenberg, & S. Granger (Eds.), *Lexis in Contrast: Corpus-based approaches* (pp. 189–214). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.7.14teu

Tirkkonen-Condit, S. (2002). Translationese – a myth or an empirical fact? *Target*, 14, 207–220. https://doi.org/10.1075/target.14.2.02tir

Wray, A. (2009). *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Computational extraction of formulaic sequences from corpora

Two case studies of a new extraction algorithm

Alexander Wahl and Stefan Th. Gries
Donders Institute for Brain, Cognition and Behaviour, Radboud University / University of California Santa Barbara & Justus Liebig University

We describe a new algorithm for the extraction of formulaic language from corpora. Entitled MERGE (Multi-word Expressions from the Recursive Grouping of Elements), it iteratively combines adjacent bigrams into progressively longer sequences based on lexical association strengths. We then provide empirical evidence for this approach via two case studies. First, we compare the performance of MERGE to that of another algorithm by examining the outputs of the approaches compared with manually annotated formulaic sequences from the spoken component of the British National Corpus. Second, we employ two child language corpora to examine whether MERGE can predict the formulas that the children learn based on caregiver input. Ultimately, we show that MERGE indeed performs well, offering a powerful approach for the extraction of formulas.

Keywords: formulaic sequences, collocation extraction, lexical association, child language, MERGE, adjusted frequency list

1. Introduction

Bolinger (1976, p. 2) famously claimed that "speakers do at least as much remembering as they do putting together", suggesting that the production of complex linguistic constituents (e.g. multiword phrases) was as often about retrieving these items from memory in pre-fabricated form as it was about constructing them online based on regular rules. While at the time such a view was seen as radical, today an increasing number of studies are examining the importance, complexity, and ubiquitousness of such *formulaic language* or *phraseology* (see Wray, 2002; Granger and Meunier, 2008 for influential discussion and overviews.)

https://doi.org/10.1075/ivitra.24.05wah

This term broadly encompasses many different types of multiword and co-occurrence phenomena. Prototypical or well-known kinds of formulaic language include idioms (*kick the bucket*), prepositional verbs (*talk about*), phrasal verbs (*pick up*), multiword prepositions (*in spite of*), and nominal compounds (*gold medal*), among others. But formulaicity operates at a more subtle level, too. Consider the well-known example of the two semantically similar words *strong* and *powerful*, where only the former is typically applied to the noun *tea*. This case demonstrates the phenomenon of *restricted exchangeability* (Erman and Warren, 2000), whereby formulaic language may be diagnosed when one or more words in a word sequence could not be substituted with synonyms without a loss in the particular meaning of that sequence. The implication is that the production of the noun phrase 'strong tea' cannot be based purely on a generative phrase structure rule agnostic to the lexical combinatorial preferences of individual words; rather, the language user must store some knowledge that circumscribes a complete phrasal unit populated with these particular lexical items. In other words, the whole is more than the sum of its parts.

While restricted exchangeability is of limited use in cases where there are no suitable synonyms (e.g. when a word sequence comprises only function words), Erman and Warren (2000) determined, primarily based on this criterion, that at least 50% to 60% of the corpora they examined comprised formulaic language. Numerous other studies have yielded formulaic sequence density estimates as well, with often wildly different results and, because of differences in diagnostic criteria, some counts of corpus formulaic language density going as high as 80% (Altenberg, 1998). This all suggests that Bolinger's historic claim, while hard to verify numerically exactly, may have essentially been correct. Ultimately, regardless of exact density, it is clear that formulaic language is an important feature of language that was ignored in much of mainstream linguistics until work from a phraseological perspective (e.g. Wray, 2002) and work from a usage-based perspective on how much is stored and how much is computed (e.g. Bybee, 2010) zoomed into what had largely been a computational-linguistic task/phenomenon.

In order to study formulaic language (or collocations), one must be able to identify these sequences in discourse. However, this is no straightforward task. One option would be to annotate sequences by hand, but then sequence identification criteria must be defined. Perhaps the most frequent approach is to simply ask annotators who have specialist-level familiarity with formulaic language to perform the task (e.g. Ellis et al., 2008). An obvious objection to such an approach would be the chance for bias, so annotations from different raters are often compared in order to arrive at a reasonably consistent set of annotations. Still, the nature of the mental criteria individual raters are applying is not necessarily clear.

Alternatively, annotators may be provided with more specific instructions in how to identify formulaic language, yet these are prone to the problem of formulaic

language definitions typically being insufficiently comprehensive. So, the aforementioned restricted exchangeability used by Erman and Warren (2000) is a succinct criterion and works well for certain sequences, but it cannot be applied in cases where, as mentioned above, there are no suitable synonyms to exchange for a given word in the sequence (to check whether that sequence thereby loses its idiomaticity under the exchange).

Finally, one could define more elaborate annotation criteria – for example questions aimed at identifying specific types of formulaic sequences (e.g. "is this sequence a nominal compound comprising two or more nouns with a non-compositional meaning?" or, "does this sequence function as a single multiword preposition?" etc.). Yet even still, certain obviously formulaic sequences can be difficult to definitively categorize, particularly in the case of sequences that do not co-extend with syntactic constituents (see Biber et al., 2004). In addition, as is particularly true of this last approach, manual annotation is slow and backbreaking work.

Ideally, one would want to be able to extract a reasonably reliable list of formulaic sequences from a corpus without an excessive amount of manpower. For this reason, a widely-used alternative to manual annotation is different collocational extraction algorithms, implemented computationally and applied to corpora. The algorithms vary in their designs, but they all return an ordered list of multiword sequences, whose ranking may be thought of as representing the confidence of the algorithm in the degree to which any sequence represents a true formulaic sequence. This ranking is assembled according to some statistical measure – which is itself based on the frequency of each sequence and the contingency/predictability of its parts – but the particular statistical measure used often varies from algorithm to algorithm; see below.

Thus, broadly speaking, automatic extraction is successful insofar as usage frequency is correlated with formulaicity. And indeed, much research has shown that the more often language users deploy a particular formulation rather than an alternative one with the same meaning, that formulation increasingly becomes (via statistical preemption) the conventionalized way of expressing oneself (and, in turn, comes to no longer mean "the same thing" as erstwhile alternatives) (e.g. Bybee, 2010, Chapter 3). At the same time, the results of automatic extraction algorithms are still noisy. The reason for this noisiness is twofold: On the one hand, this has to do with the fact that this correlation between usage and formulaicity is not perfect. On the other hand, different algorithms yield results that differ in their goals (e.g. lexicographic and translational goals differ) and their methodological implementation (statistical algorithms react differently to input frequencies), which affects their output and, thus, also their quality.

In the current study, we present an algorithm that we have developed entitled MERGE (for Multi-Word Expressions from the Recursive Grouping of

Elements).[1] We believe that our approach addresses some of the limitations of previous approaches from the literature, with regard to both these issues of counting sequences and identifying them. To investigate the degree to which this is true, we formulate the following research question:

RQ1: Does our algorithm perform better than a more conventional approach when both are compared to manual annotation?
Relatedly, remember that the ultimate goal for formulaic sequence identification is often some downstream research such as variety research, psycholinguistic processing, and L1 acquisition. For example, researchers have examined, among other things, dialectological differences on the basis of differences in formulaic language (Gries and Mukherjee, 2010); the degree to which formulaic sequences are processed more quickly than non-formulaic ones by adults (Arnon and Snider, 2010); and the degree to which formulaic sequences play a role in early child language (Lieven et al., 2009). And while many such approaches rely on manual annotation (but see Gries and Muhkerjee, 2010), if a particular corpus extraction approach is viable, it ought to be possible to put this to use in place of manual annotation. Thus, a second research question that we pursue is:

RQ2: Can an extraction algorithm be successfully employed as part of the methodology of a formulaic language-focused study?

In the next three subsections, we discuss in more detail the issues surrounding contemporary computational extraction approaches. Then, in Section 2, we define our extraction approach. The sections that follow comprise case studies: Section 3 evaluates our approach on the basis of annotated corpus data and aims to address the first research question, while Section 4 addresses the second research question by demonstrating the applicability of our approach to formulaic language research through a small case study on child language. Finally, in Section 5, we discuss conclusions and directions for future research.

1.1 Counting co-occurrences

One of the most important variables affecting the performance of different automatic extraction approaches is the statistic a particular algorithm uses to weight or merge word co-occurrences. Probably the two most popular methods, or families

1. We use the terms multiword expression (or MWE), formula, and formulaic sequence interchangeably here.

of methods, are (i) *relative frequencies*, which are simply the frequencies of a co-occurrence normalized, typically, for the frequency of the first/node word of a collocation/formulaic sequence, and (ii) *lexical association measures*. Numerous association measures have been proposed (e.g. Pecina (2009) reviews 80), and they vary mathematically and, therefore, in the precise list of results that they return. However, generally speaking, the most widely-used association measures are based on how much more or less often a particular sequence is observed than might be expected by chance. Such scores are calculated by considering not just the frequency of the target sequence, but other pieces of frequency information relevant to the occurrence as well. Depending on the specific measure, this may include the frequencies of the individual words (see above), as well as the size of the corpus (usually measured in words).

Most of these measures are based on *contingency tables*, such as the one in Table 1, which schematically represents the observed and expected frequencies of occurrence of the two constituents of a bigram (or any bipartite co-occurrence, for that matter).

Table 1. Schematic 2×2 table for bigram co-occurrence statistics / association measures

	word$_2$ = present	**word$_2$ = absent**	**Total**
word$_1$ = present	*obs.*: *a* *exp.*: $^{(a+b)\times(a+c)}/_{n}$	*obs.*: *b* *exp.*: $^{(a+b)\times(b+d)}/_{n}$	$a+b$
word$_1$ = absent	*obs.*: *c* *exp.*: $^{(c+d)\times(a+c)}/_{n}$	*obs.*: *d* *exp.*: $^{(c+d)\times(b+d)}/_{n}$	$c+d$
Totals	$a+c$	$b+d$	$a+b+c+d=n$

Based on the frequencies represented in Table 1, an association measure returns an association score for each co-occurrence type; these scores may then be used to rank the bigrams in a corpus by strength or significance. While each measure's scores represent different units, often a positive value will indicate statistical attraction between two words: that is, that the two words co-occur more often than might be expected by chance. Conversely, a negative value will indicate statistical repulsion, or that two words occur less frequently than might be expected by chance (see Evert, 2004, 2009 for comprehensive discussion).

Lexical association measures tend to offer greater sensitivity to formulaic language than relative frequency, since they can capture sequences that are infrequent though nonetheless formulaic. Consider the bigrams *San Francisco* and *in the*. While the latter sequence is clearly more frequent (and would thus be ranked more highly on a frequency list), most would agree that the former is a 'better' formulaic sequence. This is because when one of the unigrams *San* and *Francisco* does occur, there is a high probability that the other will, too, whereas when *in* and *the* occur, they may occur together but they very often occur apart as well. In other words,

San and *Francisco* embody a much greater degree of contingency than do *in* and *the*. It is this feature that most lexical association measures are designed to capture.

Of the measures that have been developed, some have emerged as more popular than others. For example, pointwise mutual information (MI) is probably the most well-known association measure. However, MI and transitional probability – which is not usually considered a lexical association measure but nonetheless measures sequence strength – exhibit a similar problem: they often rank very low-frequency, high-contingency bigrams too highly, even in the case of a bigram in which both component words are hapaxes (see Daudaravičius and Murcinkevičiené, 2004, pp. 325–326). In other words, these two measures have the opposite problem of relative frequency. Ideally, one would want a measure that 'splits the difference' between these two extremes. While alternatives such as MI^k fare somewhat better in this respect (see McEnery, 2006; Evert, 2009, p. 1225), one lexical association measure that has yielded quite good results for multiword extraction (e.g. Wahl, 2015), and does not appear oversensitive to very low frequencies is log-likelihood (Dunning, 1993), whose formula is given in (1).

(1) $$log\ likelihood = 2\Sigma_{i=a}^{d} obs \times log \frac{obs}{exp}$$

Unlike some other measures, log-likelihood takes into account observed and expected values from all four frequency cells (*a*, *b*, *c*, and *d*) of the kind of contingency table shown in Table 1. Because of the very widespread, successful adoption of log-likelihood in many studies (collocation studies, multiword extraction studies, keywords studies, etc.), log-likelihood is the measure we use in the algorithm that we develop here.[2]

The reader may note that we have not discussed in this section co-occurrences of higher-order *n*-grams. This is not an omission, but rather reflects the fact that virtually all lexical association measures are designed for two-way co-occurrences. This is of course problematic, since formulaic sequences may theoretically comprise any number of words. Some techniques for adapting lexical association to higher-order *n*-grams have been developed (see also below), but no best practice has emerged yet. In addition, while relative frequency does exhibit the insensitivity to low-frequency, high-contingency sequences as discussed above, it does not have the bigram restriction, and thus still is used by researchers today (e.g. O'Donnell, 2011). We return to these issues in a little while.

---

2. One final point that should be made is that (1) will always result in positive values. Thus, in order for log-likelihood scores to correspond to the convention in which positive values denote statistical attraction between words and negative values repulsion, the product of Equation 1 must be multiplied by –1 when the observed frequency of a bigram is less than the expected (following Evert, 2009, p. 1227).

1.2 N-Gram sizes/configurations and the problem of redundancy

Once a scoring metric has been chosen, a typical next step is to select one or more *n*-gram sizes for extraction. Furthermore, one may choose *n*-gram templates that contain one or more gaps in them. This reflects the possibility of discontinuous formulaic sequences, exemplified by the *as _ as* construction in *as tall as* or *as little as*. Next, all *n*-grams corresponding to the selected templates are extracted from a corpus, they are scored, and then they are ranked: ultimately, the higher-ranked *n*-grams are the algorithm's best hypotheses for true formulaic sequences.

However, even if one uses relative frequency for scoring or if one manages to adapt lexical association measures to co-occurrences greater than 2-grams, one still faces an issue of redundancy with this conventional approach. Specifically, if one extracts the 5-gram *as a matter of fact*, one will have also extracted the 4-grams *as a matter of* and *a matter of fact*. Because these 4-grams are at least as frequent as the 5-gram that contains them, they might be ranked higher (if ranking is based on frequency, or, in the case of a lexical association-based ranking, since strength is correlated with frequency). Of course, this effect is a problem since, in the case of *as a matter of fact*, the 5-gram is clearly a better hypothesis for a 'true' formulaic sequence than any $n<5$-grams included in *as a matter of fact*.

1.3 Recent approaches

Some recent approaches address the above-mentioned issues. For example, Daudaravičius and Marcinkevičienė (2004) develop a new lexical association measure called lexical gravity *G*. This measure computes the lexical association of two elements *x* and *y* by not only using the information in Table 1 above (i.e. the token frequencies with which *x* and *y* are observed in the corpus together and on their own), but also using the numbers of types with which *x* and *y* co-occur (i.e. the type frequencies underlying the token frequencies of cells *b* and *c* in Table 1). They then apply this measure to the identification of formulaic language by, so to speak, moving through a corpus incrementally and considering any uninterrupted sequence of bigrams with a *G*-score exceeding a threshold as constituting a formulaic sequence, or 'collocational chain' in their terminology.

In a later paper, Gries and Mukherjee (2010) develop a modification of lexical gravity for the identification of formulaic language. Specifically, they extract sequences of various lengths and score them on the basis of the *G*-score of their component bigrams, discarding those sequences with mean *G*-scores below a certain threshold. Then, they proceed through the list, discarding sequences that are contained by one or more $n+1$-grams with a higher mean *G*-score. The resulting list

constitutes their algorithm's hypothesis of the formulaic sequences in the corpus. With this pruning process, Gries and Mukherjee's approach also addresses the redundancy issue mentioned in the previous section, whereby high-scoring grams may merely be fragments of larger, true grams. However, the fact that lower-order *n*-grams are entirely discarded if a higher-order *n*-gram containing them is stronger is potentially problematic: while certain tokens of a lower-order *n*-gram may be fragmentary (*fingers crossed* in *to keep one's fingers crossed*), others may not be (*fingers crossed* in *Speaker A: "I hope we win!" Speaker B: "Fingers crossed!"*).

A recent approach by O'Donnell (2011) takes a different approach to extracting formulaic sequences of various sizes, which also avoids the problem of redundancy: Rather than adapting lexical association measures to co-occurrences beyond the bigram, O'Donnell employs frequency counts as a metric of formula strength. His Adjusted Frequency List (AFL) works by first identifying all *n*-grams up to some size threshold in a corpus. Next, only *n*-grams exceeding some frequency threshold (3, in his study) are retained in the AFL along with their frequency. Then, for each *n*-gram, starting with those of threshold length and descending by order of length, the two components *n*-minus-1-grams are derived. Finally, the number of tokens in the frequency list of each *n*-minus-1-gram is decremented by the number of *n*-grams in which it is a component. Like the approaches above, this procedure prevents the kinds of overlaps and redundancies that would result from a brute-force approach of simply extracting all *n*-grams of various sizes and then ranking them based on frequency. However, in using the AFL, there is a very real risk that low-frequency though high-contingency formulaic sequences would be ranked (too?) low, while high-frequency though non-formulaic sequences would be ranked (too?) high.

One drawback shared by all of the approaches discussed thus far is that, as implemented, they do not allow for discontinuous formulaic sequences. A recent algorithm by Wible et al. (2006) addresses this limitation. Their approach also crucially differs from these other approaches in that it does not generate a list of ranked formula hypotheses contained in a corpus. Instead, it is designed to find all of the formulaic sequences that a given node word participates in (in this way, it is more akin to a concordance). Their algorithm represents what we will call a recursive bigram approach. Upon selection of a node word to be searched, the algorithm generates continuous and discontinuous bigrams within a specified window size around each token of the node word in the corpus; these bigrams consist of all those that have the node word as one of their elements. Next, the algorithm scores the bigrams on the basis of a lexical association measure (they use MI), and all those bigrams whose score exceeds a specified threshold are 'merged' into a single representation. The algorithm then considers new continuous and discontinuous bigrams, in which one of the elements is one of the new, merged representations

and the other element is a single word within the window. The new bigrams are scored, and winners are chosen and merged. This process iterates until no more bigrams exceeding the threshold are found. Ultimately, the algorithm generates a list of formulaic sequence of various sizes that contain the original node word.

Importantly, the model never has to calculate association strengths for co-occurrences larger than two elements, since one element will always be a word, and, after the first iteration, the other element will always be a word sequence containing the node word. The obvious limitation of this approach is that it is not designed for broad-scale use on all words in a corpus. In principle, one could treat every corpus word type as a node word. However, this would result in numerous instances of redundancy, whereby partially or fully overlapping formulaic sequences would be grown from neighboring node words. And because the authors did not intend for their algorithm to be used for applications other than concordance, they do not offer a suggestion for how this might be addressed.

In the next section, we present our algorithm, which addresses all of the issues raised so far: scalability of lexical association, redundancy, discontinuity, and broad-scale use on all words in a corpus.

2. The MERGE algorithm

Similar to the algorithm developed by Wible et al. (2006), the MERGE algorithm embodies a recursive bigram approach. But unlike their work, our algorithm is designed to extract *all* formulaic sequences in a corpus – not just those that contain a particular node word. It begins by extracting all bigram tokens in the corpus. These include adjacent bigrams, and potentially bigrams with one or more words intervening, up to some user-defined discontinuity parameter (similar to Wible et al.'s use of a window). The tokens for each bigram type are counted, as are the tokens for each individual word type, and the total corpus size (in words) is tallied. Next, these values are used to calculate log-likelihood scores. The highest-scoring bigram is selected as the winner, and it is merged into a single representation; that is, it is assigned a data structure representation equivalent to the representations of individual words (this differs from Wible and colleagues' approach, wherein multiple winners were chosen at an iteration on the basis of a threshold association value). We call these representations *lexemes*. At the next stage, all tokens of co-occurring word lexemes in the corpus that instantiate the winning bigram are replaced by instances of the new, merged representation. This process by which smaller tokens are consumed by larger winners avoids the kinds of redundancy issues raised above, in which a particular word token or sequences of tokens may simultaneously participate in numerous fragmentary grams.

Frequency information and bigram statistics must then be updated. New candidate bigrams are created through the co-occurrence in the corpus of individual word lexemes with tokens of the new merged lexeme. Furthermore, certain existing candidate bigrams may have lost tokens. That is, some of these tokens may have partially overlapped with tokens of the winning bigram (i.e. they shared a particular word token). Since these word tokens in effect no longer exist, these candidates' frequency counts must be adjusted downward. Moreover, the frequency information for the individual word types found in the winner must be reduced by the number of winning bigram tokens. Finally, the corpus frequency has decreased, since individual words have been consumed by two-word sequences. After these adjustments in frequency information have been made, new bigram strengths can be calculated.

The cycle then iteratively repeats from the point at which a winning bigram is chosen above, and iterations continue until the association strength of the winning bigram reaches some user-defined minimum cut-off threshold or until a user-defined number of iterations has been completed. The output of the algorithm is a corpus, parsed in terms of formulaic sequences, and a list of lexemes, from individual words to formulaic sequences of different sizes.

Because the input to candidate bigrams at later iterations may be output from previous iterations, MERGE can grow formulaic sequences unrestricted in size (even while never considering co-occurrences larger than two items), which is similar to the Wible et al. (2006) algorithm. Another key difference, however, is that one element of their candidate bigrams must always be a single word and the other a word sequence (at least after the first iteration, where both elements are single words). In contrast, at later iterations, MERGE can choose a winning bigram that comprises two single words, a single word and a word sequence, or two word sequences. Moreover, assuming a sufficiently sized gap parameter, one element may in principal occur inside the gap of another element. Even more unusual scenarios are possible: *as _ matter* and *a _ of fact* could be interleaved to form *as a matter of fact*. Thus, there are many possible paths of successive merges that result in particular formulaic sequences, provided that the leftmost word of the two elements of a bigram never exceed the discontinuity parameter.

3. Case study 1: MERGE vs. AFL

In this case study, we address our first research question, "does our algorithm perform better than a more conventional approach when both are compared with respect to manual annotations?" To answer this question, we chose to compare the performance of MERGE to that of one of the other algorithms discussed earlier, the Adjusted Frequency List (AFL), by O'Donnell (2011).

Like MERGE, the AFL addresses the redundancy/overlap problem faced by algorithms that simply extract and rank all *n*-grams of various sizes. However, unlike MERGE, the AFL uses frequency rather than lexical association. In another study (Wahl and Gries, 2018), we show that the use of frequency reduces the quality of the formulaic sequences found by the AFL significantly, compared with those found by MERGE. However, in that study, we evaluated the performance of the two algorithms on the basis of a rating experiment conducted using naïve participants (i.e. participants who had no explicit knowledge of formulaic language and received instructions/examples describing it on the spot).

Here, we wish to see if the superior performance of MERGE holds up in a different test situation, namely a corpus already annotated for formulaic sequences. In other words, while in the previous study we assessed the performance via naïve intuitions, here we are testing performance via specialist knowledge, as those who annotated the corpus must have had some relevant lexicographic training to do so.

3.1 Materials

The corpus we use is the spoken component of the British National Corpus (BNC), which comprises approximately 10 million words. Crucially, this component of the corpus was tagged for formulaic sequences; in total, there are 436 sequence types (once tagged and all identical sequences conflated). However, a number of these sequences contain disfluencies such as *er* or *erm*. There are a total of 48 such items, and all of their 'clean' forms are also found amongst attested among the BNC's formulaic sequences. Thus, when they are removed from the list, there are only 388 total BNC items. Having worked extensively with formulaic sequences, we must point out that this estimate likely seriously underestimates the number of formulaic sequences actually present in the BNC spoken component. Consider, for example, the work of Erman and Warren (2000), who found over 50% of their corpus comprised formulaic sequences. This would mean that over 5 million words of the BNC spoken component would be distributed among a mere 388 types, which is obviously not the case – rather, the BNC annotators must have used much more conservative criteria in determining formulaic sequences than did Erman and Warren.

In order to compare the performance of the algorithms, all 388 sequence types were first obtained from the corpus. The corpus was then preprocessed so that only word strings along with utterance boundaries were retained. Next, MERGE was run for 10,000 iterations on the corpus, with the maximum gap size set to 0 (only adjacent sequences were permitted). Additionally, the AFL was run on the corpus and the top 10,000 most frequent items were selected from the list that was

generated. Note that items 9977 through 10539 in the AFL output were all tied with a frequency of 35. In order to arrive at an even 10,000 items, we randomly selected 10,000–9977 = 23 items from these 10539–9977 = 562 total tied items.[3]

The 10,000 items from each of these runs of the respective algorithms then served as the basis for comparison with respect to the 388 tagged types from the BNC.

3.2 Results

First, we checked how many of the 388 formulaic sequences from the BNC spoken were identified by the top 10000 MERGE items and by the 10000 AFL items: MERGE found 112 of the 388 formulas whereas the AFL found only 93 of the same 388 formulaic sequences. According to a one-tailed binomial test, MERGE finds a significantly higher number of formulaic sequences [binom.test(112, 388, 93/388, alternative="greater"), $p_{\text{one-tailed}} = 0.01522$); conversely, according to a second one-tailed binomial test, the AFL performs significantly worse than MERGE [binom.test(93, 388, 112/388, alternative="greater"), $p_{\text{one-tailed}} = 0.01779$].[4]

In order to more closely analyze the differing performance of MERGE and the AFL, we present Table 2, in which each column corresponds to a different category of (non-)overlap between the algorithm outputs. Thus, column A contains those items in the BNC identified by both algorithms; column B contains those identified by MERGE but not the AFL; column C contains those identified by the AFL but not by MERGE; and column D contains those BNC items identified by neither algorithm. Note that columns A and D contain only a sampling of the total number of items in those categories.

One way to explore these sets of items quantitatively is via the parameters that matter to, or are inherent to, formulaicity: frequency of occurrence, dispersion, and lexical association. Dispersion refers to how evenly tokens of a particular type are distributed in a corpus and we are using the "DP" measure of dispersion (Gries, 2008). If tokens are perfectly evenly distributed in a corpus, DP will approach 0,

---

3. In order to rule out an effect of which 23 formulas with the AFL frequency of 35 were sampled, we conducted a Monte Carlo simulation with 1,000 iterations in which the 23 formulas were replaced with 23 randomly-sampled items from all formulas with the AFL frequency of 35. The mean and 95%-confidence interval of how many of the 388 BNC-grams the AFL found were 93.03 ([93.02, 93.04], see below), which means our randomly chosen items did not skew the results in any direction (let alone in our favour).

4. We performed one-tailed tests because our first comparison of MERGE and AFL (Wahl and Gries, 2018) showed that MERGE outperformed AFL; however, even two-tailed tests proved significant for MERGE outperforming the AFL ($p_{\text{one-tailed}} = 0.02761$) and the AFL performing worse than MERGE ($p_{\text{one-tailed}} = 0.03326$).

Table 2. Comparison of attestation of BNC items among the results of the two algorithms

Column A: +M,+A (83 types)	Column B:+M,-A (29 types)	Column C: -M,+A (10 types)	Column D: -M,-A (266 types)
by way of, subject to, as usual, in case, even if, and so on, in relation to, a little, that is, next to, off of, for good, for instance, just about, for the time being, as regards, even though, each other, as it were, at once, sort of, by now, old fashioned, from time to time, of course, all round, as to, no longer, for example, kind of, in between, rather than, as opposed to, …	*in addition, whether or not, vice versa, up to date, in order, half way, depending on, up front, up until, all of a sudden, anything but, grand prix, status quo, as if, know how, per cent, in common, fed up, so as, every so often, in accordance with, as though, en suite, a great deal, less than, per annum, an awful lot, sinn fein, out of date*	*given that, in respect of, as yet, in full, for certain, in the main, near to, no matter what, with regard to, except for*	*relative to, hard up, poco a poco, now that, teeny weeny, al fresco, at large, au fait, a la, in search of, no matter how, grand mal, a la carte, as between, as from, au revoir, nom de plume, from now on, ad hominem, in return for, in place of, insofar as, as for, except for, in relation to, once more, all at once, au pairs, pate de foie gras, in vain, in proportion to, de facto, raison d'être, …*

whereas if tokens are extremely clumpily distributed (i.e. largely or even exclusively concentrated in one part of the corpus, then DP will approach 1). As a lexical association measure, we are using the MI2 measure, a version of MI that rewards *n*-grams with higher observed frequencies – log (obs a^2/exp a) – and we computed the expected frequencies on the basis of the assumption of complete independence.[5]

In order to visualize the distributional properties of the tokens in columns A-D with respect to frequency, dispersion, and MI2, Figure 1 displays empirical cumulative distribution (ECD) plots for frequency, dispersion, and lexical association respectively for all columns A-D, but our discussion will focus on the comparison of B versus C and the comparisons of both B and C with A.

At this stage of exploration of MERGE (vs. the competing algorithm), we are not yet in a position to state specific alternative hypotheses – let alone directional ones or specific effect sizes – regarding how MERGE and the AFL differ along these three parameters other than the maybe most obvious one that MERGE should behave differently with regard to MI2 because it is an algorithm whose computations involve a measure of association strength. Thus, we are restricting our discussion here to an exploratory description. With regard to the frequencies, it is obvious that

5. That means, expected frequencies were computed as they would in chi-squared tests of independence; for a 3-gram that would be ($f_{word1} \times f_{word2} \times f_{word3}$) ÷ corpus size^2; see Gries (2015, Section 2.2.1 for an example and why this can only be a first heuristic).

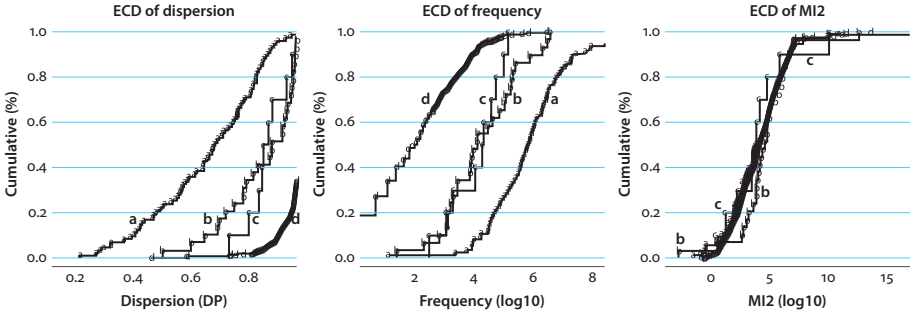


Figure 1. ECD plot of frequency (left panel), dispersion (DP, center panel), and lexical association (MI2, right panel)

(i) the formulas identified by both MERGE and the AFL are those with the highest frequencies and (ii) the formulas identified by neither MERGE nor the AFL are those with the lowest. Also, (iii) the formulas of columns B and C do not seem to differ from each other in terms of their average frequency or the variability of their frequencies, while both B and C differ from those of A (i.e. those formulas that both algorithms found). Put differently, both MERGE and the AFL agree on many high frequency collocates but the formulaic sequences that only one of them finds do not differ from each other in terms of their corpus frequencies.

With regard to dispersion, the picture changes a bit: Again, (i) the formulas identified by both algorithms are the ones with the lowest DP-values (i.e. most evenly distributed in the corpus), but it is also worth noting that the formulas found by both algorithms exhibit DP-values across the whole range of values. Then, (ii) the formulaic sequences identified by neither MERGE nor the AFL are those with the highest DP-values / clumpiness and very little variability of dispersion: 75% of the DP-values of column D are ≈ 0.96 or higher. However, (iii) while the formulaic sequences found by only one algorithm do not differ in their average dispersion, they appear to differ in the variability of their dispersion: the interquartile range of the B formulas is twice as high as that of the C formulas, which we interpret as advantageous for MERGE, because it can be seen as indicating that MERGE is better at finding formulaic sequences with diverse dispersions.

Finally, with regard to lexical association, the results are quite different: (i) the main findings are that the formulas found by at least one algorithm (i.e. those in columns A, B, and C) do not differ much from each other in terms of either central tendency or variability (with just a small effect of column A exhibiting a wider range of MI2-values). In addition, the formulas of column B do exhibit somewhat larger mean and median MI2s than those of column C, but the effect is merely suggestive at this point (in part because of the very small sample size of 10 items in column C).

Given the just-mentioned small number of cases in C, it is difficult to make a detailed qualitative comparison at this point, but it does seem to us that three of the ten column C formulas are not 'as good' examples of formulas as all of those in column B, to the extent that they seem to be incomplete or less frequent – specifically, *in respect of*, *in the main*, and *near to* – but this assessment awaits future (rating?) studies to be put on a more solid footing.

3.3 Interim conclusions

In the comparison of MERGE with the AFL in Wahl and Gries (forthcoming), we essentially employed what one might consider a kind of unsupervised approach: we ran both algorithms and then compared samples of top-ranked formulaic sequences. We found that there was a striking difference between the kinds of sequences identified by MERGE and the AFL, which patterned like the *San Francisco*/*in the* example discussed above. The present study, by contrasts, is essentially more similar to a supervised classification approach: we had a list of 388 likely positives and then the degrees to which the algorithms find them. Accordingly, we do not find the same *San Francisco*/*in the* bifurcation in the results. Rather, the results were more nuanced, with numerous items identified by both algorithms, and subtle differences in the items that were identified only by one or the other algorithm; it seems that MERGE does better in particular by being able to find formulas from a wider range of dispersion values, as well as exhibiting the tendency of identifying formulas with higher association scores.

4. Case study 2: Exploring MERGE in the context of L1 acquisition

As mentioned above, formulaic language extraction from corpora is typically a means to some other research end, used in fields as diverse as cognitive-/psycholinguistics, dialectology, digital humanities, applied linguistics, and many others. Thus, to provide evidence that an automatic extraction approach such as MERGE is powerful enough to be methodologically applicable to such downstream formulaicity research, we deploy it here in a small applied study.

Within the cognitive domain, formulaic sequences play a particularly integral role in child language. Specifically, current theories hold that they serve as a stepping stone on a child's way to more productive grammatical knowledge: children begin with stored formulaic sequences and, over time, generalize across them to acquire a mature grammar (see Tomasello, 2005 for one of the most thorough overviews); at the same time, this is not to say that representations of formulas acquired during

childhood do not endure into adulthood, nor that new formulas are not acquired beyond childhood. One question, though, is whether these early representations are truly formulaic, and not creatively constructed. One source of evidence for this would be if demonstrably formulaic structures in the adult input to the child are taken up and deployed in the child's own productions. Meanwhile, adult creative structures ought not to be reproduced by the child, at least at the same rate.

This broad style of approach, in which specific child productions are linked to specific adult inputs, has been used elsewhere in the child language literature. For example, Bod (2009) developed a parsing/grammar induction algorithm called UDOP (Unsupervised Data-Oriented Parsing). UDOP is based on a Probabilistic Context-Free Grammar (PCFG) that can store and reuse (sub)trees (including specific word terminals) that it had constructed to parse previously-encountered sentences. The lexicalized, reusable nature of the (sub)trees makes them, by definition, formulaic sequences, at least in the context of the model (whether or not they reflect true formulaic sequences known to humans is another question). The primary goal of UDOP is to demonstrate that grammatical knowledge can be induced in a bottom-up fashion, without reliance on innately-specified syntactic knowledge, contra many generative grammarians. Thus, the role of formulaic language in this model is to increase performance in the pursuit of this objective (just as a child may use formulaic language as a stepping stone).

Bod (2009) evaluated UDOP in various case studies. In one, he partitioned a longitudinal child language corpus into two sections, and then trained UDOP on the adult utterances in the earlier partition (in separate trials, he also trained the algorithm on the child utterances, and on a combination of the child and adult utterances).[6] Next, he evaluated the algorithm by seeing how well it could parse the child utterances in the later partition, based on the grammar it had acquired on the earlier adult utterances. The parses assigned were compared against manually annotated, gold standard parses for the data. Indeed, the grammar acquired based on the adult input performed well, demonstrating that a child's emergent grammatical knowledge can be modeled on concrete adult structures that the child has stored.

In another related study, Swingley (2005) examined the distributional learning of word boundaries from syllable co-occurrences. Although he did not investigate

6. A related approach is taken in Bannard et al.'s (2009) study using a Bayesian-based distributional learning algorithm that the authors had developed, as well as in Lieven et al.'s (2009) corpus-based discourse-analytic study. However, a crucial difference is that these studies use child utterances for both training and test; thus, there is no attempt to link the children's acquired structures to adult input, but rather just to account for the children's advancing linguistic development across different stages of the child's own usage.

formulaic word sequences, his design is instructive. He extracted all syllable bigrams and trigrams, scored them on the basis of MI and frequency, and ranked them. He then correlated this ranked list with how well the *n*-grams instantiated words. In other words, he examined the question of how well association strength and frequency can predict the word boundaries that children go on to learn. However, his definition of what children 'go on to learn' is mature, adult-like gold standard boundaries. Furthermore, the corpus he used was not longitudinal, but rather a collection of caregiver utterances (phonologically transcribed) from the input to a collection of different children. An interesting complementary approach would be to examine how well the ranked *n*-grams of (a) specific caregiver(s) predict the word boundaries that their child goes on to learn at the particular developmental stage of the corpus (which would be possible with a longitudinal corpus).

In the current chapter, our approach brings together techniques developed in evaluation methods from the child language studies within Bod (2009) and Swingley (2005). As in both approaches, we train the algorithm (MERGE) on a set of adult utterances. Like Bod (2009) and unlike Swingley (2005), we use longitudinal corpora, focusing on the input to/output from individual children. We compare the multiword representations generated by the model based on earlier adult utterances against the actual output of these children, as registered in later child utterances. And like Swingley and unlike Bod, we work with a list of output candidates scored and ranked on the basis of association strength, rather than best grammatical parses for whole utterances. The hypothesis is that higher-scoring formulaic sequences, extracted from the adult utterances, will go on to be learned/used by the child, while formulas that scored lower will not (at least not to the same degree).

In the next section, we discuss the corpora that we use as well as their pre-processing, and we discuss the technique for generating the stimulus items from the corpora using MERGE. After that, we turn to the results of the study. Finally, we discuss these findings.

4.1 Materials and methods

In this study, we use two longitudinal child language corpora, both of which were sourced from the CHILDES database (MacWhinney, 2000). CHILDES is an online repository for corpora of child language acquisition data. Our selected corpora are the 'Lara' corpus (Rowland and Fletcher, 2006) and the 'Thomas' corpus (Lieven et al., 2009). Both Lara and Thomas are children who have grown up in the United Kingdom (and were thus raised as native speakers of varieties of British English), and the recordings were made in the children's respective homes.

These corpora were selected for several reasons. First, they both span the early multiword speech stage of development, an ideal stage for examining the role of formulaic language in early acquisition: Lara was between the ages of 1;9.13 (i.e., 1 year, 9 months, and 13 days) and 3;3.25 when her recordings were made, and Thomas was between the ages of 2;00.12 and 4;11.20 when his were made. Second, both corpora include extensive speech from the children as well as caregivers with whom they interact (and, in the case of the 'Thomas' corpus, researcher speech as well). Finally, the corpora are relatively large/dense: while 'Lara' comprises 120 hours of transcribed audio, 'Thomas' totals 379 hours of transcribed audio.

The 'Thomas' recordings/transcriptions are in fact divided into 3 subcorpora. The first subcorpus spans the ages of 2;00.12 to 3;02.12, and recordings were made for 1 hour, 4 times per week. The second and third subcorpora span the remainder of the time, and recordings were made for 1 hour, once per week.[7] Because the first subcorpus overlaps in time most closely with the 'Lara' corpus, we only used those recordings. Even with this limitation, the first subcorpus still comprises 279 hours' worth of transcripts (i.e. more than double the size of the 'Lara' corpus). In order to make the corpora more comparable in size, the first 'Thomas' subcorpus was downsampled by including only every other corpus file. This resulted in a more comparable 140 hours' worth of transcripts.

Both corpora were transcribed according to the CHAT format (MacWhinney, 2000), so the same preprocessing procedure was used. This included the removal of metadata, transcriber commentary, punctuation, time stamps, non-speech vocalizations, and incomprehensible syllables. In addition, transcription tags were removed, which marked phenomena such as missing words, grammatically correct forms when an incorrect form appeared, and invented forms, among other things. Note that, while incomprehensible forms were removed, grammatically/phonologically incorrect and invented forms were themselves indeed included. Speaker tags were also removed, but not before they were used to separate each corpus into child and caregiver/adult utterances. Additionally, the two corpora were divided into two partitions, whereby the first two-thirds of each corpus represented partition A and the final third represented partition B.

MERGE was then run on the adult utterances of partition A, once for each corpus. No gaps in the formulaic sequences acquired were permitted, and the algorithm was allowed to run until the log-likelihood score of the top-scoring merge candidate reached 0 (Remember that positive log-likelihood values signify statistical attraction between bigram elements while negative values signify statistical repulsion. By this standard, all bigrams exhibiting a positive log-likelihood score

7. The 'Thomas' corpus additionally included video data, but this was not used in the present study.

are in theory formulaic sequences.). From the final output, all sequences of length 2 through 5 were retained.

Next, all *n*-grams from lengths 2 through 5 were extracted from the child utterances in partition B. From this group, any *n*-grams which also appeared among the child utterances in partition A were discarded in order to ensure that the group comprised only *n*-grams that were new attestations in the child's speech. Finally, the sequences from the MERGE output were compared to the *n*-grams from the partition B child utterances, and two lists were created. The first list comprised those MERGE output sequences that also appeared as *n*-grams in the child utterances. These are formulas that the child plausibly went on to learn in partition B from the input they received from the adult utterances in partition A. The second list comprised those MERGE output sequences that did not appear as *n*-grams in the child utterances. These are items that, despite being MERGE output from the adult utterances from partition A, did not later go on to be learned by the child. The hypothesis is that the log-likelihood scores on the basis of which the sequences were merged ought to be higher for the first 'learned' group than for the second 'nonlearned' group; this is because formulaic sequences with higher degrees of attraction are more likely candidates for acquisition by the child.

Finally, for each child, all of the sequences were grouped into numbered bins on the basis of log-likelihood scores – the lowest-numbered bin contained the sequences with the lowest scores and the highest-numbered bin the highest scores. Then, for each bin, the proportion of sequences that were learned by the child was calculated. In the eventual statistical model (discussed below), the proportions of sequences learned serve as the dependent variable (the variable being predicted). In contrast, the numbers of the bins (BIN) serve as (one of) the independent variables (the variable predicting). In other words, we are trying to predict the proportion of sequences learned by the children on the basis of the BIN, which is a proxy for the log-likelihood score/MERGE order.

Note that, since the 'Thomas' corpus is larger than that of Lara, the number of sequences extracted by MERGE is larger. As a result, we created many more bins for the 'Thomas' sequence scores (213 versus Lara's 75). This is because we wanted there to roughly be the same number of scores in each bin across children (99 scores per bin for Lara and 97 scores per bin for Thomas).[8] Also note that any particular score was placed into its bin only once; that is, if MERGE extracted two different sequences on the basis of the same score, this score would not be duplicated within the appropriate bin.

---

8. The final bins (i.e. the one corresponding to the highest log-likelihood scores), have slightly less than 99 and 97 items in them. Despite this, this set of bin counts and number of items per bin was chosen to ensure that the final bins came as close to possible to the other bins in terms of number of items contained.

4.2 Results

The proportions of sequences learned are plotted against the normalized bin numbers in Figure 2 for Lara (left panel) and for Thomas (right panel); bin numbers were normalized to a 0–1 range to make the values of the two children comparable. Note the consistent pattern across the two. On the right half of each plot, as one moves from mid-range log-likelihood scores to high log-likelihood scores, there is an increase in the proportion of sequences per bin that are learned by each child, which is precisely what we predicted. However, the plots also display something unexpected: Moving from the low log-likelihood scores on the left of the plots, to the mid-range scores, there is a *decrease* in the proportion of learned sequences. This pattern goes against intuition – why might this be?

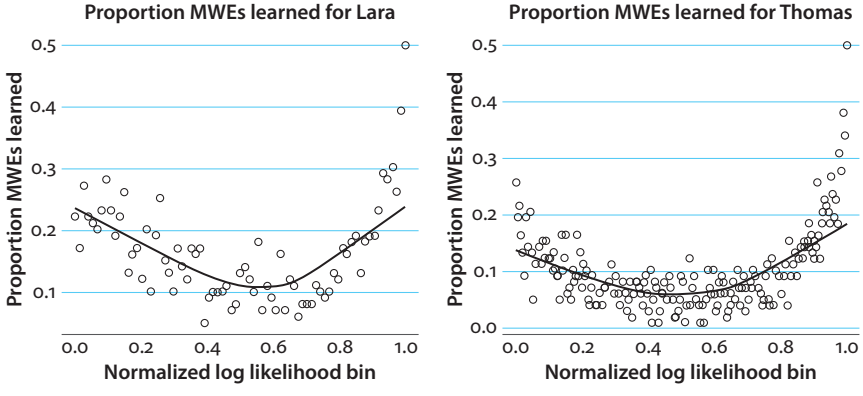


Figure 2. Proportion sequences learned as a function of normalized bin rank

One possible explanation is based on the lengths of the sequences, a factor that plays an important role in which sequences are and are not retained. Thus, in Figure 3, we show average lengths of the sequences in each bin against the bin numbers. Strikingly, the pattern is a virtual mirror image of that depicted in Figure 2, despite the fact that the *y*-axis measures a different unit: proportion of formulas learned in Figure 2 and average sequence length per bin in Figure 3. In the present context, the pattern signifies that, for both children, the average length of very low and very high scoring sequences is very short; however, sequences that were merged on the basis of a mid-range score are, on average, considerably longer.

The isomorphy between the plots in Figure 2 and Figure 3 suggests that perhaps the variable which holds all the predictive power for the proportion of sequences learned is average length, not (normalized) log-likelihood bin. Indeed, in Figure 4, we show average sequence lengths against the proportion of sequences learned for each child, and the apparent correlation between these two suggests that, somewhat

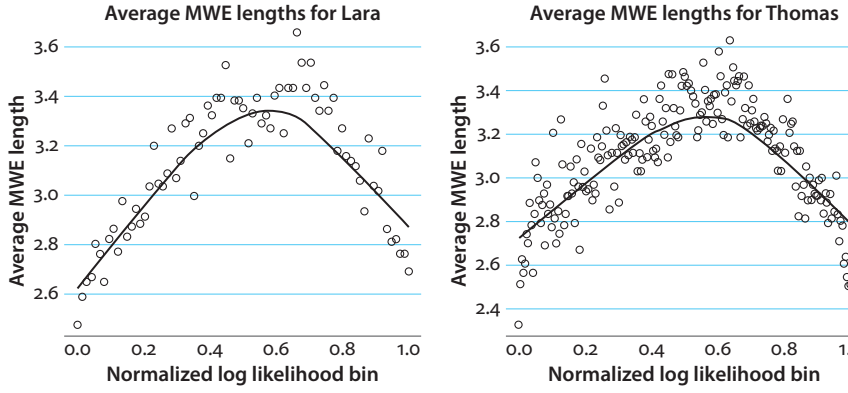


Figure 3. Average sequence lengths as a function of normalized bin rank

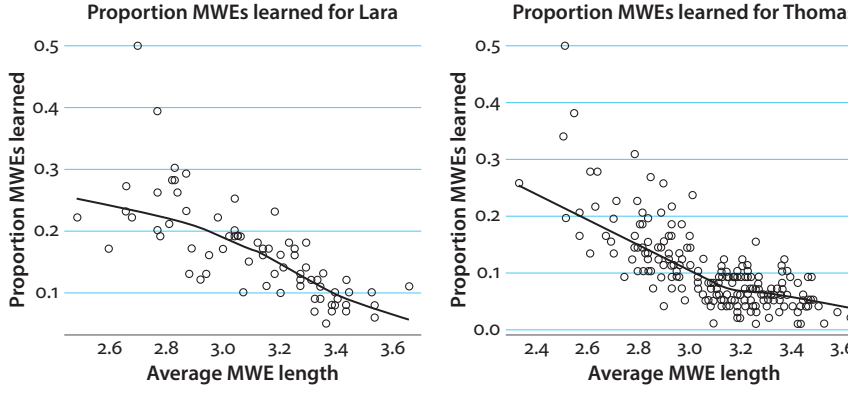


Figure 4. Proportion sequences learned as a function of average sequence lengths

unsurprisingly, average length may be strongly predictive of the dependent variable. This appears particularly true for higher average lengths, where all data points correspond to a low proportion of sequences learned. Note, however, that for shorter average lengths, there are data points which correspond to both rather high and rather low proportions of sequences learned.

To investigate this empirically, we combined the data from the two children and applied a linear model to it. Proportions of sequences learned served as the dependent variable (PROPSrt); to avoid violations of linear model assumptions, we used the square root of the dependent variable (PROPS), while child (CHILD), normalized normalised log-likelihood bin (BIN), and average sequence length (AVELEN) served as predictors. CHILD was a binary variable (Lara vs. Thomas), while all others were numeric. We began with a maximal model in which all numeric predictors were entered as a polynomial to the second degree (to allow for

curvature in the effects) and in which all predictors could interact with each other; model selection tested for the elimination of the polynomial terms and all other predictors. The final model's formula was PROPSrt ~ AVELEN * poly(BIN, 2) * KID (that three-way interaction was very significant: $p = 0.0076$) and that model was highly significant ($F_{11, 276} = 93.54$, $p<10^{-15}$) and achieved a rather high variance explanation (mult. $R^2 = 0.7885$, adj. $R^2 = 0.7801$). All regression coefficients for the model are provided in the appendix, and model checking (homoscedasticity and normality of residuals as well as autocorrelation) raised no red flags.

In Figures 5, 6, and 7, we provide visual representations of the predicted proportions from the final model. Two different perspectives are shown; let us begin with Figure 5 and Figure 6, which are contour plots in which the *x*- and *y*-axes represent the predictors AVELEN and BIN respectively, and the colours and lines displayed represent the predicted proportions of learned formulaic sequences for each combination of the two predictors; for instance, in Figure 5, the plot indicates that the model predicts (and remember that the amount of explained variance was quite high) that, when AVELEN is 3 and BIN is medium (0.5), then the proportion of learned formulas for Lara is about 40% (0.4).

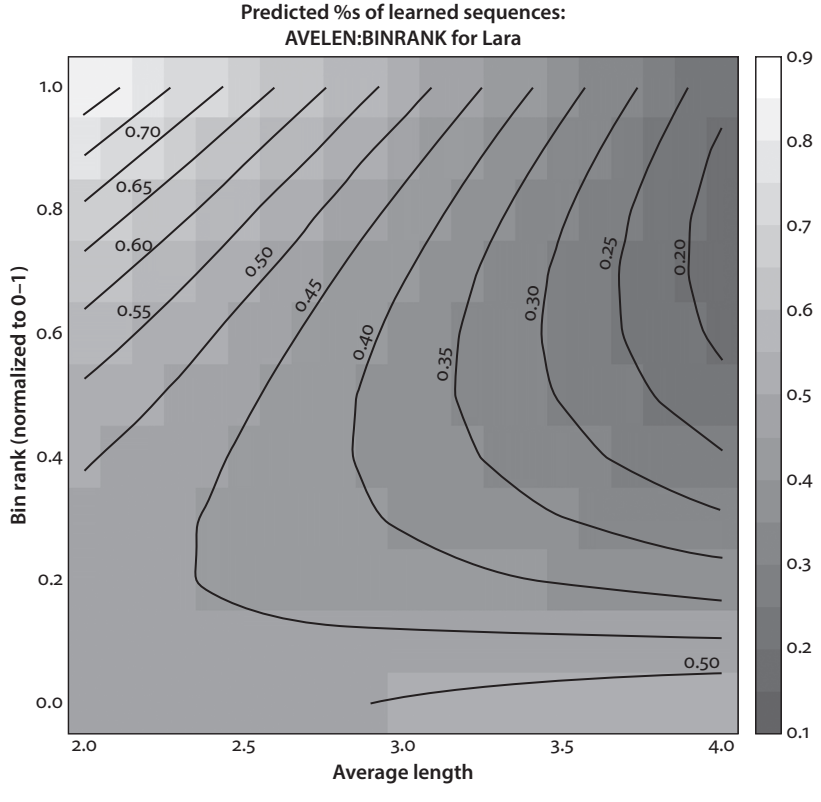


Figure 5. Contour plot of the regression surface of the final model for Lara

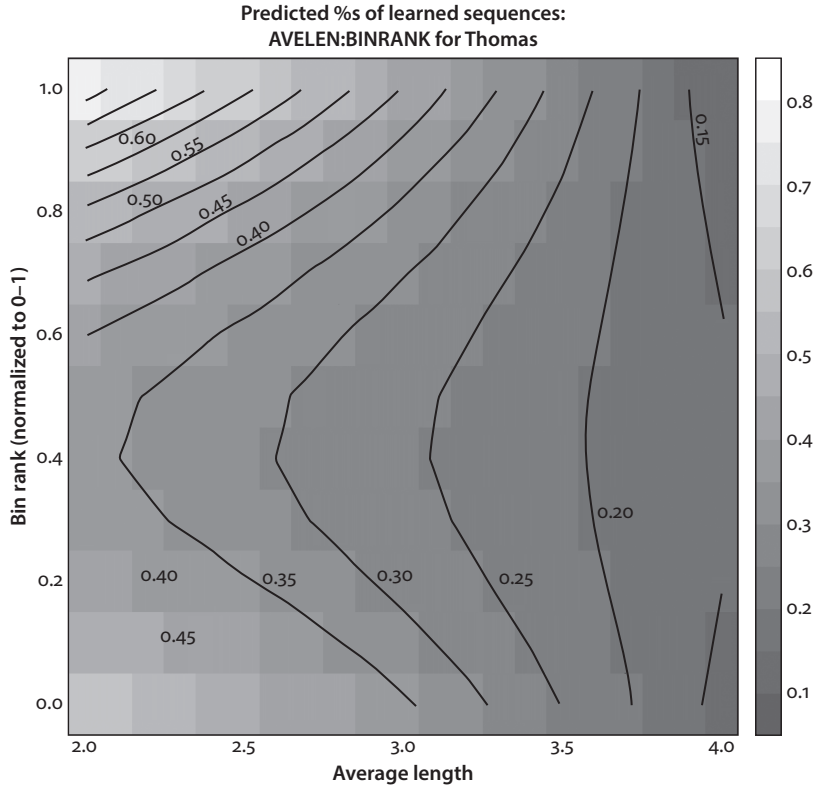


Figure 6. Contour plot of the regression surface of the final model for Thomas

These regression surfaces show that, for both children, short formulas with high log-likelihood scores are learned well/best, and long formulas with moderate to high log-likelihood scores are learned badly. The main difference between the two children is found with low log-likelihood scores: For Lara, there is an effect such that formulas with low log-likelihood scores are learned intermediately well regardless of their length, in fact with a tiny increase for the longer formulas; that finding is not compatible with a long history of research findings on child acquisition and, thus, is somewhat counterintuitive, but it has to be noted that the effect is very small (about a mere 5%) and, for instance, for the BIN-values of 0, 0.1, and 0.2 the slope of the regression surface is statistically not different from 0 (as judged from the predictions' 95%-confidence intervals). For Thomas, the results are more compatible with 'received wisdom': Across all BIN-values, longer formulas are learned less well than shorter ones, but this effect of AVELEN is weakest for intermediately high log-likelihood scores.

With the exception of the (insignificant) slope of the regression surface for low log-likelihood values of Lara, these results make sense and provide some first evidence for higher MERGE bins being learned better even when length is controlled

for. However, it needs to be borne in mind that Figure 5 / Figure 6 provide predictions for all possible combinations of AVELEN and BIN – nevertheless, most of the combinations that are mathematically possible are actually not attested in the data, which makes it useful to consider the predictions specifically for the ranges of combinations of values that *are* attested, which is what is represented in Figure 7. In each panel (one for each child), the *x*- and *y*-axes are the same as in the contour plots above, but now the predicted proportion is represented by an integer value from 0 (lowest predicted proportion) to 9 (highest predicted proportion). In other words, the integer values within the plots can be thought of as 'relative elevations' corresponding to the different predicted proportions of sequences learned, given the intersecting values of the two predictors. In addition to the number, the physical font size of the plotted number represents the predicted proportion as an additional visual clue.

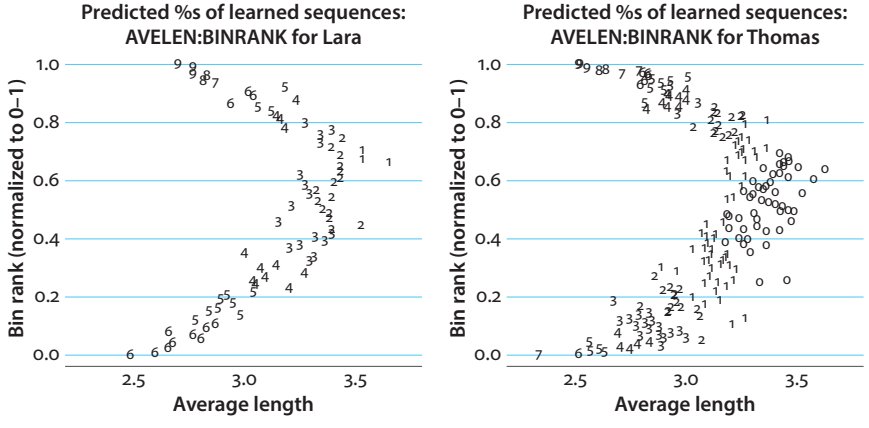


Figure 7. Regression surface of the final model for Lara (left panel) and Thomas (right panel)

This visualization of predicted values – now for those combinations of AVELEN and BIN that are attested – makes the trends even clearer: it is very apparent that there is an overall effect of AVELEN such that both children learn formulas worse the longer they are. For Lara, this effect is weaker, while for Thomas it is stronger. At the same time, one can just as clearly see that, for any observed formula length, the formulas with high BIN-values are learned better. Consider for instance the left panel for Lara, namely when AVELEN is between 2.75 and 3: in those cases, when BIN is low, the predicted values are represented with values ranging from 6 to 4, but when BIN is high, the predicted values range from 9 to 6 respectively. A similar case can be made for Thomas, if one considers AVELEN-values between 2.5 and 3.1: for every predicted value when BIN is low, the corresponding values

for when BIN is high are (sometimes considerably) higher – 6 to 1 compared to 9 to 2. In other words, and as anticipated, when BIN (i.e. MERGE values) are higher, the children learn the formulaic sequences better.

4.3 Discussion

To summarize, these children are averse to learning long sequences, regardless of the association strengths. Given that they are in the age range of 2–3, this is unsurprising, since longer multiword utterances are rare in the speech of children of this age. However, association strength indeed has an effect for all but the longest average lengths: as expected, in the case of both children, higher-strength sequences (as registered by their BIN rank) are learned at a higher rate than lower-strength sequences. In the future, it would be desirable to use longitudinal corpora from slightly older children who produce longer, more complex utterances to determine whether the same effect for short *n*-grams observed here may be likewise observed for longer *n*-grams.

More generally, we have shown that the MERGE algorithm can indeed be methodologically deployed in a theoretical application that studies formulaic language. Whether, in this particular application, automatically extracted formulaic sequences would perform better than manually annotated ones is an open question (and not one that we set out to address in this case study). However, we wish to point out that it is not clear in the first place that the formulaic sequences that children detect in caregiver input and in turn use to bootstrap their own language production are necessarily the same ones that adult annotators would identify as true formulaic sequences. Rather, it may be that the bottom-up approach of automatic extraction in general and lexical association in particular, while obviously imperfect, exhibit closer parallels to the frequency-based acquisition mechanisms employed by children than do whatever crystallized lexical knowledge that adult annotators use.

5. Conclusion

Formulaic language has become a major focus of research in linguistics, as scholars have realized how fundamental and omnipresent it is in discourse. Accordingly, techniques for its efficient identification in textual data are much in demand. While manual annotation is still considered the technique that offers the highest precision, the degree of recall it can offer is more limited given its high costs and time requirements (esp. once interrater reliability is also considered), which has led to great interest in the development of effective computational extraction algorithms. Many of the existing algorithms exhibit shortcomings, though, including the use of

statistical measures for scoring candidate sequences that are either (1) limited to bigrams, or (2) insensitive to high-frequency, low-contingency sequences. Moreover, basic approaches tend to extract many partially redundant, overlapping, and fragmentary sequences.

In this paper, we have presented and tested an algorithm that addresses various issues. Entitled MERGE (for Multi-word Expressions from the Recursive Grouping of Elements), the algorithm employs a *recursive bigram approach*, whereby it is able to grow formulaic sequences of any length in a bottom-up fashion, all while never having to calculate statistical associations for anything other than simple 2-way co-occurrences. As we have shown, MERGE stands up well against another extraction algorithm from the literature, the Adjusted Frequency List, when compared to manually annotated formulaic sequences from the British National Corpus (BNC). What is more, we have shown that MERGE can be successfully used to help predict word sequences that young children learn based on their caregiver input, lending support to the idea that automatic extraction algorithms are viable methodological tools for application in formulaic language research. But despite these successes, it is clear from case study 1 that MERGE still neglects to identify many formulaic sequences identified by the BNC annotators. Thus, further refinement of automatic techniques such as MERGE is still needed.

Along these lines of further refinement, MERGE allows for the identification of formulaic sequences that may contain one or more gaps of various sizes. However, in the present case studies, this ability was not exploited/tested. In the future, it would be desirable to investigate what benefits, if any, this built-in capacity yields. Does it improve the performance of the identification of continuous sequences by offering more paths to a particular formulaic sequence (*in spite + of* versus *in + spite of + in _ of + spite*)? Does it indeed result in the identification of true discontinuous formulaic sequences or does it not result in performance gains?

Note that paradigmatic slots within formulaic sequences (and at their edges for that matter) may be filled with constituents of different lengths in words (e.g., *as small as* versus *as vanishingly small as*). However, as it is currently implemented, MERGE would not treat, say, *as _ as* and *as _ _ as* as the same type, even though they clearly are. Again, further development of the algorithm is needed, given that formulaic sequences comprise not only frozen lexical items but they also allow for different kinds of – and varying degrees of – schematicity (see Langacker, 1987; Goldberg, 1995; 2006; or Bybee, 2010 for discussion of the many different levels of schematicity/generality of the mental lexicon/constructicon), which in turn suggests that, down the road, using an association measure, or a combination of measures that also incorporate type frequencies or type entropies, might be useful. Currently, however, we submit that MERGE offers a state-of-the-art approach to the automatic identification of formulaic sequences.

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 101–102). Oxford: Oxford University Press.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 6, 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Science* 106(41), 17284–17289. https://doi.org/10.1073/pnas.0905638106

Biber, D., Conrad, S., & Cortes, V., (2004). *If you look at …*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, 33(5), 752–793. https://doi.org/10.1111/j.1551-6709.2009.01031.x

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum* 1, 1–14.

Bybee, J. (2010). *Language, usage, and cognition*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Daudaravičius, V., & Marcinkevičienė, R. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2), 321–348. https://doi.org/10.1075/ijcl.9.2.08dau

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Evert, S. (2004). *The statistics of word co-occurrences: Word pairs and collocations*. (PhD Thesis, Universität Stuttgart).

Evert, S. (2009). Corpora and collocations. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics: an international handbook*, Vol. 2 (pp. 1212–1248). Berlin/New York: Mouton de Gruyter.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work*. Oxford: Oxford University Press.

Granger, S., & Meunier, F. (Eds). (2008). *Phraseology. An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. Th., & Mukherjee, J. (2010). Lexical gravity across varieties of English: An ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics*, 15(4), 520–548. https://doi.org/10.1075/ijcl.15.4.04gri

Gries, S. Th. (2015). Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics*, 16(1), 93–117. https://doi.org/10.1177/1606822X1455660

Langacker, R. W. (1987). *Foundations of cognitive grammar: Vol. 1: Theoretical prerequisites*. Stanford: Stanford University Press.

Lieven, E., Salomo D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3), 481–507. https://doi.org/10.1515/COGL.2009.022

MacWhinney, B. (2000). *The CHILDES project. Tools for analyzing talk*. Third edition. Mahwah, NJ: Lawrence Erlbaum Associates.

McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present*. Abington: Routledge.

O'Donnell, M. B. (2011). The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal*, 35, 135–169.

Pecina, P. (2009). *Lexical association measures: Collocation extraction*. Prague: Charles University.

Rowland, C. F., & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *The Journal of Child Language* 33(4), 859–877. https://doi.org/10.1017/S0305000906007537

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86–132. https://doi.org/10.1016/j.cogpsych.2004.06.001

Tomasello, M. (2005). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Wahl, A. (2015). Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics*, 13(1), 191–219. https://doi.org/10.1075/rcl.13.1.08wah

Wahl, A., & Gries, S. Th. (2018). Multi-word expressions: A novel computational approach to their bottom-up statistical extraction. In P. L. Cantos-Gómez and M. Almela-Sánchez (Eds.), *Lexical collocation analysis: advances and applications* (pp. 85–109). Berlin/New York: Springer

Wible, D., Kuo, C., Chen, M., Tsao, N., & Hung, T. (2006). A computational approach to the discovery and representation of lexical chunks. *TALN* 2006. Leuven, Belgium.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Appendix. Summary statistics for the linear model on the acquisition data

Predictor	***b***	***se***	***t***	$p_{\text{two-tailed}}$
Intercept	0.87926	0.18777	4.683	<0.001
AVELEN	−0.15195	0.06040	−2.516	0.012445
BIN	5.29464	1.34825	3.927	<0.001
poly(BIN, 2)	0.26540	1.29413	0.205	0.837663
CHILD $_{\text{Lara → Thomas}}$	−0.08024	0.20476	−0.392	0.695463
AVELEN : BIN	−1.72888	0.47150	−3.667	<0.001
AVELEN : poly(BIN, 2)	0.14818	0.43396	0.341	0.733017
AVELEN: CHILD $_{\text{Lara →Thomas}}$	−0.01156	0.06589	−0.175	0.860909
BIN : CHILD $_{\text{Lara →Thomas}}$	−3.26269	1.54239	−2.115	0.035296
poly(BIN, 2) : CHILD $_{\text{Lara →Thomas}}$	3.03135	1.48600	2.040	0.042309
AVELEN : BIN : CHILD $_{\text{Lara →Thomas}}$	1.19420	0.53703	2.224	0.026976
AVELEN : poly(BIN, 2) : CHILD $_{\text{Lara →Thomas}}$	−1.00916	0.49722	−2.030	0.043358

Computational phraseology discovery in corpora with the MWETOOLKIT

Carlos Ramisch
Aix-Marseille Université

Computer tools can help discover new phraseological units in corpora, thanks to their ability to quickly draw statistics from large amounts of textual data. While the research community has focused on developing and evaluating original algorithms for the automatic discovery of phraseological units, little has been done to transform these sophisticated methods into usable software. In this chapter, we present a brief survey of the main approaches to computational phraseology available. Furthermore, we provide examples of how to apply these methods using the mwetoolkit, free software for the discovery and identification of multiword expressions. The usefulness of the automatically extracted units depends on various factors such as language, corpus size, target units, and available taggers and parsers. Nonetheless, the mwetoolkit allows fine-grained tuning so that this variability is taken into account, adapting the tool to the specificities of each lexicographic environment.

Keywords: phraseological units, automatic phraseology discovery, morphosyntactic patterns, association scores, mwetoolkit

1. Introduction

Phraseological units are pervasive in human languages. They range from compound nominals (*prime minister*) to complex formulaic templates (*looking forward to hearing from you*), including idiomatic expressions (*to make ends meet*) and collocations (*heavy rain*). While easily mastered by native speakers, they pose challenges for foreign language learners, as their use confers naturalness to discourse, even though they are often unpredictable. Particularly in specialised domains, phraseological units are numerous and employing them appropriately is crucial in technical and scientific communication. For all these reasons, compiling phraseology dictionaries is an absolute need, to account for the pervasiveness of multiword phenomena in languages.

https://doi.org/10.1075/ivitra.24.06ram

Phraseology is crucial not only for lexicography, but also for computational linguistics. Indeed, the development of natural language processing (NLP) applications often relies on appropriately representing and processing phraseological units in machine-readable lexicons and grammars (Sag et al., 2002). In the NLP community, there has been much enthusiasm about so-called *multiword expressions*. Since the beginning of the 2000s, numerous algorithms and experiments have been published on the automatic processing of multiword units, covering their automatic discovery, in-context identification, syntactic and semantic analysis, as well as translation (Markantonatou et al., 2017). Synergies between NLP and lexicography are a natural consequence of this mutual interest in phraseology.

However, given the different backgrounds, goals and traditions of research communities, cooperation presents some challenges, for example due to the lack of a homogenised terminology. One question regards the overlaps among phraseological units, collocations and multiword expressions. While these terms are undoubtedly related, it is not straightforward to clearly delineate the subtle differences in the phenomena they cover.

While we will not address the issue of competing definitions, we will instead consider that phraseological units are groups of lexemes that present some idiosyncrasy with respect to ordinary word combinations (Baldwin and Kim, 2010), so that their particularities must be recorded in a lexicon (Evert, 2004). Phraseology lexicons, in turn, are useful both for humans and computers, for robust and fluent analysis and generation of language.

The manual construction of lexical resources that include phraseological units is often onerous and time-consuming. It requires not only lexicographic expertise but also corpus-based work, because many units have non-standard properties that only emerge from the study of their use in context. Computers are often employed to enhance, speed up and generally assist in lexicographic tasks, given that they can quickly process large amounts of text (Dagan and Church, 1994; Heid, 2008). Thus, computational systems play a double role in the creation of phraseological resources. On the one hand, they *use* these resources in NLP tasks and applications such as parsing, machine translation and information extraction. On the other hand, computational systems also *support* the creation of lexical resources. This chapter focuses on the latter interaction, when computers are used to help build phraseological lexicons, which we will henceforth refer to as *phraseology discovery*.[1]

In computational linguistics, much has been said about the automatic discovery of phraseological units in corpora using computational tools (Evert and Krenn,

1. Many terms have been used to denote the task of finding new phraseological units in text, including phraseology discovery, identification, extraction, and acquisition. We consistently employ phraseology discovery even when references might use a different terminology.

2005; Seretan, 2011; Ramisch et al., 2013; Ramisch, 2015). In spite of a huge literature, newcomers to the field often feel frustrated about the use of computational tools for phraseology discovery. To date, there is no simple and direct answer to the question, "What are the best freely available computational tools to help build phraseological resources?"

While solutions do exist, tools for computer-assisted phraseology are often not freely available. When they are, many are hard to use, limited to a few languages and processing systems, and do not always implement more sophisticated techniques reported in research papers. To make things even more complicated, when available tools exist, they often depend on a given syntactic formalism, data format or language-specific configurations, and are not adaptable or portable to different scenarios (language, domain, type of target phraseological unit). Users often have to choose between powerful but extremely complex systems that require computational expertise, and systems that are easy to use but do not allow fine-grained customisation and do not always implement the latest research advances.

This chapter provides a brief survey of current techniques for the automatic discovery of phraseology in monolingual corpora from a computational perspective. We will pay special attention to the availability of these techniques and their implementation in free software. Each subsection includes examples of candidate phraseological units obtained automatically using the mwetoolkit.[2]

After this introductory section, we provide a brief overview of related work in computational phraseology, focusing on tools (Section 2). Then, the main contribution of this work is to discuss a specific tool for multiword phraseology discovery in corpora: the mwetoolkit (Section 3). We present examples of successful use of the mwetoolkit for phraseology discovery, including candidate search patterns, association scores and other types of scores (Section 4). We close this chapter with a discussion of what are, in our opinion, the main bottlenecks that prevent the techniques described here from being employed in the large-scale production of lexical and phraseological language resources (Section 5).

2. Computational phraseology discovery

Although much has been published about the discovery of multiword units in corpora, not all methods and algorithms yield the publication of corresponding software. Therefore, rather than providing a comprehensive overview of phraseology discovery, this section provides an overview of the general architecture, followed by a summary of methods that have been implemented and released, and are thus

2. http://mwetoolkit.sf.net

directly applicable. Tools such as AMALGrAM (Schneider et al., 2014), LGTagger (Constant and Tellier, 2012), and jMWE (Finlayson and Kulkarni, 2011), for the identification of phraseological units in running text, to which a pre-compiled lexicon is usually provided, are not in the scope of this chapter.

For a more complete survey on phraseology discovery, the different proposed methods and their performances, we refer the reader to Evert (2004); Pecina (2008); Manning and Schütze (1999); McKeown and Radev (1999); Baldwin and Kim (2010); Seretan (2011); and Ramisch (2015). In addition to monolingual discovery, other tasks have also been investigated in computational linguistics, such as bilingual phraseology discovery (Ha et al., 2008; Morin and Daille, 2010; Weller and Heid, 2012; Rivera et al., 2013), automatic interpretation and disambiguation of multiword expressions (Fazly et al., 2009) and their integration into applications such as parsing (Constant et al., 2013) and machine translation (Carpuat and Diab, 2010). For further reading, we recommend the proceedings of the annual workshop on multiword expressions (Markantonatou et al., 2017), as well as journal special issues on the topic (Villavicencio et al., 2005; Rayson et al., 2010; Bond et al., 2013; Ramisch et al., 2013).

2.1 General architecture

Tools for corpus-based phraseology discovery use various strategies and have heterogeneous architectures. Often some *preprocessing* is applied to raw corpora before discovery, minimally by performing spurious content cleaning, sentence splitting, tokenisation, and case homogenisation. Optionally, some tools also employ automatic analysers to enrich the text with part-of-speech tags and, sometimes, automatically generated syntactic trees. The availability of taggers and parsers depends on the target language, so this is not always possible.

After preprocessing, tools extract *candidate* phraseological units from text based on recurring patterns. These patterns may be as simple as n-grams, that is, sequences of n contiguous tokens in a sentence (Pedersen et al., 2011; Silva and Lopes, 1999). Tools can also employ more sophisticated morphosyntactic patterns, such as sequences formed by a noun followed by a preposition and another noun (Kilgarriff et al., 2014). When available, syntactic information can also be used to extract candidates, for instance, focusing on verb-object pairs (Martens and Vandeghinste, 2010; Sangati et al., 2010). The choice may depend on the nature of the target phraseological units, and a mixture of these strategies can be preferable (Ramisch, 2015).

After candidate extraction, most available tools offer the possibility to *filter* the lists of candidates by using numerical scores. The idea of filtering is that true phraseological units can be distinguished from false positives by their statistical patterns. The most common filtering strategy is the use of association scores such as the candidate frequency, point-wise mutual information (Church and Hanks,

1990), Student's *t* score or log-likelihood ratio (Dunning, 1993). Scores are used to rank the candidates, assuming that those with higher scores are more likely to be kept for inclusion in a phraseological lexicon. As we will exemplify later (Section 4.2), other scores, based on contextual and contrastive information, can also help retrieve interesting units.

2.2 Freely available tools

Tools for phraseology discovery generally combine linguistic analysis and statistical information as clues for finding new units in texts. Here, we present a list of freely available tools that can be used mostly for monolingual phraseology discovery. While most of them require some familiarity with the textual command line interfaces, some also provide a graphical user interface or a web application.

The *N*-gram Statistics Package (NSP)[3] is a command-line tool for the statistical analysis of *n*-grams in text files (Pedersen et al., 2011; Banerjee and Pedersen, 2003). It provides scripts for counting *n*-grams and calculating association scores, where an *n*-gram is either a sequence of *n* contiguous words or *n* words occurring in a window of $w \geq n$ words in a sentence. While most of the measures are only applicable to 2-grams, some of them are also extended to 3-grams and 4-grams, notably the log-likelihood measure. The tool takes as input a raw text corpus and a parameter value fixing the size of the target *n*-grams, and provides as output a list of candidate units extracted from the corpus along with their counts, which can further be used to calculate association scores.

Analogously, LocalMaxs[4] is a script that extracts candidate units by generating all possible *n*-grams from a sentence. It further filters them based on the local maxima of a customisable association score distribution (Silva and Lopes, 1999), thus taking into account larger units that contain nested smaller ones. The tool includes a strict version, which prioritises high precision, and a relaxed version, which focuses on high recall. Both NSP and LocalMaxs are based purely on token counts and are completely language independent. On the other hand, there is no direct support to linguistic information such as keeping only *n*-grams that involve nouns.

Focusing on the retrieval of discontiguous units, Xtract is an algorithm that uses a sliding window of length *w* to scan the text (Smadja, 1993). It requires the input text to be tagged with parts of speech (POS), so that filters can be applied on the types of extracted candidates. Xtract first generates bigrams by calculating the average distance between words in the sliding window, as well as their standard deviation. Words that tend to occur in the same position with respect to each other

3. http://search.cpan.org/dist/Text-NSP

4. http://hlt.di.fct.unl.pt/luis/multiwords/

(i.e. have a small standard deviation) are considered as candidates, which are then expanded to larger *n*-grams. The Dragon toolkit[5] is a Java library that implements the Xtract algorithm and can be included in computational tools (Zhou et al., 2007).

The UCS toolkit[6] is a command-line package for calculating association scores (Evert, 2004). Additionally, it provides powerful mathematical tools like dispersion tests, frequency distribution models and evaluation metrics. UCS focuses on high-accuracy calculations for 2-grams, but, unlike the other approaches, it does not extract candidate units from corpora. Instead, it receives a list of candidates and their respective counts as input, relying on external tools for corpus preprocessing and candidate extraction. Then, it calculates the measures and ranks the candidates. Another tool that works in a similar way is Druid.[7] It is based on distributional similarity models which estimate to what extent a given candidate could be replaced by a single word, assuming that phraseological units convey more atomic, non-decomposable meanings than regular combinations (Riedl and Biemann, 2015).

In contrast with the above-mentioned tools, there are also some tools that are not based on word sequences, but rather work with syntactic trees. Thus, they require syntactically analysed corpora as input, generally preprocessed by an automatic parser. Such tools are especially well suited to the discovery of flexible units such as verbal idioms, formulaic phrases, and collocations. Examples of such tools include Varro (Martens and Vandeghinste, 2010),[8] DiscoDOP (Sangati et al., 2010)[9] and FipsCo (Seretan and Wehrli, 2009).[10]

The tools surveyed here are mostly developed by researchers in computational linguistics for a project or thesis. Hence, their goal is not to optimise ease of use for users who are not necessarily familiar with command-line computational tools. The idea of the Sketch Engine is to make such tools accessible and friendly by providing an intuitive web interface for corpus-based phraseology discovery.[11] Similarly to other tools, it allows the user to load corpora, preprocess them with POS tags and lemmas, and then extract co-occurrence patterns (the "sketches") based on morphosyntactic patterns and association scores (Kilgarriff et al., 2014). On the other hand, since it is not free, but commercialised by a company, special

5. http://dragon.ischool.drexel.edu/

6. http://www.collocations.de/software.html

7. http://ltmaggie.informatik.uni-hamburg.de/jobimtext/components/druid/

8. http://sourceforge.net/projects/varro/

9. https://github.com/andreasvc/disco-dop

10. http://129.194.38.128:81/FipsCoView

11. https://www.sketchengine.co.uk/

attention is given to the presentation of results, user support and the provision of useful tools for corpora work (e.g. a tool for crawling web corpora for a given language and domain).

There are also numerous freely available web services and downloadable tools for automatic term extraction, not necessarily focusing on phraseology (Drouin, 2004; Heid et al., 2010). These tools are generally language-dependent, with versions for major European languages like English, Spanish, French and Italian. Examples of such tools are TermoStat,[12] AntConc[13] and TerMine.[14] Most of them provide user-friendly graphical interfaces or direct web access. On the other hand, they do not always allow for fine-tuning of discovery parameters.

3. The mwetoolkit

In spite of the existence of a certain number of available tools for phraseology discovery, they usually only deal with part of the discovery process. For example, while UCS provides several association scores for candidate ranking, the extraction of candidates from the corpus needs to be performed externally. NSP provides support for larger *n*-grams, but it is impossible to describe more linguistically-motivated extraction patterns based on parts of speech, lemmas or syntactic relations (Ramisch et al., 2012).

In a context where existing methods only implemented part of what we needed, we wanted to conceive a *generic methodology* that would cover the whole discovery process. The mwetoolkit is a tool designed to perform automatic discovery of multiword units in both specialised and general-purpose corpora (Ramisch et al., 2010b, a; Ramisch, 2015). It implements hybrid, knowledge-poor techniques that can be applied to virtually any corpus, as long as the corpus text can be segmented into tokens.[15] The goal of the mwetoolkit is to aid lexicographers and terminographers in the challenging task of creating language resources that include multiword entries.[16]

Given the plethora of computational discovery methods available and the lack of consensus about their performances, the methodology implemented by the

12. http://olst.ling.umontreal.ca/~drouinp/termostat_web/

13. http://www.antlab.sci.waseda.ac.jp/software.html

14. http://www.nactem.ac.uk/software/termine/

15. Depending on the language, however, the corpus will require more or less sophisticated preprocessing. For instance, while text tokenization is straightforward in English, it is much harder in Chinese. For use in the mwetoolkit, a corpus must be (automatically) segmented into tokens, considered as "words".

16. http://mwetoolkit.sf.net

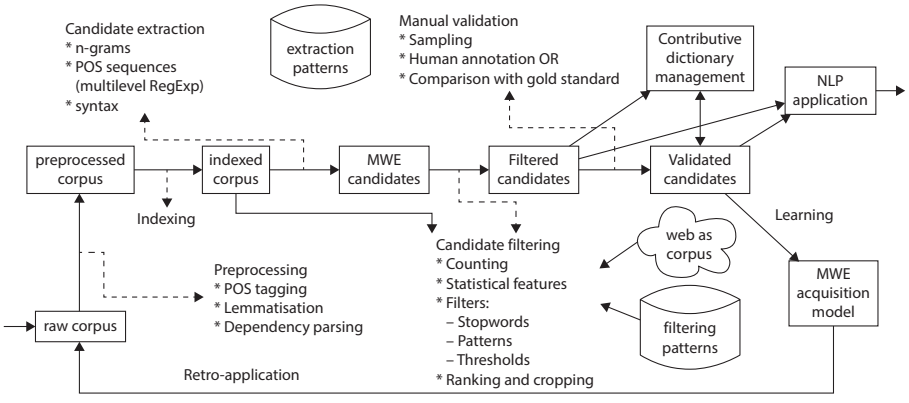


Figure 1. The mwetoolkit modules and their chaining in phraseology discovery

mwetoolkit should necessarily allow *multiple solutions* for a given sub-task. Thus, decisions such as the level of linguistic analysis, length *n* of the *n*-grams, filtering thresholds and evaluation measures should not be made by the method itself, but users should be able to choose and tune the parameters according to the needs. This implies that the tool does not provide a push-button method, but one that can be adapted and tuned to a large number of contexts, maximising its *portability*.

The mwetoolkit adopts the standard sub-task definition consisting of two phases: candidate extraction and candidate filtering. In the first phase, one acquires candidates based either on flat *n*-grams or specific morphosyntactic patterns (of surface forms, lemmas, POS tags and dependency relations). Once the candidate lists are extracted, it is possible to filter them by defining criteria that range from simple count-based thresholds, to more complex features such as association and semantic compositionality scores. Since some scores are based on corpus word and *n*-gram counts, the toolkit provides both a corpus indexing facility and integration with web search engines (for using the web as a corpus). Additionally, for the evaluation phase, it provides validation and annotation facilities.

The mwetoolkit methodology was implemented as a set of independent modules, each taking the form of a separate Python script, that handle an intermediary representation of the *corpus*, the list of *patterns*, and the list of *candidates*. Each module performs a specific task in the discovery process, from the raw corpus to the filtered list of candidates and their associated scores. Figure 1 summarises the architecture of the mwetoolkit, which is exemplified in Section 4. Examples of applications of the mwetoolkit include the discovery of nominal expressions in Greek (Linardaki et al., 2010), of complex predicates in Portuguese (Duran et al., 2011), and of idiomatic nominal compounds in French and English (Cordeiro et al., 2016a).

4. Phraseology discovery with the mwetoolkit

In the following subsections, we detail existing methods commonly employed in phraseology discovery. Each method is exemplified using the mwetoolkit. Our goal is to illustrate how each discovery method works rather than presenting original results in phraseology discovery *per se*. Therefore, our experimental set-up is simplistic: we work on a fragment of the English UKWaC that is made freely available including POS tags, lemmas and syntactic dependencies.[17] This fragment (henceforth UKWaC-frg) consists of the first 100,000 sentences of the corpus, which represent around 2.6 million tokens of general-purpose texts crawled from the British world wide web and cleaned automatically (Baroni and Bernardini, 2006).

We underline that the artificial results reported in toy experiments in the remainder of this section were not tuned for optimal performance. Moreover, the mwetoolkit is language independent and can deal with segmented/tokenised[18] corpora in any language, optionally handling any POS tagset, lemmatiser and dependency syntax format that may be applied to the corpus.

4.1 Candidate search patterns

Candidate phraseological units can be extracted from automatically POS-tagged, lemmatised and/or parsed corpora using ad hoc scripts. However, more principled corpus queries are often necessary to obtain lists of candidates to include in phraseological dictionaries. These corpus queries need to deal with multi-level annotations in the corpus, which can be very helpful in phraseology discovery. For instance, Seretan (2011) has shown that syntactic relations can be used in collocation discovery patterns to overcome the limitations of shallow POS sequence patterns.

As a consequence, phraseology discovery requires a powerful corpus query language that correctly matches user-defined search patterns with the information available in the corpus (Kilgarriff et al., 2014). The expressive power and generality of the query language is determinant for successfully adapting a candidate extraction method that works well in a given configuration to another language, domain and corpus.

---

17. The UK web as corpus: http://wacky.sslmit.unibo.it/doku.php?id=corpora

18. Prior to phraseology discovery, any corpus must be preprocessed and decomposed into sentences and tokens. Tokens are usually words, but can also be morphemes or even characters. In this article, by "words" we mean any type of tokenised unit that makes sense for a given language.

The mwetoolkit uses a multi-level regular expression language to express corpus searches. This language can be used in two modules of the system. With the first module, called 'grep.py', it is possible to explore the corpus, finding sentences that match a given search pattern. This is similar to using a concordancer enriched with an expressive query language as shown in Figure 2. For example, assuming that the corpus is contained in a file named ukwac-frg.moses, the command below allows finding and showing all sentences in the corpus that contain sequences of two contiguous common nouns (NN NN).

```
mwetoolkit/bin/grep.py -e "NN NN" --to XML ukwac-frg.moses | mwetoolkit/bin/
view.py --to PlainCorpus
```

While finding sequences of contiguous POS tags is useful, it is not always sufficient to model the target expressions. The query language of the mwetoolkit allows the use of complex operators adapted from regular expressions.[19] For instance, suppose we are interested in two-word complex nominals in English, where the modifier that precedes the noun can be either an adjective or another noun. This can be done by defining a pattern whose first member is an alternative, denoted with a pipe symbol (|) between the POS for noun and the one for adjective. We exemplify below the use of this pattern as the input for another module, 'candidates.py', which carries out the extraction of candidate units from corpora, yielding a list of candidates like the one shown in the first column of Table 1:

```
mwetoolkit/bin/candidates.py -e "(NN|JJ) NN" ukwac-frg.moses
```

Other operators are available, for instance, to indicate the repetition of elements. For example, the second column of Table 1 shows the discovered candidates corresponding to a pattern similar to the first one, but in which the first element can be repeated an arbitrary number of times, retrieving units such as *large scale show*.

Sometimes, we are interested in the collocation patterns of a specific lexical unit. For instance, we would like to know which nouns and adjectives can modify the word *show*. The second column of Table 1 exemplifies this, also showing how it is possible to limit the POS of the word *show* to nominal occurrences only, excluding adjectives and nouns which precede the verb *to show*.

In the third column, we show the results of a pattern in which the head noun was replaced by a regular expression: we are interested here in the co-occurrence pattern of words ending with the suffix *-ion*. Finally, the fourth and last column

19. Operators includes standard regular expression operators over tokens such as repetition (* and +), optional elements (?), and alternative (|), as well as specialised operators such as negation and ignored matches.

```
g_/packet ) and the recently launched ‘ Stackem ‘s’ (2.3 g per pack) .
Each used a combination of marketing_techniques specifically aimed at children and
busy parents.
These included  ; web-based promotions , such as design your own dairy-lea_movie or
an interactive web-enabled competition ; text-messages ; competitions, such as win
a year ‘s pocket_money high profile_endorsements , such as Gary Linekar cartoon_en
dorsements such as the Simpsons , in-pack_promotions , including games and colourin
g in health_claims such as high in Ca , equivalent to one glass of milk convenient
packaging , with’ ideal for lunches’ or combination_lunch packs TV_advertisements
 specifically aimed at children , vouchers for schools , discounts such buy 2 for t
he price of three multi-buy packs.
The liking for salty foods is a learned taste_preference set in childhood and so en
couraging children to eat high levels of salt sets the seeds for vascular disease ,
increasing the risk of developing stroke and heart_disease later in life.
High salt_intakes have also linked to osteoporosis , stomach_cancer , asthma and ki
dney_disease.
“_The systematic targeting of children by the food_industry who wish to habituate c
```

Figure 2. Query NN NN, matching sequences of two common nouns in UKWaC-frg, visualised in a concordancer-like command line interface

shows a syntactic pattern, in which we look for all nouns that can occur as subject of the verb *to say*.

The syntax of the query language is explained in the documentation of the mwetoolkit, but for the sake of completeness we provide below the exact queries that were used to generate the candidates presented in Table 1:

1. (JJ|NN) NN: a sequence of exactly two tokens, where the last one corresponds to a common noun (POS starts with *NN*) and the first one is a disjunction between a common noun (*NN*) or an adjective (*JJ*).[20] This captures two-word noun phrases consisting of a head noun and a modifier noun or adjective.
2. (JJ|NN){repeat=+} [pos~/NN.*/i lemma="show"]: a sequence of at least two tokens, where the last one corresponds to the common noun *show*[21] and the first one is a disjunction between a common noun (*NN*) or an adjective (*JJ*), which can be repeated more than once. This captures noun phrases where the head noun is *show*, preceded by a sequence of modifier nouns or adjectives.
3. (JJ|NN){repeat=+} [lemma~/.*ion/]: a sequence of at least two tokens, where the last one is any word ending with the suffix *-ion* and the first one is a sequence

20. This pattern uses a short-cut notation for ([pos~/JJ.*/i]|[pos~/NN.*/i]) [pos~/NN.*/i], that is, all-capital strings are internally interpreted as case-insensitive POS prefixes.

21. Here, we exemplify the explicit notation by which one can constrain different token information levels (POS, lemma, etc.) using equality (=) or regular expression approximate matching (~). The trailing i stands for case-insensitive.

Table 1. Examples of candidates extracted from UKWaC-frg using different types of patterns, shown as headers. Only a sample of the candidates that occur twice or more is are shown

Pattern 1	Pattern 2	Pattern 3	Pattern 4
more information	TV show	more information	official say
more detail	bike show	further information	spokesman say
last year	large scale show	high education	teacher say
further information	scale show	food production	spokeswoman say
first time	sci-fi show	legal information	official say
wide range	next show	above application	report say
same time	research show	additional information	government say
web site	radio show	late version	man say
local authority	successful show	special collection	mp3gadgetblog say
last week	series show	personal information	cannot say
car park	film show	voluntary organisation	minister say
high quality	few show	southern section	newspaper say
many year	trade show	trade union	signature say
next year	reality show	new version	police say
many people	radio show	other organisation	third say
long time		next generation	manager say
few year		public consultation	text say
view guestbook		good condition	
other hand		industrial action	
hard drive		data protection	

of common nouns or adjectives which may repeat once or more. This captures mainly noun phrases where the head noun is a nominalisation, preceded by a sequence of modifier nouns or adjectives.

4. [pos~/NN.*/ syndep=SBJ:v] []{*repeat*=* *ignore*=1} [pos~/VV.*/ lemma=say] {*id*=v}: a sequence of two tokens, of which the first is a common noun and the second is the main verb *to say*. The pattern is discontinuous, and allows any number of intervening words, that will not be considered as part of the matched expression (*ignore*=1). Additionally, there is a syntactic constraint saying that the first token must be the subject (syndep=SBJ:v) of the last word, identified by a unique name id=v. This captures possible nouns that can occur as the subject of the verb *to say*.

4.2 Association scores

Collocation is a linguistic phenomenon characterised in statistical terms by *outstanding co-occurrence*. That is, words in collocations have a tendency to co-occur more often than it would be expected by pure chance in a corpus, as if they attracted each other. Collocation is an important property that distinguishes phraseological units from regular word combinations.

Dozens of association scores have been proposed to model outstanding co-occurrence in texts (Pecina, 2011). Association scores estimate the strength of association between the words contained in a candidate phraseological unit. They are based on the co-occurrence count and on the individual word counts of the candidate in a large corpus.

Suppose there is a candidate phraseological unit formed by n words, w_1, $w_2 \ldots w_n$. Most association scores employed nowadays take into account the observed co-occurrence count $c(w_1 \ldots w_n)$ of the whole candidate. In Table 2, we show the top-n most frequent candidates extracted by the pattern (JJ|NN) NN in

Table 2. Examples of top-ranked candidates sorted by frequency, and by association scores, extracted from UKWaC-frg with pattern (JJ|NN) NN

frequency	PMI	t-score	Log-likelihood
young people	cerebral palsy	young people	young people
more information	metabolic acidosis	more information	wide range
more detail	john baker	more detail	last year
last year	amino acid	last year	more detail
further information	scheduled uptime	further information	further information
first time	systemic aciclovir	first time	more information
wide range	fortune teller	wide range	local authority
same time	mifid override	same time	web site
web site	injectable medicine	web site	car park
local authority	holy hierarch	local authority	view guestbook
last week	dew pond	last week	first time
car park	asylum seeker	car park	same time
high quality	rackmount cabinet	high quality	last week
next year	cupc4kes reply	next year	hard drive
many year	stainless steel	many year	high quality
many people	holdem poker	view guestbook	long term
view guestbook	carbon dioxide	many people	email address
long time	scheduled break	long time	distributive justice
few year	non-executive director	few year	mental health
other hand	wrongful pregnancy	other hand	health insurance

UKWaC-frg.[22] All words are lemmatised to neutralise variants, but this may lead to ungrammatical entries such as *many year*, which actually corresponds to occurrences of *many years*, in plural.

The problem of pure co-occurrence is that frequent word combinations can be a result of pure chance because the words involved are very frequent per se, like *and we*, *it is* and *of the* (Manning and Schütze, 1999). These are usually function words that are not necessarily interesting for phraseology discovery. Even when we restrict the set of acceptable POS tags, as we did in Table 2 to include only adjectives and nouns, very frequent words still tend to diminish the usefulness of the extracted list. For example, modifiers such as *many*, *last* and *more* are combined with frequent words like *information* and *year*, but these are regular combinations with limited phraseological interest.

The problem with co-occurrence frequency is that association scores should also consider the expected count of a word combination $E(w_1 \dots w_n)$, comparing it with the simple co-occurrence count $c(w_1 \dots w_n)$. If the appearances of words w_i in a corpus are modelled as independent events, we expect that the co-occurrence count of a group of words equals the product of their individual probabilities $\frac{c(w_i)}{N}$ scaled by the total corpus size N. Therefore, the expected count E is estimated considering the number of occurrences of individual words in the candidate $c(w_1)$ through $c(w_n)$:

$$E(w_1 \dots w_n) = \frac{c(w_1) \times \dots \times c(w_n)}{N^{n-1}}$$

Pointwise mutual information (PMI) is one of the most popular association scores using this principle. It is not only used for phraseology discovery, but also for many other tasks in computational linguistics. PMI was first proposed in multiword terminology discovery by Church and Hanks (1990). It is the log-ratio between observed and expected counts:[23]

$$PMI(w_1 \dots w_n) = \log \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)}$$

PMI indicates how well the presence of an individual word predicts the presence of the whole phrase, or, in other words, it quantifies the dependence between words. Values close to zero indicate independence and the candidate words are discarded,

22. This list is similar to the examples shown in Table 1, but sorted by descending co-occurrence frequency.

23. Notice that, since this is the logarithm of a quotient, it is equivalent to $\log c(w_1 \dots w_n) - \log E(w_1 \dots w_n)$. In other words, observed and expected counts are compared through direct subtraction, but in logarithm domain.

whereas large values indicate outstanding co-occurrence. The second column of Table 2 shows the top-ranked candidates extracted from UKWaC-frg and sorted in descending order of PMI.

If things were simple, then PMI would solve all our problems. However, corpora are often not large enough to sufficiently cover the phenomenon under study and many interesting combinations are infrequent. Data sparseness is a problem for PMI, since this score tends to over-estimate the importance of rare word combinations formed by rare words. In Table 2, we show only candidates that co-occur more than 10 times in the toy corpus, to avoid retrieving spurious rare combinations such as foreign words and typos. Even with this restriction, PMI still retrieves quite rare candidates, some of which are actually quite interesting, such as *cerebral palsy*, *amino acid*, *stainless steel* and *carbon dioxide*. Some entries, however, are simply rare words that always appear combined with the same other words, such as *cup4kes reply*, which is probably text appearing on an online forum. Variants of PMI were proposed in the literature, trying to increase the importance of the observed count, in order to avoid this kind of problem (Daille, 1995; Riedl and Biemann, 2013).

Some association scores are based on hypothesis testing. Again, assuming independence between words, we can hypothesise that in regular (non-phraseological) word combinations, the observed and expected counts should be identical, that is H_0: $c(w_1 \ldots w_n) = E(w_1 \ldots w_n)$. Using a test statistic like Student's t, large values are strong evidence to reject H_0. The third column of Table 2 shows how the t-score ranks the nominal compound candidates extracted from our toy corpus.

More sophisticated tests for two-word MWE candidates take into account their *contingency table*. Examples of such measures are χ^2 and the more robust log-likelihood ratio (Dunning, 1993). The latter is only applicable to two-word expressions, but usually provides high-quality candidates. An example of ranking by log-likelihood is shown in the fourth and last column of Table 2.

Many other association scores were proposed in the literature, and there is no "silver bullet". While most of them are useful, deciding on which score to use for a given corpus is a matter of trial and error. Therefore, the mwetoolkit calculates several scores using the module 'feat_association.py'. Afterwards, users can try different sort orders and decide on one (or several) scores to use in order to filter the candidate entries before encoding them in lexicons. It also implements other scores that are briefly surveyed below.

4.3 Other scores

While association scores are the mainstream in phraseology discovery, they provide limited information about the behaviour of the target phraseological units, especially when they are not frequently occurring in the available corpora. Other types of scores have been proposed to capture other properties of phraseological units, such as their tendency to appear in specialised texts, their limited variability, their semantic idiomaticity and their limited word-for-word translatability. Here, we provide an overview of these scores, which are also implemented and available in the mwetoolkit. However, since they require complementary resources, we do not show examples but rather provide pointers to publications in which these scores are described and evaluated in more depth.

Contrastive scores are a useful source of complementary information for terminological or, more generally, domain-specific units. Several scores have been proposed to take into account the different frequency distributions of words across domain and general-purpose (contrastive) corpora (Drouin, 2004; Bonin et al., 2010). The intuition here is that units that appear frequently in a specialised corpus but are not common in general-purpose texts are likely to be specialised terms, and this is applicable not only to single words but also to phraseological units.

A simple contrastive score consists in the ratio between the frequency of the candidate in the domain corpus and its frequency in the contrastive corpus. The higher this ratio, the more specialised the candidate is. Since contrastive corpora should be large, they can be replaced by web hit counts, that is, the number of web pages containing a given candidate phrase (Ramisch et al., 2010c). In the mwetoolkit, these scores are implemented in a module named 'feat_contrast.py'.

Another class of useful methods are those which try to predict the variability scores of candidates. Indeed, one of the properties of phraseological units is that they do not allow full morphological, syntactic, and semantic flexibility as compared to similar free word combinations. *Variability scores* are a relatively under-exploited method based on automatic variant generation and subsequent (web) corpus searches. In other words, one first generates artificial variants for a given candidate, and then verifies whether these variants are attested in a large corpus or on the web. If variants appear frequently, then it is less likely that the candidate is a frozen phraseological unit. The skewness of the variant distribution can be measured by its entropy: the higher the entropy, the more uniform the distribution is, thus a low entropy score suggests more fixed phraseological units. In the mwetoolkit, these scores are implemented in a module named 'feat_entropy.py'.

The generation of variants is language-specific, and can be performed in several ways. Probably the simplest type of variability score is *permutation entropy*, in which

candidates are randomly reordered and then looked up on the web (Villavicencio et al., 2007). A slightly more sophisticated version of this score uses the syntactic behaviour of the expressions in order to create linguistically-informed permutations (Ramisch et al., 2008). Another possibility is to introduce explicit paraphrases, for example, by replacing sequences of the type *noun1 noun2* with a corresponding structure *noun2 PREP noun1*, trying different types of prepositions that correspond to the semantics of the compound (Nakov and Hearst, 2005). Finally, it is also possible to use general-purpose synonym dictionaries such as WordNet to try replacing words in the candidate with synonyms, since fixed phraseological units will generally not accept as much replacement as a regular combination would (Pearce, 2001; Duran and Ramisch, 2011; Duran et al., 2013).

With the growing importance of distributional semantics and word embeddings (Mikolov et al., 2013), new methods have been developed to identify non-compositional combinations in texts (Salehi et al., 2015; Cordeiro et al., 2016b). The basic idea of *compositionality scores* is to measure the similarity between the candidate phraseological unit and the words that compose it. Therefore, we use distributional models to build vectorial representations for each word and phrase from raw corpora. Afterwards, we use a vector operation (such as vector addition) to create a combined vector representing the sum of the meanings of the component words. Finally, we calculate its similarity (e.g. using the cosine of the angle between the vectors) to estimate how close the vector of the whole candidate unit is to the combined vector of the component words. Compositionality scores have been successfully employed to discover non-compositional phrases such as idiomatic noun compounds in English, French (Cordeiro et al., 2016a) and German (Roller et al., 2013), and verbal expressions in English (Cook and Stevenson, 2010) and in German (Köper et al., 2016). In the mwetoolkit, these scores are implemented in a module named 'feat_compositionality.py' (Cordeiro et al., 2016b).

Phraseological units often cannot be translated word by word. Therefore, parallel corpora can be a useful source of information for phraseological units. For example, one can use automatic word alignment tools to locate phrases in which the alignment is not perfect, indicating the presence of a phraseological unit (de Medeiros Caseli et al., 2009; Tsvetkov and Wintner, 2011; Salehi and Cook, 2013). Additionally, it is possible to use a method similar to the one for variant generation for translation generation: if a candidate cannot be translated word by word, then it is probably a phraseological unit (Morin et al., 2007; Vargas et al., 2017). Translation-based scores are not available in the mwetoolkit, but this is intended as future work.

5. Conclusions and open issues

In this chapter, we have presented an overview of research-oriented computational tools for the automatic discovery of phraseological units in corpora. We have used the mwetoolkit as an example of such a tool, describing its overall architecture and exemplifying the steps of the discovery process and their results on a toy corpus of English. Given the vast literature on discovering multiword units in corpora, we believe that it is important to focus on concrete tools and examples so that new users do not feel overwhelmed by the amount of scientific literature on the topic. We believe that this chapter makes a step in this direction. The mwetoolkit is far from being perfect, and can be improved in many ways. The first and obvious limitation is the use of terminal commands rather than a visual interface. Providing its functionalities as a web application would break the access barrier for users who are not familiar with typing commands in a Unix prompt. The development of a computational tool for phraseological discovery that is both free of charge (e.g. mwetoolkit) and easy to use (e.g. Sketch Engine) should be one of the priorities for computational linguistics.

Building better tools to support the construction of phraseological resources has the potential to benefit computational tools themselves. For example, a lexicon containing multiword units can be integrated into tools that perform automatic syntactic and/or semantic analysis of texts (Savary et al., 2017). Therefore, synergies between lexicography, phraseology and computational linguistics can help create a virtuous circle in which computer engineers build better tools for lexicographers who, in turn, build better machine-readable dictionaries, with better coverage of phraseological units. The latter can then be integrated into computational linguistic software to improve their linguistic precision.

References

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya and F. J. Damerau (Eds.), *Handbook of Natural Language Processing*. 2 edition (pp. 267–292). Boca Raton, FL: CRC Press, Taylor and Francis Group.

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics* (370–381). Mexico City, Mexico. https://doi.org/10.1007/3-540-36456-0_38

Baroni, M., & Bernardini, S. (Eds.) (2006). *Wacky! Working papers on the Web as Corpus*. Bologna, Italy: GEDIT.

Bond, F., Kim, S. N., Nakov, P., & Szpakowicz, S. (Eds). (2013). *Journal of Natural Language Engineering.Special Issue on computational approaches to the semantics of noun compounds*, 19(3). Cambridge, UK:Cambridge University Press.

Bonin, F., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2010). A contrastive approach to multiword extraction from domain-specific corpora. In *Proceeginds of the Seventh LREC (LREC 2010)*, Valetta, Malta: ELRA.

Carpuat, M., & Diab, M. (2010). Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proceedings of HLT: The 2010 Annual Conference of the NAACL (NAACL 2003)* (pp. 242–245). Los Angeles, CA: ACL.

Church, K., & Hanks, P. (1990). Word association norms mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Constant, M., Roux, J. L., & Sigogne, A. (2013). Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. *ACM Transactions on Speech and Language Processing. Special Issue on MWEs: from theory to practice and use, part 2 (TSLP)*, 10(3).

Constant, M., & Tellier, I. (2012). Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. In *Proceedings of the Eigth LREC (LREC 2012)*, Istanbul, Turkey: ELRA.

Cook, P., & Stevenson, S. (2010). Automatically identifying the source words of lexical blends in English. *Computational Linguistics*, 36(1), 129–149. https://doi.org/10.1162/coli.2010.36.1.36104

Cordeiro, S., Ramisch, C., Idiart, M., & Villavicencio, A. (2016a). Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1986–1997). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1187

Cordeiro, S., Ramisch, C., & Villavicencio, A. (2016b). Mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic mwe processing. In *LREC* 2016. Portoroz, Slovenia.

Dagan, I., & Church, K. (1994). Termight: Identifying and translating technical terminology. In *Proceedings of the 4th ANLP Conference (ANLP 1994)* (pp. 34–40). Stuttgart, Germany: ACL.

Daille, B. (1995). Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *Traitement Automatique des Langues*, 36(1–2), 101–118.

de Medeiros Caseli, H., Villavicencio, A., Machado, A., & Finatto, M. J. (2009). Statisticallydriven alignment-based multiword expression identification for technical domains. In D. Anastasiou, C. Hashimoto, P. Nakov, S. N. Kim (Eds.), *Proceedings of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)* (pp. 1–8). Suntec, Singapore: ACL.

Drouin, P. (2004). Detection of domain specific terminology using corpora comparison. In *Proceedings of the Fourth LREC (LREC 2004)*. Lisbon, Portugal: ELRA.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Duran, M. S., & Ramisch, C. (2011). How do you feel? Investigating lexical-syntactic patterns in sentiment expression. In *Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference*. Birmingham, UK.

Duran, M. S., Ramisch, C., Aluísio, S. M., & Villavicencio, A. (2011). Identifying and analyzing Brazilian Portuguese complex predicates. In V. Kordoni, C. Rasmich, & A. Villavicencio (Eds.), *Proceedings of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)* (pp. 74–82). Portland, OR: ACL.

Duran, M. S., Scarton, C. E., Aluísio, S. M., & Ramisch, C. (2013). Identifying pronominal verbs: Towards automatic disambiguation of the clitic 'se' in Portuguese. In V. Kordoni, C. Rasmich, & A. Villavicencio (Eds.), *Proceedings of the 9th Workshop on MWEs (MWE 2013)* (pp. 93–100). Atlanta, GA: ACL.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. (PhD Thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany).

Evert, S., & Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language. Special issue on Multiword Expression*, 19(4), 450–466. https://doi.org/10.1016/j.csl.2005.02.005

Fazly, A., Cook, P., & Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1), 61–103. https://doi.org/10.1162/coli.08-010-R1-07-048

Finlayson, M., & Kulkarni, N. (2011). Detecting multi-word expressions improves word sense disambiguation. In V. Kordoni, C. Rasmich, & A. Villavicencio (Eds.), *Proceedings of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)* (pp. 20–24). Portland, OR: ACL.

Ha, L. A., Fernandez, G., Mitkov, R., & Corpas Pastor, G. (2008). Mutual bilingual terminology extraction. In *Proceedings of the Sixth LREC (LREC 2008)*, Marrakech, Morocco: ELRA.

Heid, U. (2008). Computational phraseology: An overview. In S. Granger, & F. Meunier. (Eds.), *Phraseology. An interdisciplinary Perspective* (pp. 337-360). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139.28hei

Heid, U., Fritzinger, F., Hinrichs, E., Hinrichs, M., & Zastrow, T. (2010). Term and collocation extraction by means of complex linguistic web services. In *Proceedings of the Seventh LREC (LREC 2010)*, Valetta, Malta: ELRA.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Köper, M., Schulte im Walde, S., Kisselew, M., & Padó, S. (2016). Improving zero-shot-learning for German particle verbs by using training-space restrictions and local scaling. In *Proceedings of *SEM* 2016 (pp. 91–96). ACL.

Linardaki, E., Ramisch, C., Villavicencio, A., & Fotopoulou, A. (2010). Towards the construction of language resources for Greek multiword expressions: Extraction and evaluation. In S. Piperidis, M. Slavcheva, & C. Vertan (Eds.), *Proceedings of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages* (pp. 31–40). Valetta, Malta.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Markantonatou, S., Ramisch, C., Savary, A., & Vincze, V. (Eds.). (2017). *Proceedings of the 13th Workshop on MWEs (MWE 2017)*, Valencia, Spain: ACL.

Martens, S. & Vandeghinste, V. (2010). An efficient, generic approach to extracting multiword expressions from dependency trees. In É. Laporte., P. Nakov, C. Ramisch, & A. Villavicencio (Eds.), *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, (pp. 84–87). Beijing, China: ACL.

McKeown, K. R., & Radev, D. R. (1999). Collocations. In R. Dale, H. Moisl, & H. Somers (Eds.), *A Handbook of Natural Language Processing* (pp. 507–553). New York, NY: Marcel Dekker.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.*CoRR*, abs/1301.3781.

Morin, E., & Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation. Special Issue on Multiword expression: hard going or plain sailing*, 44(1–2), 79–95. https://doi.org/10.1007/s10579-009-9098-8

Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2007). Bilingual terminology mining– using brain, not brawn comparable corpora. In *Proceedings of the 45th ACL (ACL 2007)* (pp 664–671). Prague, Czech Republic: ACL.

Nakov, P., & Hearst, M. A. (2005). Search engine statistics beyond the *n*-gram: Application to noun compound bracketing. In I. Dagan, & D. Gildea (Eds.), *Proceeginds of the Ninth CoNLL (CoNLL-2005)* (pp. 17-24). University of Michigan, MI: ACL.

Pearce, D. (2001). Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)* (pp. 41–46).

Pecina, P. (2008). *Lexical Association Measures: Collocation Extraction*. (PhD Thesis, Faculty of Mathematics and Physics, Charles University).

Pecina, P. (Rev.). (2011). Syntax-based collocation extraction by Violeta seretan (University of Geneva). Berlin: Springer (Text, Speech and Language Technology Series, volume 44). *Computational Linguistics*, 37(3), 631–633. https://doi.org/10.1162/COLI_r_00068

Pedersen, T., Banerjee, S., McInnes, B., Kohli, S., Joshi, M., & Liu, Y. (2011). The *n*-gram statistics package (text: NSP) : A flexible tool for identifying *n*-grams, collocations, and word associations. In V. Kordoni, C. Rasmich, & A. Villavicencio (Eds.), *Proceedings of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)* (pp. 131–133). Portland, OR: ACL.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework, volume XIV of Theory and Applications of Natural Language Processing*. Springer.

Ramisch, C., Araujo, V. D., & Villavicencio, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proceedings of the ACL 2012 SRW* (pp. 1–6). Jeju, Republic of Korea: ACL.

Ramisch, C., Schreiner, P., Idiart, M., & Villavicencio, A. (2008). An evaluation of methods for the extraction of multiword expressions. In N. Grégoire, S. Evert, & B. Krenn (Eds.), *Proceedings of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)* (pp. 50–53). Marrakech, Morocco.

Ramisch, C., Villavicencio, A., & Boitet, C. (2010a). Multiword expressions in the wild? The mwetoolkit comes in handy. In Y. Liu, & T. Liu (Eds.), *Proceedings of the 23rd COLING (COLING 2010) – Demonstrations* (pp. 57–60). Beijing, China: The Coling 2010 Organizing Committee.

Ramisch, C., Villavicencio, A., & Boitet, C. (2010b). mwetoolkit: a Framework for Multiword Expression Identification. In *Proceeginds of the Seventh LREC (LREC 2010)* (pp. 662–669). Valetta, Malta: ELRA.

Ramisch, C., Villavicencio, A., & Boitet, C. (2010c). Web-based and combined language models: a case study on noun compound identification. In C.-R. Huang, & D. Jurafsky (Eds.), *Proceedings of the 23rd COLING (COLING 2010) – Posters* (pp. 1041–1049). Beijing, China: The Coling 2010 Organizing Committee.

Ramisch, C., Villavicencio, A., & Kordoni, V. (Eds.) (2013). *ACM Transactions on Speech and Language Processing. Special Issue on MWEs: from theory to practice and use, part 1 (TSLP)*, 10(2). New York, NY: ACM

Rayson, P., Piao, S., Sharoff, S., Evert, S., & Moirón, B. V. (Eds.) (2010). *Language Resources and Evaluation. Special Issue on Multiword expression: hard going or plain sailing*, 44(1–2). Springer.

Riedl, M., & Biemann, C. (2013). Scaling to large data: An efficient and effective method to compute distributional thesauri. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 884–890). Association for Computational Linguistics.

Riedl, M., & Biemann, C. (2015). A single word is not enough: Ranking multiword expressions using distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2430–2440). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1290

Rivera, O. M., Mitkov, R., & Corpas Pastor, G. (2013). A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (Eds.), *Proceedings of the MT Summit 2013 MUMTTT workshop (MUMTTT 2013)* (pp. 18–25). Nice, France.

Roller, S., im Walde, S. S., & Scheible, S. (2013). The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In V. Kordoni, C. Rasmich, & A. Villavicencio (Eds.), *Proceedings of the 9th Workshop on MWEs (MWE 2013)* (pp. 32–41). Atlanta, GA: ACL.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd CICLing (CICLing-2002), volume 2276/2010 of LNCS* (pp. 1–15). Mexico City, Mexico: Springer.

Salehi, B., & Cook, P. (2013). Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (pp. 266–275). Association for Computational Linguistics.

Salehi, B., Cook, P., & Baldwin, T. (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 977–983). Association for Computational Linguistics.

Sangati, F., Zuidema, W., & Bod, R. (2010). Efficiently extract rrecurring tree fragments from large treebanks. In *Proc. of the Seventh LREC (LREC 2010)*. Valetta, Malta: ELRA.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., Qasemi Zadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., & Doucet, A. (2017). The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 31–47). Valencia, Spain: ACL. https://doi.org/10.18653/v1/W17-1704

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., & Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth LREC (LREC 2014)*. Reykjavik, Iceland: ELRA.

Seretan, V. (2011). *Syntax-Based Collocation Extraction, volume 44 of Text, Speech and Language Technology*. 1st edition. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-007-0134-2.

Seretan, V., & Wehrli, E. (2009) Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation. Special Issue on Multilingual Language Resources and Interoperability*, 43(1), 71–85.

Silva, J. & Lopes, G. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL6)* (pp. 369–381). Orlando, FL.

Smadja, F. A. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), 143–177.

Tsvetkov, Y., & Wintner, S. (2011). Identification of multi-word expressions by combining multiple linguistic information sources. In R. Barzilay, M. Johnson (Eds.), *Proceedings of the 2011 EMNLP (EMNLP 2011)* (pp. 836–845). Edinburgh, Scotland, UK: ACL.

Vargas, N., Ramisch, C., & Caseli, H. (2017). Discovering light verb constructions and their translations from parallel corpora without word alignment. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 91–96). Valencia, Spain: ACL. https://doi.org/10.18653/v1/W17-1711

Villavicencio, A., Bond, F., Korhonen, A., & McCarthy, D. (Eds.) (2005). *Computer Speech & Language. Special issue on Multiword Expression*, 19(4). Elsevier.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., & Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In J. Eisner (Ed.), *Proceedings of the 2007 Joint Conference on EMNLP and Computational NLL (EMNLPCoNLL 2007)* (pp. 1034–1043). Prague, Czech Republic: ACL.

Weller, M., & Heid, U. (2012). Analyzing and aligning German compound nouns. In *Proceedings of the Eighth LREC (LREC 2012)*. Istanbul, Turkey: ELRA.

Zhou, X., Zhang, X., & Hu, X. (2007). Dragon toolkit: Incorporating auto-learned semantic knowledge into large-scale text retrieval and mining. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence– ICTAI* 2007, volume 2 (pp. 197–201). Washington, DC: IEEE Computer Society. https://doi.org/10.1109/ICTAI.2007.117

Multiword expressions in comparable corpora

Peter Ďurčo
University of SS. Cyril and Methodius in Trnava

On the basis of Aranea Gigaword Web corpora, a family of comparable corpora intended for use in contrastive linguistic research, multilingual lexicography, language teaching and translation studies we discuss the pros and cons of comparable corpora in contrast to monolingual and parallel corpora for the analysis of multiword entities (MWEs). We demonstrate that by using large corpora for two or more languages, consisting of unrelated texts, yet created in a comparable manner, parallel language structures and phenomena like MWEs can be identified if the appropriate tools are employed. With the Aranea corpora, the "bilingual sketch" functionality of the Sketch Engine is one such tool which provides a new approach for analyses of similarities of (or differences between) collocation profiles (word sketches) for words and their translation equivalents.

Keywords: comparable corpora, universal tagset, compatible Sketch Grammars, multiword expressions

1. Comparable corpora: A brief survey

Text corpora with respect to language can be monolingual (texts written in a single language) or bilingual/multilingual. Of bilingual/multilingual corpora there are two basic types: parallel (mutual translations of texts) and comparable corpora (texts with common characteristics like similar domains, genres, registers, time of origin, size, strategy of annotation, etc.). All of these types of corpora can be used for a comparison of language data. We can analyse and compare similar phenomena in monolingual corpora independently and evaluate the results introspectively. Another quality of comparison is to allow bi- and multilingual corpora with different degrees of comparability. We can, for example, consider localised translations of user manuals to be totally parallel, news reports on the same events or Wikipedia articles in different languages as highly comparable, and texts from the same domain and genre but describing different events, e.g. parliamentary debates on the

https://doi.org/10.1075/ivitra.24.07dur

same issue from different countries, as somewhat comparable. Certainly, we can also use unrelated texts like Internet texts in contrastive research.

The central problem is the weakness of the concept of comparability in multilingual corpora among experts. According to the EAGLES– Expert Advisory Group on Language Engineering Standards Guidelines (1996) – "a comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora. The value of comparable corpora lies in the potential to compare different languages or varieties in similar circumstances of communication, while avoiding the inevitable distortion introduced by the translations of a parallel corpus".[1]

Within the ICE-GB (International Corpus of English) Project, we find the following statements concerning the concept of comparable corpora: "A Comparable Corpus is a collection of "similar" texts in different languages or in different varieties of a language. The criteria to define the similarity between texts are not clearly defined, but the aim of this type of corpora is to compare the languages or varieties presented in similar circumstances of communication, without the distortions which appear in translated texts of Parallel Corpora."[2]

According to SMT Research Survey "Comparable Corpora" is the main subject of 25 publications published between 2002–2015.[3] According to SMT Research Survey "Dictionaries From Comparable Corpora" is the main subject of 87 publications appeared between 2000–2015 and 9 before this period.[4]

Maia (2003) sees the main reason for constructing comparable corpora the considerable resemblance of translated text to the original. Other advantages for creating Comparable Corpora are according Maia their availability and versatility, usage in Discourse Analysis and Pragmatics, Terminology Extraction, Information Retrieval and Knowledge Engineering, where the results can be better than in parallel corpora (Maia, 2003: 27).[5]

In the SMT Research Survey we find a modified definition of what Comparable Corpora are: "A comparable corpus is a pair of corpora in two different languages,

---

1. "Comparable Corpora." Examples for Conjunction Section, www.ilc.cnr.it/EAGLES/corpustyp/node21.html.

2. https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/types/comparable.html

3. http://www.statmt.org/survey/Topic/ComparableCorpora

4. http://www.statmt.org/survey/Topic/DictionariesFromComparableCorpora

5. Maia, B. What are comparable corpora? http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.197.1279&rep=rep1&type=pdf

which come from the same domain. It means that parallel sentences may also be mined from comparable corpora such as news stories written on the same topic in different languages. The transition from parallel corpora over noisy corpora that require cleaning all the way to comparable corpora is fluent."[6]

According to J. Smith et al. (2010), there have been several approaches in recent years developed for obtaining parallel sentences from non-parallel, or comparable data, such as news articles published within the same time period, or web pages with a similar structure. Also, experiments with Wikipedia articles in different languages became popular.[7]

According to Barrón-Cedeño et al. (2015: 3) comparable corpora were first used for extracting parallel lexicons, later they were used for feeding statistical machine translation (SMT) systems. Barrón-Cedeño et al. present a model for the automatic extraction of comparable texts from Wikipedia. The authors claim that Wikipedia is a suitable source of multilingual texts with different levels of comparability.[8]

Since 2008, regular workshops on building and using comparable corpora (BUCC) have taken place.[9] The main topics here are the problems of building comparable corpora, applications of comparable corpora and mining from comparable corpora.

In the overview of research on comparable corpora (Sharoff et al., 2016), the authors describe a twenty year development of using comparable corpora for the extraction of bilingual dictionaries and statistical machine translation. There are many different algorithms identifying matching words in comparable corpora. The authors state that for further research it is not necessary to determine which algorithm is better but instead to try to combine them all together, what might be a way to weight and combine them in an optimal way to identify word translations in comparable corpora, taking the best of each world (Sharoff, Rapp, Zweigenbaum 2016: 12).

A new methodology and a system for collocation retrieval and translation is outlined by Mendoza Rivera, Mitkov and Corpas Pastor (2013) using the combination of parallel and comparable corpora. Despite the fact that working with comparable corpora is according authors not highly reliable because of its noisy nature and the translation precision of multiword expressions in comparable corpora is lower than in parallel corpora, the authors state that the way forward would be to adjust the comparable corpora algorithm (Mendoza Rivera et al. 2013: 24).

---

6. http://www.statmt.org/survey/Topic/ComparableCorpora
7. https://www.aclweb.org/anthology/N/N10/N10-1063.pdf
8. http://aclweb.org/anthology/W/W15/W15-3402.pdf
9. https://comparable.limsi.fr/bucc2016/bucc-introduction.html

2. Aranea comparable corpora

Corpora of the same size and created in the same way could arguably deserve the designation "comparable". Aranea is a family of web corpora intended for use in contrastive linguistic research, multilingual lexicography, language teaching and translation studies.

2.1 Methodology

The ARANEA project, conducted by Benko, is an on-going experiment aimed at creating a family of billion-token web corpora. The basic principles for creating comparable corpora are according Benko the same size of data gathered by crawling the web at (approximately) the same time, containing similar web-specific domains, genres and registers, further pre-processed, filtered and deduplicated by the same tools, morphologically annotated by (possibly) the same tagger and made available via Sketch Engine. The data have been downloaded from the Internet and processed by the same set of open-source and free tools (Benko, 2013, 2014a).

Language-independent processing was achieved using the same tools and included compatible tokenization, sentence-segmentation of document, paragraph and sentence-level deduplication, and POS tagging by free tools. The native tagsets have been mapped into Araneum Universal Tagset (AUT) and the Word Sketch option operates with Compatible Sketch Grammars (CSG) (Benko 2014b).

2.2 Available corpora

The Aranea family includes corpora for 14 languages: Araneum Anglicum, Araneum Anglicum Africanum, Araneum Anglicum Asiaticum, Araneum Bohemicum, Araneum Bulgaricum, Araneum Finnicum, Araneum Francogallicum, Araneum Germanicum, Araneum Hispanicum, Araneum Hungaricum Araneum Italicum, Araneum Nederlandicum, Araneum Polonicum, Araneum Portugallicum, Araneum Russicum, Araneum Russicum Russicum, Araneum Russicum Externum, Araneum Sinicum and Araneum Slovacum.

There are currently three corpora in preparation: Araneum Ucrainicum, Araneum Bohemicum and Araneum Georgianum. There are also other corpora in the planning stages: Araneum Francogallicum Africanum, Araneum Anglicum Australicum, Araneum Hispanicum Americanum, Araneum Portugallicum Ibericum and Araneum Germanicum Externum.[10] Each corpus exists in several editions, differing by their sizes. The basic (medium-sized) version, *Maius* ("greater"), contains

10. http://sketch.juls.savba.sk/aranea_about/

approximately 1.2 billion tokens (i.e., over 1 billion words). The 10% random sample of *Maius*, called *Minus* ("smaller"), is to be used for teaching purposes. A 1% sample, *Minimum* ("minimal"), is utilized in debugging the processing pipelines and tuning the sketch grammars. The largest *Maximum* ("maximal") edition will contain as much data as can be downloaded from the web for the particular language.[11]

Benko sees further work on comparable corpora through developing the Aranea Project in several directions: (1) Compare the data with other web corpora (for languages having them), as well as with traditional corpora (where data are available); (2) Include new versions of existing corpora with better filtering and annotation, alternative tagging (where tools are available), improved annotation through post-processing, size based on deduplicated data and also with feedback from the users incorporated (Benko 2014a); (3) Complete the language list with languages needed in foreign language teaching at Universities in Slovakia (Bulgarian, Romanian, Modern Greek, Swedish, Japanese, and Korean) and provide the region-specific variants for larger languages (American vs. Iberian Spanish, Canadian and African French, non-Germany German); (4) Improve the processing pipeline by incorporating user, (5) Provide additional layers of alternate morphosyntactic annotation for languages where more than one tagger and/or language model is available (Benko 2016).

2.3 Access to CC

The Aranea corpora are accessible via the free web interface at http://ucts.uniba.sk (without word sketches, however) and they are also hosted at http://kontext.korpus.cz (a free registration is required) and at http://ella.juls.savba.sk, the web page of the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences. Users who have an account with the Sketch Engine website can enjoy the full functionality of that system provided by the CSG at http://www.sketchengine.co.uk (a 30-day free trial is available).

3. Multi-word expressions in comparable corpora

Comparable corpora can be effectively used for the extraction and comparison of multiword expressions (MWEs). At the Workshop on Multi-Word Units in Machine Translation and Translation Technology (MUMTTT2015)[12] in Malaga we presented some results of our analysis and observations (Benko and Ďurčo,

11. http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_031_p_417.pdf

12. http://www.europhras2015.eu/presentation

2015: 40–42). First of all, we showed that it is possible to use the sketch difference tool for the identification of preferences in usual and fixed collocability of words in compared languages. Secondly, we can separately extract the collocational profiles and count collocations independently in compared languages and evaluate the cases of identities, inclusions and incompatibilities of the compared keywords. Another possibility is to look for appropriate or preferred equivalents.

3.1 Competition between monolingual and comparable corpora

The first question is whether the results of the analysis of MWEs in monolingual corpora are significantly different from the results obtained by use of comparable corpora. This should not be the case, if the corpora are equal in balance and representativeness and also when analysing the most salient facts of the language. Let us verify this assumption by comparing collocations with the key word *foreigner* in various corpora.

3.1.1 *Intralingual sketch in monolingual vs. comparable corpus*

A simple intralingual comparison of two different Web corpora demonstrates the crucial problem of the comparability of various corpora. Despite the fact that the corpora have different sizes, different sources and different mark-up systems, the most salient language phenomena should be almost identical in corpora of general language, but this is not the case. The following table shows the nearest left modifiers to the lemma *foreigner* based on the statistical score logDice. We can observe very different preferences in the collocability:

Table 1. Nearest left modifiers to the lemma *foreigner*

Araneum Anglicum	**enTenTen13**
1. transient	1. non-resident
2. wealthy	2. wealthy
3. unauthorized	3. diligent
4. young	4. skinned
5. rich	5. resident
6. Mexican	6. undocumented
7. white	7. english-speaking
8. bloody	8. ignorant
9. dark	9. unsuspecting
10. perpetual	10. perpetual
11. undocumented	11. non-eu
12. poor	12. kidnapping
13. illegal	13. HIV-positive
14. famous	14. untried
15. HIV-positive	15. gullible

There are just three identical collocates among the first 15 most salient items. What is the consequence of this fact for linguistic research? We have to keep in mind that there is no fully representative corpus of a language and linguists and lexicographers have to use hybrid methods, hybrid approaches and many different sources to mine representative data.

3.1.2 *Interlingual sketch from monolingual corpora*

The comparison of the English lemma *foreigner* and the German *Ausländer* in the English Web 2013 (enTenTen13) (freq. 166, 281; 7.31 per million) and German Web 2013 (deTenTen13) (freq. 254, 571; 12.78 per million) corpora, respectively, shows an astonishing uneven dispersion of the most frequent collocations in the first ranges of twenty collocates:

Table 2. Nearest left modifiers of *Foreigner/Ausländer*

modifiers of "foreigner"	**modifiers of "Ausländer"**
1. non-resident	1. kriminell
2. wealthy	2. geduldet
3. diligent	3. lebend
4. skinned	4. ausreisepflichtig
5. resident	5. straffällig
6. undocumented	6. heimatlos
7. english-speaking	7. eingebürgert
8. ignorant	8. eingereist
9. unsuspecting	9. feindlich
10. perpetual	10. arbeitslos
11. non-eu	11. unwillig
12. kidnapping	12. überzählig
13. hiv-positive	13. zugewandert
14. speaking	14. einreisend
15. untried	15. illegal
16. gullible	16. willig
17. swarthy	17. wohnhaft
18. dark-skinned	18. aufhaltend
19. enterprising	19. hochqualifiziert
20. bearded	20. eingewandert

Apart from the fact that there are practically no common collocates and the collocation profiles are in the range of the first twenty collocates disjunctive, we also observe noticeable axiological difference. While the most English collocates are either positive or neutral in connotation, nearly half of all German collocates have negative connotations.

Besides dry facts about statistically significant collocations we can also get the real picture about the axiological status of collocations with this particular word in discourse. Collocations of the node words *foreigner* and *Ausländer* with Verb Y + Noun X as object conspicuously demonstrate this fact:

Table 3. Verbs by nouns *foreigner*/*Ausländer* as objects

foreigner	**Ausländer**
1. deport	1. hetzen
2. expel	2. verprügeln
3. kidnap	3. bürgern
4. marry	4. überfremden
5. naturalize	5. wettern
6. persecute	6. schimpfen
7. detain	7. zusammenschlagen
8. bar	8. hassen
9. abduct	9. herziehen
10. blame	10. diskriminieren
11. lure	11. beschimpfen
12. domicile	12. schüren
13. forbid	13. einwandern
14. arrest	14. abschieben
15. overcharge	15. prügeln
16. despise	16. vermieten
17. target	17. heiraten
18. welcome	18. überfallen
19. prohibit	19. dulden
20. attract	20. verüben
21. hire	21. attackieren
22. permit	22. beleidigen
23. hate	23. unterrichten
24. import	24. anbelangen
25. ban	25. jagen

The striking fact is that most verbal collocates in English as well as in German can be subsumed into negative values VIOLENCE, FORCE, HATE, DISCRIMINATION.

3.1.3 *Intralingual sketch difference and collocational equivalent*

Another case of comparability is the function of "sketch difference". The comparison of collocational profiles in a monolingual corpus can aid in cases of divergent equivalence, i. e. "one to more". A big problem between Slovak and German is for example the decision on how to use, in "preposition+Noun" collocations with the Slovak preposition *na*, two corresponding German prepositions *an* or *auf*. The

profiles illustrate the preferences and differences in the usage of these interlinguistically ambiguous prepositions. This allows a clear identification of the cases of specific usage of the interlingual semantic vague and ambiguous prepositions. In addition to the "common patterns" we also get results for the "only patterns", which gives us the answer, how to disambiguate cases of poly- or semiequivalence:

Table 4. Examples for "*auf* only patterns"

PräpX SubstYDat	VerbY PräpX
1. Ebene	1. basieren
2. Internetseite	2. beruhen
3. Rang	3. belaufen
4. Erden	4. beschränken
5. Spur	5. vertrauen
6. Mio.	6. spezialisieren
7. Startseite	7. schwören
8. Sofa	8. fokussieren
9. Leinwand	9. abstimmen
10. Hochtour	10. reduzieren
11. Bundesebene	11. besinnen
12. Rückweg	12. fußen
13. Kontinent	13. beharren
14. Festplatte	14. verkürzen
15. Couch	15. lasten

Table 5. Examples for "*an* only patterns"

PräpX SubstYDat	VerbY PräpX
1. Nachmittag	1. Appellieren
2. Bord	2. grenzen
3. Hochschule	3. mangeln
4. Schluss	4. knüpfen
5. Besten	5. zweifeln
6. Vormittag	6. gewöhnen
7. Spieltag	7. versterben
8. Küste	8. kratzen
9. Institut	9. klammern
10. Sonnabend	10. anlehnen
11. Himmel	11. verfassen
12. Freitagabend	12. lehren
13. Eingang	13. basteln
14. Samstagabend	14. erkranken
15. Bedeutung	15. klingeln

We can extend and deepen the analysis and use these data in the description of grammar and lexicon. We can also use this in foreign language teaching by letting students carry out such an analysis and evaluate the results.

3.2 Data mining in comparable corpora

The Aranea Comparable Corpora offer intra- as well as interlingual comparison of different language phenomena, based on identical methodology. We can use all the functionalities of the Sketch Engine and compare the data manually or we can have recourse to the bilingual sketch functionality.

3.2.1 *Intralingual sketches in different varieties of english corpora*

One of the special options in the Aranea Corpora is the ability to compare language varieties. The results show amazing and significant differences. Let us compare the nearest left modifiers of the node word *foreigner* in Araneum Anglicum, Araneum Anglicum Africanum and Araneum Anglicum Asiaticum:

Table 6. Naerest left modifiers of *foreigner*

Araneum Anglicum Asiaticum	**Araneum Anglicum**	**Araneum Anglicum Africanum**
Naerest left modifiers of *foreigner*		
1. skilled	1. transient	1. illegal
2. resident	2. wealthy	2. skilled
3. concerned	3. unauthorized	3. white
4. white	4. young	4. wealthy
5. rich	5. rich	5. undocumented
6. wealthy	6. Mexican	6. rich
7. illegal	7. white	7. black
8. female	8. bloody	8. African
9. new	9. dark	9. poor
10. young	10. perpetual	10. greedy
11. qualified	11. undocumented	11. resident
12. non-resident	12. poor	12. legal
13. good	13. illegal	13. ignorant
14. talented	14. famous	14. compelling
15. registered	15. HIV-positive	15. young

The similarity of collocation profiles of the first 15 collocates between Araneum Anglicum and Asiaticum and also Anglicum and Africum is just 33% (*wealthy, young, rich, white, illegal*). The inclusion of the collocation profiles between

Araneum Africum and Asiaticum is 40% (*skilled, white, rich, wealthy, illegal, young*). The common collocates comprise just 20% (*wealthy, young, rich*). The consequence of this fact for linguistics and lexicography is that in case of pluricentric languages we have always to keep in mind the potential cultural and pragmatic aspects, which find their expression in noticeable differences in the preferred collocability of described words.

3.2.2 *Interlingual sketches in comparable corpora*

The Aranea Comparable Corpora offer a unified and to some extent homogeneous experimental platform for analysing such language phenomena as collocational preferences, collocational compatibility, collocational equivalence, collocational behaviour MWE and many other options.

3.2.2.1 *Collocational preferences*

Collocational preferences can serve as evidence for the different distribution of collocates by equivalent key words in compared languages caused (i) by different polysemy or (ii) by the fact of different presence in the discourse in each language. The following Comparable Word Sketch (CWS) shows the preferences in the collocability of the compared key words in German and Slovak for *Krise/kríza* (= crisis):

Among the first 50 adjective collocates there are only 13 identical (*existentiell / existenčný, humanitär / humanitárny, gegenwärtig / súčasný, ökonomisch / ekonomický, innenpolitisch / vnútropolitický, derzeitig / súčasný, tief / hlboký, global / globálny, zyklisch / cyklický, systemisch / systémový, wirtschaftlich / hospodársky, jetzig / terajší, fiskalisch / fiškálny*). Various preferences can be discovered in this CWS. For German, the preferred adjectives are *suizidal, humanitär, kapitalistisch, verschärfend, zuspitzend, existenzbedrohend, akut, überstanden*. In comparison with German adjectives expressing mental states like *psychotisch, seelisch, psychisch, psychosozial, spirituell*, the Slovak adjectives of this semantic field are either not present at all or they occur very rarely in the corpus. The explanation for the most preferred Slovak adjective *dlhový* (= in debt) is the simple fact that in German all such cases are expressed through a compound word like *dlhová služba / Schuldendienst* (= debt service), *dlhová kríza / Schuldenkrise* (= debt crisis). Such findings as the difference between cases where to use for the Slovak determinative syntagma also/or a determinative syntagma and when a compound word in German are also very important for teaching purposes and for determining the appropriate equivalents.

3.2.2.2 *Collocational compatibility*

The associative combinatorics of words from a common semantic paradigm in MWE studies is arbitrary and uneven in compared languages. CWS exposes such cases and shows the common, preferred and split distribution in cases when the

basis has two or more equivalents in the other language. For example, the Slovak verb *spáchať* has two equivalents in German: *begehen* 'to commit' and *verüben* 'to perpetrate'. CWS has shown that among the collocates common to both *begehen* and *verüben* in the sense of "to commit a bad thing" (excluding collocates *Mord*, *Straftat*, etc.), there are noticeable differences in the combinatorial preferences, e.g. for *begehen* there are preferred collocates *Delikt*, *Harakiri*, *Schweinerei*, *Verrat*, *Verbrechen*, etc., and by for *verüben* the collocates *Anschlag*, *Attentat*, *Massaker*, *Selbstmord*. Otherwise, the exchangeability is limited, e.g. *Inzest*, *Sünde*, and *einen Verstoß* are compatible only with *begehen*. This fact has direct consequences for the contrastive description.

3.2.2.3 *Collocational behaviour of MWES*

Comparable corpora also allow analysis of the collocational behaviour of MWEs, i.e. the contextual combinatorial potential of collocations. The following example illustrates the measure of identity, inclusion and disjunction of so-called minimal idiom *nach Belieben* 'at will' and the Slovak equivalent *podľa ľubovôle*.

The German MWE *nach Belieben* has two basic meanings: 1. *eigenwillig, eigensinnig* 'willfull, opinionated' and 2. *beliebig, willkürlich* 'arbitrary, optional'.

First, we observe the difference in the overall collocability of the equivalent MWEs *nach Belieben* and *podľa ľubovôľe*. The collocational potential of the German MWE with verbs is much higher:

i. German verbal collocates (106) with *nach Belieben*:
abändern, abschmecken, abwandeln, agieren, anbraten, anmalen, anordnen, anrichten, auffüllen, aufkochen, ausbeuten, ausblenden, ausrollen, ausschalten, ausstechen, aussteigen, aussuchen, austauschen, austoben, auswählen, auswechseln, ändern, bearbeiten, bedienen, befüllen, beherrschen, bekleben, bemalen, beschriften, bestäuben, bestreichen, bestreuen, beträufeln, dazugeben, dekorieren, diktieren, dirigieren, dominieren, dosieren, editieren, einfärben, einrühren, einteilen, erweitern, formatieren, formen, füllen, garnieren, gestalten, herschieben, hinzufügen, hinzugeben, individualisieren, kombinieren, konfigurieren, kontrollieren, manipulieren, mischen, mixen, modifizieren, nahezu, nachwürzen, pfeffern, platzieren, punkten, pürieren, regulieren, salzen, servieren, schalten, schälen, schneiden, skalieren, sortieren, stöbern, streuen, süßen, umspringen, untermischen, unterrühren, variieren, verändern, verbiegen, verdünnen, verfahren, verfeinern, vergrößern, verkleinern, verlängern, vermengen, vermischen, verrühren, verschieben, verschönern, verstellen, verteilen, verzieren, walten, wechseln, weiterverarbeiten, weiterverwenden, wirbeln, würzen, zoomen, zuckern, zusammenstellen

ii. Slovak verbal collocates (44) with *podľa ľubovôle*:
brať, disponovať, dosadzovať, doplniť, dosadiť, dovoliť, kombinovať, konať, meniť, nastaviť, nakladať, naložiť, narábať, odchádzať, odmeňovať, opakovať, použiť, pridávať, pripraviť, prispôsobovať, riadiť, rozhodovať, rozširovať, stavať, trestať, upraviť, upravovať, užívať, vládnuť, vyberať, vybrať, vychádzať, vymeniť, vytvárať, vytvoriť, využiť, využívať, zahrávať sa, zaobchádzať, zasahovať, zmeniť, zneužívať, zvoliť, zvyšovať.

Besides, the analysis of the collocation *nach Belieben* has shown that the cases of (in)compatibility with verbs go through the particular groups of semantically similar verbs.

In the semantic group "to dominate" we have indicated the following relations of (in)compatibility of verbal collocates (0 = not compatible, + = compatible, ++ = other preferred verbal collocate):

dominieren	*0*	*dominovať*
deklassieren	*0*	*deklasovať*
degradieren	*0*	*degradovať*
beherrschen	+	*ovládať/vládnuť/panovať*
diktieren	+	*diktovať*
kontrollieren	+	*kontrolovať*
gewinnen	*0*	*zvíťaziť/+víťaziť* (only imperfective Aspect).

In the domain "to handle freely" we have the following parallels and restrictions in the compatibility:

schalten und walten	+	*panovať a vládnuť*
verfahren	+	*postupovať*
sich bedienen	+	*obslúžiť sa*
sich austoben	*0*	*vyblázniť sa* (*do ľubovôle, do vôle*)
aussuchen	++	*vyhľadať* (*+ podľa chuti*)
stöbern	++	*prehŕňať sa* (*+ do vôle*)

In the domain "sport" we indicate only the cases of incompatibilty of parallel verbs in Slovak:

punkten	*0*	*bodovať*
treffen	*0*	*triafať*
einnetzen	*0*	*vsietiť*
kontern	*0*	*kontrovať*
scoren	*0*	*skórovať*
gewinnen	*0*	*zvíťaziť*

Very specific is the last verbal domain "cooking" where we see rather complicated relations of (in)compatibility:

garnieren	+	*obložiť*
bestreuen	+	*posypať*
variieren	+	*veriovať*
abschmecken	+	*dochutiť*
verzieren	+	*ozdobiť*
dazugeben	+	*pridať*
überziehen	+	*potrieť, natrieť*
hinzugeben	+	*pridať*
zusammenmischen	+	*zmiešať*
zuckern	0	*cukrovať*
salzen	0	*soliť*
würzen	0	*koreniť*
pfeffern	0	*koreniť*
verfeinern	0	*zjemniť*
pürieren	0	*robiť pyré*

These facts have serious consequences when it comes to determining the equivalents in Slovak. In cases of compatibility there are very often complementary equivalents and in cases of incompatibility there are many other equivalent candidates in Slovak, e.g.:

1. *podľa ľubovôle* (*nach Belieben*)
2. *do ľubovôle, do vôle* (*bis Belieben*, *bis Willen*)
3. *podľa chuti* (*nach Geschmack*)
4. *podľa uváženia* (*nach Abwägung/Erwägung*)
5. *podľa (svojej/vlastnej) vôle* (*nach (seinem/eigenem) Willen*)
6. *podľa (svojho/ vlastného) želania* (*nach (seinem/eigenem) Wunsch*)
7. *podľa slobodnej vôle* (*nach freiem Willen*).

4. Conclusion

In our paper, we have tried to illustrate different approaches to the phenomenon of comparability in various types of corpora. Modern monolingual and multilingual corpora offer new tools for the comparison of data by using large corpora for two languages, consisting of unrelated texts. We also tried to demonstrate the use of corpora created in a comparable manner, where parallel language structures and phenomena can be identified if the appropriate tools are employed. With the Aranea corpora, the "bilingual sketch" functionality of the Sketch Engine is one such tool

which allows analyses of similarities of (or differences between) collocation profiles (word sketches) for words and their translation equivalents. The comparison of results in different types of corpora, constructed in different manners with different tools, has shown the necessity of combined and hybrid approaches because of limited balance and representativeness of existing corpora today.

References

Barrón-Cedeño, A., España-Bonet, C., Boldoba, J., Màrquez, L. (2015). A Factory of Comparable Corpora from Wikipedia. In P. Zweigenbaum, S. Sharoff, & R. Rapp (Eds.), *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora* (pp. 3–13). Stroudsburg: The Association for Computational Linguistics. http://aclweb.org/anthology/W/W15/W15-3402.pdf (Accessed: 2018-06-11).

Benko, V. (2013). Data Deduplication in Slovak Corpora. In K. Gajdošová, & A. Žáková (Eds.), *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning* (pp. 27–39). Lüdenscheid: RAM-Verlag.

Benko, V. (2014a). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech and Dialogue. 17th International Conference*, TSD 2014, Brno, Czech Republic, September 8–12, 2014 (pp. 257–264). Springer International Publishing Switzerland. ISBN: 978-3-319-10815-5 (Print), 978-3-319-10816-2 (Online).

Benko, V. (2014b). Compatible Sketch Grammars for Comparable Corpora. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus 15–19 July 2014* (pp. 15–19). Bolzano/Bozen: Eurac Research. ISBN: 978-88-88906-97-3.

Benko, V. (2016). Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In N. Calzolari et al. (Eds.), *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4245–4248). Portorož: European Language Resources Association (ELRA). https://pdfs.semanticscholar.org/96f9/5b9f0a3a0b616e2551c9272a903cedf5db39.pdf?_ga=2.94800924.1804063941.1528717415-1706820565.1528717415 (Accessed: 2018-06-11).

Benko, V., & Ďurčo, P. (2015). Aranea. Comparable Gigaword Web Corpora. In G. Corpas Pastor, R. Mitkov, J. Monti, & V. Seretan (Eds.), *Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT2015)* (2nd edition) (pp. 40–42). LEXYTRAD, Research Group in Lexicography and Translation. http://www.europhras2015.eu/2mumttt2015/! (Accessed: 2018-06-11).

Maia, B. (2003). What are comparable corpora. In *Proceedings of the Corpus Linguistics workshop on Multilingual Corpora: Linguistic requirements and technical perspectives*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.197.1279&rep=rep1&type=pdf (Accessed: 2018-06-11).

Mendoza Rivera, O., Mitkov, R., & Corpas Pastor, G. (2013). A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora. In J. Monti, R. Mitkov, G. Corpas Pastor, & V. Seretan (Eds.), *Workshop Proceedings for: Multi-word Units in Machine Translation and Translation Technologies (Organised at the 14th Machine Translation Summit 2013)* (pp. 18–25). Allschwil: The European Association for Machine Translation. https://www.academia.edu/6957419/A_Flexible_Framework_for_Collocation_Retrieval_and_Translation_from_Parallel_and_Comparable_Corpora (Accessed: 2018-06-11).

Sharoff, S., Rapp, R., & Zweigenbaum, P. (2016). Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In *BUCC, 9th Workshop on Building and Using Comparable Corpora. Co-located with LREC 2016 Portorož (Slovenia) 23 May 2016*. https://comparable.limsi.fr/bucc2016/ (Accessed: 2018–06–11).

Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (pp. 403–411). Los Angeles, CA: Association for Computational Linguistics. https://www.aclweb.org/anthology/N/N10/N10-1063.pdf (Accessed: 2018–06–11).

Internet links (Last accessed on June 11, 2019)

Aranea. A Family of Comparable. Gigaword Web Corpora. (n.d.), Retrieved from http://sketch.juls.savba.sk/aranea_about/

BUCC, 9th Workshop on Building and Using Comparable Corpora. (Last modified on April 23, 2016) (Retrieved from https://comparable.limsi.fr/bucc2016/

Comparable Corpora. (n.d.) Examples for Conjunction Section, Retrieved from www.ilc.cnr.it/EAGLES/corpustyp/node21.html

Comparable Corpora. (n.d.) Retrieved from https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/types/comparable.html

Comparable Corpora. (Last modified on August 08, 2017), Retrieved from http://www.statmt.org/survey/Topic/ComparableCorpora

Learning Bilingual Dictionaries from Comparable. (Last modified on September 13, 2017), Retrieved from http://www.statmt.org/survey/Topic/DictionariesFromComparableCorpora

Wikipedia Comparable Corpora. (n.d.) Retrieved from http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

Collecting collocations from general and specialised corpora

A comparative analysis

Marie-Claude L'Homme and Daphnée Azoulay
Observatoire de linguistique Sens-Texte, Université de Montréal

Collocations are increasingly taken into account in general and specialised repositories and methodologies to collect them are heavily based on corpora. However, lexicographers and terminologists use different kinds of corpora in which combinations are likely to behave according to specific rules and/or patterns. This contribution presents a comparative analysis of the collocational behaviour of 15 lexical items found in a general language corpus and a specialised corpus on the theme of the environment. We automatically extracted large sets of collocates (three lists of 50 collocates) for each lexical item and from each corpus and analyse different facets of collocational behaviour: polysemy of lexical items, characteristics of collocates (overlap, rank and semantic classes of collocates, etc.). Our aim is to draw the attention of terminologists and lexicographers to some specific factors affecting the behaviour of collocations in specialized and general corpora.

Keywords: collocation, terminology, lexicography, specialised corpus, general corpus, semantic class

1. Introduction

Lexical combinations – and more specifically collocations – are increasingly taken into account in both general language and specialised printed or online dictionaries. The well-known general language collocation dictionary (Benson et al., 1986) and the specialised work in the field of the stock exchange (Cohen, 1986) have paved the way for the active collection of word combinations and their organisation in lexical and terminological resources (Buendia and Faber, 2014; L'Homme, 2009).

The combinations collected may vary from one reference work to another for various reasons: the types of multiword expressions considered and the way these are defined, the nature of the dictionary (general or specialised), methods

https://doi.org/10.1075/ivitra.24.08lho

for retrieving combinations, the nature of the corpus used to locate multiword expressions, etc. Furthermore, the analysis of the data itself is carried out from specific perspectives since terminologists focus on terms and their contribution to the expression of knowledge, while lexicographers consider all types of lexical units and the different meanings they convey.

In this contribution, we examine one of these factors, that is the nature of the corpus and its consequences on the collection of collocations. We assume that lexicographers refer to general language corpora (which are often already available), whereas terminologists base their analysis on corpora that contain texts linked to special fields of knowledge (which they usually compile each time they embark on a new project). This comparison is carried out so that it draws the attention of terminologists and lexicographers to some specific factors affecting the the behaviour of combinations when considering a specialised corpus as opposed to a general (or balanced) one.

More specifically, we compare collocations extracted from a general language corpus (a corpus in which the American National Corpus, ANC, and the British National Corpus, BNC, are combined) and those retrieved from a specialised corpus containing texts that deal with the environment. Collocations are extracted by means of an automated method for 15 lexical items found in both corpora.

The chapter is structured as follows. Section 2 is a short literature review on lexical combinations considered in terminology and lexicography, with a special focus on collocations. Section 3 describes the methodological steps we took to extract (candidate) collocates from the general corpus and the specialised one. In Section 4, we compare the lists of collocates extracted with special attention to the aspects of the analysis that are relevant for terminology and lexicography work. In Section 5, some concluding remarks are presented along with guidelines for lexicography and terminology work.

2. Lexical combinations in terminology and lexicography

The lexical combinations considered in this contribution, namely collocations, represent one type of multiword expression considered in dictionaries, terminological databases, and lexical or terminological resources. The types of multiword expressions that terminologists take into account and add to specialised dictionaries (L'Homme, 2000) differ from those that lexicographers consider (Iordanskaja and Mel'čuk, 2017; Moon, 2015),[1] as shown in Table 1.

1. The typologies of multiword expressions presented in Iordanskaja and Mel'čuk (2017) and Moon (2015) are much more detailed than the one reproduced in Table 1.

Table 1. Multiword expressions considered in terminology and lexicography

Terminology		**Lexicography**	
Multiword terms (typically noun phrases)	*atmospheric gas* *carbon dioxide* *greenhouse effect* *operating system*	Compounds (and in English, phrasal verbs)	*green light* *get up* *know-it-all* *setup*
Collocations	*edit a file* *endangered species* *powerful program*	Collocations	*ask a question* *heavy smoker*
		Idiomatic expressions	*bury the hatchet* *pull a fast one*

Traditionally, criteria based on compositionality have been used by lexicographers to classify multiword expressions into the categories that appear in Table 1. Compounds and idiomatic expressions are non-compositional (i.e. they cannot be understood based on the independent meanings of their components), while collocations are semi-compositional. Semi-compositionality is defined as a phenomenon according to which the base or key word conveys a meaning that it carries in other combinations. The collocate, on the other hand, conveys a specific meaning within a collocation (Haussman, 1979).

It should be noted that this traditional definition of collocations has been challenged and revised. According to Meľčuk (1996), a collocation is composed of: 1. a key word freely selected by a speaker on the basis of its meaning; and 2. a collocate whose use is imposed by the key word. In *heavy smoker*, for instance, *smoker* is selected freely. However, in order to express an intensification with this specific lexical unit, *heavy* must be used. With other key words, intensification would be expressed with different collocates. Furthermore, in other languages, *heavy*, when used with *smoker*, is not necessarily translated by its usual equivalent. In French, for instance, the common translation for *heavy* is *lourd*. However, the translation for *heavy smoker* is *fumeur invétéré* or *gros fumeur*.

Terminologists approach multiword terms and collocations differently. Although the "(semi-)compositionality" criterion may apply to some multiword terms or specialised collocations (such as *greenhouse effect* and *operating system*), most combinations placed in either categories show that terminologists do not consider compositionality as a necessary condition to determine whether a linguistic sequence qualifies as a multiword term or a collocation. *Atmospheric gas*, which appears as a multiword term, and *powerful program*, defined as a collocation in Table 1, both have a compositional meaning.

Regarding collocations per se, in addition to the fact that compositionality is not a central property for terminologists, other differences have been reported in the literature. For instance, terminologists usually focus on collocations whose key

words are nouns (L'Homme, 2000). Since most entries in specialised dictionaries are devoted to nominal terms, it seems only natural that terminologists list collocates that combine with nouns (collocates will thus be verbs, adjectives and other nouns)[2] as shown in Table 2.

Table 2. Collocations in specialised resources

Key word	Collocate	Dictionary
action	(verb) ~ *bondit, ~ s'envole* (noun) *montée d'une ~, regain d'une ~* (adjective) ~ *élevée, ~ haute*	Cohen (1986)
achat	(verb) *effectuer un ~, procéder à l'~* (noun) *comportement d'~, incitation à l'~,* (adjective) ~ *régulier, ~ occasionnel*	Binon et al. (2000)
file (verb)	*create a ~, edit a ~* (noun) *installation ~, text ~* (adjective) *corrupted ~, empty ~*	DiCoInfo (2016)

Identifying specialised collocations in running text raises another challenge. Since compositional combinations can be defined either as multiword terms or collocations (or even as free expressions for that matter), these two types of groupings can hardly be distinguished according to purely linguistic criteria. For instance, *atmospheric gas*, *endangered species*, *edit a file*, and *powerful program* can all be understood based on the individual meanings of their components. The first sequence (*atmospheric gas*) was defined as a multiword term and the last two as collocations in Table 1. *Endangered species* could be considered either as a multiword term or as a collocation depending on the terminologists.

In this contribution, the approach to lexical units and multiword expressions is the following:

– First, for a lexical sequence to qualify as a multiword term, its meaning must not be compositional. Hence, according to this definition, the examples in Table 1 need to be re-examined: *greenhouse effect*, *carbon dioxide*, and *operating system* correspond to multiword terms. However, *atmospheric gas*, *endangered species* and *powerful program* contain two distinct terms and could potentially qualify as collocations. A consequence of this approach is that most terms considered (and all those that are taken into account in this analysis – see Section 3.2) are single-word terms.

2. These structural patterns apply to languages such as English, French, and Spanish. In other languages, patterns for term formation and collocations could differ quite drastically.

– Secondly, we comply with terminology methods with respect to the parts of speech of key words and collocates. All lexical items selected for this analysis are nouns. Collocates will thus belong to the parts of speech of verb, adjective and noun.
– Thirdly, since we use an automated method to extract candidate collocates (explained in Section 3.3), we take a flexible approach to the notion of "collocation" and consider that a statistical calculation provides an objective basis for a comparative analysis between specialised and general corpora. We also assume that the automated extraction represents an excellent starting point from which both lexicographers and terminologists can benefit. Of course, we are aware that a lexicographer or terminologist using similar techniques would probably be much more selective and resort to linguistic and usage criteria similar to those mentioned in this section.

3. A comparative analysis

Our comparative analysis is based on sets of candidate collocates linked to 15 lexical items[3] found in a general language corpus and a specialised corpus on the environment. This section describes the methodological steps we took to select the items and extract the collocates. We also present the corpora briefly.

3.1 Corpora

The general language corpus (NC) used in this analysis results from a combination of the well-known American National Corpus (ANC) and British National Corpus (BNC). The number of tokens (lowercased and part-of-speech (POS) tagged with TreeTagger (Schmid, 1994)) amounts to 142,716,339.

The specialised corpus contains texts that are assumed to be related to climate change and was compiled in a previous project using an iterative method that consists in retrieving texts in PDF format from the web (Azoulay, 2017).

The specialised corpus contains 10,800 documents, which correspond to 139,600,662 tokens. Prior to collocation extraction, the corpus was lemmatised and part-of-speech (POS) tagged with TreeTagger (Schmid, 1994), and characters were lowercased and normalised.

3. *Lexical item* is used here, since some of them are polysemic and correspond to more than one lexical unit.

3.2 Lexical items selected

We selected 15 lexical items defined by the term extractor TermoStat[4] (Drouin, 2003) as some of the most specific units in our environment corpus (EC). The extractor carries out a comparative evaluation of the frequency of lexical items in an analysis corpus (in this case, the analysis corpus is the EC) and a reference corpus (the reference corpus is a combination of the BNC and the ANC).[5] Candidates are classified according to their degree of statistical specificity in the analysis corpus. They can also be sorted by descending order of specificity. In this work, we are interested in candidates with high specificity scores assuming that they correspond to terms of the domain of the environment.

The lexical items were chosen among the most specific ones in the specialised corpus. Using the candidates placed at the top of the list, we selected non-predicative nouns (most of them denote entities) based on the assumption that they would often appear as relevant key words in collocations. Nouns denoting events or activities were ignored in our selection. The resulting list of key words is reproduced in Table 3. Lexical items are sorted according to their rank as assigned by the term extractor once other units were discarded.

Table 3. Lexical items selected

1. climate	6. area	11. land
2. energy	7. system	12. source
3. water	8. temperature	13. air
4. gas	9. model	14. resource
5. carbon	10. soil	15. policy

3.3 Automated extraction of collocations

The next step consisted in extracting from the EC and the NC candidate collocations for each lexical item that appears in Table 3. The candidates were retrieved applying the same automated procedure to both corpora.

The proximity between words was measured using a window-based approach that limits the scope to the surrounding words of a target word within a certain span. Symmetric windows of different sizes –1, 2 and 3 – were defined in order to focus on the immediately adjacent words of a target (each selected lexical item).

4. http://termostat.ling.umontreal.ca/

5. The reference corpus is the same as the one used to extract collocations.

A "stop list" of 960 words and characters was applied, which excludes from the list of candidate collocates function words, and recurrent typos resulting from the conversion of PDFs.[6]

A Python program was written to extract collocations, applying a commonly-used association measure, simple log-likelihood (Evert, 2008).[7] This measure uses the expected frequency of co-occurrence (*E*) which is obtained by multiplying the individual frequency of each word (f_1, f_2) and the span-size (k) and dividing the product by the total number of tokens in the corpus (*N*). The program then computes a score for each co-occurrence by comparing the observed co-occurrence frequency (*O*) with the expected co-occurrence frequency (*E*). If *O* is much greater than *E*, pairs are considered collocations.

$$E = \frac{K \cdot f1 \cdot f2}{N}$$

Expected frequency of co-occurrence

$$\text{Simple-ll} = 2\left(O \cdot log\frac{O}{E} - (O - E)\right)$$

Simple log-likelihood

The procedure generated six different lists of candidate collocates for each of the 15 lexical items. Three different lists were retrieved from the EC (one associated with each window size); the other three lists were extracted from the NC (again, one list per window size). The candidate collocates are sorted according to their level of association with the lexical item as a result of the simple log-likelihood calculation. For the analysis, we considered the first 50 candidates in each list.

Examples are shown in Table 4 for the lexical items *temperature* and *climate*. Column 1 and 2 show the first 50 candidates obtained for *temperature* for a window size of 1. Columns 3 and 4 present the top 50 candidates for *climate* for a window size of 3. The figures that appear next to the candidates correspond to the scores produced by the simple log-likelihood measure.

---

6. For instance, function words such as *this* and *while* were discarded. Other kinds of strange character strings (e.g. yearNUM, NUM-cm) were also removed.

7. Authors are fully aware that other statistical measures are used for collocation extraction (Evert, 2004; Pecina, 2009) and could have been used in this work. Our selection was based on purely practical reasons (availability of the script) and this choice should not interpreted as a preference of this measure over others. As was mentioned above, statistical analysis is used here as an objective entry point to each corpus and the extracted candidates must then be analysed carefully by lexicographers or terminologists. This being said, it would be interesting to apply other measures to extract collocations from each corpus and examine how they can highlight similarities and differences.

Table 4. Collocates extracted for *temperature* and *climate*

EC_F1	NC_F1	EC_F3	NC_F3
8. temperature	**8. temperature**	**1. climate**	**1. climate**
surface (98559.84),	room (6118.99),	change (1376201.79),	change (2335.32),
air (36510.13),	high (2027.99),	impact (97506.17),	economic (1344.29),
average (36222.91),	low (1557.17),	intergovernmental (65297.87),	current (815.26),
increase (32495.58),	water (1384.39),	panel (59787.88),	political (679.22),
global (28444.30),	body (1142.49),	global (49316.77),	temperate (582.57),
high (28403.20),	surface (664.91),	variability (39568.15),	warm (522.79),
rise (23745.58),	rise (624.47),	adaptation (34726.46),	global (495.32),
mean (20559.92),	ambient (574.71),	model (28184.53),	create (457.03),
change (11234.51),	annual (402.18),	convention (23676.90),	opinion (397.11),
maximum (10179.35),	drop (317.53),	sensitivity (22546.10),	soil (365.12),
minimum (9904.01),	melting (304.73),	action (18645.79),	mild (330.55),
low (8650.62),	range (303.21),	framework (16448.18),	cold (329.81),
anomaly (8283.77),	global (299.92),	effect (15037.11),	present (305.37),
ambient (8148.82),	sub-zero (289.68),	response (13114.64),	favourable (304.45),
warm (7630.90),	average (279.82),	mitigation (11885.37),	intergovernmental (292.53),
record (6813.56),	gradient (238.81),	science (11622.84),	hot (260.84),
gradient (6773.32),	transition (229.55),	future (11485.97),	humid (219.05),
room (5723.68),	non-permissive (226.68),	system (11017.70),	tropical (187.22),
annual (5131.72),	I (223.90),	policy (11005.74),	panel (174.31),
trend (5126.54),	control (213.00),	projection (10553.58),	prevailing (168.10),
difference (4667.45),	freezing (194.46),	scenario (10304.90),	@card@ (167.67),
range (4509.29),	air (193.30),	adapt (9393.16),	convention (165.60),
sea-surface (4336.98),	maximum (192.30),	regional (9144.56),	harsh (154.43),
extreme (4226.70),	fall (189.53),	weather (8686.24),	warming (150.28),
precipitation (4189.34),	constant (187.98),	affect (8398.29),	cool (146.85),
ocean (3849.83),	increase (184.15),	risk (7889.53),	vegetation (134.85),
water (3847.68),	dependence (183.13),	emission (7336.68),	moral (110.92),
summer (3840.35),	sensitive (170.10),	address (7085.65),	arid (101.39),
daily (3682.60),	elevated (157.74),	plan (6942.83),	dry (99.87),
equilibrium (3558.33),	minimum (156.69),	prediction (6181.79),	extreme (98.98),
variation (3482.06),	extreme (142.25),	extreme (5976.21),	orbiter (97.96),

Table 4. (*continued*)

EC_F1	NC_F1	EC_F3	NC_F3
8. temperature	**8. temperature**	**1. climate**	**1. climate**
profile (3190.52),	oven (141.51),	vulnerability (5339.47),	equable (97.12),
diurnal (3082.76),	melt (137.70),	earth (5249.62),	affect (96.60),
inside (3042.44),	change (135.86),	implication (5180.78),	intellectual (91.49),
atmospheric (2877.94),	core (131.70),	consequence (4934.43),	improve (89.98),
inversion (2533.09),	critical (129.66),	year (4929.93),	earth (89.44),
winter (2286.71),	optimum (125.20),	fuel (4904.05),	sensitivity (88.65),
humidity (2272.49),	difference (123.99),	feedback (4787.18),	weather (88.36),
soil (2013.26),	gauge (116.44),	use (4346.08),	influence (88.17),
climate (1955.43),	dependent (115.08),	area (4286.09),	social (85.86),
emission (1860.81),	daytime (111.16),	influence (4212.72),	winter (85.75),
dependence (1825.45),	soar (111.05),	respond (3992.70),	seasonal (83.40),
vine (1680.32),	sea-surface (108.12),	strategy (3965.18),	kyoto (83.02),
cold (1639.61),	variation (107.64),	plant (3926.13),	world (76.09),
energy (1635.67),	reach (107.47),	issue (3698.08),	worsen (75.40),
optimum (1556.46),	subzero (106.65),	waste (3693.91),	climate (73.33),
drop (1497.79),	correct (105.23),	force (3686.23),	conference (69.75),
elevated (1492.06),	normal (96.88),	abrupt (3686.09),	difficult (69.58),
use (1489.38),	atmospheric (93.70),	assessment (3681.63),	mediterranean (68.18),
fluctuation (1467.94),	humidity (92.17),	past (3656.21),	topography (67.21),

Of course, lists of candidate collocates that were extracted automatically contain errors, as can be seen in the results reproduced in Table 4. The most common ones are listed below:

- Segmentation problems: e.g. *current-year* was retrieved as a single collocate for *climate* instead of two words;
- Use of special characters: e.g. *@card@* was retrieved as a collocate;
- Lemmatisation problems: e.g. *data* and *datum* were retrieved as two separate collocates;
- Collocate of another term: e.g. *abrupt* was extracted as a candidate collocate for *climate*. Although it can be found in the vicinity of *climate*, it modifies *change* (*abrupt climate change*).

We took into account the entire list of candidates when analysing the data quantitatively. In qualitative observations, we did not take errors into consideration.

4. Observations on the lists of candidate collocations

4.1 Overlap of candidate collocates

The first observation that can be made about the lists produced from each corpus is that the candidate collocates vary significantly. Of course, some candidates were suggested for both corpora: for instance, *change*, *extreme*, and *global* appear as candidate collocates for *climate* in both lists. However, other collocates only appear in the lists of 50 candidate collocates specific to a corpus: for example, *regional*, *system*, and *variable* were extracted from the EC corpus (again, for *climate*), whereas *cultural*, *economic*, and *favourable* were retrieved from the NC corpus.

Figure 1 presents the degree of overlap between the lists obtained for each lexical item (regardless of the window size and rank of candidates) in the two corpora.[8] Although the overlap varies between 54% and 19% depending on the lexical item considered, it averages at 35%, which is surprisingly low.

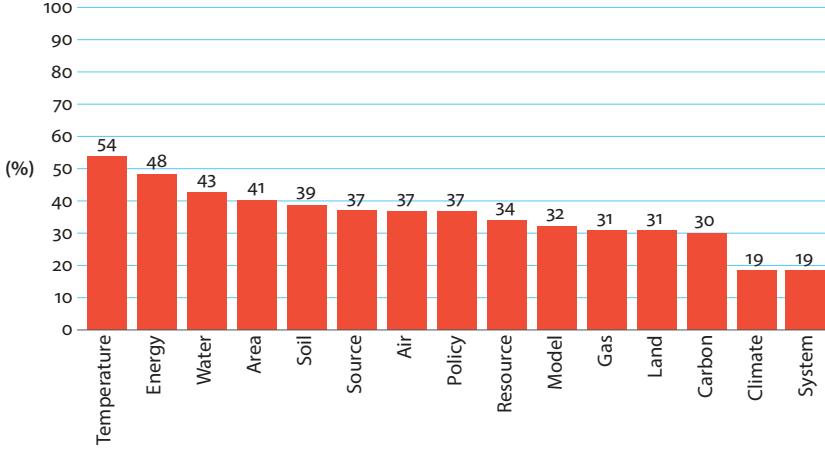


Figure 1. Overlap of collocates for 15 lexical items in the specialised and general corpora

We can suggest explanations for this low degree of overlap and will examine some of them more carefully in the following subsections. The nature of the corpus and the topics it deals with can explain why differences are observed between the two corpora. The general corpus contains texts that address a variety of topics, while the specialised corpus focuses on topics related to the environment and probably takes a more regular perspective on these topics. This inevitably has an effect on the use of terms and of their collocates. In a specialised corpus, lexical variety is likely to be reduced when compared to that of a general language corpus. Hence, lexical

8. At this stage, we did not take into account the rank of the candidate collocates. As we will see in the next section, the ranking itself introduces even more differences between lists.

items will appear much more frequently and are also likely to be used more often with a reduced set of collocates.

4.2 Rank of candidates

Section 4.1 showed that the degree of overlap between lists of collocates extracted from the environmental corpus (EC) and the general corpus (NC) is rather low. Taking into account the rank of candidate collocates increases discrepancies even more. Table 5 gives the ranks of collocates that are common to both corpora for *climate* (the figures for all lexical items appear in the appendix).The figures are given for a window size of 1.

Table 5. Ranks of common collocates retrieved for *climate*

Climate		
Collocates	**Rank_EC_F1**	**Rank_NC_F1**
change	1	1
global	2	5
sensitivity	5	17
extreme	14	30
warm	30	6
current	50	7

Table 6 gives the ranks of collocates that are common to both corpora for *temperature*. Again, the window size considered is 1. This time, 25 common collocates were extracted.

Table 6. Ranks of common collocates retrieved for *temperature*

Temperature		
Collocates	**Rank_EC_F1**	**Rank_NC_F1**
surface	1	6
average	3	15
increase	4	26
global	5	13
high	6	2
rise	7	7
change	9	34
maximum	10	23
minimum	11	30
low	12	3
ambient	14	8
gradient	17	16

(*continued*)

Table 6. (*continued*)

	Temperature	
Collocates	**Rank_EC_F1**	**Rank_NC_F1**
room	18	1
difference	21	38
range	22	12
sea-surface	23	43
extreme	24	31
water	27	4
variation	31	44
atmospheric	35	49
humidity	38	50
dependence	42	27
optimum	46	37
drop	47	10
elevated	48	29

As can be seen in Tables 5 and 6 as well as in the appendix, common collocates may appear at very different ranks in the general and specialised corpora. This added to the fact that a sizable number of collocates are specific to each corpus (from 46% for *temperature* to 81% for *system* and *climate*, see Figure 1) shows that there are large differences between corpora as far as collocational behavior is concerned.

A few collocates are shared by both corpora and given the same rank. In fact, for most keywords, collocates that were assigned high ranks (1 or 2) are the same in both corpora: *change* was assigned rank 1 both in the EC and in the NC for the lexical item *climate*; *greenhouse* was assigned rank 1 in both corpora for the item *gas*; *renewable* was given rank 1 in the EC and rank 2 in the NC for the item *energy*; *dioxide* was given rank 1 in both corpora for the item *carbon*.

However, this situation does not apply in all cases. Most collocates appear at very different ranks in the lists we analysed. In some cases, the NC seems to favour collocates that appear at a lower rank in the EC: for the lexical item *soil*, for instance, *fertile* was assigned the rank 3 in the NC, but 39 in the EC; *clay* was assigned the rank 8 in the NC, but 40 in the EC. Conversely, the EC gives priority to collocates that the NC ranks lower: for instance, *organic* (as a collocate for *soil*) was assigned rank 1 in the EC, but 29 in the NC; *emission* (as a collocate for *gas*) was assigned rank 2 in the EC, but 21 in the NC. These are but a few examples: many other differences can be observed (see the appendix).

These figures confirm a statement made at the end of Section 4.1 according to which the nature of the corpus and the topics it addresses has consequences on terms used and on their collocates. Even common collocates can be profiled quite differently in general and specialised corpora.

4.3 How collocates reveal specific meanings of items

One possible explanation for the low degree of overlap between collocates extracted from the EC and those retrieved from the NC is linked to the number of different meanings conveyed by lexical items in each corpus. Polysemy is much higher in the NC than in the EC and this has an impact on the diversity of collocates in lists associated with lexical items retrieved from each corpus.

Lexical items are likely to have a single meaning in a specialised corpus.[9] Conversely, if a lexical item is polysemic, most of its meanings are likely to appear in a general language corpus. This, of course, has an effect on the collocates produced from a specialised corpus and those retrieved from a general one since given sets of collocates are linked to specific meanings of lexical items.

Climate is an interesting case that seems to corroborate the observation made in the previous paragraph (Figure 2). Most collocates extracted from the EC are linked to *climate* defined as "the prevailing meteorological conditions of a region observed during a given period of time" (e.g. *change, cool, dry, variability*). In the lists of candidates produced from the NC, some collocates are indeed attached to *climate* defined as meteorological conditions (e.g. *change*, *cool*, *global*, *sunny*). However, others are linked to a different meaning, i.e. "the usual or most widespread mood or conditions in a place" (Merriam-Webster, 2016) (e.g. *economic*, *financial*, *intellectual*, *social*).

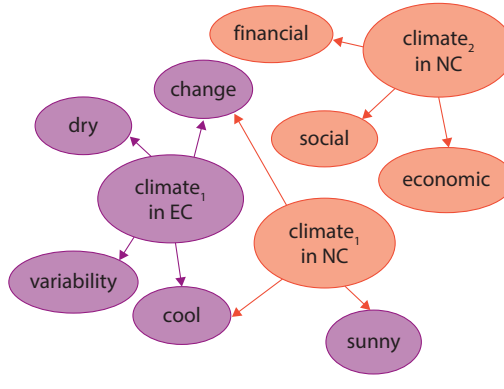


Figure 2. Collocates according to the meanings of *climate*

9. Although lexical items are likely to convey a single meaning in a specialised corpus, it is not always the case. For instance, the lexical unit *system* that we studied here is polysemic in the EC. In can be found in combinations such as *agricultural system* and *air-conditioning system*. In the first sequence, *system* refers to a set of rules, while it denotes an entity in the second.

Given the topically focused nature of the specialised corpus, it is also likely that collocates are semantically related. We identified in the lists of collocates retrieved from each corpus those collocates that are linked semantically. Relationships taken into account are synonymy and near synonymy (e.g. *prediction, projection, scenario*), and antonymy (e.g. *absorb, release*). We also considered other cases where terms shared a significant meaning intersection, such as morphological relations accompanied by semantic relationships (e.g. *forest*, *forested*, *forestry*), terms that belong to the same paradigm (e.g. *combustion, emission*).

Figure 3 shows how collocates retrieved for the term *source* can be grouped in semantic classes in both corpora.

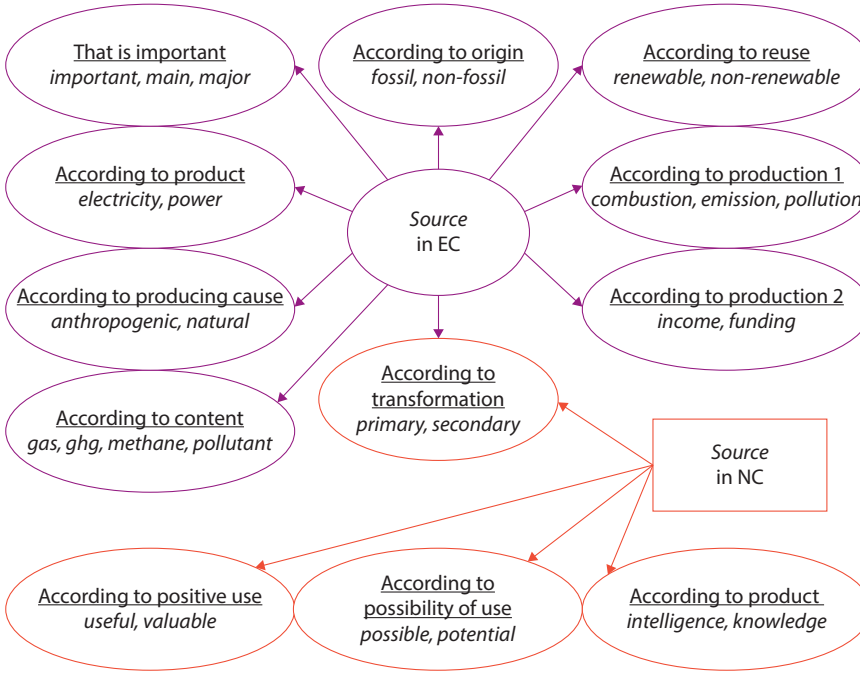


Figure 3. Some collocates of *source* retrieved from the NC and the EC grouped in classes

Some of the groups we identified highlight different perspectives on the meaning of lexical items or different meanings altogether. For instance, the groups "*intelligence:knowledge*" in the NC and "*income*:*funding*" in the EC, for the lexical item *source*, show that we are no longer dealing with *source* as "the point of origin of a substance or energy".

For the 15 lexical items analysed, we could observe more groupings (such as the ones for *source*) in the EC than in the NC. Figures in Table 7 give, for each lexical item, the proportion of collocates that could be grouped into semantic classes in

Table 7. Proportion of collocates that belong to a semantic class

Terms	EC	NC	Difference
1. climate	78%	42%	36%
2. energy	68%	38%	30%
3. water	64%	38%	26%
4. gas	66%	30%	36%
5. carbon	58%	20%	38%
6. area	58%	36%	22%
7. system	64%	36%	28%
8. temperature	62%	34%	28%
9. model	44%	28%	16%
10. soil	76%	46%	30%
11. land	76%	38%	38%
12. source	60%	28%	32%
13. air	54%	32%	22%
14. resource	62%	30%	32%
15. policy	62%	32%	30%
AVERAGE	**63%**	**34%**	**29%**

each corpus (here we considered all window sizes). They clearly show that more collocates are related semantically in the specialised corpus than in the general one.

These figures tend to confirm observations that were made previously. Lexical items are more likely to be polysemic in general corpora resulting in a higher diversity of collocates. Conversely, the perspective taken on items is much more focused in a specialised corpus and this is likely to result in semantic cohesiveness of units in general (lexical items as well as collocates).

5. Concluding remarks: Summary and guidelines for terminologists and lexicographers

In this contribution, we examined sets of collocates associated with 15 lexical items extracted from a specialised corpus (containing environmental texts) and a general one (a combination of the British National Corpus and the American National Corpus). The same method was used to extract collocates: 50 candidate collocates were retrieved automatically using three window sizes. Our comparison took into account: the overlap of lists of collocates, the ranks of shared collocates, polysemy of lexical items, and semantic relationships between the collocates retrieved.

As could be expected, differences were observed in lists of candidate collocates produced for each corpus. However, some differences were surprisingly high. First, the degree of overlap of collocates (for all three window sizes and regardless of the

rank at which collocates were extracted) averaged at 35%. Additionally, when considering ranks that were assigned to common collocates, although a small number of collocates were assigned close ranks, most of them appeared in very different parts of the lists retrieved from each corpus.

Collocates extracted are linked to different meanings of the lexical items chosen. Polysemy has an evident impact on the lists of candidates produced. Although some lexical items studied are polysemic in the specialised corpus, this was found to be a more prevailing characteristic of the general language corpus. Polysemy results in a higher variety of collocates retrieved from the general corpus. This was confirmed by looking at semantic relationships between collocates in lists from both corpora. For all the lexical items analysed, more semantic classes could be built from the specialised corpus.

Some of these differences can be explained by the nature of the corpora that provide different perspectives on the lexical units they contain. The topically focused nature of the specialised corpus has a number of linguistic consequences. First, lexical variety in the environment corpus is inevitably lower. Terms are less likely to convey multiple meanings in a specialised corpus and are more likely to combine recurrently with a smaller set of collocates. The higher variety of topics that are discussed in a general corpus will, on the other hand, result in a more pronounced linguistic diversity.

Our study leads to a series of observations regarding the collection of collocations in reference works:

1. One of the first steps taken by lexicographers using methods similar to the one we applied (automatically extract collocates from corpora) will probably be to distinguish multiple meanings key words may convey and separate collocates based on these distinctions. Terminologists, on the other hand, might only make meaning distinctions when necessary. Even for items that are polysemic, they might only consider one of their meanings since they will focus on specialised ones.
2. The low degree of overlap between lists extracted from the general corpus and the specialised one in addition to the different ranks assigned to common collocates suggest that lexicographers using general corpora such as the BNC or the ANC and terminologists referring to specialised corpora (such as the EC used in this study) are likely to select different collocates for inclusion in reference works. Some collocates might appear more important than others in different corpora even for lexical items that carry the same meaning.
3. Given the larger variety of collocates in the general corpus, lexicographers might need longer lists of candidates to ensure they cover most cases for the different meanings of lexical items. Although terminologists will probably also

use longer lists of candidates, they can already have a clear picture of the collocational behaviour of lexical items and start building classes of collocates with 50 candidates.

Although this study was carried out for a small sample of 15 lexical items and lists of 50 candidate collocates extracted with a specific statistical measure, we believe that it provides valuable insights for understanding the impact of the corpus on the collection of collocations. Overall, our study shows that corpora have a significant impact on this task and might influence what lexicographers and terminologists consider to be relevant collocations. Of course, more in-depth analyses are required to understand the extent to which these differences influence the compilation of reference works, but our contribution highlights the fact that they cannot be overlooked.

Acknowledgements

This research work is supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada. The authors would like to thank Gabriel Bernier-Colborne for his help in designing the Python program to extract collocations. We also extend our thanks to the reviewers and editors who made useful suggestions to clarify parts of our contribution.

Funding

Funded by Social Sciences and Humanities Research Council (SSHRC) of Canada.

References

Azoulay, D. (2017). Frame-Based Knowledge Representation Using Large Specialized Corpora. In *Proceedings of the AAAI Spring Symposium on Computational Construction Grammar and Natural Language Understanding*. Stanford University, CA.

Benson, N., Benson, E., & Ilson, R. (1986). *The BBI Combinatory Dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.bbi1(1st)

Binon, J., Verlinde, S., Van Dyck, J., & Bertels, A. (2000). *Dictionnaire d'apprentissage du français des affaires. Dictionnaire de compréhension et de production de la langue des affaires*. Paris: Didier.

Buendía, M., & Faber, P. (2014). Collocation dictionaries: a comparative analysis. *MonTi: Monografías de Traducción e Interpretación*, 6, 203–235. https://doi.org/10.6035/MonTI.2014.6.7

Cohen, B. (1986). *Lexique de cooccurrents. Bourse–conjuncture économique*. Montréal: Linguatech.
DiCoInfo. *Dictionnaire fondamental de l'informatique et de l'Internet*. (2016). http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi.
Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology*, 9(1), 99–117. https://doi.org/10.1075/term.9.1.06dro
Evert, S. (2004). *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. (Thesis presented at the University of Stuttgart, Germany).
Evert, S. (2008). Corpora and collocations. In A. Ludeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
Haussmann, F. J. (1979). Un dictionnaire des collocations est-il possible? *Travaux de linguistique et de littérature*, 17(1), 187–195.
Iordanskaja, L., & Mel'cuk, I. (2017). *Le mot dans le lexique et le mot dans la phrase*. Paris: Hermann.
L'Homme, M. C. (2000). Understanding Specialized Lexical Combinations. *Terminology*, 6(1), 89-110. https://doi.org/10.1075/term.6.1.06hom
L'Homme, M. C. (2009). A Methodology for Describing Collocations in a Specialized Dictionary. In S. Nielsen, & S. Tarp (Eds.), *Lexicography in the 21st Century In honour of Henning Bergenholtz* (pp. 237–256). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/tlrp.12.18hom
Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (Ed.), *Lexical Functions in Lexicography and Language Processing* (pp. 37–102). Amsterdam/Philadelphia: Benjamins. *Merriam-Webster Dictionary*. (2016). http://www.merriam-webster.com/dictionary/
Moon, R. (2015). Multiword Items. In J. Taylor (Ed.), *Handbook of the Word* (pp. 121–140). Oxford: Oxford University Press.
Merriam-Webster Dictionary. 2016. (http://www.merriam-webster.com/dictionary/).
Pecina, P. (2009). Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation*, 44(1–2), 137-158.
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44–49). Manchester, UK.

Appendix

1. climate	Rank_EC_F1	Rank_NC_F1
Change	1	1
Global	2	5
Sensitivity	5	17
Extreme	14	30
Warm	30	6
Current	50	7

2. energy	Rank_EC_F1	Rank_NC_F1
Renewable	1	2
efficiency	2	1
consumption	3	6
Source	4	5
Solar	5	7
Efficient	9	32
Demand	10	27
Supply	11	17
Conservation	12	4
Save	13	16
thermal	16	23
Kinetic	19	3
Production	20	43
Nuclear	21	8
Policy	22	9
Geothermal	23	19
Electrical	27	33
Resource	29	44
Alternative	36	34
total	50	24

3. water	Rank_EC_F1	Rank_NC_F1
Supply	3	2
Vapour	4	18
Quality	5	16
Surface	6	17
Drinking	7	4
Fresh	8	7
Hot	9	1
level	12	42
Drink	13	36
Heater	16	19
Deep	20	13
Pollution	25	29
Flow	32	23
Salt	33	22
Cold	35	3
Warm	36	14
Clean	39	33
Temperature	41	11
Shallow	43	9
Distilled	45	10
Pump	49	25

4. gas	Rank_EC_F1	Rank_NC_F1
Greenhouse	1	1
Emission	2	21
Flue	5	8
Pipeline	15	22
Oil	17	28
Exploration	22	37
Methane	26	13
chromatograph	28	45
Petroleum	32	35
Natural	39	2
chromatography	40	26

5. carbon	Rank_EC_F1	Rank_NC_F1
Dioxide	1	1
Organic	5	17
Monoxide	7	2
Sink	8	32
Tax	10	8
Sequester	12	46
Atmospheric	15	13
Emission	16	6
Taxis	27	30
Global	30	24
Isotope	41	42
Tetrachloride	43	4
Atom	46	3
Reduce	49	9

6. area	Rank_EC_F1	Rank_NC_F1
Urban	1	2
Rural	2	1
Coastal	4	33
Metropolitan	9	8
large	10	5
Designate	12	31
Surface	13	15
Catchment	16	3
Residential	17	4
Surround	18	14
Geographic	24	40
Geographical	25	7
Conservation	26	12
built-up	27	11

6. area	Rank_EC_F1	Rank_NC_F1
Subject	43	9
Remote	47	18
Sensitive	48	21
Populated	49	38

7. system	Rank_EC_F1	Rank_NC_F1
Management	2	3
Heating	7	22
Information	11	7
Solar	15	18
Rating	26	40
Nervous	30	2
Transport	31	26
Control	33	28
Immune	36	4

8. temperature	Rank_EC_F1	Rank_NC_F1
Surface	1	6
Average	3	15
Increase	4	26
Global	5	13
High	6	2
Rise	7	7
Change	9	34
Maximum	10	23
Minimum	11	30
Low	12	3
Ambient	14	8
Gradient	17	16
room	18	1
Difference	21	38
Range	22	12
sea-surface	23	43
Extreme	24	31
Water	27	4
variation	31	44
Atmospheric	35	49
Humidity	38	50
Dependence	42	27
Optimum	46	37
Drop	47	10
Elevated	48	29

9. model	Rank_EC_F1	Rank_NC_F1
Simulation	3	9
Computer	4	8
Mathematical	6	3
Numerical	8	46
Simple	9	20
Predict	12	28
Regression	16	2
Conceptual	19	23
Mental	20	39
Predictive	25	38
Economic	33	14
Statistical	40	13
System	42	27
Econometric	43	9

10. soil	Rank_EC_F1	Rank_NC_F1
Organic	1	29
Moisture	2	14
Erosion	3	1
Fertility	5	7
Type	9	21
Degradation	14	13
Surface	16	10
Water	19	18
Profile	21	32
Sandy	23	4
Forest	24	50
Sample	26	23
Conservation	32	2
Structure	33	11
Fertile	39	3
Clay	40	8
Nutrient	41	45
Condition	47	36

11. land	Rank_EC_F1	Rank_NC_F1
use	1	5
Agricultural	2	2
Arable	4	3
Area	5	37
Grazing	7	19

11. land	Rank_EC_F1	Rank_NC_F1
Forest	9	34
Tenure	13	16
Dry	14	13
Private	19	33
Reclamation	22	8
Irrigate	26	35
Mass	32	21
contaminated	33	15
Vacant	34	45
Adjacent	39	38

12. source	Rank_EC_F1	Rank_NC_F1
Energy	1	4
Renewable	4	10
Major	5	2
Datum	7	30
Main	8	1
Data	11	33
Primary	12	5
Funding	16	44
Important	19	6
Include	20	49
Food	26	19
Different	32	27
Principal	34	22
Secondary	35	16
Light	41	31
Additional	50	43

13. air	Rank_EC_F1	Rank_NC_F1
Quality	1	7
Pollution	2	1
Pollutant	3	17
Conditioning	4	3
Clean	6	14
Ambient	7	28
Conditioner	8	11
Warm	12	15
Cold	15	10
Travel	17	13
Hot	18	8
Pressure	19	42

13. air	Rank_EC_F1	Rank_NC_F1
Cool	20	35
Dry	21	36
Traffic	24	6
Transport	35	34
Fresh	37	2
Compressed	40	18

14. resource	Rank_EC_F1	Rank_NC_F1
Natural	1	1
Water	2	44
Management	3	9
Renewable	5	19
Financial	8	3
Mineral	9	17
Energy	10	30
Human	16	7
non-renewable	17	21
Genetic	19	43
Use	20	32
Scarce	28	4
Limited	33	6
Physical	40	40
Allocation	43	5
Additional	44	10
Valuable	49	12

15. policy	Rank_EC_F1	Rank_NC_F1
Maker	1	9
Environmental	2	21
Energy	4	23
Public	7	6
Planning	8	32
Implication	11	34
Development	12	47
National	13	24
Statement	17	25
Objective	22	39
Agricultural	23	11
Making	24	36
Change	27	45
Issue	42	29
Initiative	43	41
Transport	46	37
Document	48	33

Résumé

Les ressources générales et spécialisées accordent une place de plus en plus importante aux collocations. Les méthodologies pour les recueillir reposent principalement sur des corpus. Toutefois, les lexicographes et les terminologues font appel à des corpus de naturedifférente et dans lesquels les combinaisons sont susceptibles d'obéir à des règles spécifiques. Cette contribution présente une analyse comparative du comportement collocationnel de 15 formes lexicales apparaissant dans un corpus général et un corpus spécialisé portant sur l'environnement. Nous avons extrait automatiquement de longues listes de collocations (trois ensembles de 50 collocations) pour chaque item lexical de chacun des deux corpus. Nous observons différentes facettes du comportement collocationnel: polysémie des formes lexicales, caractéristiques des collocatifs (convergence, rangs et classes sémantiques des collocatifs, etc.). L'objectif est d'attirer l'attention des terminologues et des lexicographes sur des aspects particuliers du comportement des collocations dans des corpus de nature différente.

Mots clés: Collocations, terminologie, lexicographie, corpus spécialisé, corpus général, classe sémantique

What matters more: The size of the corpora or their quality?

The case of automatic translation of multiword expressions using comparable corpora

Ruslan Mitkov and Shiva Taslimipoor
University of Wolverhampton

This study investigates (and compares) the impact of the size and the similarity/quality of comparable corpora on the specific task of extracting translation equivalents of verb-noun collocations from such corpora. The comprehensive evaluation of different configurations of English and Spanish corpora sheds some light on the more general and perennial question: what matters more – the quantity or quality of corpora?

Keywords: multiword expressions, automatic translation, comparable corpora, size of corpora, vector representations

1. Rationale

Parallel corpora are the natural and most obvious choice of data to be used in Machine Translation and other multilingual NLP applications. Unfortunately, parallel corpora are not widely available and do not cover all domains. An alternative and more promising approach would be to use comparable corpora as these corpora can be compiled from the web in a relatively straightforward way, making use of available purpose-built tools.

Due the scarcity of parallel corpora, comparable corpora are now increasingly used as an alternative resource in a number of multilingual applications which include but are not limited to Machine Translation (Smith et al., 2010; Rapp et al., 2016), word translation (Rapp, 1999; Gaussier et al., 2004; Gamallo and Pichel, 2007; Pekar et al., 2008; Vulić and Moens, 2012), term extraction (Fung and McKeown, 1997; Daille and Morin, 2005; Saralegi et al., 2008), bilingual document similarity (Sharoff et al., 2015; Jagarlamundi and Daumé, 2010), crosslingual coreference resolution (Green at al., 2011), Name entity transliteration (Udupa et al., 2008;

https://doi.org/10.1075/ivitra.24.09mit

Klementiev and Roth, 2006), automatic identification of cognates and false friends (Mitkov et al., 2008), testing the validity of translation universals (Corpas et al., 2008) and tracking language change (Štajner et al., 2013).

The size of the corpora, whether monolingual, parallel or comparable, is often regarded as a decisive factor for the performance of NLP tasks or applications the expectation being that the larger the corpora used, the better the performance of the tasks or applications exploiting them. However, to the best of our knowledge, no study has addressed the question of whether size is the decisive factor that always matters. It would be noteworthy to establish whether the size of the corpora is an important factor irrespective of their quality and whether even sufficiently large data of inferior quality could deliver better results than smaller data of better quality. This study seeks for the first time to shine a light on this fundamental question. In order to answer this question, we studied the task of automatic translation of multiword expressions (MWEs) using comparable corpora, experimented with corpora of different sizes and quality and compared the results. In this project, our premise is that quality of comparable corpora is directly related to their comparability: the more comparable they are, the better their quality is deemed to be. For the purpose of this study we operationalise the concept of comparability and quality through similarity: the more similar two corpora are, the more comparable they are and in other words the comparable corpus built on these corpora is of higher quality.

The rest of the paper is structured as follows. In Section 2 we describe our original approach to finding translation equivalents from comparable corpora. In Section 3 we describe the data used in our experiments, provide detail on the experiments and outline the preparation of the gold standard used in the evaluation. Section 4 presents the evaluation results, but most importantly, discusses these results from the point of view of the quality of the corpora and their size. It compares and discusses the performance of the task of automatic translation of MWEs depending on the different size and quality of the evaluation data thus seeking to answer the question whether it is the size or the quality of the corpora the one which matters more.

2. Our methodology for translating multiword expressions

We have developed an original general methodology for extracting and translating MWEs from any pairs of languages. This methodology represents a knowledge-poor approach and does not use any bilingual grammar nor does it depend on translation resources such as dictionaries, translation memories or parallel corpora[1] which can

1. We only use a list of loosely aligned word pairs automatically extracted from parallel corpora.

be time-consuming to develop or difficult to acquire, being expensive or proprietary. The only supporting information comes from comparable corpora, inexpensively compiled. The first proof-of-concept stage of this project covers English and Spanish and focuses on a particular subclass of MWEs: verb-noun expressions (collocations) such as *take advantage*, *make sense*, *prestar atención* and *tener derecho*.

The translation of multiword expressions is viewed as a two-stage process. The first stage is the extraction of MWEs in each of the languages; the second stage is a matching procedure for the extracted MWEs in each language which proposes the translation equivalents. In this paper we focus on the second stage: the proposal of translation equivalents. The automatic extraction of MWEs has already been described in Taslimipoor et al. (2017).

In order to identify translation equivalents for collocations, we use an extended version of the word embedding approach (Mikolov et al., 2013b). Following the distributional similarity premise (Haris 1954), the method draws on the patterns of word cooccurrences within a small window to predict similarities among words. The idea is to represent each word as a dense vector by way of neural-network modelling.

The new word embedding approach, also known as *word2vec*, learns low-dimensional word vectors from raw (monolingual) text. We adapt the model to the task of automatic translation of MWEs by defining bilingual contexts derived from a core set of known translation pairs. Specifically, we define context in a bilingual space by pairing words from the two languages with the help of an automatic word alignment tool.

A standard word2vec model which employs a window of size *k* around a target word *w* produces *2k* context words: *k* words before *w* and *k* words after *w*. Following Taslimipoor et al. (2016), we use this standard approach to model context but consider only specific words from a predefined bilingual lexicon rather than all the words in the context window. More specifically, we focus on nouns. Nouns are less ambiguous and most of the previous word vector representations have focused on nouns for evaluating word vectors (Mikolov et al., 2013a, Zhang et al., 2017). In this study we use a core lexicon of paired English-Spanish nouns as our bilingual context terms.

Another difference in our implementation of vector representation is that we construct vectors for verb + noun expressions (rather than single words). We consider translations of verb + noun expressions to be either verbs or verb + noun expressions. To this end we extract context words around expressions and transfer our data to expression-context pairs.

The generalised word2vec model (called *word2vecf*) (Levy and Goldberg, 2014) can then be trained on expression-context pairs with the vectors of the two languages which are defined over the same space and which can be compared via cosine similarity.

Given a target collocation *s* from the source language (e.g. Spanish), our goal is to find the best translation equivalent, *t*, in the target language (e.g. English). For each collocation *s*, we examine all target language documents which are paired with the source language documents containing *s*. All verb combinations appearing in these documents are examined for their similarities to *s*. Verb combinations that we consider are of the forms: single verb, verb + noun (bigram combination) and verb + something + noun (trigram combination) as possible translations of a verb + noun combination *s*.

The candidate with the highest vector similarity to the collocation *s* is selected as its translation.

3. Data and experiments

3.1 Comparable corpora

The main objective of this study is to compare the performance of extracting translation equivalents over various configurations of comparable corpora. We experiment with corpora of different size and quality in order to establish their impact on the performance.

Two comparable corpora are used for our experiments. One is a collection of aligned documents from English and Spanish Wikipedia.[2] It includes around 673,000 document pairs with 456.6 million English and 316.2 million Spanish tokens. The documents are aligned one by one using the language links in Wikipedia pages; therefore, they are accurately aligned based on their contents and regarded as high-quality corpora in terms of comparability.

The other comparable corpus is compiled from various news sources on the Web. We collected news feeds from a variety of news sources in both Spanish and English from July 2015 to February 2016. The ACCURAT toolkit (Pinnis et al., 2012; Skadina et al., 2012; Su and Babych, 2012a) was employed to automatically compile the comparable corpora for this study. News articles from the web from the RSS feeds of ABC news, Yahoo news, CNN news, Sport news and Euronews in both Spanish and English were downloaded. In addition, RSS feeds of Ultimahora and Europapress for Spanish were also added to ensure the Spanish data is more balanced. The downloaded data from online news (1.5 GB) consisted of 200,000 documents in English and 112,000 documents in Spanish. These documents were classified with a view to building a corpus of English texts and another of Spanish texts which are comparable. Each monolingual corpus was designed to feature

2. http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

documents paired with documents in the other language in terms of the similarity between them. For the purpose of this study we operationalised comparability via similarity. Similarity was automatically computed with the help of the ACCURAT tool, *DictMetric*, which compares documents by employing cosine similarity. More specifically, DictMetric converts text into index vectors and then computes a 'comparability score' of document pairs by applying a cosine similarity measure on the index vectors. In order to measure the comparability of two documents in different languages, one of the documents is translated into the language of the other. DictMetric translates non-English texts into English by using lexical mapping from the available GIZA++ based bilingual dictionaries.

By varying comparability thresholds, we generated comparable corpora of different size and quality. Recall that quality in terms of comparability is modelled through similarity in this study. It was expected that higher comparability thresholds would result in more accurate alignments between documents but also in the generation of smaller corpora. To this end, we set the comparability threshold to five values from 0.5 to 0.1. The number of paired documents in each of these five sets are reported in Table 1.

Table 1. The number of paired documents in the news comparable corpora in both directions

	CC 0.5	**CC 0.4**	**CC 0.3**	**CC 0.2**	**CC 0.1**
es-en	6,544	15,417	31,856	63,703	114,313
en-es	9,337	22,123	46,798	94,896	195,175

3.2 Data

We experimented with the most productive and widely used verbs in verb + noun combinations. To this end, eight highly frequent verbs occurring before nouns in English and six such verbs in Spanish were identified[3] and all such occurrences were extracted from our paired documents / comparable corpora. We only selected occurrences with frequencies higher than 3 in the aligned documents of similarity threshold 0.5; this process resulted in a list of 220 English and 210 Spanish verb + noun collocations.

3. English verbs: *take*, *have*, *make*, *give*, *get*, *find*, *pay*, *lose*; Spanish verbs: *tener*, *dar*, *hacer*, *formar*, *tomar*, *poner*.

3.3 Vector representations

For learning vectors, the monolingual English and Spanish components of the Europarl corpus and the English and Spanish components of our News corpora were used to obtain co-occurrence statistics. All English and Spanish verb combinations (unigrams, bigrams, trigrams) were indexed according to their occurrences with the context word pairs. Specifically, words that exist in our bilingual context pairs are identified within a context window of length 10 around a target expression. As a result, expression-context pairs are generated.

Two approaches were experimented with to construct vectors:

bi-word2vec: bi-word2vec is our proposed approach for vector representation of expressions and is detailed in Section 2. In this experiment the *word2vecf* software is used to train vectors on the indexed corpora.
co-occurrence Jaccard: given an expression from the source language and another from the target language, their similarity is measured by comparing their corresponding sets of (bilingual) context pairs (using a context window of size 10). More specifically, the Jaccard similarity coefficient is employed to measure this similarity.

Experimenting with both types of vectors, we apply our methodology to find translations for collocations in both directions: Spanish to English, and English to Spanish.

Note that we focus on finding translations for verb + noun combinations. We assume that for most such expressions, the translation equivalent is either a verb (unigram), a verb + noun (bigram), or a verb + noun with an intervening word, such as a determiner or an adjective (trigram). For every expression from the source language, our goal is to find the five most similar verb or verb + noun combinations (bigram or trigram) in the target language.

3.4 Gold standard

For the purpose of the evaluation (see the discussion of the evaluation results in Section 4), we prepared a list of correct translations for the candidate collocations from online dictionaries such as Wordrefernce, Linguee, Spanish Central and Reverso Dictionary. We also asked a human expert to examine and rate the top-ranked translations of four sample result lists and mark the correct translations. We extended the gold standard list with the correct translations marked by the annotator. Note that our approach might identify correct translations which are not on the list.

4. Comparable corpora and translation of mwes: Size vs. quality

This is the first study seeking to establish the impact of different thresholds which control the quality of the selected documents in terms of comparability and which also lead to the generation of different sizes of paired documents. The threshold values are directly proportional to the quality in terms of comparability and inversely proportional in terms of size. A higher comparability threshold implies better quality but also means smaller corpora. Lower comparability thresholds generate larger corpora of inferior quality. The performance of the task of finding translation equivalents has been evaluated by applying the two distributional similarity approaches on the five groupings of paired documents (referred to as CC0.5, CC0.4, CC0.3, CC0.2, CC0.1).

Note that we use a similarity measure to rank the candidate translations of each expression. By setting different threshold values for this similarity, we obtained ranked lists of varying sizes. The higher this threshold, the smaller the number of the resulting translation candidates, and hence the higher the number of expressions for which we may not have any good translations. In other words, we trade accuracy for coverage. In our experiments we set the similarity thresholds to different values in order to measure accuracy for three various degrees of coverage (20%, 50% and 80%). These different configurations offer a meaningful picture of the overall performance of each method on each of the comparable corpora.

Table 2 displays the accuracy and coverage values for finding translations of both Spanish (*es*) and English (*en*) expressions. As illustrated in Table 2, for Spanish expressions the choice of lower comparability threshold yields better results provided that the threshold is not lower than 0.2. The larger size of the corpora 'matters' up to that point and as long as the corpora exhibit minimal quality (e.g. comparability 0.2). It can be seen that the accuracies drop when the threshold is set to 0.1. This trend holds for both distributional similarity approaches. However, in the case of bi-word2vec the optimal threshold for comparability is 0.3 for lower coverages. For English expressions in the case of *biword2vec*, the accuracy drops when we use CC0.2 rather than CC0.3, but not when we use *co-occurrence Jaccard*. A general conclusion from these results is that size indeed matters and the larger the size, the better the performance as long as the quality is above a minimal comparability threshold.

The performance in translation equivalent identification is further evaluated on the accurately aligned Wikipedia comparable corpora and is reported in Table 3. In terms of accuracy, *bi-word2vec* does better than the simple *co-occurrence Jaccard* at establishing the translations of Spanish expressions (es). On the other hand, the simple *co-occurrence Jaccard* fares better at finding the translations of English expressions (en).

Table 2. The accuracies compared on different sets of comparable corpora

coverage		20%		50%		80%	
		es	en	es	en	es	en
Co-occurrence Jaccard	CC 0.5	37%	36%	15%	15%	10%	9%
	CC 0.4	42%	52%	24%	24%	13%	14%
	CC 0.3	63%	65%	29%	31%	16%	15%
	CC 0.2	**64%**	**67%**	**40%**	35%	**20%**	20%
	CC 0.1	45%	58%	38%	**40%**	20%	**23%**
Bi-word2vec	CC 0.5	38%	28%	25%	17%	11%	14%
	CC 0.4	32%	34%	34%	23%	24%	18%
	CC 0.3	**37%**	**48%**	**36%**	**31%**	28%	24%
	CC 0.2	37%	44%	34%	30%	**31%**	25%
	CC 0.1	31%	43%	24%	29%	27%	**27%**

Table 3. Accuracies (%) of finding translations from aligned wikipedia comparable corpora

coverage	20%		50%		80%	
	es	en	es	en	es	en
Co-occurrence Jaccard	64%	84%	58%	68%	40%	46%
Bi-word2vec	74%	78%	63%	55%	48%	43%

The Wikipedia-aligned corpus is almost seven times bigger than our news corpora. To compare these two, we focus on a sample of the Wikipedia documents which are of comparable size with our news corpora (specifically, on 96,193 document pairs). Figure 1 shows that the translation results from the Wikipedia-aligned corpora (*wikiJ* and *wikiW*) significantly outperform the results from our automatically-paired comparable corpora[4] (*ccJ* and *ccW*) for both English and Spanish expressions. As the Wikipedia-aligned corpus is deemed to be of better quality in terms of comparability, here the quality makes it point – that it does matter.

Finally, according to the obtained results, the simple *co-occurrence Jaccard* approach performs very well at finding translations for English expressions. It appears that this approach delivers promising results for highly frequent expressions (e.g. *have time*) for which the *biword2vec* approach proposes semantically-related but incorrect translations (e.g. *tomar tiempo*, *haber tiempo*, *pasar mucho tiempo*).

4. We experiment with a sample of the news documents which are paired with similarity threshold 0.2 and return the best results of all samples from the comparable corpora.

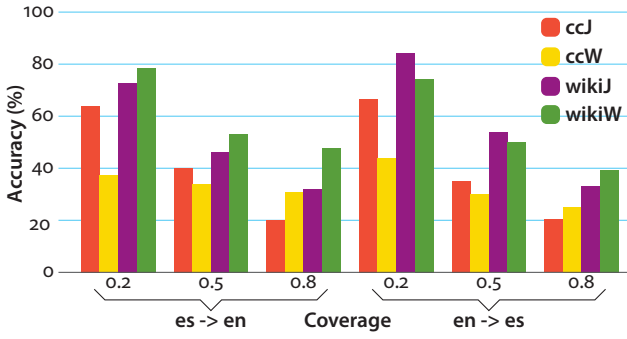


Figure 1. Accuracies of translation equivalents using: CCJ (co-occurrence Jaccard on our comparable corpora), ccW (bi-word2vec on our comparable corpora), wikiJ (co-occurrence jaccard on wikipedia comparable corpora), wikiW (bi-wordevec on wikipedia comparable corpora)

5. Conclusion

To the best of our knowledge, this is the first study which seeks to answer the fundamental question, 'What matters– the size or the quality of comparable corpora?'. We study the particular task of automatic translation from comparable and show that the employment of larger aligned corpora results in identifying translation equivalents with higher accuracy. At the same time, the importance of the quality of the corpora cannot be underestimated. If the quality of the comparable corpora is under a specific 'minimal' threshold, the performance deteriorates. Therefore, we can conclude that both quantity and quality matter with comparable corpora of larger size delivering better performance as long as comparable corpora are of 'minimal quality'.

References

Corpas Pastor, G., Mitkov, R., Afzal, N., & Pekar, V. (2008). Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the AMTA' 2008 conference* (pp. 75–81). Honolulu, Hawaii.

Daille, B., & Morin, E. (2005). French-English terminology extraction from comparable corpora. In *Proceedings of 2nd International Joint Conference on Natural Language Processing* (pp. 707–718).

Fung, P., & McKeown, K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora* (pp. 192–202).

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., & Hervé, D. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 527–534). Barcelona, Spain.

Gamallo, P., & Pichel, J. R. (2007). Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Procesamiento del Lenguaje Natural*, 39, 241–248.

Green, S., Nicholas, A., Gormley, M., Dredze, M., & Manning, C. D. (2011). *Cross-lingual Coreference Resolution: A New Task for Multilingual Comparable Corpora*. Technical Report 6, Johns Hopkins University.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(23), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Jagarlamudi, J., & Hal, D. III. (2010). Extracting multilingual topics from unaligned comparable corpora. In *ECIR* (pp. 444–456).

Klementiev, A., & Roth, D. (2006). Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-06)* (pp. 817–824), Sydney, Australia.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 302–308). Baltimore, Maryland.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mitkov, R., Pekar, V., Blagoev, D., & Mulloni, A. (2008). Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1), 29–53. https://doi.org/10.1007/s10590-008-9034-5

Pekar, V., Mitkov, R., Blagoev, D., & Mulloni, A. (2008). Finding Translations for Low-Frequency Words in Comparable Corpora. *Machine Translation*, 20(4), 247–266. https://doi.org/10.1007/s10590-007-9029-7

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., & Bogdan B. (2012). ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 91–96).

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 519–526). https://doi.org/10.3115/1034678.1034756

Rapp, R., Sharoff, S., & Zweigenbaum, P. (Eds.) (2016). Special Issue: Machine Translation using comparable corpora. *Journal of Natural Language Engineering*, 22(4). https://doi.org/10.1017/S1351324916000115

Saralegi, X., San Vicente, I., & Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of LREC 2008 Workshop of Building and Using Comparable Corpora* (pp. 27–32). Basque Country.

Sharoff, S., Zweigenbaum, P., & Rapp, R. (2015). BUCC shared task: cross-language document similarity. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora* (pp. 74–78). Beijing, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3411

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Lestari Paramita, M., & Pinnis, M. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)* (pp. 438–445).

Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403–411). Los Angeles, CA: Association for Computational Linguistics.

Su, F., & Bogdan, B. (2012). Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi)Parallel Translation Equivalents. In *Proceedings of the EACL'12 Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)* (pp. 10–19).

Štajner, S., Mitkov, R., & Leech, G. (2013). Natural Language Processing Methodology for Tracking Diachronic Changes in the 20th Century English Language. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 1(1), 71–112. https://doi.org/10.1558/jrds/720788885881

Taslimipoor, S., Mitkov, R., Corpas Pastor, G., & Fazly, A. (2016). Bilingual Contexts from Comparable Corpora to Mine for Translations of Collocations. In *Lecture Notes in Computational Linguistics. Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2016)*. Springer.

Taslimipoor, S., Rohanian, O., Mitkov, R., & Fazly, A. (2017). Investigating the opacity of verb-noun multiword expression usages in context. In *Proceedings of the 13th workshop on multiword expressions (MWE 2017)* (pp. 133–138). Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1718

Udupa, R., Saravanan, K., Kumaran, A., & Jagarlamudi, J. (2008). Mining Named Entity Transliteration Equivalents from Comparable Corpora. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1423–1424).

Vulić, I., & Moens, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 449–459).

Zhang, M., Peng, H., Liu, Y., Luan, H., & Sun, M. (2017). Bilingual Lexicon Induction From Non-Parallel Data With Minimal Supervision. In *Proceedings of the Thirty-First {AAAI} Conference on Artificial Intelligence* (pp. 3379–3385). San Francisco, CA.

Statistical significance for measures of collocation strength (WP3)

Michael P. Oakes
University of Wolverhampton

Of the commonly-used measures of lexical association or collocation strength, only some directly relate to statistical significance: the t-score, chi-squared, log-likelihood, the z-score and Fisher's exact test. We describe each of these tests, and also describe a computer simulation by which we can derive confidence limits, and hence the statistical significance, of any measure of lexical association which is derived from the contingency table. We illustrate this approach using pointwise mutual information (PMI). We also describe how the Poisson distribution enables us to find the statistical significance of the raw frequency with which a collocation is found. We compare all these methods using collocates of "take", namely "take up", "take place", "take advantage" and "take stock".

Keywords: collocation strength, statistical significance, Monte Carlo Methods, Poisson Distribution

1. Introduction

Following the seminal work of Church and Hanks (1989), a large number of measures of collocation strength (or "lexical association measures") have now been developed. The most comprehensive summary of these is given in Pecina (2008). Many of these measures are derived from the contingency table, which consists of four values: (a) is the number of times a pair of words such as "doctors" and "dentists" appear within the same window in a corpus, where a window is an arbitrary span of words, typically five, within which both words must occur; (b) is the number of times the first word occurs in the corpus but is not accompanied by the second in the same window; (c) is the number of times the second word appears in the corpus without the first being in the same window; and (d) is the number of contexts in the corpus in which neither word occurs, and we also need to know N, the total number of contexts in the corpus. An example of a contingency table is given in Table 1. This is a 2 by 2 table, where the cells are labelled 'a' to 'd'.

https://doi.org/10.1075/ivitra.24.100ak

Table 1. The Contingency table

	Word 2 present	**Word 2 absent**
Word 1 present	a	b
Word 1 absent	c	d

'a' is the number of "windows" (such as within a span of five words, or occupying immediately adjacent positions) in which word 1 and word 2 both occur; 'b' is the number of windows where word 1 is found, but word 2 is absent; 'c' is the number of windows where word 1 is absent, but word 2 is found; and 'd' is the number of windows where neither word is found.

Statistical significance measures how likely we would have been to achieve a value (such as the PMI value of 11.3 obtained by Church and Hanks for "honorary" and "doctor", which co-occurred 12 times in the AP corpus of 15 million words) had there been no real association between the two words, i.e. they were randomly scattered throughout the corpus. If the value of statistical significance (known as the p-value) is less than 0.05 (or 5%), then we can be 100% – 5% = 95% confident that the words "honorary" and "doctor" are not randomly distributed, but where one occurs, there is genuinely a tendency to find the other.

2. The chi-squared test (X^2)

In order to describe the chi-squared test, we will use the example of the word pair "if only" (in that order, with no intervening words) as it occurs in the LOB corpus. This word pair is found 21 times so cell 'a' in the contingency table is given the value 21. The total frequency of "if" is 2,479. Since on 21 of these occasions it forms part of the collocation "if only", the word "if" must occur without "only" on the other 2,479–21 = 2,458 occasions. Similarly the total frequency of "only" is 1,815, so in the LOB corpus it is found without "if" preceding it on 1,815–21 = 1,794 occasions. Finally, there are a million words in LOB, so there are 999,999 adjacent overlapping word pairs, of which 999,999 – (21 + 2,458 + 1,794) = 995,705 contain neither the word "if" in the first position nor the word "only" in the second position. These values are placed in the contingency table as shown in Table 2:

Table 2. Contingency table for the collocation "if only"

	"only" present in position 2	**"only" absent in position 2**
"if" present in position 1	21	2,458
"if" absent in position 1	1,794	995,726

For the chi-squared test, the contingency table is extended to show the totals of each row and each column. The grand total of all four values, N, is the total number of contexts or positions in which the pair of adjacent words could fit.

Table 3. Row, column and grand totals for the contingency table

	"only" present in position 2	**"only" absent in position 2**	**Row total**
"if" present in position 1	21	2,458	2,479
"if" absent in position 1	1.794	995,726	997,520
	1,815	998,184	Grand total = 999,999

We call the counts in the original contingency table "observed" values, because these are the counts we actually see in the corpus. These must now be compared with a corresponding set of "expected" values, which are the counts we would expect to find were there no particular association between the word "if" in the first position of a bigram and the word "only" in the second. These expected values are calculated for each cell of the table using the formula:

Expected Counts = (Row Total × Column Total) / Grand Total.

Table 4. Expected values for the collocation "if only"

	"only" present in position 2	**"only" absent in position 2**
"if" present in position 1	4.5	2,474.5
"if" absent in position 1	1,810.5	995,709.5

For example, the value for cell a is (1,815 × 2,479) / 999,999 ≈ 4.5.

To determine whether the number of occurrences of the collocation "if only" is higher than we would expect if there were no particular affinity between the two words, we work out the differences between the observed and expected values. If these differences are small, there is probably no real affinity between the words of the pair, but if they are large, we may have found a statistically significant collocation. The differences for each cell are first squared, and then divided by the expected number of counts. For example, for cell 'a', its contribution to the overall chi-squared value is $(21–4.5)^2 / 4.5 = 60.5$. This value and the contributions of the other three cells are shown below:

Table 5. Individual contributions to the overall chi-squared value by each of the cells in the contingency table

	"only" present in position 2	"only" absent in position 2
"if" present in position 1	60.5	0.1
"if" absent in position 1	0.2	0.0

The square roots of these contributions to the overall chi-squared value are called Pearson residuals. Although we do not need them directly for the calculation of our final chi-squared value, we will meet them again in the section on the z-score.

To find the overall chi-squared value, we simply add together the individual contributions of each of the cells, which for "if" and "only" in the LOB corpus gives 60.5 + 0.1 + 0.2 + 0 = 60.8. The statistical significance or p-value corresponding to a chi-squared value may be looked up in a table which is given as an appendix in most statistics textbooks. To do this, we need to know the number of "degrees of freedom", which is the number of rows in the contingency table minus one, times the number of columns minus one. For a 2 × 2 table, this is (2–1) × (2–1) = 1. In this case, the p-value is extremely small, much less than the arbitrary cut-off point of 0.05, so we can say that "if only" is a statistically significant collocation.

All the steps described in this section can be calculated on the computer using the R statistical programming package. Firstly we "bind" together the numbers (the observed values) in our original contingency table:

```
con = cbind(c(21,1794),c(2458, 995726))
```

and then perform all the above steps with a single command:

```
chisq.test(con)
```

This gives the following output:

```
Pearson's Chi-squared test with Yates' continuity correction
data: con
X-squared = 57.146, df = 1, p-value = 4.046e-14
Warning message:
In chisq.test(con) : Chi-squared approximation may be incorrect
```

In the third line we see the overall chi-squared value of 57.146 and the corresponding p-value which is vanishingly small. Strictly speaking, for 2 × 2 contingency tables we should employ Yates' correction, which is to reduce the difference between the observed and expected value by 0.5 if it is positive, and increase it by 0.5 if it is negative, which is why the value calculated by R is slightly less than the one we calculated by hand. The warning message is given because strictly speaking (see Section 3), all four expected values should be at least 5. However, since only one value is less than 5, and that one is very close to 5, we are probably safe in using the test.

We can store the results of the test in a variable called output as follows:

```
output = chisq.test(table)
```

and then view the tables generated at each step of the test. The expected values may be seen by the command:

```
output$expected
```

The Pearson residuals are shown by:

```
output$residuals
```

and the individual contributions by each of the cells are shown by:

```
output$residuals^2
```

The disadvantage of using the chi-squared test (and other tests which require the calculation of observed and expected values) to find collocations is that the very large values in cell 'd' (because the majority of positions in the corpus contain neither of the two constituent words) means that the expected value of cell 'a' tends to be very much less than the observed value of cell 'a'. Thus the expected value of cell 'a' will only be 5 or more for high-frequency collocations.

3. The log-likelihood test (G^2)

An alternative to Pearson's chi-squared test is the log-likelihood test, sometimes called 'G^2'. It is given by the formula

$$G^2 = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

Where O_i and E_i are the observed and expected values for each cell of the contingency table in turn, calculated in exactly the same way as for the Pearson chi-squared test. Both the chi-squared test and the log-likelihood test have the disadvantage that they cannot be used when the cell frequencies are very small. In his section on the "effect of small samples on X^2 and G^2", Agresti (2002, p. 396) reports that Koehler (1986), Koehler and Larntz (1980) and Larntz (1978) showed that X^2 applies with smaller sample sizes and more sparse tables than G^2. Paul Rayson has created an online log-likelihood calculator, available at http://ucrel.lancs.ac.uk/llwizard.html.

We can also calculate log-likelihood in R, first creating the contingency table and running the chi-squared test as we did before:

```
> con = cbind(c(21,1794),c(2458, 995726))
> output = chisq.test(con)
```

The observed (o) and expected (e) frequencies are the same for both tests, so we calculate:

```
> o = output$obs
> e = output$exp
> g2 = 2 * sum(o * log(o/e))
> g2
[1] 31.96456
```

The resulting value of G^2 is about 31.96. To convert a G^2 value (or a chi-squared value) to a p-value, we can use the command:

```
>p = 1 – pchisq(g2, df)
```

As was the case for the chi-squared test, the degrees of freedom (df) for a 2 × 2 contingency table = 1.

4. Fisher's exact test

Fisher's exact test arose from an informal experiment to test whether a lady could taste whether a cup of tea had been prepared by pouring the milk first or the tea first (Agresti, 2002, p. 92). As we have seen, there are problems when using the chi-squared test and the log-likelihood test with small expected values. For tables with small expected frequencies, Fisher's exact test can be used.

Imagine our corpus analysis reveals the following co-occurrence pattern between "kith" and "kin". We found three occasions where the two words appeared close together (so $a = 3$), no occasions where "kith" was found without the word "kin" (so $b = 0$), just one occasion where "kin" was found without "kith" (so $c = 1$), and six places in the corpus where both words were absent (so $d = 6$). This gives the following contingency table:

Table 6. Imaginary contingency table for "kith and kin"

	"kin" present	**"kin" absent**
"kith" present	$a = 3$	$b = 0$
"kith" absent	$c = 1$	$d = 6$

These values are also found in the bottom row of Table 7. The other values in Table 7 are those in all other possible contingency tables where the row totals and the column totals are the same: in this case (a + b) = 3, (a + c) = 4, (b + d) = 6, and (c + d) = 7. Although we do not have to do this directly, we could count for each contingency table the number of ways (combinations) of sharing out ten (the grand

total) items among the four cells so that each received the stipulated number. If we do this for every contingency table in Table 7, and find the total number of combinations over all four tables, the probability of obtaining an individual table would be the number of combinations producing that table, divided by the total number of combinations over all the tables. If the null hypothesis that there is no association between "kith" and "kin" (no particular tendency for them to appear together), then the probability of each of these 2 × 2 tables which have the same column and row totals can be found using the following formula (Altman, 1991: 256):

$$\frac{(a+b)!(a+c)!(b+d)!(c+d)!}{N!a!b!c!d!}$$

The '!' or "factorial" symbol means that we multiply together the number itself, the number below, the number below that and so on down to 1. For example, 5! = 5 × 4 × 3 × 2 × 1. We also assume 0! = 1. For our "kith" and "kin" example, this gives

$$\frac{3!4!6!7!}{10!3!0!1!6!} = \frac{1}{30} = 0.033$$

Such calculations are simplified by taking such steps as 7! / 10! = 1 / (10 × 9 × 8). These probabilities are calculated for every row in Table 7.

Table 7. All possible contingency tables with the same row and column totals as the table for "kith and kin"

A	B	C	D	probability
0	3	4	3	1/6 = 0.167
1	2	3	4	1/2 = 0.5
2	1	2	5	3/10 = 0.3
3	0	1	6	1/30 = 0.033

Notice the total of the probabilities is 1. We add together the probability of the contingency table we have actually observed to the probabilities of any of the rows which have higher counts in cell 'a'. Altogether this will give the probability of encountering either the frequency we found of the collocation or a higher frequency, which is a "one-tailed" test. Since there are no contingency tables with higher cell 'a' counts than the one we observed, we take that probability as our p-value for the test. Since this is less than 0.05, we can say that the association between "kith" and "kin" is statistically significant. The commands for running the Fisher test in this case are as follows:

```
> con = cbind(c(3,0),c(1,6))
> fisher.test(con)
```

Unlike the chi-squared and log-likelihood tests, Fisher's test is not an approximation, but gives exact probabilities of encountering a certain number of probabilities given the frequencies of the individual words and the corpus size. At one time Fisher's test was less often used as it was more computationally intensive, but with modern programming languages such as R this is no longer a problem. Its main advantage is that it can be used with very small values in the contingency table (Moore, 2004).

5. The z-score

The z-score was first used as a measure of collocational strength by Berry Rogghe (1973). The form of the z-score used by Seretan is the Pearson residual for cell a of the contingency table we encountered in the section on the chi-squared test:

$$z = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

O_{11} is the observed value in cell a (the raw frequency of the collocation) and E_{11} is the expected value of cell a if there were no relation between the pair of words in the collocation. Seretan (2011, p. 42) points out that the problems with the z-score and the t-test described in Section 6 is that they make the assumption that language data is "normally distributed". For example, we might see very few highly-frequent or very rare collocations, and most collocations would be of medium frequency, giving a characteristic bell-shaped curve. However, in reality, most language data is skewed – most collocations occur just a very few times, while only a few collocations are extremely frequent.

6. The t-test

Manning and Schütze (1999, pp. 163–166) describe the use of the t-test, which is often used for collocation discovery. In general, the value of *t* is related to the probability of finding a collocation which occurs a certain number of times (or more), given the frequencies of the original words and the expected variation in those frequencies between samples.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

μ is the number of occurrences of the collocation we would expect given the frequencies of the original words and the size of the corpus. In a fictitious example, we might have 100 occurrences of the word "bitten" and 200 occurrences of the word "bug" in a million-word corpus, the probability of an individual word being "bitten" would be 100 / 1,000,000 = 0.0001, and the probability of an individual word in the corpus being "bug" would be 200 / 1,000,000 = 0.0002. We can then estimate the probability of "bitten" and "bug" being in a common context as 0.0001 * 0.0002 = 0.00000002 = 2 e−8 in scientific notation. Now imagine that we find the pair of words in collocation 5 times in the 1,000,000 words (actually 999,999 spaces for bigrams) of the corpus. Thus the probability p of an individual pair of adjacent words consisting of this collocation is 5 / 999,999 which is about 0.00005 or 10 e−5 in scientific notation. The variance, $\sigma^2 = p(1\text{-}p)$, is a standard property of the binomial distribution, which is related to the "bell-curve" seen in normal distributions. Since p is very small, σ^2 is approximately p = 10 e-5. Finally, N is our corpus size (minus 1). Taking all these into account,

$$t = \frac{0.00005 - 0.00000002}{\sqrt{\frac{0.00005^2}{999999}}} \approx 999.6$$

Church, Gale, Hanks and Hindle (1991) originally used the t-test to examine collocations in a slightly different way – to find collocates which distinguish near synonyms such as "strong" and "powerful". For example, "strong tea" is a more typical collocation than "powerful tea", but "powerful computers" is a more typical collocation than "strong computers". In this chapter the focus is on using statistical measures to find the strength of collocation between individual word pairs. The following R commands were used to calculate the t-value for the collocation between "take" and "up" which occurs 30 times in the LOB corpus. The individual frequencies for "take" and "up" are 654 and 1,975 respectively. The final command converts the t-score into a p-value for a two-tailed test. In this command, N is used as the degrees of freedom.

```
> mu = (654 / 1000000) * (1975 / 1000000)
> barx = 30 / 999999
> s2 = barx
> N = 999999
> t = (barx– mu) / sqrt(s2/N)
> t
[1] 5.241404
> 2 * (1– pt(t, N))
[1] 1.593917e-07
```

7. Pointwise mutual information

The earliest use of statistics as measures of collocational strength was that of Church and Hanks (1989) who proposed the use of pointwise mutual information (PMI). If two words, *x* and *y*, have probabilities *p(x)* and *p(y)*, and are found to co-occur in the same window with probability *p(x,y)*, then the pointwise mutual information *I(x,y)* between them is given by the relation:

$$I(x, y) = log_2 \frac{p(x, y)}{p(x), p(y)}$$

To find the probability of a word we simply divide its frequency by the number of the words in the corpus, so using the values in the contingency table we get $p(x) = (a + b) / N$; $p(y) = (a + c) / N$, and $p(x,y) = a / N$. The formula above can be implemented in R as follows:

```
> pxy = a / N
> px = (a + b) / N
> py = (a + c) / N
> MI = log2( pxy / (px * py))
```

If a pair of words such as "doctors" and "dentists" are relatively rare in general, but occur relatively frequently together within a short span of words, their mutual information will be high and positive, showing that the two words are associated. If PMI is close to 0, the pair of words are not "attracted" to each other at all. Finally, if the words tend to avoid each other, mutual information will be negative. As suggested by Church and Hanks (1989), statements about mutual information can be made more precise by the reporting the confidence levels and statistical significance of the mutual information for a given word pair in a given corpus. Their solution was to examine word pairs using both mutual information and the t-score, and throw out word pairs which did not have significant t-scores as being uninteresting. Of course this filtering method can be used with any measure of lexical association measure to remove statistically-insignificant word pairs, but here we will look at how to determine the statistical significance inherent within each lexical association measure. In the following section we will look at a method of estimating statistical significance by computer simulation. Although we use PMI as an example, the method is suitable for any other measure of collocation strength.

8. Computer simulations to estimate statistical significance

A related question to statistical significance is that of "confidence limits". Confidence limits are the upper and lower bounds of the confidence interval, which is "a range of values which we can be confident includes the true value" (Altman, p. 162). For example, if we find that the 95% confidence limits for PMI for a pair of words are between 5 and 15, we can be 95% confident that the true value of PMI for the association between these words is in the range 5 to 15. In the case of PMI, since this range does not contain the value of 0, we can be 95% confident that there is a positive association between these two words. Since the *p-value* = 1 – confidence, our collocation has a p-value of less than 5%, and so is statistically significant. Following on from this definition of confidence limits, we suggest a method for estimating them based on a computer simulation, using data originally gathered by Church and Hanks (1989).

If we run many repeated experiments by creating imaginary corpora of the same size as the AP corpus, with the number of occurrences of "honorary" (111) and "doctor" (621) as found by Church and Hanks, assigned randomly to positions in the corpus, within what range would the resulting mutual information be in a given proportion of these simulation experiments? This was the idea of our computer simulations. We ran 100,000 such simulations, and recorded the mutual information each time. To find the range of values in which mutual information for the random association would lie 95% of the time, we took the 2,500th highest value and the 2500th lowest value obtained. In this way the 5,000 most extreme values would lie outside the range, and the other 95,000 (or 95%) would lie inside. Similarly, we could estimate the 99% confidence limits, which were the 500th highest and 500th lowest values obtained. The 99% confidence limit was from minus infinity (the value obtained for simulations where the two words never appeared in the same window) to 8.766. The 95% confidence limits were minus infinity to 7.766. Note that the value of 11.3 obtained for the real experiment was outside the 99% confidence limits of the simulation, showing that the lexical association between "honorary" and "doctor" in the AP corpus was statistically significant, with a p-value of less than 0.01. In fact the "real" value of 11.3 exceeded our very highest simulated mutual information value of 9.351 (obtained when the words co-occurred three times), giving a p-value of less than 1 in 100,000. This computer simulation is based on the method of Koehn (2000), whose technique was originally developed for the BLEU measure of machine translation quality.

9. The Poisson distribution

In the previous section, we saw that the number of co-occurrences of "honorary" and "doctor" in Church and Hanks' (1989) "real" experiment (12) greatly exceeded the number of co-occurrences in any of the computer simulations in which the words were randomly distributed throughout the corpus. This observation suggests the use of the cumulative Poisson probability to determine the probability of obtaining 12 or more co-occurrences in any simulation where "honorary" and "doctor" were randomly distributed throughout a 15-million-word corpus. The formula to find the probability p of encountering k events where the probability of each event is λ is as follows:

$$p(k;\lambda) = \frac{\lambda^k \times e^{-\lambda}}{k!}$$

For example, if we know that the average number of goals seen in a World Cup football match is 2.5, we can work out the proportion of matches which end up goalless, the proportion where just one goal is scored, where two goals are scored, and so on. For example, we can work out the proportion of games where exactly 4 goals are scored. (This example is taken from the Wikipedia page on the Poisson distribution[1]). Here k is 4, λ is 2.5, and e is the mathematical constant, about 2.71828 (note e^x can be worked out using the 'exp(x)' button on a calculator) and the '!' ("factorial") symbol was explained in the section on Fisher's exact distribution.

Thus:

$$p(4;2.5) = \frac{2.5^4 \times e^{-2.5}}{4!} = \frac{39.0625 \times 0.082085}{24} \approx 0.133$$

So we would expect that about 13% of World Cup football matches will have exactly four goals scored in them. A related question is what proportion of matches will have four or more goals scored in them? For this we need to use the cumulative Poisson distribution. One way is to use the formula above to work out the proportion of games with no goals (0.082), one goal (0.205), two goals (0.257) and three goals (0.213). The proportion of games where *fewer than* four goals are scored is the total of these, 0.082 + 0.205 + 0.257 + 0.213 = 0.757. All other games must have at least four goals scored, so we subtract the value for fewer than four goals from 1 to find the proportion of games with four goals or more, i.e. 1–0.757 = 0.243.

Going back to the example of the association strength between "honorary" and "doctor", our value of k is the number of occurrences of the collocation seen

1. https://en.wikipedia.org/wiki/Poisson_distribution

in the corpus, which is 12. We want to know the probability of encountering 12 or more occurrences of "honorary" and "doctor" in a common window. Next we need to calculate λ, which is the most likely number of times the words would be in collocation if the two words had been randomly assigned to the corpus. Since the frequency of "honorary" was 111, the probability of any word in the corpus being "honorary" was 111 divided by the corpus size, which is 15 million. Similarly, the probability of any given word being "doctor" is 621 / 15 million. While the most likely probability of finding the word "honorary" immediately before "doctor" is the product of the probabilities of the two words, we must also take into account a window size of 10. This means that for each occurrence of "honorary", there are ten positions in the corpus (five earlier and five later) which "doctor" can take in order for the word pair to be counted as a collocation. Taking all this into account,

$$\lambda = \frac{111}{15{,}000{,}000} \times \frac{621}{15{,}000{,}000} \times 10 = 3.0636 \times 10^{-10}$$

It would be possible to calculate $p(k; \lambda)$ for every value of k up to 11, sum these together and subtract the total from 1 (as in the goals example above). However, we can also use the online calculator for the cumulative Poisson probability[2]. In the box for "Poisson Random Variable" we insert the value 12, and in the box for "Average Rate of Success" we insert the value for λ in either the form 0.00000000030636 or 3.0636 e-10, then click on the "Calculate" button. We are interested in knowing the probability of obtaining 12 or more occurrences of our collocation in a random corpus, which will be displayed in the bottom box labelled "P($X > 12$)". Using this particular data, the probability is given as 0, so the collocation between "honorary" and "doctor" in the AP corpus is highly significant.

An advantage of the Poisson distribution is that it "is appropriate for studying rare events" (Altman, p. 68). It gives good results when the data is skewed, in the sense that the incidence of "honorary doctor" is very much less than the total incidence of all the other bigrams in the AP corpus put together.

10. Confidence limits of the mean and standard deviation

The emphasis of this chapter has been on measures of collocational strength, but in their seminal paper, Church and Hanks (1989) also write about measures which characterise the quality of collocations: the mean and variance of the distance between the two words making up the collocation. In fixed collocations like "bread

2. http://stattrek.com/online-calculator/poisson.aspx

and butter" and "drink and drive", the two key words are always separated by a fixed number of words – two in these cases. Since they are always exactly 2 words apart when they occur in collocation, their mean separation is 2, and the variance of this separation is 0. Variance is a measure of variability in a set of values, and is 0 if they are all the same, positive otherwise. Closely related to variance is its square root, standard deviation. For example, we might find three instances of the idiom "bitten … bug", as in "bitten by the bug", "bitten by the travel bug" and "bitten by the theatre bug". The separation between the two main words is 2, 3, and 3 words. The mean, variance and standard deviation can be easily calculated using R:

```
bug = c(2,3,3)
> mean(bug)
[1] 2.666667
> var(bug)
[1] 0.3333333
> sd(bug)
[1] 0.5773503
```

Other "semantic" word relations like "man" and "woman" are less fixed. This example has a value close to 0 for the mean separation, since either word is similarly likely to appear before the other as after. Since this word relation is not fixed, there is a high variance of the mean separation. The actual values found by Church and Hanks (1989) for "man" and "woman" in the AP corpus were mean separation = 1.46, variance = 8.07.

For both the mean and the standard deviation of the length, confidence limits may be calculated. These are useful, because for example we might have found the mean length of an idiom over a few examples in a small experiment, but we would like to estimate a range of lengths within which we are 95% certain that the true mean, which we would find in an infinitely large experiment, must lie. For numeric data such as mean length, we first calculate a quantity called the standard error (standard deviation over the square root of *n*, the number of examples of the idiom we study) (Altmann, 1991, pp. 183–4). If we want to find the 95% (or 0.95) confidence limits, we next find the quantity *pr*, for which the R command is

```
pr = 1 - ((1-0.95) / 2)
```

We then find a value of *t*, the same value that we found in the section on the t-test:

```
t = qt(pr, n)
```

and the 95% confidence limits are the mean plus or minus (*t* times the standard error). For example, we found 33 variants of the multiword expression "grasp/clutch at straws", which had a mean length of 2.213 with a standard deviation of 1.623.

Thus the standard error was 2.213 / √33 = 0.385. To find the 95% confidence limits, *pr* is 0.975, and *t* is 2.034. The lower limit is then 2.213 – (2.034 × 0.385) = 1.430 and the upper limit is 2.213 + (2.034 × 0.385) = 2.996. This means that we can be 95% confident that the true mean length is in the range 1.430 to 2.996.

Confidence limits can also be found for proportions, such as the ratio of (idiomatic instances of a MWE) divided by (all instances of the MWE), which we call the idiomaticity ratio (Hanks et al., 2017). The standard error of a proportion is $p = \sqrt{p(1-p)/n}$ (Altman, 1991, p. 230). *t* is found in the same way as when we calculated the confidence limits of the mean, and once again the 95% confidence limits are the idiomaticity found in the experiment plus or minus (*t* times the standard error). Using the "clutch/grasp at straws" example again, we have an idiomaticity of 0.892 over the 33 found examples. Now the standard error is $\sqrt{0.892(1-0.892)/33} = 0.054$. *t* is again 2.034, and so the lower confidence limit is 0.892 – (2.034 × 0.054) = 0.789 and 0.892 + (2.034 × 0.054) = 1.002. Thus we can be 95% confident that the true idiomaticity lies between these two limits.

Less well-known is that it is possible to calculate confidence limits for a standard deviation. Although we can measure the standard deviation in a small experiment, the data may have been unusually bunched or spread out in that small sample. What is the probable range in which the true standard deviation would like in a very large experiment where every example of the idiom ever used was included? An online calculator for the confidence limits of a standard deviation is available from MathCelebrity.com[3]. Note that the value to input is not the standard deviation itself, but the variance, which is standard deviation squared. The calculator helpfully displays a step-by-step working-out of how the confidence limits were calculated. The "clutch/grasp at straws" example had 33 instances with a standard deviation of the lengths (a measure of flexibility) of 1.623, which can be squared to give a variance of 2.6341. In the three boxes of the MathCelebrity.com calculator, we should enter 33, 2.6341, and 95%. The result of the calculation is a lower limit of 1.3052 and an upper limit of 2.1467.

11. Experimental comparison of measures

Four collocates of the word "take" are considered, which occur in the LOB corpus as follows: "take up" (30 times), "take place" (19 times), "take advantage" (ten times) and "take stock" (twice). In each case, the collocation is considered to consist of two words, the first being "take", and the second immediately following the first.

3. http://www.mathcelebrity.com/chiconf.php?n=+16&variance=+4.84&conf=+99&pl=Standard+Deviation+Confidence+Interval#sthash.tBZLAKj5.dpuf

The individual words making up these collocations have the following frequencies: "take" (654), "up" (1,975), "place" (499), "advantage" (71) and "stock" (90). There are a million words in the LOB corpus. These values allow the construction of the following set of contingency tables:

Table 8. Contingency tables for collocates of "take"

	"up" present	**"up" absent**
"take" present	30	624
"take" absent	1945	997400
	"place" present	**"place" absent**
"take" present	19	635
"take" absent	480	998865
	"advantage" present	**"advantage" absent**
"take" present	10	644
"take" absent	61	999284
	"stock" present	**"stock" absent**
"take" present	2	652
"take" present	88	999257

The results we obtained for each of the measures of collocation strength are given in Table 9. The chi-squared, G^2 and z-scores all show that the first three collocations ("take up", "take place" and "take advantage" yield p-values of virtually 0, showing that the strength of these collocations is highly significant. The significance of "take stock" is also high, but not as high as for the other three as "take" and "stock" can each occur in many contexts other than in collocation with each other. However, the significance levels yielded by three measures may not be accurate (hence the warning given by R's *chisq.test()* command) since they make us of expected values which are less than the required 5. The t-test also finds the leftmost three collocations to be highly significant, but not "take stock" which has a p-value of 0.17, which is greater than the arbitrary cut-off point of 0.05. In general the t-test was more conservative than the other measures, yielding lower p-values. A problem with the t-test and the z-score is that they assume that the underlying data is normally distributed, when in reality it is highly skewed with few instances of the collocates compared with all other bigrams in the corpus. The Poisson test is well suited to such skewed data. It does not show that the collocational strength for "take stock" is less significant that the others. The computer simulation of the significance levels of pointwise mutual information showed that all the collocations were equally significant. To distinguish between them, it would be necessary to run the simulation for many

more iterations, which would be very time consuming. The Fisher test, which can be used with a very small sample size, clearly shows that although all four collocations are significant, "take stock" is less so than the others.

Table 9. Results of the experiments on the strength of collocations involving "take"

	"take up"	**"take place"**	**"take advantage"**	**"take stock"**
Chi-squared	$X^2 = 617.66$; $p < 2.2$ e-16; warning	$X^2 = 1013.2$; $p < 2.2$ e-16; warning	$X^2 = 1912.5$; $p < 2.2$ e-16; warning	$X^2 = 35.04$; $p = 3.22$ e-9; warning
G^2	$G^2 = 133.00$; $p = 0$	$G^2 = 118.34$; $p = 0$	$G^2 = 89.01$; $p = 0$	$G^2 = 10.24$; $p = 0.0014$
Fisher	$p < 2.2$ e-16	$p < 2.2$ e-16	$p < 2.2$ e-16	$P = 0.0017$
z-score	$z = 25.26$; $p = 0$	$z = 32.69$; $p = 0$	$z = 46.03$; $p = 0$	$z = 7.97$; $p = 1.55$ e-15
t-test	$t = 5.24$; $p = 1.59$ e-7	$t = 4.28$; $p = 1.84$ e-5	$t = 3.15$; $p = 0.0016$	$t = 1.37$; $p = 0.17$
MI (simulation)	$MI = 4.538$; $p < 0.001$	$MI = 5.863$; $p < 0.001$	$MI = 7.751$; $p < 0.001$	$MI = 5.087$; $p = 0.001$
Poisson	$p = 8.34$ e-13	$p = 5.40$ e-14	$p = 9.99$ e-16	$p = 3.00$ e-15

12. Conclusion

In this chapter we have considered the statistical significance of a number of measures of collocational strength. Chi-squared, G^2 and Seretan's formulation of the z-score all have the requirement that the expected values should be at least 5 (according to a rule of thumb), in order to calculate the statistical significance of these measures, which means that we can only find the significance of the more frequent collocates. However, these three measures can still be used for ranking collocations by strength, and the quality of these evaluations can be measured using the method of Daille (1994). Moore (2004) showed the near equivalence of G^2 and mutual information, which is closely related to pointwise mutual information. The t-test and the z score assume that the data is normally distributed, although this is not generally the case in corpus linguistics. The statistical significance of mutual information can be estimated by computer simulation, but this process is time consuming for frequently occurring collocates. The Poisson distribution is appropriate for studying rare events, and the Fisher test can deal with very small sample sizes. The Fisher test can easily be performed using the R statistical programming languages, and should be recommended for determining the statistical significance of collocational strength.

References

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition. Heboken, NJ: John Wiley. https://doi.org/10.1002/0471249688

Altman, D. G. (1991). *Practical Statistics for Medical Research*. Boca Raton, FL: Chapman and Hall/CRC.

Berry-Rogghe, G. L. M. (1973). The Computation of Collocations and their Relevance in Lexical Studies. In A. J. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The Computer and Literary Studies* (pp. 103–112). Edinburgh: Edinburgh University Press.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). In U. Zernik (Ed.), *Exploiting Online Resources to Build a Lexicon* (pp. 115–164). Hillsdale, NJ: Lawrence-Erlbaum.

Church, K., & Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 26–29 June 1989, Vancouver* (pp. 76–83).

Church, K.W. & Hanks, P. (1989). Word association norms, mutual Information, and Lexicography. *Proceedings of the Annual Meeting of Association for Computational Linguistics*, Vancouver, 76-83.

Hanks, P., El-Maarouf, I., & Oakes, M. (2017). In M. Sailer, & S. Markantonatou (Eds.), *MWE: Insights from a Multi-lingual Perspective*. Berlin: Language Science Press.

Koehler, K. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, 81, 483–493. https://doi.org/10.1080/01621459.1986.10478294

Koehler, K., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, 75, 336–344. https://doi.org/10.1080/01621459.1980.10477473

Larntz, K. (1978). Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, 73, 253–263. https://doi.org/10.1080/01621459.1978.1048156

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: Massachusetts Institute of Technology.

Moore, R. C. (2004). On Log-Likelihood ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)* (pp. 333–340). Barcelona, Spain.

Pecina, P. (2008). *Lexical Association Measures: Collocation Extraction*. (PhD Thesis, Charles University in Prague).

Seretan, V. (2011). *Syntax-based Collocation Extraction*. Berlin: Springer. https://doi.org/10.1007/978-94-007-0134-2

Verbal collocations and pronominalisation

Eric Wehrli, Violeta Seretan and Luka Nerima
University of Geneva

Precise identification of multiword expressions (MWEs) is an important qualitative step for several NLP applications, including machine translation. Since most MWEs cannot be translated literally, failure to identify them yields, at best, inaccurate translation. While some expressions are completely frozen and thus can be listed as compound words, others display a sometimes very large degree of syntactic flexibility.

In this chapter, we argue not only that structural information is necessary for an adequate treatment of collocations, but also that the detection of collocations can be useful for the parser. For instance, it is very useful for solving part-of-speech ambiguities and also some attachment ambiguities. We therefore claim that collocation identification and parsing are interrelated processes.

Section 2 describes the two processes of parsing and collocation detection and their interaction, (i) when and how the collocation identification process is triggered during parsing, and (ii) how the identification of a collocation helps the parser. In Section 3 we describe how anaphora resolution has been implemented in our parsing system, to handle cases where the antecedent and the pronoun are within the same sentence or in adjacent sentences. Section 4 focuses on more intricate cases of verbal collocations where their nominal element has been pronominalised, in the form of a relative pronoun or a personal pronoun. Verb-object collocations with a relative pronoun are extremely frequent and relatively easy to handle for a "deep" parser. In most cases, the relative clause is directly attached to the noun which is part of the collocation. Collocations in which the nominal element takes the form of a personal pronoun are much harder to deal with, as they depend on the process of anaphora resolution, a very challenging task. The last section describes an evaluation of the collocation detection procedure, enhanced with anaphora resolution using a corpus of newspaper articles of about 10 million words.

Keywords: collocation, multiword expressions, anaphora resolution, pronominalisation, deep parsing

https://doi.org/10.1075/ivitra.24.11weh

1. Introduction

Proper identification of multiword expressions (MWEs) is an important qualitative step for several NLP applications. Consider, for instance, the case of machine translation. Since most MWEs cannot be translated literally, failure to identify them yields, at best, inaccurate translation. As an illustration, consider the following German example:

(1) a. *Paul **kam** gestern abend.*
'Paul came last night'
b. *Paul **kam** gestern abend **um**.*
'Paul died last night'

Although both sentences contain the word *kam*, their translation is radically different. This is due to the fact that *kam* is the past tense of the lexeme *kommen* ('to come') in (1a), but of the particle verb ***um**kommen* ('to die') in (1b)[1].

While some expressions are completely frozen and thus can be listed as compound words, others display a sometimes very large degree of syntactic flexibility. The first case comprises so-called "words with spaces" (e.g. *little by little*, *by and large*, *bull fighting*, *close call*). Verbal collocations are good examples of the second case, as illustrated in (2), where the collocation constituents are in boldface:

(2) a. *The Bangkok stockmarket plunged 4.5% in a single day after **news** of the possible human-to-human transmission **broke**.*
b. *The top 500 listed firms **made** about 45% of the global **profits** of all American firms*
c. *That **gave** the thoroughbred industry a needed **boost**.*
d. *Judges deserve great credit for **holding** Brazil's mightiest businessmen and politicians to **account**.*
e. *John Kerry said **progress** was being **made** on a truce in Aleppo.*
f. *Sceptics will wonder if the **money** will be efficiently and honestly **spent**.*
g. *This **record** will be hard to **break**.*

Example (2a) shows a subject-verb collocation, (2b)–(c) verb-object collocations and (2d) a collocation of verb-prepositional object type. In each of these examples, several words separate the two constituents of the collocation. Examples (2e)–(g) display the verb-object collocations *to make progress*, *to spend money* and *to break a record*, but due to syntactic processes – the passive transformation in (2e)–(f), the *tough*-movement in (2g) – the two constituents are in reverse order. Such examples

1. In German, separable particles occur at the end of main clauses with simple tense, but attached to the verb in other cases.

clearly show the need for fine-grained syntactic knowledge for a precise treatment of collocations.

In this chapter, we go one step further and claim not only that structural information is necessary for an adequate treatment of collocations, but also that the detection of collocations can be useful for the parser. For instance it is very useful for solving part-of-speech ambiguities and also some attachment ambiguities. We therefore claim that collocation identification and parsing are interrelated processes.

In the first part, we describe the two processes and their interaction, (i) when and how the collocation identification process is triggered during parsing, and (ii) how the identification of a collocation helps the parser.

In the second part, we turn to more intricate cases, such as the ones illustrated in (3), where the nominal element of a verbal collocation (*to face a challenge*, *to spend money*, *to make a case*, *to make a decision*) has been pronominalised, in the form of a relative pronoun (3a)–(b) or a personal pronoun (3c)–(d).

(3) a. *He will survey the **challenges** we all **face** together.*
b. *Yet the sea walls are using up **money** that could be better **spent** elsewhere.*
c. *Every Democrat is making this **case**. But Mr Edwards **makes** it much more stylishly than Mr Kerry.*
d. *The **decision** to leave behind a child is a hard one. Why do so many migrants **make** it?*

Verb-object collocations with a relative pronoun are extremely frequent and relatively easy to handle for a "deep" parser. In most cases, the relative clause is directly attached to the noun which is part of the collocation. Collocations in which the nominal element takes the form of a personal pronoun (3c)–(d) are much harder to deal with, as they depend on anaphora resolution (AR), a process known to be particularly challenging (cf. Mitkov, 2002). We will describe how AR has been implemented in our parsing system as an adaptation (and simplification) of the Lappin and Leass (1994) algorithm.

2. Parsing and collocation detection

The issue of the interaction between collocation detection and parsing has been discussed in numerous papers. While the usefulness of structural information has been long recognised by lexicographers (see for instance Heid, 2004), which implies that parsing must (at least in part) precede the identification of collocations, computational linguists in their vast majority consider that collocation detection – and more generally MWE detection – occurs prior to parsing (e.g. Butt et al., 1999).

This view, which probably can be explained by the fact that the MWEs considered are mostly of the "words with spaces" (Sag et al., 2002) variety, quickly shows its limits when more challenging cases are considered, such as verbal collocations like the ones described above.

In this section we briefly describe how our Fips parser,[2] a multilingual grammar-based parser, handles MWEs. As explained above, we assume that MWEs must be "known", that is they are listed in the lexical database used by the parser. Compounds (and listed named entities) can be recognized during the lexical analysis of a sentence, just like plain words. As for the other types of MWEs, since their identification requires syntactic knowledge (cf. Seretan, 2011), it should happen during the parse, as soon as the last term of the association (collocation or expression) is attached to the structure.[3]

A collocation database has been added to our monolingual lexical databases, using a collocation extraction system developed by Violeta Seretan and others at LATL (cf. Seretan and Wehrli, 2009; Seretan, 2011). This system extracts candidate collocations from a corpus, filters those candidates using standard association measures and then lets the linguist/lexicographer validate the best candidates, which are entered in the collocation database. The current content of the database for six European languages is shown in Table 1.

Table 1. Number and types of collocations in the Fips lexical database

Collocation Type	English	French	German	Italian	Spanish	Greek
adjective-noun	3,083	9,087	490	1,325	1,621	20,154
noun-noun	5,729	485	2,486	131	66	485
verb-object	866	1,623	197	250	1,098	454
subject-verb	13	22	4	4	5	13
noun-prep-noun	567	9,955	22	1,246	988	12
others	933	2,991	329	205	587	130
total	11,191	24,163	3,528	3,161	4,365	21,248

The collocation detection component integrated in the Fips parser works as follows. It is triggered, during the parse, by the application of a right (or left) attachment rule. Governing nodes of the attached element are iteratively considered, halting

2. See Wehrli (2007) and Wehrli and Nerima (2015) for a description of the Fips parser.

3. Alternatively, one might consider that the identification could be delayed until the end of the parsing process. This, however, would prevent the parser from exploiting collocational knowledge, for instance as heuristics to rank alternatives (cf. Wehrli, 2014).

at the first node of each major category (NP, VP, AP, AdvP)[4]. Then, the procedure checks whether the pair [governing item + governed item] corresponds to an entry in the collocation database. This procedure will be illustrated by means of a simple example. We will return and refine it to handle more complex cases below.

Consider as first example sentence (4a) with the verb-object collocation *to take up a challenge*. The structure, as assigned by Fips, is given in (4b) in the labelled-bracketing form, as well as in the more familiar phrase-structure representation in Figure 1.

(4) a. *Paul took up a new challenge*
b. [$_{TP}$ [$_{DP}$ Paul] [$_{VP}$ took up [$_{DP}$ a [$_{NP}$ [$_{Adj}$ new] challenge]]]]

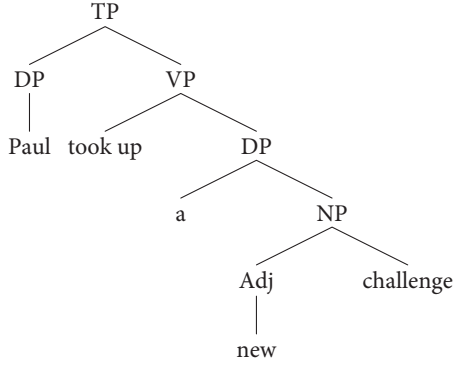


Figure 1. Phrase-structure representation of sentence (4a)

When Fips reads the word *challenge*, finding an adjective on its left, a left-attachment rule will create the noun phrase [$_{NP}$ [$_{Adj}$ new] challenge], which can be attached as complement to the determiner phrase headed by the indefinite determiner *a*, itself governed by the verb *took up*. Given the strategy for collocation detection described above, going up the phrase-structure representation from the noun phrase, first the DP node is found and then the VP node. The latter being a major category node, the procedure halts and checks whether the pair [*take up* + *challenge*], with *take up* as a verb and *challenge* as a direct object, constitutes an entry in the collocation database. This is indeed the case, so the collocation reading is assigned to the verb phrase.

Let us turn now to more complex cases, such as the ones involving syntactic movement, as in Examples (5a)–(e).

4. NP stands for 'noun phrase', VP for 'verb phrase', AP for 'adjectival phrase' and AdvP for 'adverb phrase'. The Fips grammar also uses the labels TP for 'tense phrase' and DP for 'determiner phrase'.

(5) a. *wh*-interrogatives
*Which **record** did Paul **break**?*
[$_{CP}$ [$_{DP}$ which record]$_i$ did [$_{TP}$ Paul [$_{VP}$ break [$_{DP}$ e]$_i$]]]
b. relative clauses
*The **record** that Paul has just **broken** was very old.*
c. *tough*-movement
*This **record** seems difficult to **break**.*
d. *wh*-interrogative + *tough*-movement
*Which **record** did Paul consider difficult to **break**?*
e. passive + small clause + tough-movement
*This **record** was considered very difficult to **break**.*

How can the parser identify the collocation *to break a record* in such sentences? The answer is surprisingly simple if one takes into account the fact that Fips assumes Chomsky's *wh*-movement analysis for such sentences (cf. Chomsky, 1977). According to this view, *wh*-phrases (e.g. interrogative phrases or relative pronouns) bind an empty category in the position corresponding to their interpretation[5]. For instance, a *wh*-phrase interpreted as a direct object binds an empty category in that position. In the analysis of sentence (5a) returned by Fips, the empty category is represented as [DP e] and is co-indexed with the *wh*-phrase.

Given this analysis, to handle collocations involving *wh*-objects, the identification procedure must be slightly modified in order to be triggered not just by the attachment of a direct object, but also by the attachment of an empty category (a trace) in the direct object position of a verb. Then the procedure will check whether the verb and the antecedent of the trace constitute a verb-object entry in the collocation database.

Finally, consider the case where the direct object of a verb-object collocation – or the subject of a subject-verb collocation – has been pronominalised, as in Examples (6a)–(b).

(6) a. *Paul set a new **record** last year and he hopes to **break it** this year.*
b. *Paul set a new **record** last year. He hopes to be able to **break it** again.*

In such cases, to identify the collocation one has to consider the antecedent (the referent) of the pronoun. Fips uses an anaphora resolution component which tries to connect a pronoun with a preceding noun phrase in the sentence as in Example (6a) or in the preceding sentence as in (6b). This procedure is described in the next section. The collocation identification procedure has been updated, again, in order

5. In Chomsky's view, the *wh*-phrase moves from its "original" position to the initial position, leaving a trace (the empty category) behind.

to be triggered by a pronoun attached, for instance, in the direct object position. First the anaphora procedure attempts to identify the antecedent of the pronoun and then the collocation procedure verifies whether the verb and the antecedent of the pronoun correspond to a collocation in the database.

3. Anaphora resolution

Anaphora resolution (AR) – restricted here to the detection of antecedents for third person personal pronouns – has been a hot topic in theoretical as well as computational linguistics since the 1970s[6]. Within the framework of generative grammar, Chomsky's (1981) binding theory and the computational implementations of Hobbs (1978) and Lappin and Leass (1994) are arguably the most significant contributions to AR. Chomsky's binding theory is not intended to be an AR method *per se*, but a set of contraints on the distribution of reflexive/reciprocal pronouns (called anaphors) and of referential pronouns (called pronouns) which are extremely useful to filter out the set of potential antecedents. In a nutshell, binding theory states (i) that reflexive/reciprocal pronouns must be bound in their minimal governing category, and (ii) that referential pronouns cannot be bound in their minimal governing category, where 'bound' means governed and coindexed[7]. Roughly speaking, we can define the minimal governing category of a pronoun as the minimal clause that includes it. It follows from the binding principles (i) and (ii) that anaphors and pronouns (to use Chomsky's terminology) have complementary distribution.

In our AR implementation, we drastically simplify the two relevant binding principles, in such a way that (i) reflexive/reciprocal pronouns must refer to the subject of their minimal clause and (ii) referential pronouns (restricted to third-person personal pronouns) may not have their antecedent in their minimal clause.

The implementation of our AR procedure, which roughly follows the Lappin and Leass (1994) algorithm adapted to the specificities of the grammatical representations of the Fips parser, consists of three steps. The first step deals with the distinction between (third-person) anaphoric and non-anaphoric pronouns. In English, this concerns primarily the pronoun *it*, which can have an impersonal reading (often called pleonastic) as in the following examples:

---

6. See Mitkov (2002) for a comprehensive overview.

7. See Haegeman (1994) for a discussion of the notions of governement, governing category, etc. Notice that Chomsky's binding theory has a third principle, which states that referring expressions (lexical noun phrases) cannot be bound. This principle does not concern AR.

(7) a. ***It** is snowing.*
b. ***It** is well-known that Paul is a reckless driver.*
c. ***It** is easy to prove this theorem.*
d. ***It** was claimed that Paul had cheated.*

Impersonal pronouns are identified by means of lexical and/or syntactic features. Thus, the *it* subject pronoun of meteorological verbs (e.g. *rain*, *snow*, *hail*) is pleonastic, as in example (7a). Examples (7b)–(d) illustrate cases of extraposed sentential subjects – of adjectives in the b and c sentences, of a passive verb in the d sentence – a process which triggers the insertion of a pleonastic pronoun *it* in the subject position. Notice that the distinction between referential and pleonastic *it* is not always easy to make, even when rich lexical information is available. Consider for instance sentence (8), which is truly ambiguous, allowing both a pleonastic reading of the pronoun with the intransitive verb *win* ('winning is easy') and a referential reading with the transitive verb *win* ('something is easy to win').

(8) ***It** is easy to win.*

The latter sense is an example of the so-called *tough*-movement construction, in which the subject is understood as the direct object of the infinitival verb governed by a specific class of adjectives (eg. *easy*, *difficult*, *tough*).

Turning now to third-person referential pronouns, we follow the (much simplified) binding principles described above. In the case of reflexive/reciprocal pronouns, our procedure will verify that they refer to the subject of their minimal clause. In English, that means that the pronoun must agree in number and gender with the subject, as in the following examples:

(9) a. *Paul$_i$ is talking to himself$_i$.*
b. *Mary persuaded Paul$_i$ to do it himself$_i$.*
c. *Mary$_i$ promised Paul to do it herself$_i$.*

Example (9a) is straightforward. The reflexive pronoun agrees with the subject of its minimal clause. In the other two examples, the problem looks a bit harder, at least at first sight. To explain how the reflexive pronoun can refer to the direct object of the main clause in (9b), but to the subject in (9c), one must take into account the fact that infinitival complements are clauses, not just verbal complements, and as such, have their own subject, which is not lexically realised. This subject, called PRO, is an abstract pronoun which, when the infinitival sentence is a sentential argument as in our examples, is controlled by an argument (called 'controller') of the verb governing the infinitival clause. In the case of the verb *persuade*, as in (9b), the controller is the direct object, but for *promise* the controller is the subject. The choice of the controller is an inherent (i.e. lexical) property of the relevant verbs,

which derives from their semantics. With this theoretical background, the mystery is now cleared, and we can represent the relevant structures as (10), where clauses are marked with the preceding TP category and PRO denotes the abstract subject of the infinitival clause. As can be seen, principle (i) of the binding theory is respected: in both structures, the reflexive pronoun refers to the subject of its minimal clause.

(10) a. *[$_{TP}$ Mary persuaded Paul$_i$ [$_{TP}$ PRO$_i$ to do it himself$_i$]].*
b. *[$_{TP}$ Mary$_i$ promised Paul [$_{TP}$ PRO$_i$ to do it herself$_i$]].*

By far, the hardest task for AR comes with third-person referential pronouns such as *it*, *they*, *them*, etc., for which our much simplified binding principle (ii) only states that they cannot have their antecedent in their minimal clause. For such pronouns, the antecedent is usually found in a preceding clause or in a preceding sentence. Often, the context that the procedure must consider provides more than one possible candidate (i.e. noun phrases which agree with the pronoun). Thus, the antecedent of the pronoun *them* could be just about any plural noun phrase preceding the clause in which the pronoun occurs. A good illustration is given by Wilks (1975) with the following example:

(11) *Give the bananas to the monkeys although they are not ripe, because they are very hungry*.

This simple sentence, which is not considered ambiguous by native speakers, displays two occurrences of *they* and two possible antecedents (i.e. plural noun phrases not in the minimal clauses containing the pronouns). What makes such a sentence unambiguous is the knowledge that 'being ripe' is a property of fruit and not of animals, while 'being hungry' is a property of animals and not fruit. Such knowledge is clearly outside the scope of syntactic parsers, for which sentences like (11) will remain ambiguous. In an attempt to limit the high level of ambiguity in real sentences, we order the potential antecedents by means of heuristics inspired by the Centering theory (Grosz et al., 1995; Kibble, 2001), which gives preference first to subjects then to direct objects, etc.

4. Verbal collocations and pronominalisation

We can now turn to the case of verbal collocations in which the nominal constituent has been pronominalised, as discussed in the introduction section. This will concern mostly collocations of types subject-verb and verb-object. Although a majority of such cases concern relative pronouns, as illustrated with Examples (3a),(b), we will focus here on cases involving personal pronouns, for which the AR procedure can provide a reference. Consider the following examples:

(12) a. *The explosion of the IT business and its offshoots has helped produce a new breed of young professionals with* ***money*** *in their pockets and their own ideas on how to* ***spend it****.*

b. *As the outlook for growth brightens, so it becomes more and more likely that* ***interest rates*** *hit bottom when* ***they*** *were* ***cut*** *in July to 3.5%.*

Sentence (12a) shows an occurrence of the verbal collocation *to spend money*, where the term *money* has been pronominalised. When processing this sentence, Fips will attach the pronoun *it* as direct object of the verb *spend*. The AR procedure is activated, searching the (ordered) list of preceding noun phrases for one which both agrees with the pronoun and satisfies the binding principles, as discussed in the previous section. The antecedent (*money*) is then used instead of the pronoun for the collocation detection task. Again, this sequence of events shows the interaction between the attachment procedure of the parser, the AR procedure and the collocation detection procedure.

A relatively similar chain of events occurs in Example (12b), but this time the pronoun (*they*) occurs in the subject position. However, since the verb is in the passive mood, the subject is itself linked to the direct object position. Hence, when the AR procedure returns *interest rates* as antecedent of the pronoun, it replaces the direct object for the collocation detection task, which identifies the verb-object collocation *to cut interest rates*.

Anaphoric pronouns and their antecedents can often be found in the same sentence (modulo the binding principles) as in the above examples. There are, however, a large number of cases where the antecedent occurs in a previous sentence, usually in the immediately preceding sentence, as in the following examples[8]:

(13) a. *Africa has, to put it mildly, a lot of* ***problems****; even a hyperpower cannot* ***solve them*** *all.*

b. *Lots of EU* ***money*** *is owing to Poland and the rest.* ***It*** *must be* ***spent*** *fast.*

In order to handle these cases, the Fips parser was augmented with a mechanism recording the noun phrases of the preceding sentence. When the AR procedure cannot find an antecedent within the current sentence, it considers the recorded noun phrases of the preceding sentence. Since this process can be done iteratively, we found examples such as (14). In (14a) the antecedent (*headscarf*) of the direct object pronoun *it* of the verb *wear* occurs several sentences before. The link between this antecedent and the verb with which it constitutes a verb-object collocation are highlighted. Similarly, in (14b), the last sentence displays an occurrence of a pronominalised verb-object collocation *to set limits*, where the antecedent of the pronoun *them* occurs three sentences before.

8. Laurent (2001) reports that in a large corpus of French, 67% of the antecedents were found in the same sentence as the anaphor and 22% in the immediately preceding sentence.

(14) a. *Is a **headscarf** –or yarmulke or turban– significantly different? Yes, evidently **it** is, but not in a way that supports the true secularist's argument. **It** is different because of the many and various symbolisms **it** bears. Feminists dislike **it** because they see those who **wear it** as victims of male-chauvinist brothers and fathers, who militantly foist their views about women on their sisters and daughters.*

b. *But some **limits** are real, hard and enforced. **They** define the extent of individual liberty and delineate the intrusive powers of the state. **They** are the tricky arithmetic of democracy. **Setting them** is a test; it is not always passed.*

5. Experimental results

The examples provided in the previous section are actual cases of pronominalisation in collocations that were identified in a corpus by our AR collocation detection procedure. More examples of successful identification are presented in Table 2. They illustrate the potential of our procedure to identify "intricate" cases of verbal collocations, where the components appear in different clauses or sentences. To systematically assess the performance of our procedure, we carried out a medium-scale evaluation experiment, which is described in this section.

Table 2. Examples of pronominalised collocation instances detected by our procedure

cause – disease	*Prion **diseases** are odd. Although they can sometimes be passed from one individual to another as an infection, they do not appear to be **caused** by an organism that has genes, unlike all the other infections that are known.*
do – job	*For all its mistakes, modern finance is worth saving – and the **job** looks as if it is still only half **done**.*
exacerbate – problem	*The **problem** is not new, but is getting worse in many parts of the country. And it is **exacerbated** by the return of millions of war refugees and by decades of upheaval that have left land tenure in chaos.*
fault – logic	*But the **logic** of building defences before they are needed rather than after the event is hard to **fault**.*
implement – plan	*Great **plans** are in place to resuscitate South Asia's biggest city. As ever, the difficulty lies in **implementing** them.*
implement – policy	*The department's **policy** was ruled unlawful last November, meaning that although the government appealed against the judgment, it was once more unable to **implement** it in 2008.*
shake – hand	*"We've always held a **hand** out towards the Poles; now maybe it will be **shaken**," says a German official.*
sustain – growth	*Beneath the good cheer, though, are two worries: that **growth** will not be strong enough to produce many jobs and that firms will not invest enough to **sustain** it.*

5.1 Evaluation methodology

The corpus considered in the experiment is an English corpus of newspaper articles from the online version of "The Economist". It contains 10,537 files manually collected from the 1995–2017 issues of the journal, for a total of approximately 10 million words and 520,000 sentences. The corpus was processed using the two-stage collocation extraction method presented in Seretan (2011):

i. In the first stage, potential collocations of various syntactic types (see Table 1) were identified using the AR-enhanced collocation detection procedure presented in this chapter.
ii. In the second stage, these were scored according to the log-likelihood ratio (LLR) (Dunning, 1993), a lexical association measure which is typically employed in computational lexicography for predicting unithood.

The total number of candidate pairs identified in the corpus, all syntactic types considered, is 2,136,433 (corresponding to a total of 1,118,555 distinct pairs). Statistics on the number of subject-verb and verb-object pairs are shown in Table 3. Pronominalised instances are found in a small fraction of the results (see columns 3 and 4).

Table 3. Statistics on collocation extraction results (number of candidates identified)

Collocation type	All	Pronominalised	Pronominalised (Percentage)
Subject-verb			
types	264,173	44,865	17.0%
tokens	338,255	53,862	15.9%
Verb-object			
types	257,029	19,732	7.7%
tokens	468,630	22,278	4.8%

The test set considered in our evaluation experiment was created as follows. For each of the two types of verbal collocation we are interested in – subject-verb and verb-object, as discussed in Section 4 – we considered the pronominalised instances, which we filtered according to the following criteria:

1. LLR score ≥ 20. We focused our evaluation on those pairs presenting a higher chance to constitute a collocation, as predicted by the LLR measure.
2. Lexicographic interest. We manually browsed the pairs satisfying the first criterion and selected 100 pair types that are deemed to comprise valid collocation on the basis of the lexicographic inclusion test (*Is the pair worth storing in a lexicon?*). In some cases, the Oxford collocation dictionary was consulted in order to decide whether a pair was relevant or not.

3. Between 2 and 10 instances. In order to balance the test set according to frequency, for each of the 100 pair types we selected a minimum of two and a maximum of ten instances to evaluate.

After applying the filter, we ended up with a test set which contains 345 pronominalised instances of type subject-verb and 295 instances of type verb-object. The total size of the evaluation set is 640 instances, corresponding to 200 distinct collocations.

The 640 instances were manually evaluated by one of the authors using an in-house concordancer, which displays each collocation pair in context. The evaluation categories were:

– CORRECT, if the instance is indeed a case of pronominalisation of the verbal collocation;
– INCORRECT, otherwise (the collocation procedure failed – either for parsing or AR-related reasons – or the pair, even if syntactically correctly identified, cannot be considered as an instance of the collocation type expected; for instance, it is a regular combination happening to have the same form as the collocation – see Table 6 for examples).

The evaluation task is relatively easy to perform, in the sense that the evaluation criteria are objective. Therefore, the results of a single annotator can be considered as reliable, with no need to appeal to additional annotators. However, about a dozen examples of each syntactic type (subject-verb and verb-object) were unclear due to the inherent ambiguity and complexity of language. Those cases were discussed and settled by the authors and a third linguist.

Table 4 lists some of the unclear examples and the evaluation category assigned after discussion.

Table 4. Examples of collocations instances in the evaluation set and their annotation

Collocation	Instance	Annotation
shareholder–own	*None of this should much concern anybody other than the insurers' **shareholders**, customers and an increasing number of lawyers, were it not for a broader worry. One consequence of the insurers' enthusiasm for equities is that **they** now **own** about a quarter of the British stockmarket.*	INCORRECT
side – win	*A war that neither **side** dares to lose and both believe **it** can **win** is a perilous thing.*	CORRECT
have– word	*HOLLYWOOD calls it star quality. In politics, the **word** is charisma. Nobody can define it; everybody wants it; and France's President Nicolas Sarkozy seems to **have it**.*	INCORRECT

(*continued*)

Table 4. (*continued*)

Collocation	Instance	Annotation
have – access	*The regime's **access** to western Aleppo and its 1.1m people is now a bridgehead less than a kilometre wide. Rebel snipers **have it** in their sights and are battling to breach the siege.*	INCORRECT
ratify – agreement	*Still to clear all the legislative hurdles in America, the **agreement** is also under fire from both right and left in India. Mr Singh has emerged as an unofficial convenor of opponents of the deal in the Indian Parliament, which, unlike America's legislature, does not have to **ratify it**.*	INCORRECT

5.2 Evaluation results

Following the token- and context-based evaluation process described above, the performance of our AR-enhanced collocation detection procedure can be reported in terms of precision, as shown in Equation 1:

$$P = \frac{number\ of\ instances\ annotated\ as\ CORRECT}{total\ number\ of\ instances\ in\ the\ test\ set}$$

The evaluation results are reported in the table below. They are in line with previous results we obtained in a previous experiment (Nerima and Wehrli, 2013) carried out on "The Economist" data (52 verb-object instances corresponding to 31 types; $P = 98\%$).

Table 5. Precision results by collocation type

Collocation type	Precision
Subject-verb	84.3%
Verb-object	80.3%

The higher performance reported in the past is explained by experimental design factors: Nerima and Wehrli considered a smaller test set and limited their evaluation to lexicalised collocations (i.e. collocations which are already in the lexical database of the parser) and which influence the attachment decisions of the parser, as explained in Section 2.

Examples of false positives, i.e. pairs that were erroneously retrieved by our procedure, are shown in Table 6.

Table 6. Examples of false positives (instances erroneously detected by our procedure)

Collocation type	Instance	Issue
make – deal	*He introduced bold and sensible tax reforms, encouraged people to start making private pension provision, and gave immigrants a better* ***deal*** *by* ***making it*** *easier for them to get German nationality.*	Non-referential pronoun
mall– open	*The* ***Mall*** *of America, in Minnesota, has three rollercoasters and more than 500 shops arranged in "streets" designed to appeal to different age groups. Every morning* ***it opens*** *early to accommodate a group of "mall walkers" who trudge around its 0.57-mile perimeter for exercise.*	non-idiomatic instance (compare to: *Canada's West Edmonton Mall, which opened in 1982, …*)
have– word	*What are the sounds in a language, and how do they combine? What* ***words*** *does* ***it have****, and how do they combine in sentences?*	non-idiomatic instance (compare to: *have a word with someone*)
make – deal	*A lasting* ***deal*** *eluded the previous government; Miss Suu Kyi has* ***made it*** *her central ambition, (…)*	non-idiomatic instance (compare to: *roughly half of all deals are now confirmed on the day they are made*)
make – money	*Central banks now generally see broad* ***money*** *as passive, responding to the economic weather, not* ***making it****.*	linguistic analysis
system– work	*That changed in 1997 when Mr Pepy took charge, introducing a dose of modern marketing and a yield-management* ***system*** *from American Airlines to fill seats and take on the low-cost airlines.* ***It worked****: sales and profits immediately started to rise.*	linguistic analysis

The results of our medium-size evaluation experiment confirm the high performance of our new collocation extraction procedure and the interest to integrate an anaphora resolution module. The combined approach makes it possible to identify some of the most intricate cases of verbal collocations that would otherwise escape a less-informed extraction approach (e.g. a window-based, chunk-based or dependency-based one). The following examples show the importance of properly identifying collocations in a text before submitting it to further computer processing, such as machine translation.

Table 7. Impact of collocation identification on machine translation

Original	*And if the monsoon **meets expectations**, the country may produce a record crop of grains this year.*	Collocation identified
Translation	*Et si la mousson répond aux attentes, le pays peut produire une récolte record de céréales cette année.*	Good translation: *répondre aux attentes* 'reply to expectations'
Original	***Expectations** for his first debate with them on October 9th were low. To meet **them**, "[a]ll he has to do is not drool," reckoned Roger Simon of the Politico, a Beltway newspaper.*	Collocation not identified
Translation[a]	*Les attentes pour son premier débat avec eux le 9 octobre étaient faibles. Pour les rencontrer, "[il] il doit faire n'est pas drole", a estimé Roger Simon du Politico, un journal Beltway.*	Wrong translation: *rencontrer attentes* 'encounter expectations'

a. Translations obtained with Google Translate, April 2017.

6. Conclusion

In this chapter, we described a collocation detection procedure enhanced with an anaphora resolution module that we designed and implemented as part of a syntax-based collocation extraction system. We conducted an evaluation experiment which showed the high performance achieved by the procedure at the task of identifying cases of pronominalisation in collocations. Despite the admittedly low applicability of the procedure – the phenomenon of pronominalisation, if we exclude relative pronouns, is relatively rare in collocations – our work is a step toward advancing collocation identification technology, and filling the gaps which have been identified by theoretical studies on collocations. Stone and Doran (1996) gave the example, "*Their **escape** had been lucky; Bill found it uncomfortably **narrow***" to illustrate the limitations of existing technology, which fail to establish the link between *narrow* and *escape* and thus to detect the collocation *narrow escape*. To their example, we now answer: "*Pronominalisation is a **challenge**; we **took** it*".

As far as future work is concerned, the remaining challenges (illustrated in Table 6) are: telling apart idiomatic and non-idiomatic instances, referential and non-referential pronouns and complete collocations from mere fragments; resolving cataphora; and, more generally, improving the linguistic analysis procedure in order to deal with more of the many issues caused by the complexity, ambiguity and intricacies of language.

References

Butt, M., King, T. H., Niño, M.-E., & Segond, F. (1999). *A Grammar Writers Cookbook*. Stanford: CSLI Publications.

Chomsky, N. (1977). On wh-movement. In P. Culicover, T. Wasow, & A. Akmajian (Eds.), *Formal Syntax*. Academic Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris Publications.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2), 203–225.

Haegeman, L. (1994). *Introduction to Government and Binding Theory*. Blackwell

Heid, U. (2004). On the presentation of collocations in monolingual dictionaries. In *Proceedings of the eleventh EURALEX International Congress Vol. II* (pp. 729–738). Lorient, France.

Hobbs, J. (1978). Resolving pronoun references. *Lingua*, 44, 311–338. https://doi.org/10.1016/0024-3841(78)90006-2

Kibble, R. (2001). A Reformulation of Rule 2 of Centering Theory. *Computational Linguistics*, 27(4), 579–587. https://doi.org/10.1162/089120101753342680

Lappin, S., & Leass, H. (1994). An algorithm for pronominal anaphora resolution, *Computational Linguistics*, 20(4), 535–561.

Laurent, D. (2001). *De la résolution des anaphores*. Internal report. Synapse Developpement.

Mitkov, R. (2002). *Anaphora Resolution*. Longman.

Nerima, L., & Wehrli, E. (2015). Résolution d'anaphores appliquée aux collocations: une évaluation préliminair. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles* (pp. 772–778). Les Sables d'Olonne, France.

Sag I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbukh (Ed.), *CICLING02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 1–15). Springer.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Springer Verlag. https://doi.org/10.1007/978-94-007-0134-2

Seretan, V., & Wehrli, E. (2009) Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation. Special Issue on Multilingual Language Resources and Interoperability*, 43(1), 71–85. https://doi.org/10.1007/s10579-008-9075-7

Stone, M., & Doran, C. (1996). Paying heed to collocations. In *Proceedings of the Eighth International Workshop on Natural Language Generation* (pp. 91–100). Herstmonceux, Sussex, England.

Wehrli, E. (2007). Fips, a "deep" linguistic multilingual parser. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing* (pp. 120–127). Prague, Czech-Republic.

Wehrli, E. (2014). The relevance of collocations for parsing. In *Proceedings of the Workshop on Multiword Expressions* (pp. 26–32). Gothenburg: EACL.

Wehrli, E., Seretan, V., & Nerima, L. (2010). Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Application* (pp. 27–35). Beijing: Coling.

Wehrli, E., & Nerima, L. (2015). The Fips multilingual parser. In N. Gala, R. Rapp, & G. Bel-Enguix (Eds.), *Language Production, Cognition, and the Lexicon. Series Text, Speech and Language Technology*, 48 (pp. 473–489). Springer.

Wilks, Y. (1975). Preference Semantics. In E. Keenan (Ed.), *The Formal Semantics of Natural Language* (pp. 329–350). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511897696.022

Empirical variability of Italian multiword expressions as a useful feature for their categorisation

Luigi Squillante
Sapienza – Università di Roma

In contemporary linguistics the definition of those entities which are referred to as multiword expressions (MWEs) remains controversial. It is intuitively clear that some words, when appearing together, have some "special bond" in terms of meaning (e.g. black hole, mountain chain), or lexical choice (e.g. strong tea, to fill a form), contrary to free combinations. Nevertheless, the great variety of features and anomalous behaviours that these expressions exhibit makes it difficult to organise them into categories and gives rise to a great amount of different and sometimes overlapping terminology.

So far, most approaches in corpus linguistics have focused on trying to automatically extract MWEs from corpora by using statistical association measures, while theoretical aspects related to their definition, typology and behaviours arising from quantitative corpus-based studies have not been widely explored, especially for languages with a rich morphology and relatively free word order, such as Italian.

This contribution attests that a systematic analysis of the empirical behaviour of Italian MWEs in large corpora, with respect to several parameters, such as syntactic and lexical variations, is useful for outlining a categorisation of the expressions in homogeneous sets which approximately correspond to what is intuitively known as multiword units ("polirematiche" in the Italian lexicographic tradition) and lexical collocations. The importance of this kind of approach is that the resulting categorisation of MWEs is grounded on empirical data rather than relying on intuitive and not-always-coherent linguistic definitions.

The variational features taken into account are (1) the possibility for the expressions to be syntactically transformed, and (2) the possibility for one of the component to be replaced with a synonym. These features can be automatically and quantitatively investigated using *ad hoc* designed tools, whose methodology is fully explained, if an annotated corpus and a list of expressions are provided. It is possible to show that the kind of attested variations and the magnitude of variation appear highly correlated to the grammatical structure of a given phrase,

https://doi.org/10.1075/ivitra.24.12squ

indicating that the bond between the components for a multiword unit or a lexical collocation can be formed by activating different kinds of restrictions, depending on the considered grammatical pattern.

Keywords: collocation, categorisation, multiword expressions, PAISÀ corpus, semantic variation

1. Introduction

In the linguistic tradition, the definition of *word* is still problematic because of the complexity involved in the identification of its defining features. The concept of a 'word' is intuitively present in the speaker's consciousness, yet it seems impossible to provide a univocal definition because of the different levels of analysis that words can undergo. In fact, the idea of unity that stands behind this concept is relative, since the same linguistic material can form a *unicum* or a set of analysable parts, depending on the perspective that one considers.[1] The difficulty in delimiting the concept of a word becomes that much more significant when one approaches phraseologisms, which show how expressions formed by two or more graphic words[2] can operate as a unit or show specific bonds, placing certain constraints onto the component words so that they are not completely free. In general, one refers to these kinds of entities as multiword expressions (MWEs). According to the most comprehensive definition given by Calzolari et al. (2002), a MWE is "a sequence of words that acts as a single unit at some level of linguistic analysis".

In Italian, examples of this kind of phenomena include common expressions with unitary or idiomatic meaning such as *luna di miele* 'honeymoon' or *pollice verde* 'green thumb', technical terms such as *Presidente del Consiglio* 'Prime Minister' or *anidride carbonica* 'carbon dioxide', syntactic constructions as *fintanto che* 'as long as', adverbs as *alla bell'e meglio*, lit. 'at the beautiful and better' meaning 'in a mediocre way', and also particular associations between lexemes that a native

1. Voghera (1994) states, though, that the debate on the concept of words is based on a number of shared assumption, according to which some reliable criteria allowing the identification of words do exist, although some of them are more effective than others: namely, uninterruptibility (it is not possible to insert linguistic material inside a word), impossibility to move the components (it is not possible to change the order of the morphemes), potential isolation (it is possible to construct a statement by only using one word), potential break (it is always possible to insert a pause, before or after the word). Among these criteria, only the first two seem to guarantee a wide reliability in word identification. For a deeper analysis on the definition of word, see Di Sciullo and Williams (1987), Lepschy (1989), Simone (1990), and Ramat (1990, 2005, 2016).

2. In our written tradition a graphic word is any sequence of characters between two blank spaces.

speaker would recognise, such as *prestare attenzione*, lit. 'to borrow attention' translating 'to pay attention'. Italian linguistic and lexicographic tradition (De Mauro and Voghera, 1996; Voghera, 2004; De Mauro, 2007; Urzì, 2009; Lo Cascio, 2011; Tiberii, 2012) has generally divided this indistinct set of expressions into two main classes of entities that represent two opposite poles of a continuous spectrum.[3] On one hand we have the *unità polirematiche* (multiword units), which include expressions that *need* the co-occurrence of the components in order to convey the expressions' overall meaning (which is not inferable from the components taken in isolation), and are often further characterised by opacity of meaning. Examples of this kind of entity are the already-mentioned *luna di miele*, but also *buco nero* 'black hole', *catena montuosa* 'mountain chain' or *fai da te* 'do it yourself'. On the other hand, we have expressions that exhibit some "special bond" in terms of lexical choice and are characterised only by a *preference* for the co-occurrence of the components. Such expressions are generally compositional and examples are *prestare attenzione* 'pay attention', *compilare un modulo* 'to fill out a form', and *capelli castani* 'brown hair', where *castani* is a version of brown only used for hair and eyes in Italian. The continuum between these two prototypical poles is full of expressions that show different levels of cohesion or lexical preferences and it is often very difficult to allocate a given expression to one or the other of the two categories. The present work aims at outlining a methodology that might be used to discriminate between these expressions on the basis of their empirical behaviour as attested in large corpora.

2. Anomalous behaviours of Italian Multiword Expressions

It is empirically evident (Burger, 1998; Wermter and Hahn, 2004) that MWEs exhibit anomalous behaviours with respect to free combination of words, such as *beautiful house*. While the latter are generally grammatically formed and undergo all the transformations that grammar and lexicon allow, MWE anomalies include several restrictions. In the case of Italian, the main features of anomalies are listed below.

Non-grammaticality. Some expressions appear not to follow standard grammatical rules in their structure, such as the already mentioned *alla bell'e meglio*, but also *essere in forse*, lit. 'to be in maybe' meaning 'to be doubtful', or *prendersela a male*, lit. 'to take it at bad' meaning 'to be offended'.

3. Historically, the terminology that has emerged in relation to the concept of phraseologisms and MWEs is wide and ambiguous. For an overview of the English tradition see Williams (2003) and Bartsch (2004). As for the Italian tradition, Masini (2007) provides an excellent survey.

Unsubstitutability. For some expressions it is not possible to replace a component with one of its synonyms, unless the original meaning is lost, as in the case of *colonna sonora* / **pilastro sonoro*, lit. 'sonorous column / *sonorous pillar' meaning 'soundtrack', where *colonna* and *pilastro* are generally synonyms.

Fixed order. In some expression it is not possible to change or invert the order of the components in all the cases where this transformations would be natural for free expressions, unless we lose the original meaning of the expression or create a sequence which is unusual for the speakers. One example is *botta e risposta* / **risposta e botta* 'tit for tat / *tat for tit'.

Uninterruptibility. Some expressions do not allow for the insertion of words between the components, such as in *luna di miele meravigliosa* / **luna meravigliosa di miele* 'wonderful honeymoon': although in Italian the unmarked adjective always follows the noun and precedes the prepositional phrase, in this case it is shifted to after the entire MWE sequence.

Inflection frozenness. Some expressions keep their idiomatic meaning only when they appear in a specific inflectional form, such as *alti e bassi* / **alto e basso* 'highs and lows / *high and low'.

Restrictions on syntactic transformations. For verbal MWEs it is common that some of the standard syntactic transformations are not allowed. In the case of *tirare le cuoia* 'to kick the bucket', for example, the passive transformation is not permitted.

Non-compositionality. For some expressions it is intuitively clear that their meaning is not a function of the meanings of its parts and of the way they are syntactically combined and thus they are not compositional, according to the definition of Katz and Fodor (1963). Semantic opacity is evident in cases such as *tirare le cuoia*, *pan di Spagna*, lit. 'Spanish bread' meaning 'sponge cake', *camera oscura*, lit. 'dark room' meaning 'camera obscura'.[4]

It is important to note that the abovementioned features of anomaly are not necessary and sufficient conditions to identify MWEs. In fact, non-standard features are exhibited either together or individually and at different levels of frozenness. For example, considering the feature of interruptibility, it is possible to have *punto debole* / *punto <u>più</u> debole* 'weak point / weaker point' but not *carro armato* / **carro <u>grande</u> armato* 'tank / big tank'. Also, we can replace *guerra* with *conflitto* in *guerra mondiale* / *conflitto mondiale* 'world war/conflict' but the replacement of *colonna* with *pilastro* in *colonna sonora* is not allowed as already seen above.

4. Nevertheless, the criterion of non-compositionality appears to be very problematic in a formal theory, since it is difficult to explicitly define what the parts, the rules, the absolute meaning of a word are, etc. (Casadei, 1996, p. 16).

3. A quantitative approach to MWEs

3.1 Reasons to go beyond statistics

The unpredictability of restrictions and the great variability of behaviours make it challenging to identify MWEs with automatic processes and, above all, to clearly establish features that are suitable to categorise them in homogeneous sets.

So far, statistical associations have proved to be effective in the automatic identification of MWEs from texts (Evert, 2004), since they rely on a quantitative feature that all expressions show, regardless of their differences: namely, the frequency or statistic information about the co-occurrence of their components. In fact, it is possible to attest that MWE component words appear together more than they appear with other words. The association between components can be quantitatively measured thanks to association measures (AMs), which, in numerous approaches, combine the word frequency information extracted from an empirical basis (a corpus of texts) with statistics. Every AM is able to identify MWEs comparing the expected frequency, computed under the hypothesis that the components are casually combined, with the observed frequency of the expression in the corpus (Evert, 2004, 2008). The greater the difference between expected and observed frequency, the more the expression is likely to be a MWE.

Nevertheless, it is important to note that the set of expressions including components that show statistically relevant associations (which we can define as *empirical collocations* according to Evert, 2008) is not coincident with the set of MWEs. As Bosque (2004) states, the fact that *read* and *book* appear as an empirical preferential combination does not imply that we must consider it as a MWE. According to quantitative evidence, a book could be *read*, *browsed*, *opened*, *closed*, or *borrowed* even if, in particular, *open* and *close a book* may not be considered lexical collocations. In these cases, the potential appearance of this kind of expressions in the set of empirical collocations will attest the possibility for AMs to bring up false positives in the process of identification, depending on the *a priori* hypotheses according to which one should accept the identification of a MWE. Empirical collocativity can be seen as an *epiphenomenon* brought about by phraseological causes: "idioms, lexical collocations, clichés, cultural stereotypes, semantic compatibility and many other factors are hidden causes that result in the observed associations between words" (Evert, 2008, p. 1218).

Moreover, statistical associations that identify empirical collocations do not seem to place them on an axis that reproduces the intuitive polarisation between the more cohesive and opaque multiword units and the more compositional lexical collocations: in fact, there is no correlation between the association score and the level of idiomaticity of the expressions. In other words, AMs are not able to

Table 1. List of the first eight lemmatised noun-adjective bigrams extracted from the PAISÀ corpus and sorted according to their log-likelihood score

Expression	English translation	Log-likelihood score
guerra mondiale	world war	149509.86
colonna sonora	soundtrack	67671.19
anno successivo	following year	61530.51
evoluzione demografica	demographic evolution	42625.52
essere umano	human being	41729.75
serie televisiva	TV series	41324.95
chiesa cattolica	catholic church	36355.75
sistema operativo	operating system	32940.96

discriminate MWEs on the basis of their different nature, since this is not inferable only from mere co-occurrence information. As an example, Table 1 shows the top eight lemmatised noun-adjective bigrams extracted from the Italian corpus PAISÀ (Lyding et al., 2014) using the well known log-likelihood measure (Dunning, 1993).

As one can see, expressions as *anno successivo*, whose components show a high grade of freedom and transparency, appear alongside more cohesive and institutionalised expressions such as *colonna sonora* and *sistema operativo* (which has a much lower score). The use of statistics alone, then, does not seem to be sufficient to account for the categorical continuum of MWEs.

3.2 Reasons for an empirical, quantitative approach to MWEs

So far, theoretical aspects related to the definition of MWEs, their typology and behaviours arising from quantitative corpus-based studies have not been widely explored. In this scenario, it would thus be interesting to identify other possible procedures, apart from statistics, that could help to shed light on the nature and categories of MWEs, while mantaining a quantitative approach.

Analyses performed on very few English patterns, such as phrasal verbs or verb-object constructions, proved the usefulness of linguistic information in automatic procedures, since it has been shown that using AMs in combination with semantic information can improve the automatic identification of MWEs (Lin, 1999; Bannard et al., 2003; McCarthy et al., 2003; Baldwin et al., 2003). Moreover, a study by Wermter and Hahn (2004) showed how the anomalies in the behaviour of MWEs are empirically relevant with respect to the behaviour of standard phrases: in a corpus, indeed, the number of modified MWEs is much lower than the number of modified standard phrases. Finally, Fazly and Stevenson (2007), in their case study on the verb-object pattern in English, argued that syntactic and semantic

modifications seem to be the most relevant axes of variation that can empirically suggest a categorisation in different types of MWEs.

If we focus on linguistic information only, and accept the hypothesis according to which MWEs have at least one anomalous behaviour in terms of modifications they can undergo, then it is possible to examine: (1) the kinds of variations they undergo, and (2) the magnitude of the modification in terms of percentage of modified expressions.

In order to perform such an analysis, it is necessary to have a large corpus available, a list of expressions to be analysed and that can be extracted by the corpus itself, and a system which is able to automatically check, for each expression in the list, the occurrence of the expression in its basic form and that of its several possible variants.

The corpus needs at least part-of-speech labelling for its tokens and must include references to lemmas for each of them. In case the corpus has been also syntactically parsed, this annotation level can be very useful since it will allow the user to search for expressions whose components can freely move within the sentence but are linked together anyway.

4. Methodology

In order to analyse the modifications that each expression allows, I used an *ad hoc* built tool (Squillante, 2015), whose methodology is described below and which can automatically count the number of modified expressions attested in the corpus with respect to the number of expressions in their basic form. The study takes into account three kinds of modifications: syntactic variations (answering the question: *is it possible to move the components?*), lexical variations (*is it possible to replace one of the components with a synonym?*) and inflectional variations (*does the expression allow for inflection?*). For reasons connected to the computational load, the tool is only able to investigate expressions formed by two words or three words, of which two are content words.

4.1 Syntactic variations

The tool can interrogate the corpus using queries that count the number of expressions matching the search criteria. If the corpus is syntactically parsed we can search for lemmas of the expressions that appear distant in the sentence but are connected by dependency links. If the corpus has no syntactic parsing we should only consider surface co-occurrences of the lemmas and then search for any variation of the expressions by queries based only on sequences of words. The basic idea

is that we can compute an index quantifying the magnitude of variation through the following formula:

$$I_{syn} = \frac{n_i}{n_i + n_{bf}}$$

where n_i is the number of syntactically modified expressions and n_{bf} is the number of expressions in their basic, lemmatised and unmarked form. In this way, the higher the n_i, the higher the value of I_{syn}. The underlying assumption is that if the modified expression is highly attested in the corpus, then the expression is syntactically modifiable; if it is not, then the expression is syntactically frozen.

Concerning syntactic variation, I take into consideration the following modifications.

Interruptibility test. The tool checks whether the expressions are attested in a form in which the components are separated by one or more words, as in the case of *prestare attenzione* / *prestare molta attenzione* 'to pay attention / to pay much attention'.

Fixed order test. The tools checks if it is possible to invert the order of the content words forming the expression, as in the case of *giorno e notte* / *notte e giorno* 'day and night / night and day'.

Syntactic verbal transformations. For verbal patterns, the tools checks if the basic expression is attested in four variants: (1) topicalisation, where the object is in first position (*il campionato ho disputato* 'the championship I contested'); (2) anaphora, where there is a double object construction with the insertion of a pronoun (*il campionato l'ho disputato* 'the championship, I contested it'); (3) passive form, where the expression is inflected in its passive form (*il campionato è stato disputato*, 'the championship has been contested'; (4) relativisation, where the object is linked to a subordinate clause by a relative pronoun (*il campionato che disputi* 'the championship that you compete for').

4.2 Lexical variations

Lexical variations for the components of an expression are investigated by checking within the corpus to see whether they can be replaced by one of their synonyms. It is clear that there is a need for an external resource as a database of synonyms where the tool can consider, for each word, all the replacement options. In my study, the resource is the OpenOffice Thesaurus for the Italian language,[5] a database of 26,823

5. http://linguistico.sourceforge.net/pages/thesaurus_italiano.html, accessed October 19, 2016.

lemmas whose synonyms are divided by word senses. The choice of this resource is due to the fact that it is one of the major current synonym databases freely available for Italian and it is structured in raw text, which makes it easily accessible to computational tools.

Given an expression, the tool searches for the synonyms of each content word in the thesaurus and, if present, it generates a lemmatised modified expression, whose frequency must be checked in the corpus. The greater the presence of modified expressions, the more the original expression will be considered empirically modifiable. The index quantifying the level of lexical variability is given by the following formula

$$I_{sub} = \frac{n_s}{n_s + n_{bf}}$$

where n_s is the number of modified expressions and n_{bf} is the frequency of the expression in its original form. The number of modified expressions is better explained using the following formula:

$$n_s = \sum_i n_{syn1,i} + \sum_i n_{syn2,i}$$

where it is possible to see that n_s is the sum of the number of all the expressions where only the first content word is replaced with all the expressions where only the second content word is. Similarly to the index of interruptibility, the higher the n_s, the higher the value of I_{sub}.

As an example, we can consider *colonna sonora*, for which the synonyms in the database are shown in Figure 1.

colonna|6
(s.f.)|pilastro|sostegno
(s.f.)|cariatide|cippo|obelisco|stele
(s.f.)|aiuto|appoggio|cardine|fondamento|perno|sostegno
(s.f.)|elenco|fila|serie
(s.f.)|carovana|coda|compagnia|drappello|fila|formazione|schiera
(s.f.)|banda|pista
sonoro|1
(agg.)|acustico|altisonante|enfatico|forte|risonante|roboante|rumoroso|squillante

Figure 1. Example of the lemmas *colonna* 'column' (noun) and *sonoro* 'sonorous' (adjective) in the Italian OpenOffice Thesaurus

Colonna sonora generates a total of 24 + 8 = 32 possible modified expressions, whose frequency is checked and summed up to obtain n_s.[6]

Nevertheless, empirical evidences showed that when a component is replaced by a synonym, the resulting expression is not always a synonym of the original one. For instance, if we consider the case of *braccio destro* 'right arm' and replace *braccio* with *ala* 'wing' which is a possible synonym, we obtain the expression *ala destra* 'right wing', which is not a synonym of *braccio destro*[7]. If *ala destra* is highly attested in the corpus, its occurrences would indicate that *braccio destro* is substitutable, even if the original and the modified expressions do not share the same meaning.

The tool can take this into account and go beyond this problem by considering a distributional semantic approach. According to Miller and Charles' definition (Miller & Charles, 1991), the more similar the linguistic context of two words, the more likely it is that the words have similar meanings. This concept can be also applied to MWEs, in order to verify their level of synonymity. Indeed, the words with which any lexical entity co-occurs are able to create an information structure, defined as a vector, which can be compared to that of another lexical element. If one is able to translate the context information into vectors, then it is possible to easily compare the vectors by applying mathematical procedures and obtain the degree of similarity between two expressions, as exposed below.

Let me now show the tool's methodology by considering the expression *guerra mondiale* 'world war'. The tool is able to retrieve all the sentences in the corpus where the expression is present, thus creating a subcorpus. It is then possible to count the occurrences of all the content words in the subcorpus and create the information vector, which assigns frequency values to each of these words.

The tool is able to automatically reproduce the same procedure for all the modified expressions created by the replacement of *guerra* or *mondiale* with their synonyms. Figure 2 shows the vector structures for the first 15 more frequent

---

6. The choice of replacing only one component at a time in the expression, instead of considering simultaneous substitutions of both the content words, is mainly due to two reasons. First, if n and m are respectively the number of synonyms for the first and the second content word, our procedure generates $n+m$ expressions to be checked, while with simultaneous substitution the total number would increase to $n \cdot m$ and this would generate a huge computational load without any further optimisation processes. Secondly, tests performed on sample expressions, as reported in Squillante (2015), showed how simultaneous substitution led to a great dispersion of the expression's original meaning, as in the case of *via d'uscita* 'way out', for which expressions as *inizio di pubblicazione* 'start of publication' and *itinerario d'apertura* 'opening itinerary' were created.

7. This is an emblematic case, since *braccio destro* has the idiomatic meaning of "right-hand person", while *ala destra* has, in turn, a totally different idiomatic meaning, defining the role of a soccer player.

guerra mondiale		conflitto mondiale		spedizione mondiale	
lemma	**freq.**	**lemma**	**freq.**	**lemma**	**freq.**
essere	17,287	essere	1,337	partecipare	4
avere	4,035	avere	425	avere	3
venire	3,533	venire	302	tuttavia	2
suo	3,415	suo	269	perdere	2
anno	2,568	anno	258	parte	2
militare	2,486	non	225	infortunio	2
tedesco	2,268	più	205	impedire	2
fine	2,201	anche	193	già	2
più	2,135	parte	189	fare	2
parte	2,073	fine	188	edizione	2
non	2,001	ultimo	146	arrivare	2
anche	1,957	scoppio	128	giungere	1
italiano	1,671	italiano	124	girone	1
aereo	1,477	italia	124	giocatore	1
scoppio	1,375	tedesco	120	ginocchia	1

Figure 2. Vector structures for the first 15 most frequent words in the case of *guerra mondiale* 'world war', *conflitto mondiale* 'world conflict' and *spedizione mondiale* 'world expedition', with frequency information extracted from the corpus PAISÀ

words in the case of *guerra mondiale*, *conflitto mondiale* 'world conflict' and *spedizione mondiale* 'world expedition', with frequency information extracted from the PAISÀ corpus.

For all the expression vectors to be compared, I order the co-occurring words alphabetically, such that every word has a component value in each vector, which is equal to its frequency, as shown by Figure 3. Then the tool is able to evaluate the *geometrical proximity* of the vectors using the values of their components by calculating the cosine distance, whose formula is reported below:

$$\cos\theta = \frac{\vec{A}\cdot\vec{B}}{\left\|\vec{A}\right\|\left\|\vec{B}\right\|} = \frac{\sum_{k=1}^{n} A_k B_k}{\sqrt{\sum_{k=1}^{n} A_k^2}\sqrt{\sum_{k=1}^{n} B_k^2}}$$

A_k and B_k represent the k-th component (which means frequency of the k-th co-occurring word) for the expressions A and B; n is the frequency rank threshold for the considered co-occurring words, which was set to 50 in my study: this means that the vectors consider only the 50 most frequent co-occurring words for each expression.

	guerra mondiale	conflitto mondiale	spedizione mondiale
aereo	1,477	0	0
anche	1,957	193	0
anno	2,568	258	0
arrivare	0	0	2
avere	4,035	425	3
edizione	0	0	2
essere	17,287	1,337	0
fare	0	0	2
fine	2,201	188	0
Già	0	0	2
ginocchia	0	0	1
giocatore	0	0	1
girone	0	0	1
giungere	0	0	1
impedire	0	0	2
infortunio	0	0	2
italia	0	124	0
italiano	1,671	124	0
militare	2,486	0	0
non	2,001	225	0
parte	2,073	189	2
partecipare	0	0	4
perdere	0	0	2
Più	2,135	205	0
scoppio	1,375	128	0
Suo	3,415	269	0
tedesco	2,268	120	0
tuttavia	0	0	2
ultimo	0	146	0
venire	3,533	302	0

Figure 3. Vectors of the 15 most frequent content words co-occurring with the expressions *guerra mondiale* 'world war', *conflitto mondiale* 'world conflict' and *spedizione mondiale* 'world expedition' extracted from PAISÀ. The number of total words is more than 15, but less than 45, since many of the most frequent words are shared between expressions

The results of the cosine distance are values in the range between 0 and 1 and thus we can interpret them as a percentage of similarity between the expressions. In the case of Figure 2, the similarity between *guerra mondiale* and *conflitto mondiale* is 97%, while the similarity between *guerra mondiale* and *spedizione mondiale* is 11%.

We can use the cosine distance as a weight in the computation of the occurrences of the modified expressions, and thus calculate a corrected value for n_s, as shown by the following formula:

$$n_s = \sum_i \cos\theta_{1,i} n_{syn1,i} + \sum_i \cos\theta_{2,i} n_{syn2,i}$$

In this way, if there is a high number of attested modified expressions, but they are not synonyms of the original expression, the weight will rule out these occurrences or at least reduce them proportionally to the level of synonymity of the modified expression with respect to the original one.

4.3 Inflectional variations

Finally, for each of the given expressions, the tool is able to check if there is a concentration of their occurrences in a specific inflected form. The index is given by the following formula:

$$I_{infl} = \frac{n_{bf} - n_{prev}}{n_{bf}}$$

where n_{bf} is the number of expressions in their basic form and n_{prev} is the number of attested expressions in the most prevalent form. In this way, the higher the value of n_{prev} is, the lower the value of I_{infl} will be.

5. Analysis and results

In the analysis performed on the Italian language I used the tool to study nine grammatical patterns among the most common MWE generators in Italian,[8] which are noun + adjective (NA, *casa editrice* 'publishing house'), adjective + noun (AN, *libero arbitrio* 'free will'), noun + preposition + noun (NPN, *gioco d'azzardo* 'bet'), noun + preposition combined with determinative article + noun (NPdN, *vigile del fuoco* 'firefighter'), noun + preposition + infinitive verb (NPVinf, *macchina da scrivere* 'typewriter'), noun + noun (NN, *sala giochi* 'penny arcade'), noun + conjunction + noun (NCN, *punto e virgola* 'semicolon'), verb + conjunction + verb (VCV, *gratta e vinci* 'scratch card'), verb + determinative article + noun (VDN, *dare i natali*, lit. 'to give the Christmases' meaning 'to give birth').

8. The nine patterns include eight patterns that are typical generators of nominal MWEs and only one sequence (VDN) to test our methodology also on verbal patterns.

The corpus that we considered is the already-mentioned PAISÀ, which is a wide collection of Italian texts extracted from the web, including 250 millions tokens labelled with lemmas, grammatical categories and dependency structures. It was possible to extract from the corpus the 500 most frequent expressions for each of the patterns, regardless of their nature of standard expression or MWE: in this way, high frequency works as a guarantee for a good sample of sentences in order to investigate variations.

The first result highlighted by empirical evidences is that the inflection index does not seem to add relevant information to the categorisation of the expressions. It is possible to analyse this feature separately for nominal and verbal patterns.

Nominal expressions can only have two inflection variants: singular and plural. Some expressions that appear unitary and highly idiomatic, such as *cartone animato* 'cartoon', show a high score for the inflection index, attesting that they are present in both singular and plural form, but since these expressions are functionally substitutable by simple nouns, it is reasonable that they acquire both forms. Secondly, there are some expressions that are not attested in the plural form such as *felicità ultraterrena* 'afterlife happiness' or *anidride carbonica* 'carbon dioxide'. In these cases, though, the absence of the plural is easily explained by semantic reasons, such as the abstract nature of the expression, rather than by their MWE status. Finally, there are some expressions that are only or mainly attested in their plural form, such as *giochi olimpici* 'olympic games', but this inflection restriction is proved to be always associated with other syntactic or lexical restrictions.[9] Also when it comes to verbal expressions, the inflection index is a secondary information: in fact, with verbs, the loss of conjugation implies the creation of nominal multiword units only, as already noted by De Mauro and Voghera (1996).

Because of this, the inflection variations are not considered as a relevant feature that is able to generate a categorisation of the MWE continuum. Consequently, my study focuses on the analysis of how the considered expressions arrange themselves across the plane defined by the two axes of syntactic and lexical variations. The distribution of the expressions is pattern-specific, as shown in Figures 4, 5 and 6, although we can find regularities when analysing what kinds of expressions cluster together in specific areas.

Patterns such as NA and NPN, as well as AN (to a lesser extent), show a concentration of expressions in the area of frozen syntactic variations, evidencing that the sequence itself is more likely to oppose a higher resistance with respect to interrupting or moving its constituents. The NPdN pattern, instead, shows a concentration of expressions in a lower middle area of frozen syntactic variations and thus reveals a greater flexibility in terms of modification, which is inferable from the presence of

9. In the case of the pattern NA, for example, 100% of the expressions whose prevalent form is the plural have values lower than 4% for syntactic variations.

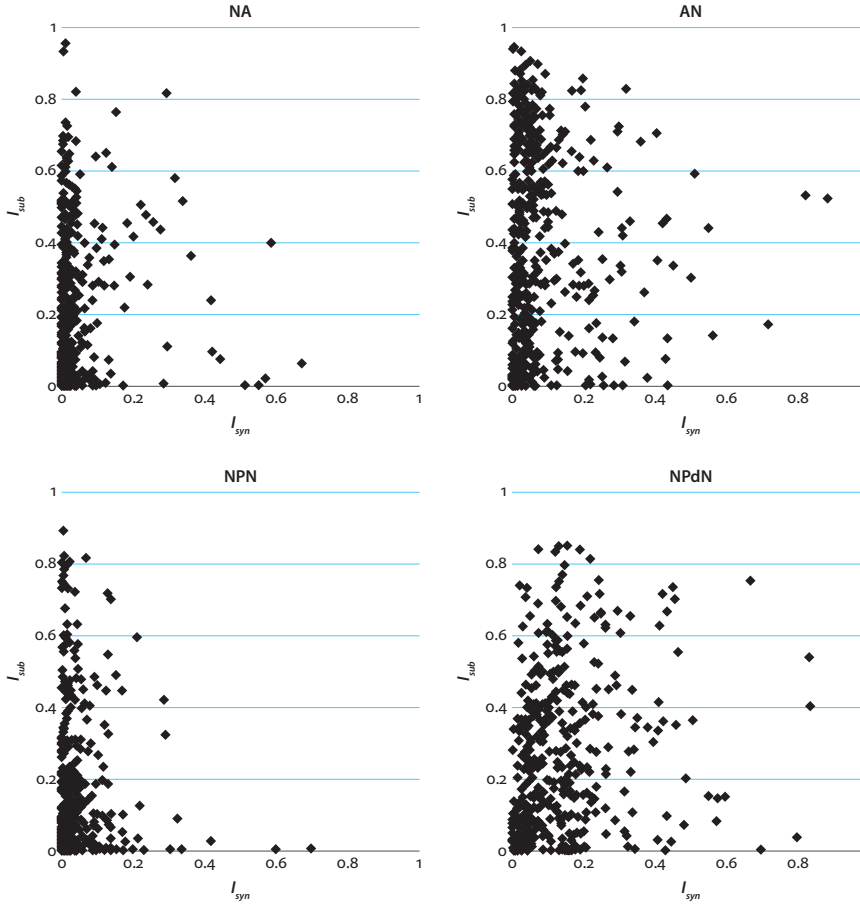


Figure 4. Distribution across the plane defined by the indexes of syntactic (I_{syn}) and lexical (I_{sub}) variations of the 500 most frequent expression belonging to the patterns NA, AN, NPN and NPdN extracted from the PAISÀ corpus

the determinative article.[10] As for the NPVinf pattern, the strong presence of low syntactic variability but high lexical variation depends on the fact that almost the totality of the considered expressions are fragments of the sequence *[in] grado di* + infinitive verb 'able to + infinitive', which can be associated with a broad variety of verbs and shares the same pattern with the well-known crystallised MWEs such as *macchina da scrivere* 'typewriter', *gomma da masticare* 'chewing gum', *associazione per delinquere* 'criminal conspiracy', etc. The NN pattern, which is not a sequence

10. As noted by Masini (2008), the presence of the determinative article, which is the natural syntactic contour to nouns, let them integrate into standard rules of syntax and grammar. His absence, instead, can relate to phenomena such as composition or incorporation.

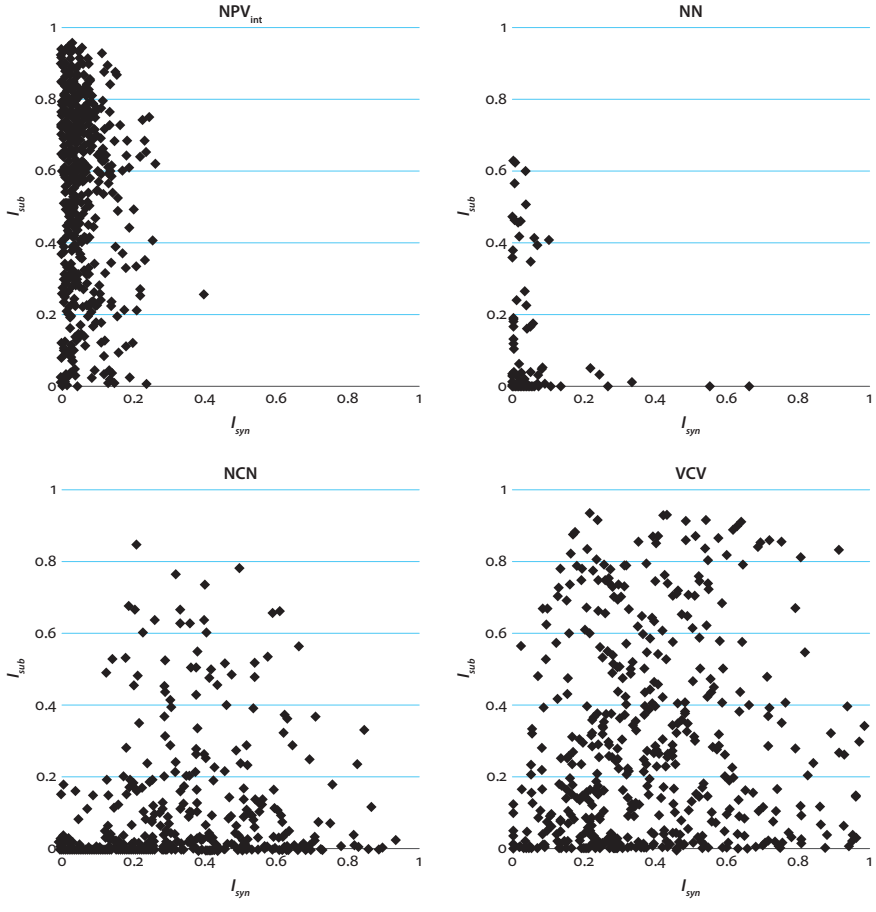


Figure 5. Distribution across the plane defined by the indexes of syntactic (I_{syn}) and lexical (I_{sub}) variations of the 500 most frequent expression belonging to the patterns NPVinf, NN, NCN and NVN extracted from the PAISÀ corpus

integrated in standard Italian syntax since it is more associated with composition, shows only a few expressions out of the area of total frozenness. Finally, the NCN, VCV and VDN patterns show a distribution of the expressions that spreads over the entire plane, due to the higher syntactic independence of their constituents.

For every pattern, the bottom-left corner is the place where multiword units gather, that is, expressions that show a high unity of meaning or a strong terminological nature.[11] On the contrary, for every pattern, the upper-right area, i.e. the

11. NN is the only pattern to include entities recognisable as multiword units also when lexical modification is allowed for one of the components, which can be strictly replaced only by one

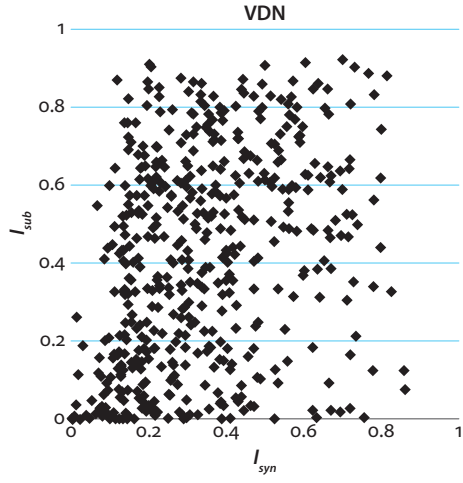


Figure 6. Distribution across the plane defined by the indexes of syntactic (I_{syn}) and lexical (I_{sub}) variations of the 500 most frequent expression belonging to the pattern VDN extracted from the PAISÀ corpus

area where both syntactic and lexical variations are allowed, is the place where free expressions gather. The fact that the upper-right area is often almost empty is due to the fact that taking the 500 most frequent expressions for each pattern implies a high proportion of MWEs. Tables 2 and 3 show examples of the expressions found in these areas.

Table 2. Examples of expressions extracted from PAISÀ which are in the bottom-left corner of the plane defined by I_{syn} and I_{sub} and are thus identified with low values for both indices

Pattern	Expression	English translation	I_{syn}	I_{sub}
NA	*carro armato*	tank	0.00127	0.00371
AN	*pronto soccorso*	first aid	0.00097	0.00003
NPN	*opera d'arte*	artwork	0.00355	0.00232
NPdN	*forze dell'ordine*	law enforcement agency	0.00273	0.00969
NPV_{inf}	*ragion d'essere*	raison d'être	0	0.08080
NN	*centro benessere*	spa	0.00551	0.00195
NCN	*botta e risposta*	tit for tat	0.00671	0
VCV	*tira e molla*	hesitation	0	0
VDN	*cessate il fuoco*	cease-fire	0.00189	0.00468

specific synonym as in the cases of *sito internet/web* 'internet/web-site', *servizio passeggeri/viaggiatori* 'passenger/traveller service'.

Table 3. Examples of expressions extracted from PAISÀ which are in the upper-right area of the plane defined by I_{syn} and I_{sub} and are thus identified with high values for both indices

Pattern	Expression	English translation	I_{syn}	I_{sub}
NA	*famiglia nobile*	noble family	0.58510	0.39928
AN	*importante ruolo*	important role	0.81944	0.53419
NPN	*serie di eventi*	series of events	0.29160	0.31903
NPdN	*zona della città*	area of the city	0.66253	0.75384
NPV_{inf}	*scelta di fare*	choice to do	0.24546	0.74917
NN	//	//	//	//
NCN	*amici e familiari*	friends and relatives	0.50714	0.78417
VCV	*saccheggiare e devastare*	to ransack and devastate	0.48889	0.70458
VDN	*avere l'effetto*	to have the effect	0.39463	0.66223

Depending on the pattern, we can find in the upper-left to bottom-right diagonal those expressions that do show some phraseological cohesion and can be found in collocation dictionaries,[12] according to the schema of Table 4.

Table 4. Examples of expressions extracted from PAISÀ which are in the upper-left to bottom-right diagonal of the plane defined by I_{syn} and I_{sub} and are recognisable as lexical collocations

Pattern	Syntactic variation	Lexical variation	Example	En. translation	I_{syn}	I_{sub}
NA	–	+	*crescita economica*	economic growth	0.00583	0.67480
AN	//	//	//	//	//	//
NPN	+	–	*olio d'oliva*	olive oil	0.30246	0
	–	+	*via di fuga*	way out	0.00317	0.48524
NPdN	–	+	*scopo del gioco*	goal of the game	0.05914	0.40996
NPV_{inf}	//	//	//	//	//	//
NN	//	//	//	//	//	//
NCN	+	–	*morti e feriti*	dead and wounded	0.67103	0
VCV	+	–	*ridere o piangere*	to smile or cry	0.35897	0
VDN	+	–	*coniare un termine*	to coin a term	0.46188	0.03227

12. The set of expressions extracted from the PAISÀ corpus inevitably also brings up expressions that are not included in the three main collocation dictionaries available for Italian language (Urzì, 2009; Lo Cascio, 2011; Tiberii, 2012) as a consequence of the empirical methodology used in my analysis. In fact, although the study of the variations is performed through a quantitative approach, the analysis of the expressions gathering in certain areas of the plane is inevitably performed with a qualitative approach. The fact that Italian collocation dictionaries are compiled on the basis of intuition or not fully explained methodologies makes them just a first reference in outlining the nature of a certain expression found in the corpus, whose status must inevitably be validated by the author.

The rest of the expressions that are present in the remaining areas can be considered free.[13]

6. Conclusion

The study of the variational behaviour of Italian expressions formed by two or three words has shown how it is possible to shed new light on the categorisation of MWEs. Considering only two axes of variations that identify syntactical and the lexical modifications, it is possible to empirically outline homogeneous groups of expressions thanks to a tool that quantifies their level of flexibility with respect to standard modification tests. Our results show that multiword units ("*polirematiche*"), which are historically related to strong unitary meanings or technical terminology, are characterised by total frozenness over syntactic and lexical variability. On the contrary, preferential or lexical combinations are empirically defined by two opposite and complementary mechanisms, depending on the considered pattern, which involve, in turn, only one of the two possible modifications. The strategy preferred by nominal phrases is a restriction on syntactic modifications while lexical variations are allowed, although for the NPN pattern the opposite mechanism is evidenced for some expressions. Verbal and coordinated phrases, on the other hand, have restrictions on lexical modifications while they generally allow for syntactic variations. This can be explained by the fact that these kinds of phrases, when they are not crystallised, are more easily included in standard syntax transformations and lexical frozenness becomes the only guarantee of a recognisable bond between the components, attesting the presence of a phraseologism.

13. The fact that even some free expressions exhibit some restriction, as Figures 4, 5 and 6 show, is mainly due to the nature of the considered pattern. AN, for example, is a marked sequence in Italian, since the adjective is typically found in the post-nominal position. If the adjective is placed before the noun, it tends to acquire a metaphorical or subjective meaning that is not present when it follows the noun. *Grande maestro* 'great teacher' is different from *maestro grande* 'adult/tall teacher' and while the first is highly attested in PAISÀ (912 occurrences), the latter appears with only two occurrences since it has a more specific and unusual meaning. As a consequence of this, the expression results as syntactically frozen not due to its MWE status, but due to intrinsic standard strategies of the language. For the coordinated patterns NCN and VCV, the impossibility of changing the order of the sequence is often due to the semantic consequentiality of the two content words rather than their MWE status, as in the case of *infanzia e adolescenza* 'childhood and adolescence' or *ideare e progettare* 'to conceive and design'. For a more detailed analysis of the many cases of this kind, see Squillante (2015).

References

Baldwin, T., Colin B., Takaaki, T., & Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (pp. 89–96).

Bannard, C., Timothy, B., & Lascarides, A. (2003). A Statistical Approach to the Semantics of Verb-Particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (pp. 65–72).

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English*. Tübingen: Narr.

Bosque, I. (2004). *REDES, Diccionario Combinatorio del Español Contemporaneo*. Milan: Hoepli.

Burger, H. (1998). *Phraseologie: Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.

Calzolari, N., Filmore, C., Grisham, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

Casadei, F. (1996). *Metafore ed espressioni idiomatiche. Uno studio semántico sull'italiano*. Roma: Bulzoni.

De Mauro, T., & Voghera, M. (1996). Scala mobile. Un punto di vista sui lessemi complessi. In P. Benincà, G. Cinque, T. De Mauro, & N. Vincent (Eds.), *Italiano e i dialetti nel tempo* (pp. 99–131). Roma: Bulzoni.

De Mauro, T. (2007). *GRADIT – Grande Dizionario Italiano dell'Uso*. Torino: UTET.

Di Sciullo, A. M., & Williams, E. (1987). *On the Definition of Word*. Cambridge, MA: MIT Press.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61–74.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. (PhD Thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart).

Evert, S. (2008). Corpora and collocations. In A. Lüdeling, & M. Kyto (Eds.), *Corpus Linguistics. An International Handbook* (pp. 1212–1248). Berlin: Mouton De Gruyter.

Fazly, A., & Stevenson, S. (2007). Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Jun 2017. Prague, Czech Republic. ACL (Association for Computational Linguistics). 9-16. (Retrieved from: https://www.aclweb.org/anthology/W07-1102).

Katz, J. J., & Fodor, J. A. (1963). The structure of a Semantic Theory. *Language*, 39(2), 170–210. https://doi.org/10.2307/411200

Lepschy, G. C. (1989). *Sulla linguistica moderna*. Bologna: Il Mulino.

Lin, D. (1999). Automatic Identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 317–324). https://doi.org/10.3115/1034678.1034730

Lo Cascio, V. (2011). *Dizionario Combinatorio Italiano*. Amsterdam: John Benjamins.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., & Pirrelli, V. (2014). The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as a Corpus Workshop of the Association for Computational Linguistics* (pp. 36–43).

Masini, F. (2007). *Parole sintagmatiche in italiano*. (PhD Thesis, Università degli Studi di Roma Tre, Roma).

Masini, F. (2008). Binomi coordinati in italiano. In E. Cresti (Ed.), *Prospettive nello studio del lessico italiano* (pp. 563–571). Firenze: Firenze University Press.

McCarthy, D., Keller, B., & Carroll, J. (2003). Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (pp. 73–80).

Paolo, R. (2005). *Pagine Linguistiche*. Roma-Bari: Laterza.

Miller, G. A., & Walter, G. C. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1), 1–28. https://doi.org/10.1080/01690969108406936

Ramat, P. (1990). Definizione di parola e sua tipologia. In M. Berretta, P. Molinelli & A. Valentini (Eds.), *Parallela 4. Morfologia*. Tübingen: Gunter Narr.

Ramat, P. (2016). What's in a word? In *SKASE Journal of Theoretical Linguistics*, 13(2), 106–121.

Simone, R. (1990). *Fondamenti di linguistica*. Roma-Bari: Laterza.

Squillante, L. (2015). *Polirematiche e Collocazioni dell' Italiano. Uno Studio Linguistico e Computazionale*. (PhD Thesis, Sapienza- Università di Roma).

Tiberii, P. (2012). *Dizionario delle Collocazioni. Le combinazioni delle parole in italiano*. Bologna: Zanichelli.

Urzì, F. (2009). *Dizionario delle Combinazioni Lessicali*. Luxembourg: Convivium.

Voghera, M. (1994). Lessemi complessi: percorsi di lessicalizzazione a confronto. *Lingua e Stile*, XXIX(2), 185–214.

Voghera, M. (2004). Polirematiche. In M. Grossmann, & F. Rainer (Eds.), *La formazione delle parole in italiano* (pp. 56–69). Tübingen: Niemeyer.

Wermter, J., & Udo, H. (2004). Collocation Extraction Based on Modifiability Statistics. In *Proceedings of the 20th International Conference on Computational Linguistics* (pp. 980–986).

Williams, G. (2003). Les collocations et l'école contextualiste britannique. In F. Grossmann, & A. Tutin (Eds.), *Les Collocations: analyse et traitement* (pp. 33–44). Amsterdam: De Werelt.

Too big to fail but *big enough to pay for their mistakes*

A collostructional analysis of the patterns [*too* ADJ *to* V] and [ADJ *enough to* V]

Anatol Stefanowitsch and Susanne Flach
Freie Universität Berlin / Université de Neuchâtel

In this paper, we illustrate the usefulness of the family of methods collectively known as Collostructional Analysis for phraseological research. Investigating two patterns, [*too* ADJ *to* V] and [ADJ *enough to* V], we show how a technique originally developed for the investigation of words and constructions can be fruitfully applied to issues pertinent to phraseology, such as the co-existence of compositional and idiomatic semantics and the analysis of semantically complementary patterns more generally. To this end, we use the three conventional methods (Simple, Distinctive and Co-varying Collexeme Analyses) and propose a novel extension (Distinctive Co-varying Collexeme Analysis) particularly suitable for the investigation of complementary patterns. We show that collostructional analysis is suitable for confirming hypotheses derived from qualitative analyses, as well as uncovering subtle differences that are otherwise inaccessible for non-empirical research.

Keywords: Collostructional Analysis, Simple Collexeme Analysis, Distinctive Collexeme Analysis, Co-Varying Collexeme Analysis, Distinctive Co-Varying Collexeme Analysis, association, collostructions, collocations

1. Introduction

Over the past decade, collostructional analysis – a set of extensions of traditional collocation analysis to associations between lexical items and grammatical structures – has established itself as part of the quantitative corpus-linguistic toolkit.

In this paper, we will apply this family of methods to two grammar patterns of English, [*too* ADJ *to* V] (as in *The banks were too big to fail*) and [ADJ *enough to* V] (as in *The banks are big enough to pay for their mistakes*). These patterns have

https://doi.org/10.1075/ivitra.24.13ste

relatively compositional general meanings that have been described in the formal semantics literature (roughly, they encode respectively excess and sufficiency of a particular property for a particular event to occur). However, they also show a substantial degree of idiomaticity that has largely been ignored.

We will show that the various types of lexical association of the two patterns yield evidence for the general meanings as well as for the idiomatic exceptions of interest in the context of research on phraseology. We will make use of the three main variants of collostructional analysis, as well as one extension that we are sketching out here for the first time. Section 2 of this paper provides the minimal descriptive (Section 2.1) and methodological (Section 2.2) background necessary for our investigation. Section 3 provides some general background concerning our corpus and our data extraction and coding (Section 3.1), and then presents four case studies demonstrating the following variants of collostructional analysis:

- *simple collexeme analysis* to determine which adjectives are statistically significantly associated with each of the two patterns when compared against the language as a whole (Section 3.2);
- *distinctive collexeme analysis* to determine which adjectives statistically differentiate between the two patterns (in Section 3.3), or some of their sub-patterns (in Section 3.4) when compared directly with each other;
- *co-varying collexeme analysis* to determine which adjective-verb pairs are statistically associated with each other in each of the two patterns (in Section 3.4); and
- *distinctive co-varying collexeme analysis*, an extension of co-varying collexeme analysis, to determine which verbs statistically differentiate between the two patterns for a given adjective (in Section 3.5).

2. Background

2.1 Descriptive background

The two constructions investigated here appear straightforward, but this straightforwardness is somewhat deceptive, so a brief discussion is warranted before we look at the empirical results.

Let us begin with the pattern [*too* ADJ *to* V], which we will refer to as the *too*-pattern for expository compactness. Consider the examples in (1a)–(1f) (unless otherwise noted, all examples are from the ENCOW14 corpus; see Section 3.2 below for further details):

(1) a. [F]ireworks bursts can be *too big to capture* with a telephoto.
b. DNA molecules are *too small to study* directly under the microscope.
c. [T]he majority of meteoroids are *too small to penetrate* the hull of spacecraft.
d. All matter […] is made of particles […] that are *too small to be seen* by the unaided eye.
e. [W]hen the chassis is sending dynamic info to the driver, it omits nothing – *no pavement irregularity is too small to overlook*… [www.caranddriver.com]
f. The darker dressed man was *too hard to distinguish* from the trees.

Examples (1a)–(1d) are the most straightforward (and typical) cases of this pattern. The infinitival clause encodes a potential event that is implied to be possible (in other examples: admissible, justifiable, etc.) only as long as one of its central participants falls within a certain range along the dimension referred to by the adjective. The pattern presents this event as impossible (inadmissible, unjustifiable, etc.) by stating that this participant falls outside of this range; in the case of positive polarity adjectives, it is presented as falling above the upper bound of this range (as in 1a), and in the case of negative polarity adjectives, it is presented as falling below (or failing to fall above) the lower bound of this range (as in 1b) (see Meier (2003) for a detailed discussion and formalisation of this analysis).

The central participant whose properties are at issue may be the logical object (or other complement) of the infinitival clause, as in (1a, b), or the logical subject (as in 1c): as is typical of infinitival clauses, their implicit syntactic subject is controlled by a noun phrase in the matrix clause, which may correspond logically to any participant of the verb in the infinitival clause. In (1b), for example, a fuller paraphrase would be "DNA molecules$_i$ are too small *for someone* to study *them*$_i$ under a microscope" – *DNA molecules* is co-referential with the logical object of *study*; the same is true of (1d), where the infinitival clause is in the passive (a fuller paraphrase would be "The particles$_i$ are too small *for them*$_i$ to be seen by the unaided eye", with *particles* being co-referential with the logical object of *see*). In contrast, in (1c), the paraphrase "The meteoroids$_i$ are too small *for them*$_i$ to penetrate the hull" shows that *meteoroids* is co-referential with the logical subject of *penetrate*.

With respect to the analysis of the pattern as coding insufficiency or excess of some property, Example (1e) is anomalous in that the utterance meaning is actually the opposite of the propositional meaning. This is difficult to notice, but it becomes clear if we paraphrase the example with positive polarity: *This pavement irregularity is too small to overlook* would mean that pavement irregularities must *exceed* a certain size in order to be justifiably overlooked; what is meant, however, is that they must be *below* a certain size before it becomes justifiable to overlook them. Such examples do occur in language use (Liberman (2009) presents authentic examples

including the one in (1e)); a plausible explanation for the reversal in interpretation is presented by Wason and Reich (1979). Since such examples do not occur in our (rather large) sample, they do not seem to be very frequent, however, and we will not discuss them further here.

Finally, the example in (1f) differs in that the adjective does not encode a range within which a central participant of the event must fall, but a range within which the event itself can be brought about by an unspecified external participant: it is not the man, who is *hard*, but the seeing event itself. This is, of course, the so-called "raised" variant of the more general pattern [NP *be* ADJ *to* V], distinguishable from the "non-raised" variant by the fact that it has a paraphrase with [*it be* ADJ *to* V NP] (*It was too hard to distinguish the darker man from the trees*). It may be useful to distinguish the two variants of the construction in an empirical analysis; we will briefly discuss how this could be done in the context of our first case study in Section 3.2, below.

Note that in all its variants, the pattern itself is capable of coercing a scalar interpretation of the adjective occurring in it – non-scalar adjectives (such as *pregnant*, *female* or *impossible*) will be re-conceptualised as referring to a range with upper and lower bounds when they are used in the pattern [*too* ADJ *to* V] (cf. Jensen (2014a) for an empirical study confronting subjects with sentences like *I was too female to be ruthless* or *Her mother had been too pregnant to make the trip*).

The second pattern, [ADJ *enough* to V], is, essentially, the opposite of the first, encoding an event and presenting one of its central participants as falling *inside* the range that makes the event possible (admissible, justifiable, etc.). Consider the following examples, which are parallel in type to those in (1a)–(f) above:

(2) a. It usually takes a week or two for the [fungal] colonies to be *big enough to see*.
b. [T]he telescope's continuous tracking motion is driven by a 1/25th horsepower electric motor, *small enough to hold* in the palm of one's hand.
c. Our eyes aren't *big enough to see* all of nature's beauty.
d. [He] succeeded in breaking down chromatin to fragments which are *small enough to be studied* by X-ray diffraction.
e. The story's narrative is *easy enough to follow*.
f. In three portions of the walls, there are small looped windows… *small enough to see out of*, and perhaps fire a weapon, but too small to gain entry.

Again, (2a)–(d) are the most straightforward and typical examples. As (2a) and (2b) show, the relation between the adjective and the central participant of the event is the reverse of the corresponding Examples (1a) and (1b): with positive polarity adjectives, the pattern presents the central participant as falling *above the lower bound*, and with negative polarity adjectives as falling *below the upper bound* of

the range referred to by the adjective (see, again, Meier (2003) for discussion and formalisation).

Again, the infinitival clause can be controlled by the logical object (or other complements) of the verb (as in 2a, b, or by the logical subject (as in 2c), and, again, the infinitival clause can be a passive (as in 2d). Finally, there is, again, a raised variant of the pattern: in (2e), it is not the narrative that is inherently easy, it is the action of following the narrative that is easy *for some unnamed external participant* (compare the non-raised *It is easy enough (for someone) to follow the narrative*). In the case of [ADJ *enough to* V], we are not aware of literature discussing the type of reversed interpretation seen in (1f) above, but (2f) is such an example: what is meant is not that the windows are "small enough" to see out of, but that they are "not too small" to see out of. As far as we were able to tell, such examples are exceptional and very rare in our sample, so we will ignore them here.

As in the case of the [*too* ADJ *to* V] pattern, the [ADJ *enough to* V] pattern will coerce a scalar interpretation of non-scalar adjectives (one example in our corpus is *When and how do you decide that I must be pregnant enough to have to take your required test*). Note that this is not surprising in either case, since such adjectives are easily reconceptualised as scalar in all kinds of grammatical contexts; however, it is a non-trivial question how big a role reconceptualisation plays in specific adjectival patterns such as the ones discussed here.

While there is an excellent theoretical discussion of the two constructions (Meier (2003), mentioned above), there are few corpus-based studies confronting the general characterisation of the construction with authentic usage data. We are aware only of two rather exploratory studies, Jensen Jensen (2014b), which presents a collostructional analysis of the *too*-pattern, and Jensen (2015), which does the same for the *enough*-pattern. The results of these studies will be discussed in the appropriate places in the case studies below.

2.2 Methodological background

The three main variants of collostructional analysis have been described extensively in a series of individual treatments (Stefanowitsch and Gries, 2003; Gries and Stefanowitsch, 2004a; Stefanowitsch and Gries, 2005) and a number of overviews (Stefanowitsch and Gries, 2009; Stefanowitsch, 2011, 2013; Hilpert, 2014); given spatial limitations, we will outline them only briefly here and trust our case studies to demonstrate their working in more detail.

Collostructional analysis is, fundamentally, a variant of collocation analysis, and is, as such, concerned with identifying the strength and direction of association between linguistic items based on their frequency of co-occurrence in language

use. In collocation analysis, the items in question are words co-occurring within a given span in a corpus, whose observed frequency of co-occurrence is evaluated statistically against their expected (or chance) frequency of co-occurrence.

Table 1 shows the research design underlying collocation analysis in schematic form (based on Evert, 2004, pp. 36–37). *U* and *V* are two positions in a span (for example, directly adjacent to each other, forming a *bigram*), and *u* and *v* are two specific word forms occurring in these positions. Their observed frequency of co-occurrence is then compared to their expected frequency of co-occurrence and evaluated statistically using any of a range of test statistics or effect size measures appropriate to two-by-two tables.

Table 1. Schematic collocation contingency table

		V		**Total**
		v	***!v***	
U	***u***	O_{11}	O_{12}	R_1
	!u	O_{21}	O_{22}	R_2
	Total	C_1	C_2	N

To take Firth's (1957) classic example of a collocation, *silly ass* (in the British, i.e. equine sense of the latter): in the British National Corpus, these words *u* and *v* co-occur in this sequence seven times (O_{11}); *silly* directly precedes other word forms (*!v*) 2,632 times (O_{12}); conversely, *ass* is preceded by words other than *silly* (*!u*) 295 times; finally, there are 110,683,384 bigrams in the corpus (N), i.e. 110,680,445 bigrams that neither begin with *silly* nor end with *ass*. On the basis of these four frequencies we can calculate the expected (i.e. chance) frequency of co-occurrence for *silly ass* – it is the total number of words N multiplied by the product of the probabilities of *silly* and *ass*, i.e. 0.0072. The observed frequency is thus much higher (about a thousand times higher) than the expected frequency, suggesting that *silly ass* is indeed a collocation in (British) English. The test statistic of the log-likelihood test, G2, which is commonly used as an association measure in corpus linguistics, is 82.51 ($p < 0.001$), showing that the two words are statistically significantly associated with each other.

In collostructional analysis, the same quantitative reasoning is applied to associations between words and a particular slot of a given grammatical structure, or between words in two slots of a given grammatical structure. There are three variants: (i) simple collexeme analysis, which looks at the co-occurrence of lexical items with a particular grammatical structure or pattern (Stefanowitsch and Gries, 2003); (ii) distinctive collexeme analysis, which looks at the co-occurrence of a lexical item with one grammatical structure or pattern compared to a different, related structure or pattern (Gries and Stefanowitsch, 2004a), and (iii) co-varying

collexeme analysis, which is very close to collocational analysis as introduced above, except that it looks at the co-occurrence of two words not in a given span, but in specific positions of a particular grammatical structure or pattern (Gries and Stefanowitsch, 2004b). There are also multifactorial variants of collostructional analysis, for example, looking at lexical items co-occurring with a particular construction in different registers (Stefanowitsch and Gries, 2008) or at two lexical items occurring inside and outside a particular construction (Stefanowitsch and Gries, 2005; see Section 3.4 below for further details).

3. Case studies

3.1 Data: Source, extraction, cleaning

Our data come from the publicly available ENCOW14AX03, a 600m+ token subset of a 16bn token corpus of English web data (Schäfer and Bildhauer, 2012). ENCOW14AX03 was indexed for use with the CQP language as implemented in the *Corpus Workbench* (Evert 2010a, 2010b). Using the queries {"too"%c[pos="(JJ|RR|VBN)"] "to"%c[pos="VB"]} and {[pos="(JJ|RR|VBN)"] "enough"%c "to"%c [pos="VB"]}, we extracted 31,003 instances of the *too*-pattern and 33,634 instances of the *enough*-pattern. The form of these queries was chosen based on the consideration of two problems that needed to be solved in order to maximise precision and recall of the query.

First, the adjective slot in both constructions is most often occupied by clear cases of adjectives (that are tagged as such with a high degree of accuracy), but there is nevertheless a substantial number of less clear cases, namely (i) items that outside of the patterns tend to be used in non-adjectival contexts (e.g. *soon*, which is usually tagged as an adverb), and (ii) items that are ambiguous between adjectives and participles (*scared*, *complicated*, *terrified*, tagged, with a less-than-satisfying accuracy, sometimes as adjectives, sometimes as verbs). While the former comprises only a handful of high-frequency types that could be dealt with semi-manually, the latter affects an entire class of mostly low-frequency (derived) adjectives.

One option to maximise recall is to leave the adjective slot unspecified in the query. As the data set is too large for manual post-processing, this would have decreased precision considerably, including in the results cases where *too* comes between the subject and (mostly) semi-modals (*I too had to go to Paris* or *whose excellence too ought to have been reverenced*). Also, it would have left us with a more general problem: in a simple collexeme analysis (as in collocation analysis), the frequency of a word in a specific context is compared against its frequency in the corpus as a whole; if we had left the adjective slot unspecified, the frequencies would have had to be compared against the frequencies of all word forms in the

corpus, so, for example, the frequency of the word *cold* in the pattern *too cold to* V (e.g. *The water is too cold to go in*) would have to be compared with the frequency of the word form *cold* in all its uses, including the nouns $cold_1$ ("condition of low temperature") and $cold_2$ ("illness", as in *common cold*) and the adverb $cold_1$ ("with finality", as in *This stopped them cold*) and $cold_2$ ("without preparation", as in *She performed the piece cold*).

The option we chose instead is to specify the word-class of the tag such that the query would include adjectives, adverbs and past participles, and to perform the frequency comparison against a frequency list of words with the same tags. This does not get rid of adverbial uses like those just mentioned, but it does exclude nouns and it gives us a more principled standard of comparison in general. It also means that, for example, the frequency of word forms that are partially mistagged as participles inside these patterns is compared to their frequency with both adjective and participle tags outside the respective pattern.[1]

The second problem of data extraction pertains to the raised variant of the pattern mentioned in the preceding section. Again, there are two potential solutions: First, separate the two manually before running a collostructional analysis, which, given the size of our samples would be a monumental task that should be reserved until such a time that there is a research question that makes this task seem worthwhile; second, accept the ambiguity and keep it in mind when interpreting the results. This is the option we chose, although we complemented it with a step intended to estimate the extent and quality of the problem, which we discuss in the first case study in Section 3.2.

3.2 Case study: Simple collexeme analysis (SCA)

Method. This section illustrates the simplest variant of collostructional analysis, designed simply to investigate which lexical items are strongly (and significantly) associated with a particular pattern (introduced in Stefanowitsch and Gries (2003) and retroactively named *simple collexeme analysis*). Here, we apply it to the adjective slot of the two patterns since this seemed to be the lexically more restricted, and hence linguistically more relevant slot (i.e. the slot most likely to yield information about the semantics and function of these patterns). Table 2a shows the research design in schematic form.

1. This underestimates association strengths for items with higher frequencies as participle or adverb uses than as adjectives (cf. Flach, 2015); it is necessary since items would be left with any frequency in O_{21}, and would either have to be excluded altogether (e.g. *soon*) or have grossly overestimated association values (e.g. *complicated*).

Table 2a. Schematic contingency table for simple collexeme analysis

		L		Total
		l	*!l*	
C	*c*	O_{11}	O_{12}	R_1
	!c	O_{21}	O_{22}	R_2
	Total	C_1	C_2	N

Table 2b illustrates the research design using the most frequent adjective in the [*too* ADJ *to* V] pattern. The word form *late* (*l*) occurs 2,122 times in the pattern [*too* ADJ *to* V] (*c*) and 95,824 times in contexts other than *c*. The pattern's overall frequency was given above, subtracting 2,122 cases means that 28,881 instances of the pattern contain an adjective other than *l*. The last remaining cell contains the frequency of all adjectives that are not *l* in contexts that are not *c*. In our case, this meant the total number of words tagged as adverbs, adjectives or participles, minus the combined frequencies of the other three cells.

Table 2b. Contingency table for *late* in the pattern [*too ADJ to V*]

		L		Total
		late	*!late*	
C	*too* ADJ *to* V	2,122 (30.6)	28,881 (30,972.4)	31,003
	! too ADJ *to* V	95,824 (97,915.4)	99,240,796 (99,238,705)	99,336,620
	Total	97,946	99,269,677	99,367,623

On the basis of this table, the expected frequencies and the association measure can now be calculated. The expected frequencies are shown in parentheses, and the association score G2 is a very high 14003.10 ($p < .0001$). In other words, *late* is not only the most frequent adjective in the pattern [*too* ADJ *to* V], it is also a very strongly associated collexeme.

Repeating this analysis for every word form tagged as an adjective, adverb or participle in the *too*-pattern and the *enough*-pattern respectively yields the data discussed in the next subsection. All subsequent calculations (except for those in Section 4.5) were performed using the R package {collostructions} (Flach, 2017). Note that in collostructional analyses, the *p*-value of the exact test after Fisher and Yates is typically used as an association measure (cf. Pederson, 1996). We use the test statistic of the log-likelihood test here, first, because its use is more widespread in phraseological research and second, because it allows us to use the same association measure for all case studies (including the one in 4.5, where the exact test could not be applied).

Results and discussion. Table 3 presents the top 25 attracted adjectival collexemes of [*too* ADJ *to* V] (all items shown are significantly attracted at $p < .0001$, as all G2 > 19.51). It shows the their corpus frequencies (F_{corp}), their observed and expected values in the construction (O:E), and the collostruction strength (G2).

Table 3. Top 25 attracted collexemes in [*too ADJ to V*]

ADJ	F_{corp}	O:E	G2
late	97946	2122:30.6	14003.10
early	177191	1247:55.3	5442.44
easy	115942	942:36.2	4363.41
lazy	7416	488:2.3	4291.80
hard	160166	949:50	3820.98
big	169934	953:53	3737.59
young	154274	912:48.1	3667.03
small	237445	1046:74.1	3629.67
busy	30652	516:9.6	3119.64
weak	20811	433:6.5	2798.89
scared	8627	356:2.7	2790.06
long	311211	925:97.1	2538.81
difficult	105916	626:33	2511.54
good	596596	1136:186.1	2240.91
stupid	21104	361:6.6	2192.23
tired	16131	330:5	2120.92
happy	82327	468:25.7	1840.85
expensive	40523	367:12.6	1770.75
old	223518	575:69.7	1424.98
eager	8095	166:2.5	1066.82
numerous	36412	242:11.4	1022.36
afraid	24195	208:7.5	981.58
dangerous	34467	227:10.8	954.94
embarrassed	3873	128:1.2	944.80
ill	19359	187:6	924.66

The first observation is that all top collexemes are "scalar" adjectives, suggesting that, indeed, the function of the pattern is one of relating (participants of) events to a range (placing them outside this range) (see also Jensen, 2014b).

Second, there is a balanced mix of marked and unmarked members of antonymic pairs, with a slight preference for unmarked cases. Often, both members of a given pair are present (*early*/*late*, *young*/*old*, *small*/*big*); the rest falls fairly evenly into unmarked (*long*, *good*, *happy*) and marked (*weak*, *stupid*). This suggests that the pattern has no clear preference with respect to the polarity of adjectives. A major part of the adjectives are negatively connoted (*late, lazy, hard, weak, scared,*

difficult, stupid, tired, expensive, old, afraid, dangerous, embarrassed, ill), although there are clearly positive adjectives, too (*easy*, *good*, *happy*).

The final question to return to is that of the "raised" and "unraised" variants. Intuitively, some of the adjectives on the list are likely to occur exclusively or predominantly in the raised variant (such as the textbook cases *easy* and *hard*). Instead of relying on intuition in order to decide which collexemes are likely to occur in the raised variant, we conducted a second simple collexeme analysis on the general pattern [*it* BE ADJ *to* V NP] which should give a good indication of which adjectives are strong candidates for raising (the CQP query used for extraction was {it"%c [lemma="be"] [pos="JJ"] "to"%c [pos="VB"] [pos="(DT|JJ.*|N.*)]}). The most strongly attracted adjectives in this pattern are (in descending order of attraction): ***important***, *possible*, ***hard***, *impossible*, ***easy***, *necessary*, ***difficult***, ***nice***, *like*, *advisable*, ***good***, ***great***, *essential*, *possible*, *interesting*, *possible*, ***likely***, *ok/okay*, ***tempting***, *useful*. From this list, the bold items are also significantly attracted collexemes in the *too*-pattern, while the rest are either completely absent from, or even repelled (less frequent than expected) by the *too*-pattern.

With the exception of *hard*, *easy*, *difficult*, and *good*, none occur among the top 30 in [*too* ADJ *to* V]. However, the adjectives *late* and *early* must be counted as raising adjectives too; they were not identified by our second analysis because they cannot occur in the non-raised pattern in the absence of expressions like *too* or *enough* (cf. **It was late to stop*). In contrast, *good* is not a raising adjective, its occurrence in the pattern [*it* BE ADJ *to* V NP] has a different reason that we will return to below. Thus, five of the top 30 adjectival collexemes are raising adjectives, among them the top collexeme *late*. This shows that while the raised variant of the pattern is important in that it contributes some of the most strongly associated types in the *too*-pattern, these represent only a few highly frequent conventionalised expressions, so it does not seem very productive overall.

Let us turn to the *enough*-pattern. Table 4 presents the top 25 attracted adjectival collexemes of this pattern.

Again, all top collexemes are scalar, showing that this pattern serves to relate (the participant of) an event to a property conventionally thought of as scalar range (in this case, placing it inside the range, cf. again Jensen, 2014b).

The next observation is that there is systematic overlap between the top collexemes of the two patterns. Five adjectives occur on both lists (*big, small, good, old, stupid* and *easy*). In addition, there are another ten adjectives that form antonym sets that are systematically distributed across the patterns; the *enough*-pattern has *strong/powerful*, *brave/bold*, *confident* and *old*, the *too*-pattern has *weak*, *scared/afraid*, *embarrassed* and *young* (it has *old* in addition). Both types of overlap are expected based on the general semantic characterisation of these patterns, which makes them complementary. In some cases, a particular property may enable one

Table 4. Top 25 attracted collexemes in [ADJ *enough to* V]

ADJ	F_{corp}	O:E	G2
lucky	21348	2783:7.2	28200.21
fortunate	7669	1133:2.6	11727.70
strong	103276	1728:35	10209.34
large	198648	1517:67.2	6629.51
old	223518	1532:75.7	6378.06
smart	20825	811:7	6140.07
big	169934	1314:57.5	5766.39
good	596596	2003:201.9	5693.96
brave	8225	564:2.8	4917.26
stupid	21104	436:7.1	2741.81
powerful	49859	529:16.9	2633.83
easy	115942	650:39.2	2442.28
flexible	17433	350:5.9	2180.22
foolish	5272	234:1.8	1829.65
small	237445	650:80.4	1589.33
unfortunate	8976	232:3	1561.22
unlucky	2496	160:0.8	1370.85
confident	19504	254:6.6	1364.41
clever	11022	212:3.7	1301.69
high	229032	544:77.5	1194.40
mature	9056	179:3.1	1108.59
dumb	5817	158:2	1078.60
bold	8976	172:3	1054.61
daft	2100	112:0.7	917.22
intelligent	15125	176:5.1	906.24

event and prevent another; for example, there are events that require a participant to fall above or below a certain age, so we can get both [*old enough to* V] and [*too old to* V]. In other cases, a particular property is more likely to enable events while its opposite is more likely to prevent them. For example, strength is more likely to be necessary for an event while weakness would prevent it, so we get [*strong enough to* V] and [*too weak to* V].

However, this overlap is only partial – in addition, there are collexemes that are unique to each construction. Very strikingly, there are two larger semantic clusters among the top collexemes of the *enough*-pattern that are absent from the *too*-pattern: first, the cluster "luck" with *lucky, unlucky* and *unfortunate*, second, the even larger cluster "intelligence", accounting for almost a quarter of the top twenty-five collexems: *intellligent, smart*, *clever*, *stupid*, *dumb* and *daft*. This suggests a higher degree of idiomaticity for the *enough*-pattern than for the *too*-pattern: idiomatic phrases often have a set of synonyms in one of their slots (think of the pattern

drive someone {*crazy, mad, insane, bonkers, batty, nuts, mental*, …}, cf. Stefanowitsch and Flach, 2016). Such a higher degree of idiomaticity is also suggested by the fact that the top three collexemes of the *enough*-pattern are roughly twice as strongly associated with their pattern than the top collexemes of the *too*-pattern, suggesting that the pattern has fewer types but these types have a higher token frequency. Note the difference in meaning between, for example, (3a) and (3b):

(3) a. Name calling, what you do when you are *too stupid to make* an articulate argument.
b. I needed a job, and I was *stupid enough to give* them all of my bank account and personal information.

The semantics of (3a) is straightforwardly that described in Section 3 above: in order for someone to make an articulate argument, they need to fall into a certain range with respect to intelligence, and (3a) states that some people fall below the lower bound of that range and must hence resort to a different discussion strategy. Example (3b) may be superficially equivalent: one might argue that for someone to give a person your bank account and personal information requires them to fall into a certain range of intelligence characterised by the adjective *stupid*, and that (3b) states that the person referred to by *I* here falls within that range. However, this is not really what the sentence conveys pragmatically. Instead, it conveys that the speaker gave someone their information and that this turned out to have been a stupid thing to do. The same is true of other examples with adjectives from the two clusters just mentioned. For example, in (4) the speaker is not claiming that one needs to fall within a certain range of luckiness in order to attend the conference in question, but that they attended it and consider themselves lucky because of this:

(4) A few weeks ago I was *lucky enough to attend* the Flash on the Beach conference in Brighton.

Generally, the adjectives associated with the *enough*-pattern tend to be positively connoted, with the clear exception of the idiomatic clusters mentioned above (*stupid*, *foolish*, *unfortunate*, *unlucky*, *dumb*, *daft*); in this, the *enough*-pattern is complementary to the *too*-pattern.

Finally, note that there is only a single raising adjective, *easy*, among the top collexemes of the *enough*-pattern. This is unexpected, since the general semantic characterisation of the pattern does not lead us to expect a difference with respect to preferences for raising and non-raising adjectives. Again, this could be an indication of a higher degree of idiomaticity of the *enough*- versus the *too*-pattern.[2]

2. Note also that the *enough*-pattern, but not the *too*-pattern, may take nouns in the adjective slot that are then coerced to a scalar adjectival reading (e.g. *be man/gentleman enough to* V, cf. Jensen, 2015).

3.3 Case study: Distinctive collexeme analysis (DCA)

Method. In the preceding section, we compared the *too*-pattern and the *enough*-pattern based on two individual simple collexeme analyses, finding both similarities and differences. In order to get a better sense of the differences between two patterns, it may be desirable to compare them directly against each other rather than comparing them individually against the corpus as a whole. This is what distinctive collexeme analysis (DCA; see Gries and Stefanowitsch (2004a), building on Church et al. (1991)) is meant to achieve.

Table 5a shows the research design in schematic form. It is similar to that of a simple collexeme analysis, with the crucial difference that an item's occurrence in construction c_1 is assessed relative to its frequency in c_2, rather than to its corpus frequency. Hence, the table total consists of the overall frequency of both constructional alternatives; corpus size is not relevant for this method.

Table 5a. Schematic contingency table for distinctive collexeme analysis

		L		**Total**
		l	*!l*	
C	c_1	O_{11}	O_{12}	R_1
	c_2	O_{21}	O_{22}	R_2
	Total	C_1	C_2	N

Table 5b illustrates the research design using the adjective *early* (*l*) in the pattern [*too* ADJ *to* V] (c_1) and [ADJ *enough to* V] (c_2).

Table 5b. Contingency table for *early* in [*too* ADJ *to* V] and [ADJ *enough to* V]

		L		**Total**
		early	***!early***	
C	*too* ADJ *to* V	1,247 (607.3)	29,756 (30395.75)	31,003
	ADJ *enough to* V	19 (658.7)	33,613 (32973.25)	33,632
	Total	1,266	63,369	64,635

The observed frequency is much higher than the expected one for the *too*-pattern, and, correspondingly, much lower than expected in the *enough*-pattern. The G2 value is 1685.71 ($p < .0001$), which means that *early* is a *significantly attracted distinctive collexeme* of [*too* ADJ *to* V], or, conversely, a *significantly repelled distinctive collexeme* of the alternative construction [ADJ *enough to* V].

Results and discussion. Table 6 shows the results of a DCA between the two patterns (c_1 and c_2, respectively), listing the top 20 distinctive collexemes for each construction, with observed and expected values and the association measure (G2) for each collexeme.

Table 6. Top 20 distinctive collexemes for [*too* ADJ *to* V] and [ADJ *enough to* V]

too ADJ *to* V (c_1)				ADJ *enough to* V (c_2)			
ADJ	c_1(O:E)	c_2(O:E)	G2	ADJ	c_1(O:E)	c_2(O:E)	G2
late	2122:1020.2	5:1106.8	3131.86	*lucky*	0:1334.9	2783:1448.1	3751.48
early	1247:607.3	19:658.7	1685.71	*strong*	116:884.5	1728:959.5	1599.56
hard	949:497.4	88:539.6	919.78	*fortunate*	0:543.5	1133:589.5	1498.96
busy	516:253.7	13:275.3	657.43	*large*	335:888.3	1517:963.7	742.80
lazy	488:239.4	11:259.6	629.60	*brave*	2:271.5	564:294.5	717.80
long	925:530.5	181:575.5	619.61	*smart*	71:423.1	811:458.9	677.83
young	912:531	195:576	573.35	*flexible*	4:169.8	350:184.2	421.07
difficult	626:333.4	69:361.6	565.67	*powerful*	57:281.1	529:304.9	404.14
weak	433:217.3	20:235.7	501.35	*old*	575:1010.6	1532:1096.4	388.76
expensive	367:179.4	7:194.6	481.03	*been*	0:127.6	266:138.4	348.56
scared	356:174.1	7:188.9	465.15	*confident*	1:122.3	254:132.7	321.17
tired	330:161.6	7:175.4	427.70	*unfortunate*	0:111.3	232:120.7	303.89
afraid	208:100.2	1:108.8	294.97	*done*	0:100.7	210:109.3	275.01
dangerous	227:111.8	6:121.2	286.45	*foolish*	7:115.6	234:125.4	253.41
ill	187:89.7	0:97.3	275.36	*mature*	1:86.3	179:93.7	223.42
short	305:162.6	34:176.4	272.99	*bold*	3:83.9	172:91.1	199.20
happy	468:279.2	114:302.8	263.08	*unlucky*	1:77.2	160:83.8	198.72
numerous	242:123.3	15:133.7	261.72	*good*	1136:1505.7	2003:1633.3	186.14
eager	166:82	5:89	205.71	*clever*	17:109.8	212:119.2	181.38

The results are roughly comparable to those of the two simple collexeme analyses, but they are even clearer, as DCA ignores similarities and focuses on differences. In particular, the DCA shows even more clearly that the *too*-pattern is associated more strongly with negatively connoted adjectives and the *enough*-pattern with positively connoted ones. In particular, the adjective *good*, which is associated strongly with both patterns when compared against its corpus frequency, is clearly associated with the *enough*-pattern when the two patterns are compared directly. The exceptions are, on the one hand the cluster of adjectives of stupidity and unluckiness already discussed above, and, on the other hand, the adjective *happy*, which is sill attracted to the *too*-pattern even in the DCA (a point we will return to below).

3.4 Case study: Co-varying collexeme analysis (CCA)

Method. As the third major member of the family of collostructional analysis, co-varying collexeme analysis (CCA) is an explicitly syntagmatically-oriented method and thus especially useful for the analysis of phraseological patterns. It identifies items that co-occur significantly in two specific slots of a pattern (Stefanowitsch and Gries, 2005).

Table 7a shows the research design in schematic form. Note that this design is very close to traditional collocation analysis, except that it is based on structural rather than sequential co-occurrence (although some variants of collocation analysis also take structure into account, cf. Stefanowitsch and Gries, 2009).

Table 7a. Schematic contingency table for a co-varying collexeme analysis

		L_2 IN SLOT B		Total
		l_2 *in* S_B	$!l_2$ *in* S_B	
L_1 IN SLOT A	l_1 *in* S_A	O_{11}	O_{12}	R_1
	$!l_1$ *in* S_A	O_{21}	O_{22}	R_2
	Total	C_1	C_2	N

In the present context, the obvious slots to investigate are, of course, the adjective and verb slots in each pattern, since they are the slots that vary. Also, we have so far ignored the verbs, so including them into the analysis might help us explain some of the unexpected results mentioned in the preceding sections.

Table 7b illustrates this design for the adjective *big* and the verb *fail*, which co-occur in the notorious (but fully compositional) phrase *too big to fail*.

The combination is vastly more frequent than expected; the association measure G2 is 2865 ($p < .0001$). Note that the measure is much lower than in the above analyses; this is mainly due to the fact that the sample (N) is much smaller (it encompasses only the 31,003 instances of the pattern, ignoring the rest of the corpus).

Table 7b. Contingency table for *big* and *fail* in [*too* ADJ *to* V]

		L_2 IN SLOT B		Total
		fail	*!fail*	
L_1 IN SLOT A	*big*	415 (13.8)	538 (939.2)	953
	!big	33 (434.2)	30,017 (29,615.8)	30,050
	Total	448	30,555	31,003

Results and discussion. Table 8 lists the top 25 co-varying collexemes for the *too*-pattern, in descending order of attraction. It provides the overall frequencies of the adjective (f(adj)) and the verb in (f(verb)), as well as their combined observed and expected frequencies (O:E), together with the direction of association and the collostruction strength (G2).

Table 8. Top 25 adjective-verb-combinations in the pattern [*too* ADJ *to* V]

ADJECTIVE	VERB	F(ADJ)	F(VERB)	O:E	G2
good	*be true*	1136	766	703:28.1	4770.40
big	*fail*	953	448	415:13.8	2865.16
numerous	*mention*	242	264	124:2.1	917.63
early	*tell*	1247	334	191:13.4	811.17
early	*say*	1247	567	235:22.8	808.93
happy	*help*	468	232	113:3.5	658.42
young	*remember*	912	162	100:4.8	504.07
late	*save*	2122	152	120:10.4	497.96
important	*be left*	294	86	59:0.8	455.70
close	*call*	122	155	54:0.6	426.57
good	*miss*	1136	93	65:3.4	321.72
late	*change*	2122	148	91:10.1	302.27
hot	*handle*	209	171	45:1.2	264.62
large	*fit*	335	222	52:2.4	240.79
stupid	*understand*	361	240	52:2.8	223.83
late	*stop*	2122	136	74:9.3	220.39
poor	*afford*	207	48	28:0.3	219.50
easy	*forget*	942	46	37:1.4	215.05
young	*have*	912	608	103:17.9	212.00
poor	*pay*	207	132	34:0.9	197.00
dark	*see*	89	421	36:1.2	193.90
big	*fit*	953	222	63:6.8	187.38
cheap	*meter*	76	15	15:0	183.51
early	*judge*	1247	61	40:2.5	181.45
hard	*find*	949	217	61:6.6	180.57

As a general observation, note that the adjective-verb pairs tend to make sense intuitively – many of them encode typical combinations of properties and events that are prevented by these properties (good examples are *numerous/mention*, *young/remember*, *stupid/understand*, *poor/afford* or *big/fit*). This confirms the abstract meaning of the pattern as characterised in Section 2.1 above.

Many of the adjective-verb pairs that fit this characterisation are (semi-)fixed phrases, for example, the already mentioned *too big to fail*, as well as *too early to tell/say/judge*, *too close to call*, *too hot to handle* and *too cheap to meter*. The strong association between these adjectives and verbs demonstrates nicely that even compositional instances of a particular structure may become conventionalised if they serve a useful communicative purpose (i.e. if there are enough salient situations where they can be applied).

In addition, there are adjective-verb pairs that do not fit the abstract meaning of the construction very well. The top-ranked *too good to be true*, a highly conventionalised expression, is not intended to communicate that an event must fall within a specified range of goodness in order to be true; rather, it is used to express disbelief in contexts where an event is highly desired by the speaker. The high incidence of this fixed expression explains why the positively connoted adjective *good* is strongly associated with the otherwise rather negatively connoted *too*-pattern.

The other surprising exception to the negative semantic prosody of this pattern, *happy*, is particularly interesting in this respect. As Table 8 shows, this adjective is strongly associated with the verb *help*; however, we do not usually say someone is *too happy to help* – we say they are ***only*** *too happy to help*. Moreover, we don't intend this to communicate that that person falls above the range of positive emotions required to help (which would entail that they will refrain from helping), but that they fall *within* the range of positive emotions that would predispose them to help. Could it be that there is a systematic pattern of the form [*only too* ADJ *to* V] that overlaps structurally with the *too*-pattern but is semantically distinct? This is the type of question for which distinctive collexeme analysis, as introduced in the preceding section, is very useful.

Table 9 shows the results of a DCA of the adjective slot in the pattern [*only too* ADJ *to* V] compared with the pattern [*¡only too* ADJ *to* V] (i.e. the pattern [*too* ADJ *to* V] preceded by anything other than *only*).

A very clear pattern emerges: [*only too* ADJ *to* V] is exclusively associated with adjectives describing positive dispositions. Thus, this is indeed a separate pattern, formally distinguished by the presence of *only*. Semantically, this pattern means something like "fall within the range of positive dispositions necessary to help" (other verbs associated with *happy* are *assist, oblige, offer*, etc.). Note that this is actually the opposite of the compositional meaning, so the [*only too* ADJ *to* V] seems

Table 9. Significantly distinctive collexemes for the patterns [*only too* ADJ *to* V] and [*!only* ADJ *to* V]

***only too* ADJ *to* V (c_1)**				***[!only]* ADJ *enough to* V (c_2)**			
ADJ	**c_1(O:E)**	**c_2(O:E)**	**G2**	**ADJ**	**c_1(O:E)**	**c_2(O:E)**	**G2**
happy	347:10.4	121:457.6	2320.55	*late*	1:47.2	2121:2074.8	88.95
pleased	71:2.0	21:90.0	449.97	*early*	0:27.7	1247:1219.3	57.22
glad	58:1.6	15:71.4	372.95	*good*	0:25.2	1136:1110.8	52.03
willing	71:4.2	118:184.8	302.53	*small*	0:23.2	1046:1022.8	47.84
ready	23:2.3	81:101.7	69.49	*hard*	0:21.1	949:927.9	43.33
anxious	15:0.8	21:35.2	66.54	*long*	0:20.6	925:904.4	42.22
eager	24:3.7	142:162.3	52.54	*young*	1:20.3	911:891.7	33.49
keen	15:1.6	55:68.4	44.2	*old*	0:12.8	575:562.2	26.09
delighted	7:0.2	4:10.8	39.12	*big*	3:21.2	950:931.8	25.49
thankful	2:0.0	0:2.0	15.23	*busy*	0:11.5	516:504.5	23.39
				lazy	0:10.8	488:477.2	22.11
				difficult	1:13.9	625:612.1	21.08
				weak	0:9.6	433:423.4	19.6

to be a conventionalised case of the reversed interpretation discussed in connection with Example (1e) in Section 2.1 above.[3]

Turning to the *enough*-pattern, Table 10 lists the top 25 co-varying adjective-verb collexemes.

Again, many adjective-verb pairs confirm the general meaning of the pattern in that they encode typical combinations of properties and events that are enabled by these properties – including pairs that correspond almost exactly to those mentioned for the *too*-pattern above, like *old/remember*, *smart/know*, *rich/afford* and *small/fit*; additional examples are *honest/admit*, *large/accommodate, hot/melt* and many others.

In fact, almost all of the top 25 pairs instantiate the general semantics; the only systematic exception are cases with *lucky*: *lucky enough to have/get/live/find*. At first glance, this seems to contradict the suggestion made in Section 3.2 above, that the *enough*-pattern has more idiomatic instantiations than the *too*-pattern. But this contradiction is only an apparent one: the idiomatic behaviour is simply restricted

3. Note that Jensen (2014a) also notices the oddity of adjectives like *happy* in the *too*-construction and attempts to deal with them semantically by positing an "enablement" subsense in addition to the general sense (which he characterises as "prevention"). He tries to motivate the "enablement" subtype via general force-dynamic schemas in the sense of Talmy (2000), but he comments neither on the fact that this subtype cannot be derived compositionally, nor on the specific formal properties associated with it.

Table 10. Top 25 adjective-verb-combinations in the pattern [ADJ *enough to* V]

ADJECTIVE	VERB	f(adj)	f(verb)	O:E	G2
old	*remember*	1532	298	272:13.6	1555.14
lucky	*have*	2783	1690	574:139.8	969.67
small	*fit*	650	294	135:5.7	694.50
fortunate	*have*	1133	1690	272:56.9	497.99
honest	*admit*	140	124	54:0.5	446.65
lucky	*get*	2783	1067	305:88.3	394.26
smart	*know*	811	545	125:13.1	381.65
large	*accommodate*	1517	205	101:9.2	357.38
hot	*melt*	211	64	42:0.4	352.78
naive	*think*	107	242	40:0.8	261.07
serious	*warrant*	193	287	48:1.6	251.34
large	*hold*	1517	373	101:16.8	220.42
sensitive	*detect*	89	45	22:0.1	204.79
lucky	*live*	2783	185	88:15.3	201.42
good	*win*	2003	195	79:11.6	199.13
good	*play*	2003	202	79:12	192.86
naive	*believe*	107	189	30:0.6	189.67
lucky	*find*	2783	275	102:22.8	178.06
strong	*withstand*	1728	128	58:6.6	177.05
rich	*afford*	165	54	23:0.3	174.54
big	*hold*	1314	373	84:14.6	173.77
strong	*resist*	1728	57	40:2.9	170.66
broad	*cover*	104	174	25:0.5	152.96
cold	*freeze*	66	17	13:0	146.28
smart	*figure*	811	64	31:1.5	145.01

to the adjectives from the two clusters mentioned above, but these adjectives are not in turn associated strongly enough with specific verbs to be instantiated heavily among the top co-varying collexemes.

In other words, the idiomatic uses in the *enough*-pattern have an open verb slot: [*lucky/fortunate enough to* V], [*stupid/dumb enough to* V]. In contrast, the idiomatic uses in the *too*-pattern specify the verb slot either lexically, as in [*too good to be true*], or in terms of a semantic class, as in [*only too happy/willing/eager/… to* V_{HELP}].

3.5 Case study: Distinctive co-varying collexeme analysis (DCCA)

Method. As mentioned in the introduction, collostructional analysis has been extended to designs with three variables. One of these extensions (Stefanowitsch and Gries, 2005) combines co-varying collexeme analysis with simple collexeme analysis by comparing the co-occurrence of words in two slots of a pattern (the *into*-causative, as in *My mother **talked** me into **trying** yoga*) inside this pattern and in the corpus as a whole. This was an attempt to include the general frequency of co-occurrence of two words as a baseline.

When dealing with two related patterns, like the *too*- and the *enough*-patterns, it might be insightful to extend this design one step further to a combination of co-varying collexeme analysis and distinctive collexeme analysis, comparing the co-occurrence of words in one as opposed to the other. This would be a "distinctive co-varying collexeme analysis" that would allow us to group the results according to one of the words and then compare that word's co-varying collexemes in the two constructions. We performed such an analysis by submitting all triples of [adjective × verb × pattern] to a configural frequency analysis (using the R {cfa} package (Funke, 2007)), and then calculating the G2 test statistic for each triple (the {cfa} package outputs the chi-square test statistic, which is less useful as an association measure).

Results and discussion. As already observed in Section 2.2 above, there is only a limited overlap between the adjectives in the two patterns, which makes it all the more interesting to see how the patterns differ in their associations where they do share an adjective. Essentially, there are two different situations that we find in such cases, each of which will be briefly illustrated (an exhaustive analysis of all overlapping adjectives being beyond the scope of this paper).

The first situation is that expected on the basis of the general semantics of the two patterns: the *too*-pattern is associated with verbs referring to events that are impossible (inadmissible, etc.) if a central participant falls above (with positive polarity adjectives) or below (with negative polarity adjectives) a range referred to by the adjective, whereas the *enough*-pattern is associated with events that are possible (admissible, etc.) only if a central participant falls within this range. As an example, consider the adjective *hot* in Table 11. Its top distinctive co-varying collexemes in the *too*-pattern are *handle*, *touch* and *sleep*, i.e. events that are conventionally thought of as being made impossible by too much heat; in contrast, its top distinctive co-varying collexemes in the *enough*-pattern are events that are conventionally thought of as being made possible by heat – *melt*, *burn*, *fry*, etc.

Table 11. Distinctive co-varying collexemes of *hot* in [*too* ADJ *to* V] and [ADJ *enough to* V]

***too hot to* V**			***hot enough to* V**		
VERB	**O:E**	**G^2**	**VERB**	**O:E**	**G^2**
handle	45:1.19	269.52	*melt*	42:0.23	416.95
touch	23:0.22	180.49	*burn*	15:0.11	131.86
sleep	6:0.16	43.64	*fry*	7:0.03	76.38
walk	10:0.57	42.23	*kill*	4:0.32	60.82
bake	3:0.02	25.48	*cook*	8:0.15	56.58
wear	4:0.30	21.27	*ignite*	5:0.04	42.84
play	4:1.06	16.25	*destroy*	3:0.19	42.18
eat	5:0.74	12.48	*evaporate*	4:0.03	36.27
think	4:1.35	10.13	*weld*	4:0.03	33.55
stay	2:0.50	9.66	*forge*	4:0.05	33.18
			scald	2:0.01	31.12
			glow	3:0.02	30.85
			soften	3:0.02	28.24
			boil	3:0.02	26.31
			be auctioned	2:0.01	23.47
			fuse	2:0.01	23.47
			deform	2:0.02	21.10
			put	6:1.37	18.85
			flow	2:0.02	15.88
			set	3:0.28	12.21
			guarantee	2:0.09	11.78

The second situation is one brought about by idiomatic uses of certain adjectives. As discussed in Sections 2.2–2.3 above, these seem to be pattern-specific – adjectives used idiomatically in one of the two patterns are not used idiomatically in the other.

For example, the adjective *good* was seen to be associated with the *too*-pattern due to the idom *too good to be true*. As Table 12 shows, this pattern, together with a variant *too good to miss/pass/refuse* is also responsible for the distinctive co-varying collexemes of *good* in the *too*-pattern. In contrast, in the *enough*-pattern, *good* is associated with verbs that encode events that literally require a central participant to fall within a particular range of goodness.

Table 12. Distinctive co-varying collexemes of *good* in [*too* ADJ *to* V] and [ADJ *enough to* V]

***too good to* V**			***good enough to* V**		
VERB	**O:E**	**G^2**	**VERB**	**O:E**	**G^2**
be true	703:17.94	6047.27	*win*	79:5.61	517.78
miss	65:2.73	560.13	*play*	79:8.59	388.98
pass	51:5.36	369.34	*beat*	38:1.84	353.61
refuse	10:0.70	230.85	*make*	97:39.07	298.05
be threw	4:0.09	218.49	*eat*	43:6.04	278.66
			get	107:46.37	268.98
			compete	20:2.38	235.38
			be re-used	3:0.08	206.98

The situation is reversed, for example, with the adjective *dumb*. As Table 13 shows, this adjective is associated in the *too*-pattern with mental activities that are literally made impossible if a central participant falls above an upper limit of stupidity. In contrast, it is associated in the *enough*-pattern idiomatically with events that do not require a particular level of stupidity, but that are often *evaluated* as stupid after the fact.

Table 13. Distinctive co-varying collexemes of *dumb* in [*too* ADJ *to* V] and [ADJ *enough to* V]

***too dumb to* V**			***dumb enough to* V**		
VERB	**O:E**	**G^2**	**VERB**	**O:E**	**G^2**
notice	3:0.18	85.36	*believe*	10:0.59	64.92
understand	11:0.90	53.84	*think*	14:0.89	64.35
realize	4:0.33	26.80	*fall*	8:0.23	51.81
succeed	2:0.05	22.71	*try*	8:0.59	43.80
figure	3:0.25	21.94	*buy*	7:0.52	33.33
			plagiarize	2:0.01	31.81
			sign	4:0.10	31.57
			vote	3:0.19	29.65
			be caught	2:0.03	23.97
			point	2:0.10	23.57
			put	4:0.83	23.33
			pay	4:0.51	23.03
			install	2:0.06	22.84
			send	3:0.26	20.50
			walk	2:0.38	18.76

In sum, the distinctive co-varying collexeme analysis helps to identify (or, if already known, confirm) both the abstract semantic characterisations of two patterns such as the *too*- and the *enough*-pattern – as manifest in the complementing lists of adjectives in Table 11. It also helps identify idiomatic expressions on the basis of asymmetrical, non-complementing lists of adjectives as in Tables 12 and 13.

4. Summary

This paper illustrates the contribution that the family of collostructional methods can make to questions of phraseological interest. In contrast to much of the collostructional literature which conventionally addresses constructional alternations, the two patterns we analysed here are quite typical for phraseological investigations, in that the [*too* ADJ *to* V] and [ADJ *enough to* V] patterns are complementary (rather than 'synonymous'). Thus, while our discussion was not meant to furnish an exhaustive analysis of the two patterns (and many interesting aspects had to be left unexplored for the present), by the very design of collostructional analysis for the investigation of associations between words and patterns, the discussion provided linguistic and methodological pointers for the usefulness of the collostructional method as part of the phraseologists' methodological toolkit.

First, the method enables an objective confirmation of observations that could be – or have previously been – arrived at intuitively by qualitative analysis, such as the general characterisation of the two patterns as encoding events made impossible because a participant falls outside a particular range (in the *too*-pattern) or made possible because a participant falls within a particular range (for the *enough*-pattern).

Second, the focus on statistical similarities and differences *in the actual usage* of these patterns reveals pattern-specific properties that are difficult to detect by intuition-based qualitative analyses. A case in point is the systematic distribution of antonym pairs across the patterns in addition to their often only partial overlap, which – while not unexpected given the patterns' general semantic characterisation – would remain largely subjective in non-empirical analyses.

Third, some insights are *only* available through systematic quantitative investigation. The one single finding that stands out here is the tendency of the *enough*-pattern towards a much greater degree of idiomaticity, evidenced in its tendency towards fewer, but internally more coherent semantic clusters, revolving around luck and intelligence, as well as a general skew towards stronger associations of a small(er) number of types in high frequency, conventionalised expressions.

Finally, the statistical analysis of usage data yielded findings that contradict the patterns' general semantics, thus forcing the analyst to confront findings that might

have been convenient to ignore in an intuition-based analysis. These were shown here to represent either highly conventionalised expressions that are compatible with, but not exhaustively characterised by the general meaning of the respective pattern (e.g. *too good to be true* or *dumb enough to* V) or that constitute fully idiomatic sub-patterns (*only too happy/eager/willing/etc. to* V).

Quantitative corpus-linguistic methods like collostructional analysis cannot, of course, replace qualitative analyses drawing, among other things, on the researcher's introspection. However, they can, and should, provide a rigorous methodological frame for identifying those aspects of a linguistic phenomenon that need to be accounted for in the first place.

References

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 115–164). Hillsdale, NJ: Lawrence Erlbaum Associates.

Evert, S. (2004). *The statistics of word cooccurrences. Word pairs and collocations*. (PhD Thesis, Universität Stuttgart). http://www.stefan-evert.de/PUB/Evert2004phd.pdf.

Evert, S. (2010a). *Corpus Encoding Tutorial*. http://cwb.sourceforge.net/.

Evert, S. (2010b). *CQP query language tutorial (CWB Version 3.0)*. http://cwb.sourceforge.net/.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Flach, S. (2016). *Collostructions: An R implementation for the family of collostructional methods. Version 0.0.7*. http://bit.ly/sflach.

Flach, S. (2015). Let's go look at usage: A constructional approach to formal constraints on go-VERB. In T. Herbst, & P. Uhrig (Eds.), *Yearbook of the German Cognitive Linguistics Association (Volume 3)* (pp. 231–252). Berlin: De Gruyter Mouton.

Funke, S. (2007). *The cfa Package. Version 0.9–3*. https://cran.r-project.org/web/packages/cfa/.

Gries, S. Th., & Stefanowitsch, A. (2004a). Extending collostructional analysis: A corpus- based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1), 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Gries, S. Th., & Stefanowitsch, A. (2004b). Covarying collexemes in the into-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225–236). Stanford, CA: CSLI.

Hilpert, M. (2014). Collostructional analysis: Measuring associations between constructions and lexical elements. In D. Glynn, & J. A. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy (Human Cognitive Processing*, vol. 43) (pp. 391–404). Amsterdam: John Benjamins. https://doi.org/10.1075/hcp.43.15hil

Jensen, K. E. (2014a). *Too female to be ruthless* and *too pregnant to argue*: Semantic conflict and resolution in the [*too* ADJ *to* V]-construction. *Suvremena Lingvistika*, 40(77), 1–26.

Jensen, K. E. (2014b). This construction is *too hot to handle*: A corpus study of an adjectival construction. In *Proceedings of the 14th Annual Meeting of the Japanese Cognitive Linguistics Association* (pp. 740–748). Kyoto: Japanese Cognitive Linguistics Association.

Jensen, K. E. (2015). Adjectives and usage-patterns in the [X enough to VERB]-construction. *International Cognitive Linguistics Conference (ICLC-13)*. Northumbria University, Newcastle.

Liberman, M. (2009). No detail too small. *Language Log*. http://languagelog.ldc.upenn.edu/nll/?p=1924.

Meier, C. (2003). The meaning of *too, enough*, and *so… that*. *Natural Language Semantics*, 11(1), 69–107. https://doi.org/10.1023/A:1023002608785

Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference* (pp. 188–200). Austin, TX: South-Central SAS Users Group.

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 486–493). Istanbul: ELRA.

Stefanowitsch, A. (2011). Cognitive linguistics meets the corpus. In M. Brdar, S. Th. Gries, & M. Ž. Fuchs (Eds.), *Cognitive Linguistics: Convergence and expansion (Human Cognitive Processing*, vol. 32) (pp. 257–290). Amsterdam: John Benjamins. https://doi.org/10.1075/hcp.32.16ste

Stefanowitsch, A. (2013). Collostructional analysis. In T. Hoffmann, & G. Trousdale (Eds.), *The Oxford handbook of construction Grammar* (pp. 290–306). Oxford/New York: Oxford University Press.

Stefanowitsch, A., & Flach, S. (2016). The corpus-based perspective on entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning. How we reorganize and adapt linguistic knowledge*. (Language and the Human Lifespan). Berlin/New York: De Gruyter Mouton.

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Stefanowitsch, A., & Gries, S. Th. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1), 1–43. https://doi.org/10.1515/cllt.2005.1.1.1

Stefanowitsch, A., & Gries, S. Th. (2008). Channel and constructional meaning: A collostructional case study. In G. Kristiansen, & R. Dirven (Eds.), *Cognitive Sociolinguistics. Language variation, cultural models, social systems* (pp. 129–152). Berlin/New York: Mouton de Gruyter. https://doi.org/10.1515/9783110199154.2.129

Stefanowitsch, A., & Gries, S. Th. (2009). Corpora and grammar. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics: An international handbook, Vol II (Handbooks of linguistics and communication science* 29) (pp. 933–952). Berlin/New York: Mouton de Gruyter.

Talmy, L. (2000). *Toward a cognitive semantics, vol. 1: Concept structuring systems*. Cambridge, MA: MIT Press.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology*, 31(4), 591–597. https://doi.org/10.1080/14640747908400750

Multi-word patterns and networks

How corpus-driven approaches have changed our description of language use

Kathrin Steyer
Institut für Deutsche Sprache

This paper discusses a theoretical and empirical approach to language fixedness that we have developed at the Institut für Deutsche Sprache (IDS) ('Institute for German Language') in Mannheim in the project Usuelle Worterbindungen (UWV) over the last decade. The analysis described is based on the Deutsches Referenzkorpus ('German Reference Corpus'; DeReKo) which is located at the IDS. The corpus analysis tool used for accessing the corpus data is COSMAS II (CII) and – for statistical analysis – the IDS collocation analysis tool (Belica, 1995; CA). For detecting lexical patterns and describing their semantic and pragmatic nature we use the tool lexpan (or 'Lexical Pattern Analyzer') that was developed in our project. We discuss a new corpus-driven pattern dictionary that is relevant not only to the field of phraseology, but also to usage-based linguistics and lexicography as a whole.

Keywords: pattern-based lexicography, German reference corpus, phraseology, language fixedness, multiword expressions

1. Introduction

This paper discusses a theoretical and empirical approach to language fixedness that we have developed at the *Institut für Deutsche Sprache* (IDS) ('Institute for German Language') in Mannheim in my project *Usuelle Worterbindungen* (UWV) over the last decade.[1] Our research has always had two main areas on focus: (i) corpus linguistic exploration of phraseological phenomena, and (ii) new forms of online

1. Special thanks to Annelen Brunner and Marcas Mac Coinnigh for reading this manuscript and for giving valuable advice concerning the correctness and comprehensibility of this English version.

https://doi.org/10.1075/ivitra.24.14ste

lexicographic representation. From the beginning our work has been driven by the core question, "How can we interpret the results of quantitative analyses in a qualitative way?". Following in the comprehensive work of John Sinclair (cf. Herbst et al., 2011, Granger and Meunier, 2008), of Elena Tognini-Bonelli (2001) and many other scholars of corpus linguistics, we also distinguished between 'corpus-based' and 'corpus-driven' approaches and considered ourselves "corpus-driven human beings".[2] Drawing such strict lines is undoubtedly important in the initial development of a new research field. But nowadays we know that the distinction corpus-based vs corpus-driven seems to focus solely on the degree by which we allow ourselves to be led by the data – and the borders are fuzzy. We use the term 'corpus-driven' in a broader sense: a bottom-up corpus linguistic approach that allows us to find typical patterns by collecting many similar cases of usage. Looking at many cases of usage does not mean describing what is already known and visible, it means discovering hidden structures. Not only do we find more data, but we also detect new interrelations, unusual cross-connections, and surprising relationships and networks.

The following analysis is based on the *Deutsches Referenzkorpus* ('German Reference Corpus'; DeReKo)[3] which is located at the IDS. The corpus analysis tool used for accessing the corpus data is *COSMAS II* (CII) and – for statistical analysis – the IDS collocation analysis tool (Belica, 1995; CA). For detecting lexical patterns and describing their semantic and pragmatic nature we use the tool *lexpan* (or 'Lexical Pattern Analyzer') that was developed in our project (see Section 4.3).

Despite our focus on German resources, the principles of qualitative interpretation that we discuss should be transferable to other corpora and languages and other tools like the collocation analysis function in Sketch Engine (SkE).

2. The rocky road of qualitative interpretation

We imagined that with corpora everything would be better, faster, larger and – most importantly – more accurate in regard to understanding and describing language use. We thought that we could compile new corpus-based dictionaries and lexical information systems that would capture all linguistic aspects. Collecting and restructuring mass data with machine-aided methods in the last 30 years allowed linguists to discard their old-fashioned slip boxes and to look for samples of real language use in new quantitative dimensions – an empirical revolution. The early

2. This phrase was coined by Patrick Hanks in his plenary speech at the Malaga EUROPHRAS conference in 2015.

3. For this paper, I used the largest DeReKo subcorpus of written language, the W archive with a size of about 8 billion word forms in 2015 (DeReKo 2015-II) (Institut für Deutsche Sprache).

euphoria was short-lived, however, as it soon became clear that the empirical road would be more tricky to navigate than initially expected. The borders between 'langue' and 'parole' become more and more fuzzy. If we look at the data, we find that syntactical phenomena are deeply motivated by pragmatics; we have to distinguish between ad hoc constructions and frequent usage, etc. Computational linguists gave us extensive databases and sophisticated tools. But the human researcher, the linguist and the lexicographer, cannot always keep pace in the face of mass data. If one gets, for instance, a search result list of 100,000 occurrences or more, one will need a navigation system that structures the results and shows the most typical clusters of usage. Another reason for this growing gap is the fact that the results of computer processes are often a black box for the human interpreter. To give an example: When interpreting results of statistical collocation analyses like the IDS CA we should not overestimate the ranking of a specific lexical item in a single collocation profile. Such profiles are always just snapshots at a particular time based on a specific corpus. A second analysis can elevate other lexical items up to the higher ranks while previous results are relegated. Nevertheless, the important point is that the underlying pattern is stable, despite the correlation of a specific lexical unit to another lexical unit.[4] The designation 'underlying pattern' means that collocations have to be interpreted first and foremost as correlations between groups or classes of units, not between specific words. Without knowing this fundamental principle a user could run into danger of misinterpreting the results or questioning the method as a whole.

Patterns are the key to why speakers understand each other in everyday communication in spite of ever-changing lexical material, syntactic variance and strong idiosyncrasies. The idea that language production and learning work by repetition, recurrence, and imitation on the basis of patterns, templates, and schemas is not new. It has been discussed at length, especially in cognitive linguistics and language acquisition research, as well as in research fields like pattern grammar (Hunston and Francis, 2000) and corpus pattern analysis (Hanks, 2013) – I won't go into the long history of terms like 'multiword units', 'fixed expression', 'formulaic language', 'schemes', 'patterns' and so on. For this I recommend the very readable paper by Hans-Jörg Schmid (2014). Of course, automatic pattern retrieval is not a new invention either, e.g. it is used extensively in data mining.

But regarding a *qualitative* reconstruction of hidden patterns in language use and their applications in lexicography and second language teaching, we are just at the beginning.

This chapter discusses this central pattern-based change of perspective in phraseology and beyond from the point of view of corpus linguistics.

---

4. The only exception is a fixed correlation between specific words that means words cannot be substituted by other words. Those units are strongly lexicalised.

3. Kinds of lexical fixedness

3.1 From multiword expressions to patterns

Phraseology as a sub-discipline of lexicology has a long tradition.[5] The following central research question runs through the history of phraseology until today: Under which circumstances does a sequence of words become a *holistic unit*, a *lexical item*, a *lexicon entry*? The answers have been varied and depended to a significant degree on the predominant linguistic paradigms.

The rise of corpus linguistics fundamentally expanded the subject area of phraseology and the external perception of this discipline. It became more and more evident how essential multiword expressions (MWEs) are for understanding language use itself. In addition to this it became apparent that all – sometimes even competing – concepts of MWEs are based on the same fundamental principle of language, namely linguistic frozenness and fixedness. Compositional collocations and idioms, for example, differ in their degree of lexical fixedness and semantic opacity, their recognisability and prototypicality. Nevertheless, they all share one important characteristic: they are autonomous units that fill a specific role in communication. There is no core and no periphery. The difference is only in the degree of recognisability for the observer. These word clusters did not become fixed expressions by chance, but because speakers required an economical way to complete communicative tasks. Against this background we proposed the term 'usuelle Wortverbindungen' ('multiword expression in common use'; UWV)[6] in Steyer (2000). UWVs are conventionalised patterns of language use that manifest themselves in recurrent syntagmatic structures. This includes not only idioms and idiosyncratic structures, but all multiword units that have acquired a distinct function in communication. Our focus is on real-life usage, pragmatics and context. The central characteristic is the autonomous status as a communicative and entrenched cognitive unit[7] (see Section 3.2).

---

5. Concerning phraseology as an independent discipline of linguistics I would like mention the two volumes of the *International Handbook of Phraseology* edited by Burger et al. (2007), the *Einführung in die Phraseologie* ('Introduction to Phraseology') edited by Burger (2015 in its 5th edition) and the two volumes of the De Gruyter International Bibliography edited by Mieder in 2009. For some years, a special journal for phraseology, the *Yearbook of Phraseology*, has been published by De Gruyter Mouton. A reference book for English phraseology is Moon (1998) and – of course – Sinclair (1991). I also mention Herbst et al. (2011), Granger and Meunier (2008) and Gries (2008). Information about the "phraseological community" can be found on the website of the European Society of Phraseology (EUROPHRAS) (www.europhras.org).

6. English word-for-word translations of German are put in single quotation marks and in brackets.

7. The term 'entrenchment' is one of the central concepts of cognitive grammar, first and foremost in Langacker's research (e.g. Langacker 1987).

In recent times, there has been a perceivable shift not only in phraseology and multiword research but also in usage-based linguistics as a whole (cf. Steyer, 2015). Due to the ability to detect invisible structures based on linguistic mass data, it has been shown that phrasemes, idioms, and frozen sentences like proverbs are not as singular and unique as was often assumed in phraseology in the past. Usually, they are specific lexical realisations of templates, more noticeable and more fixed than ad hoc formulations, but not unique. Such templates or patterns emerge from repeated usage and can be instantiated with ever-changing lexical elements, both phraseological and non-phraseological. In addition to corpus linguistics, construction grammar (CxG) also played a part in this paradigm shift in phraseology and it was against this backdrop that Dobrovol'skij introduced, for example, the term 'phraseme construction' in 2011.

Despite the fact that we share many commonalities with the CxG we prefer the term 'Wortverbindungsmuster' ('multiword patterns[8]; MWP) (cf. Steyer, 2013, 2016) as a subtype of the general term 'lexikalisch geprägte Muster 'lexical patterns'; LP) (Steyer 2018).[9] Our pattern concept focuses much more on structures and interrelations of lexical items and wants to contribute to a usage-based theory of lexis. This approach arose from the tradition of phraseology as a genuine discipline of lexicology. That does not connote the negation of the syntax level. Naturally, our explorations are based on syntactic structures. But the dominance of the syntactic view can induce us to overlook the complexity of lexical phenomena. Probably this is a heuristic problem of analysis: One cannot observe all phenomena with the same intensity but has to fade some of them into the background (e.g. syntactical phenomena) for a much clearer observation of others (lexical structures and networks).[10]

Multiword patterns are conventionalised lexical schemes that are frozen by recurrent use. Recurrence is defined as the repeated appearance of similar linguistic structures in comparable contexts (cf. "geprägte komplexe Ausdrucksmuster" by Feilke, 1996, p. 187). Multiword patterns consist of fixed lexical components as well as obligatory slots that can be filled with specific entities (Renouf and Sinclair, 1991). These fillers have similar semantic and/or pragmatic characteristics, but do not necessarily belong to the same morpho-syntactic category. Sometimes all that they have in common are functional characteristics, which cannot be captured by traditional ontologies. Speakers are able to recall those schemes as lexicon entries

8. Biber also use ous the term 'multi-word pattern', but much more in the sense of multiword formulaic sequences like *it should be noted* or *as we have seen* (Biber, 2009).

9. Another type of lexical patterns is the so called 'sentence pattern', primarily proverb patterns like *There is more than one way to* VP (*skin a cat*) or *He who* V V (*He who want reap must sow*).

10. The discussion between phraseologists and construction grammarians to find out a common ground has been gathering momentum over the last years.

and fill the gaps in a specific communicative situation in a functionally adequate way. The MWP concept focuses on the semantic and functional restrictions of the slot fillers much more than any other scheme or pattern theory.

3.2 MWPS as autonomous units

The main criterion for a multiword pattern is that it has a *holistic quality* that gives it a status as an autonomous unit. 'Holistic quality' does not necessarily mean that it is idiomatic. The MWP can just have a specific function, even in a very abstract sense. Example (1) illustrates which components are mandatory for the autonomous status:

(1) [*in* ADJ *Zeit*]
['in ADJ time']
ADJ fillers: *absehbarer* ('foreseeable') / *kurzer* ('short') / *nächster* ('next')

The adjective slot is mandatory for the holistic meaning, 'forthcoming'. In German, we cannot reduce this pattern to a binary MWE **in Zeit* ('in time'). By contrast, the MWE *mit Genugtuung* ('with satisfaction') is a binary autonomous lexical item (P+N) with the meaning 'positively perceived'. Its recurrent internal adjective fillers, e.g. *großer* ('great') → *mit großer Genugtuung* ('with great satisfaction') only modify the core meaning as specific context markers (see Section 3.3).

Example (2) shows the functional nature of slot types and the distinct context restriction of MWPs:

(2) [*allen* N *zum Trotz*]
['despite all N']

The most frequent N filler group includes speech acts in a broader sense like the fillers *Vorhersagen* ('predictions'), *Prognosen* ('forecasts') or *Einwänden* ('objections') (pattern: [*allen Vorhersagen* / *Prognosen* / *Einwänden zum Trotz*] '*despite all predictions* / *forecasts* / *objections*'). Some of these have explicit negative connotations: *Unkenrufen* ('cries of naysayers'), *Getöses* ('hullabaloo') or *Horrormeldungen* ('horror stories'). Sometimes they are extended by adjectives, e.g. *anderslautenden Gerüchte* ('contrary rumours'), *düsteren Prognosen* ('dark predictions') or *vollmundigen Versprechungen* ('overblown promises'), e.g. *despite all overblown promises*. Another filler group includes references to speakers, like *Zweifler* ('sceptics'), *Kritiker* ('critics') or *Pessimisten* ('pessimists'), typically with some degree of negative connotation. Despite the variance of the fillers the MWP [*allen* N *zum Trotz*] has a holistic meaning: 'although something has been expected to go a certain way, it has turned out differently', from positive to negative and vice versa.

3.3 Extended context patterns (ECPS)

ECPs are recurrent – but not mandatory – context extensions. The context partners can appear inside an autonomous lexical unit as well as in its immediate surroundings (external ECP). Let's take a look at the example from the previous section again: [*mit* X *Genugtuung*] is an ECP of *mit Genugtuung* ('with satisfaction') with the following three frequent extension types (among others)[11]:

(3) *mit großer Genugtuung*
tiefer
grimmiger
('with great / deep / grim satisfaction')

(4) *mit Stolz und Genugtuung*
Freude und
Häme und
('with pride and / joy and / scorn and satisfaction')

(5) *mit einem Hauch von Genugtuung*
einem Anflug von
einem Schuss
einer Prise
('with a hint of / a touch of / a shot of / a pinch of satisfaction')

The extensions in (3) and (4) have two functions: intensification and/or connotation. (5) illustrates a third group of internal extensions, so called syntagmatic connotative quantifiers. These extensions express the observation of a rather hidden emotional reaction from the speaker who perceives something positively.

A recurrent external ECP in postposition of the MWE *mit Genugtuung* is the combination with verbs that refer to communicative or cognitive acts embedded in a '*that*-clause', e.g. [*mit Genugtuung zur Kenntnis nehmen* ('take note') / *feststellen* ('see') / *registrieren* ('notice'), *dass* ('that')…].

All these functional restrictions cannot be predicted *a priori* and do not always follow rules. They can only be discovered by an inductive reconstruction based on large corpora and sophisticated automatic methods.

---

11. The English equivalent behaves very similarly and also has comparable recurrent ECPs [*with* X *satisfaction*].

4. Corpus-linguistic methodology and interpretation

We now explain our iterative corpus linguistic methodology using examples of qualitative interpretation. Our empirical approach includes several steps, each with a specific explanatory potential: (a) complex phrase searches and reciprocal analyses are using *COSMAS II*; (b) IDS CA; (c) slot-filler analysis with *lexpan*.

4.1 Corpus searches

We look at the data, and then we hypothesise about the fixedness, variance and function of the MWE or MWP. Based on these findings we then configure the search strategy and return to the corpus. This can be repeated for several cycles. Thus, we study the nature of MW patterns by exploring KWIC concordances of multiword units. As a result of our focus on syntagmatic word surface structures, our approach is guided by the following two principles: First, we made the decision that searching without grammatical annotations follows our firm conviction that many MWPs cannot be found based on tagged corpora. MWPs often don't conform to syntactical phrases like NPs or PPs and traditional word classes change their function in a MWP. MWPs are primarily *functional* lexical buildings blocks.[12]

Our second principle is that we always use the word form – not the lemma – in our searches. As Sinclair already emphasised:

> There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity. (Sinclair, 1991, p. 8)

We will explain the distinctive use of word forms in the next section.

4.2 Collocation profiles

The IDS CA can be used to detect significant word pairs and MWEs as well as recurrent syntagmatic context patterns. This method enables us to identify typical aspects of meaning and usage of a MWE or a MWP – the extension of the principle of contextualism to multiword phenomena.

Figure 1 shows small snippets of collocation profiles of the word forms *Grund* (1,646,568 hits), *Grunde* (198,390 hits) and *Gründen* (559,751 hits):

12. Therefore, the untagged DeReKo corpus is our preferable resource.

Kookkurrenzpartner	LLR	Frequenz	Syntag pattern	Kommentar	Kwics
warum	1569	325	99% ein |der Grund […] warum	why	Kwics
gutem	901	137	100% aus |Aus mit gutem […] Grund	good	Kwics
genug	809	213	97% Grund […] genug für …	enough	Kwics
Feiern	714	107	99% Grund zum Feiern	celebrating	Kwics
Freude Doppelten	516	4	100% Doppelten Grund zur Freude	joy double	Kwics
Freude	516	133	96% Grund zur Freude	joy	Kwics
Boden	481	139	98% in Grund [und] Boden	bottom	Kwics
weshalb	422	89	98% ein |der Grund […] weshalb die …	wherefore	Kwics
Jubeln	315	36	100% Grund zum Jubeln	cheering	Kwics
Kookkurrenzpartner	**LLR**	**Frequenz**	**Syntag pattern**	**Kommentar**	**Kwics**
genommen	8163	1142	100% im |Im Grunde […] genommen	taken	Kwics
ist egal	683	45	51% ist mir |es im Grunde […] egal	doesn't matter	Kwics
ist	683	1406	54% im |Im Grunde […] ist	is	Kwics
Herzens	550	54	92] im |Im Grunde ihres |seines Herzens	of the heart	Kwics
gelegt	392	94	100% zu Grunde […] gelegt	laid	Kwics
ja anderes	377	4	50% ja im Grunde … anderes	indeed different	Kwics
ja	377	136	60% Im |im Grunde […] ja	indeed	Kwics
Kookkurrenzpartner	**LLR**	**Frequenz**	**Syntag pattern**	**Kommentar**	**Kwics**
gesundheitlichen	6697	611	99% aus gesundheitlichen […] Gründen	health	Kwics
beruflichen	4069	445	100% aus beruflichen […] Gründen	professional	Kwics
finanziellen	3130	430	99% aus finanziellen […] Gründen	financial	Kwics
ungeklärten bislang	1142	16	100% Aus |aus bislang […] ungeklärten Gründen	as yet unkown	Kwics
unerfindlichen	1007	70	100% aus unerfindlichen Gründen	mysterious	Kwics

Figure 1. CA profiles *Grund – Grunde – Gründen* (snippets) (CII random sample of 10,000)[13]

13. These CA snippets are visualisations created by *lexpan* after exporting the data from COSMAS II. English translations for the collocation partners have been added in the column "Kommentar". The main principles of the CA are described on the CA Website of the IDS project 'Methoden der Korpusanalyse und –erschließung' http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html.

As we can see, there are significant differences in the highest LLR-ranks of the CA profiles: The common singular form *Grund* has a wide range of syntagmatic patterns, e.g. *Grund, warum* ('reason why'), *aus gutem Grund* ('with good reason'), *Grund genug* ('reason enough'), [*Grund zum|zur* N (*Feiern / Freude / Jubeln*)] ('cause for N (celebrating / joy / cheering)'). There are also fixed MWEs like the intensifier *in Grund und Boden* ('in ground and bottom') which means 'totally'.

In contrast to this, the collocation profile of the – much less common – singular form *Grunde* is strongly focused on realisations of the preposition noun combination *im* plus *Grunde*: (i) as adverbial MWEs *im Grunde* and *im Grunde genommen*. Both MWEs can be translated as 'basically', (ii) as a MWP with the function of intensification [*im Grunde* PRON *Herzens*] ('at the bottom of PRON heart'). These three entities can also take the meaning 'eigentlich' ('actually').[14] We can see significant context partners like *ist egal -> im Grunde ist es egal* ('doesn't matter -> actually it doesn't matter') or *nichts anderes als* ('nothing other than -> actually it is nothing other than').

The third snippet shows the CA profile of the plural form *Gründen*. It is again very different. The context partners also indicate a strong restriction. Adjective collocation partners like *gesundheitlichen* ('health'), *beruflichen* ('professional') or *finanziellen* ('financial') are very dominant and they all come from realisations of the recurrent syntagmatic pattern [*aus* ADJ *Gründen*] ('for ADJ reason') (see Section 4.2).[15] To verify this observation we configure our search query for *Gründen* to yield only those occurrences without the preposition *aus* (in a range of up to five words before *Gründen*). Another useful method is to compare automatically the CA profiles of a MWE with the profile of a single lexeme with a similar meaning and function (see Figure 2). Using this strategy we can figure out the overlapping contexts and the differences in usage between these lexical units. Figure 2 illustrates a comparison of the CA profiles of the MWE *im Grunde* and the single lexeme *eigentlich* which can be translated as 'actually, basically'.

When comparing the CA profile of *eigentlich* with the CA profile of *im Grunde* we can see that although many clusters are similar, some contexts are strongly preferred by *eigentlich* and not highly ranked for *im Grunde*.

For example, there are adjectives in upper case which commonly appear before *eigentlich*, e.g. *Schade* ('Pity') (*Schade eigentlich* 'Pity actually'), *Komisch* ('Odd') (*Komisch eigentlich* 'Odd actually'), *Merkwürdig* ('Strange, Curious') (*Merkwürdig*

14. Another significant partner is the form *gelegt* ('laid'). This indicates an inflected form of the functional verb *zu Grunde legen* ('to take as a basis').

15. Most of these adjective collocation partners do not appear in the profiles for the singular forms.

Kookkurrenzpartner	LLR	Frequenz	Syntag pattern	Kommentar	Kwics
genommen	11287	1473	99% im Grunde genommen	taken	Kwics
ist	1189	1974	72% ist […] im Grunde …	is	Kwics
Herzens	719	88	100% im Grunde seines |ihres Herzens	of the heart	Kwics
gar nicht	479	128	89% im Grunde gar […] nicht	not at all	Kwics
ja	456	257	59% ist ja […] im Grunde …	indeed	Kwics
anderes	411	122	96% ist im Grunde nichts anderes als	different	Kwics
egal	280	73	97% ist im Grunde … egal	doesn't matter	Kwics
Kookkurrenzpartner	**LLR**	**Frequenz**	**Syntag pattern**	**Kommentar**	**Kwics**
gar nicht	789	212	93% eigentlich […] gar […] nicht	not al all	Kwics
sollte	670	303	67% sollte […] eigentlich	should	Kwics
Warum	543	156	100% Warum […] eigentlich nicht	Why	Kwics
ja	487	265	80% ja […] eigentlich	is indeed	Kwics
wollte	457	200	65% wollte […] eigentlich	would	Kwics
ist egal	395	26	65% ist [mir] eigentlich […] egal	never mind	Kwics
Schade	146	33	96% Schade […] eigentlich	Pity	Kwics
Wieso	75	21	100% Wieso […] eigentlich	Why	Kwics

Figure 2. CA profiles of the MWE *im Grunde* and the single lexeme *eigentlich* (snippets) (CII random sample of 10,000)

eigentlich 'Strange|Curious actually'). The corpus data show that the adverb *eigentlich* cannot be substituted by the MWE *Im Grunde* in those patterns: **Schade im Grunde*; **Komisch im Grunde*; **Merkwürdig im Grunde*. The capitalisation of the adjectives in the *eigentlich* profile indicates that this pattern appears at the beginning of a sentence. The KWIC (see Example (6)) and a selected citation (see Example (7)) show that these ECPs are elliptical constructions with the pragmatic function of an anaphoric comment.

(6) *P14 aber eine Schreibmaschine gibt es nicht mehr.* ***Schade, eigentlich.***
'but there isn't a typewriter anymore. ***Pity actually.***'

(7) *Die Redaktion beherbergt geschätzte 150 Computer, dazu Großdrucker, Scanner und sogar noch Fax-Geräte, aber eine Schreibmaschine gibt es nicht mehr.* ***Schade, eigentlich****.*
Die Presse, 27.07.2014, S. 36,37; Erika, Olympia und Remington: *Meine Schreibmaschine und ich*:
'The desk hosts approximately 150 computers, plus larger printers, scanners and even fax machines, but there isn't a typewriter anymore. ***Pity actually****.*'

This example is a good argument for an analysis without lemmatising the collocation partners. We wouldn't have detected this special use of elliptical constructions in the CA profile if the partner word forms were conflated to the lemma *schade*.

Furthermore, the qualitative comparison of the corpus citations of *im Grunde* versus *eigentlich* shows that speakers use the MWE *im Grunde* (*genommen*) much more indirectly and give the communicative partner the chance to agree or disagree. In contrast, the adverb *eigentlich* directly expresses a truth claim. So, even if a quasi-synonymous single lexeme exists, the MWE shows differences in usage which become apparent when studying large quantities of data.

4.3 KWIC bundles and slot-filler analysis

Our central goal is the corpus-driven detection of fixedness and variance to learn about the nature of multiword and extended context patterns. For this purpose, we have developed the language-independent pattern matching tool *lexpan*[16] that can be used for bundling large quantities of KWICs by search patterns based on qualitative hypotheses or on a specific research question. In addition, it supports the exploration of lexical patterns with variable slots and the qualitative annotation of CA profiles and the fillers of pattern slots.

Since the beginning of our project this tool has not only been developed to support the semi-automatic detection of patterns – as a heuristic analysis instrument – but also to serve as a working environment for the corpus-driven lexicographer. So, it is possible to export and visualise the results (systemised as KWIC bundles, as qualitative annotated CA profiles and filler tables) and to use them as part of a new form of lexicographic representation of MWEs and MWPs (see Section 5).

16. We offer *lexpan* to all interested users; the program is available for download on the *lexpan* website (lexpan).

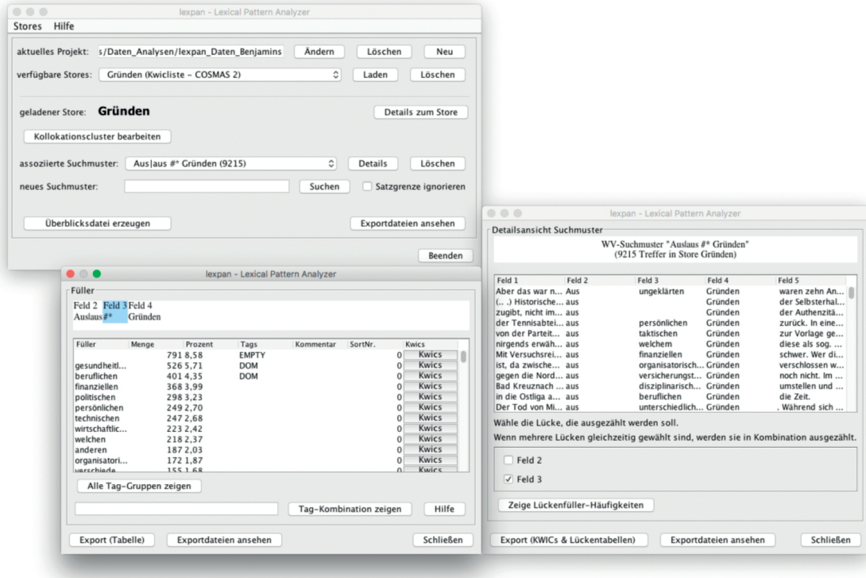


Figure 3. *lexpan* – user interface

Let's take a closer look at how to work with *lexpan*. With its help one can formulate search patterns to capture specific surface characteristics and to bundle the KWIC lines accordingly, e.g. *Gründen*:

KWIC-Überblicksdatei

- Aus|aus #* Gründen (9215 Treffer
- Mit|mit #* Gründen (326 Treffer
- ohne|Ohne Angabe* von #* Gründen (87 Treffer
- Reihe von #* Gründen (31 Treffer
- Restkwics (521 Treffer)

9215 KWIC-Treffer zu "Aus|aus #* Gründen"

Die Gesamttreffermenge entspricht 92,15 Prozent der KWICs im durchsuchten Store.

Sigle		Aus|aus	#*	Gründen	
S8L08	Aber das war noch nicht alles.	Aus	ungeklärten	Gründen	waren zehn Angehörige der afghanischen Armee (ANA) beim Rückzug
R87	(. . .) Historische Umstände haben uns	aus		Gründen	der Selbsterhaltung gezwungen, uns in kleine Gruppen
T01	zugibt, nicht immer ganz einfach zu bewerkstelligen sei, aber	aus		Gründen	der Authenzität unbedingt sein müsse. Die männlichen Darsteller
RHZ98	der Tennisabteilung der Spvgg Nahbollenbach	aus	persönlichen	Gründen	zurück. In einer außerordentlichen Mitgliederversammlung
U92	von der Parteitagsregie allerdings ohnehin nur	aus	taktischen	Gründen	zur Vorlage gebracht. Tsongas machte die öffentliche
WJ3D11	nirgends erwähnt welche Wissenschaftler/Forscher	aus	welchem	Gründen	diese als sog. Pseudowissenschaft beschimpfen. Würde man diese
T97	Mit Versuchsreihen Anerkennung zu finden ist auch	aus	finanziellen	Gründen	schwer. Wer die Wirksamkeit einer Tablette untersucht, kann in
NKU00	ist, da zwischenzeitlich die Türen	aus	organisatorischen	Gründen	verschlossen werden müssen. In Templin werden übrigens am selben
NLJZ08	gegen die Nordamerikaner kam es	aus	versicherungstechnischen	Gründen	noch nicht. Im Gegenteil: Nachdem der 2,13-Meter-Hüne die erste
RHZD8	Bad Kreuznach 0:1 (0:1). "Ich musste	aus	disziplinarischen	Gründen	umstellen und hatte nur zwölf Leute, plus einen Feldspieler im
O99	in die Ostliga aber nicht mitgehen: Dazu fehlt ihm	aus	beruflichen	Gründen	die Zeit.
NZZ06	Der Tod von Milosevic bewegt die Serben	aus	unterschiedlichen	Gründen	. Während sich die Mehrheit in ihrer Beurteilung des
PNI	es der Vermessungsund Katasterverwaltung	aus	den von mir genannten	Gründen	möglich, mehr und schneller Personal abzubauen, als im Rahmen
BRZ07	Entscheidung offiziell seiner Fraktion mit. "	Aus	persönlichen	Gründen	lege ich das Amt nieder", sagte er auf Nachfrage unserer
Z05	in Hamburg und ein Wohnhaus in Stuttgart kommen -	aus	welchen	Gründen	auch immer - unter den Hammer. Termine und Objekte für

87 KWIC-Treffer zu "ohne|Ohne Angabe* von #* Gründen"

Die Gesamttreffermenge entspricht 0,87 Prozent der KWICs im durchsuchten Store.

Sigle		ohne|Ohne	Angabe*	von	#*	Gründen	
NZZ00	mit ihnen allen".",". Einige hundert sind	ohne	Angabe	von		Gründen	zu verhalten - und ab die Post, meine Herren!" Das erklärte Ziel
NLN98	Auszubildende vor dem Landgericht verantworten.	Ohne	Angabe	von		Gründen	legte nun Rechtsanwalt Wolfgang Schwemmer aus Bayreuth, der
U05	Erst im Februar war Finanzvorstand Rudolf Winning	ohne	Angabe	von		Gründen	ausgeschieden. Daraufhin hatte der Unternehmensberater und
NZZ14	Fraktion in schwerwiegender Weise schaden, auch	ohne	Angabe	von		Gründen	ausgeschlossen werden können. Die Gründe will Tuena nicht näher
I99	auch andere Sorgen. Völlig unerwartet und vorerst	ohne	Angabe	von		Gründen	trat Klubbesitzer Massimo Moratti als Präsident zurück.

Figure 4. Example for KWIC bundles for *Gründen* (*lexpan* snippet)
(CII random sample of 10,000)

In the case of the *Gründen* KWICs, only one search pattern covers 92.15% of all occurrences, namely the preposition *aus* plus an arbitrary number of unspecified words plus *Gründen* (aus|Aus #* Gründen) (see KWIC snippet in Figure 4). The other three search patterns capture the MWP [*mit* X *Gründen*] ('with/for X reasons') and the MWEs *ohne Angabe von Gründen* ('without giving reasons') (see KWIC snippet in Figure 4) and *eine Reihe von Gründen* ('a number of reasons'). The remaining KWICs (5.21%) feature occurrences that are not bound to *Gründen* patterns, e.g. *zu den Gründen äußerte er sich nicht* ('he didn't comment on the reasons') (verb pattern: *sich zu* etw. *äußern* 'to comment on sth').

Once the KWIC lines have been captured by a search pattern, *lexpan* can count the slot fillers and present them in filler tables.[17] Such tables show the absolute and relative frequencies of the lexical fillers for one or more slots. Furthermore, *lexpan* allows us to incorporate qualitative annotations (tags and comments) for grouping the fillers according to semantic and pragmatic characteristics, e.g. in this table for the pattern *aus* X *Gründen* ('for X reasons') we use the column "Kommentar" ('comment') for English translations of the fillers:

Lückenfüller	Anzahl	Prozentanteil	Tag	Kommentar	Kwics
	791	8,58	[EMPTY]		Kwics
gesundheitlichen	526	5,71	[DOM]	health	Kwics
beruflichen	401	4,35	[DOM]	professional	Kwics
finanziellen	368	3,99	[DOM]	financial	Kwics
politischen	298	3,23	[DOM]	political	Kwics
persönlichen	249	2,70	[DOM]	private	Kwics
guten	114	1,24	[VAL]	good	Kwics
taktischen	90	0,98	[DOM]	tactical	Kwics
unerfindlichen	49	0,53	[VAL	mysterious	Kwics
bisher unbekannten	14	0,15	[VAL]	as yet unknown	Kwics

Figure 5. Filler table of *aus|Aus* 0 … n slots *Gründen* (*lexpan* snippet) (CII random sample of 10,000)

The highest ranked slot is annotated as a zero-gap which indicates cases where no lexical item appears between the fixed elements of the pattern: in Figure 5 this zero-gap stands for the lexical core *aus Gründen* ('for reasons'), a MWE with the meaning 'because of' with typical NP extensions like *der Verkehrssicherheit* / *des*

17. Methodologically, our slot-filler approach bears some similarities to Stefanowitsch and Gries' 'collostructions' (Stefanowitsch, 2013). As mentioned in Section 3.1., we feel closer to the holistic perspective of phraseology.

Datenschutzes / ('for reasons of traffic safety / data protection'). The other fillers can be grouped into functional groups, like adjectives referring to a specific domain [DOM], e.g. *politisch* ('political'), *persönlich* ('personal'), *finanziell* ('financial'); *technischen* ('technical') or to the validity of reasons or motifs [VAL] with an evaluative potential of the adjectives, e.g. *nachvollziehbar* ('understandable'), *gut* ('good') or *triftig* ('valid').

Particularly interesting is the first group of adjectives referring to domains like politics, privacy, finance and many others. The ADJ realisations seem to be much more neutral than the other evaluative ADJ fillers like *good* or *understandable*. Speakers say, "The reason is X". But in these cases we can also observe a specific pragmatic aspect: When somebody steps back *aus persönlichen Gründen* ('for personal reasons') the reason is really quite unspecified, we are left to speculate whether there were family problems or possibly internal pressure to step down.

The speakers use the plural form of the pattern to express a certain degree of vagueness and avoid a direct interpretation or evaluation.

It is also very interesting to examine those fillers that occur only once; these are often compound words.

(8) *strömungs**politischen*** ('trend-**political**')

(9) *interview**taktischen*** ('interview-**tactical**')

(10) *wetter**technischen*** ('weather-**technical**')

(11) *nebulösen* ('nebulous') / *schikanösen* ('vexatious')

Traditionally, as linguists, we would interpret such hapax legomena as ad hoc realisations of the MWPs, as phenomena that can be disregarded for the lexicographic description. But when exploring the filler tables it becomes clear that they as well contribute to forming the emergent pattern and follow the same predispositions: The examples in (8), (9) and (10) are compounds which use words from the highest ranks as their basis. The examples in (11) are evaluative adjectives which are also a dominant filler group in the higher ranks.

lexpan also enables to detect n-grams.[18] This is particularly useful for describing context patterns like typical chunks in the environment of the MWE *aus Gründen* ('for reasons'):

18. For the examples in Figure 6 and 7 I extended the random sample to 100,000.

Lückenfüller	Anzahl	Prozentanteil	Kommentar	Kwics
nicht … nur	66	0,87	not only	Kwics
vor … allem	53	0,70	first and foremost	Kwics
und … zwar	36	0,47	in fact	Kwics
aber … auch	23	0,30	but also	Kwics
nicht … zuletzt	22	0,29	last but not least	Kwics

Figure 6. Filler table of ECP two slots + *aus Gründen* (*lexpan* snippet) (CII random sample of 100,000)

Frequent chunks (see Figure 6) like *nicht nur* ('not only'), *vor allem* ('first and foremost'), *und zwar* ('in fact') or *nicht zuletzt* ('last but not least') can be classified as functional chunks that indicate a specific argumentation frame used by the speakers.

In other cases, the slot-filler analysis reveals a high degree of restriction, e.g. for the internal extension of the syntagma *mit … Grund* ('with … reason'):

Lückenfüller	Anzahl	Prozentanteil	Kommentar	Kwics
ein	10912	49,04	a	Kwics
gutem	7365	33,10	good	Kwics
	1856	8,34	'zero gap'	Kwics
dem	317	1,42	the	Kwics
einigem	118	0,53	some	Kwics
festem	4	0,02	concrete	Kwics
goldenem	4	0,02	golden	Kwics

Figure 7. Filler table *mit* 0–1 slots *Grund* (*lexpan* snippet) (CII random sample of 100,000)

In Figure 7, we can observe that the first two fillers *ein* ('a') and *gutem* ('good') have a disproportionately high frequency (prototypical fillers). The third rank is taken by a zero gap. The large number of remaining fillers are distributed with low frequencies. The zero gap indicates the MWE *mit Grund* ('with reason') with the meaning 'legitimately'. The frequent adjective filler *gutem* ('good') points to the extended pattern of this MWE and has the function of intensification. Due to its frequency, the ECP *mit gutem Grund* can also be considered an autonomous MWE that also means 'legitimately'. The other frequent filler, the indefinite article *ein* ('a'), constitutes the interesting syntagma *mit **ein** Grund* ('with **a** reason'); *with* in this case means 'among others' so the pattern can be paraphrased as 'a reason among others'. At first glance it seems that the components of this syntagma are the word

form *mit* ('with') and the NP *ein Grund* ('a reason'). But the NP slot is filled in a very systematic manner with nouns that refer to causative concepts. Beside *Grund* there are e.g. *Ursache* ('cause') -> ***mit eine*** *Ursache* ('a cause among others') or *Motiv* ('motive') -> ***mit ein*** *Motiv* ('a motive among others'). From the pattern perspective, the fixed lexical element is *mit ein/eine* ('with a'). That seems to be non-grammatical (the regular grammatical unit is the NP (indefinite article + noun)), but *mit ein/eine* ('with a') can be regarded as a functional chunk as a component of the LP [*mit ein/eine* NOUN].

5. A new type of corpus-driven, pattern-based MW dictionaries

Parallel to the development of the MWP model and the corpus-driven methodology, we have been investigating how these results could lead to new forms of description of common language use, e.g. for a change of perspective in foreign language acquisition and teaching (Steyer, 2009; Steyer and Brunner, 2014). The fundamental insight that only a small number of expressions are saved as isolated entries in the mental lexicon whereas the majority bases on lexical patterns as part of a network calls for new forms of lexicographical representation. A learner needs information about the status of a lexical unit: Is it a strongly fixed MWE? If so, then the item must be memorised like a word. Or: Is this MWE only a typical realisation of a pattern? If so, then it is necessary to understand the semantic or functional principle underlying this pattern and to learn which realisations are typical in common use.

Another innovation related to developments in language technology is that mass data not only can used as a basis for empirical analyses but also as a form of lexicographic information.[19]

A prototype is our corpus-driven, pattern-based pilot study "Wortverbindungsfelder" ('Multiword fields'), especially Version 3, "Grund" (http://wvonline.ids-mannheim.de/wvfelder-v3/index.html) that contains the following four data types:

19. Pioneering work in this area is also done by Hanks and his colleagues in their Pattern Dictionary of English Verbs (PDEV). This online dictionary presents usage patterns of English verbs with their KWIC lines, systematised according to semantic properties. Thus, users can understand and learn these patterns not on the basis of idealised example sentences but in their authentic usage contexts.

a. automatically compiled corpus data (KWICs, CA profiles, filler tables)
b. semi-automatically compiled corpus data that illustrates specific aspects of usage; this is original corpus data that has been selected and/or annotated manually, including
   – qualitatively selected KWIC bundles and citations
   – qualitatively annotated and systematised collocation fields
   – qualitatively annotated and systematised filler groups
c. narrative lexicographic comments
d. hierarchies of MWE and MWP nodes on several levels of abstraction which contain several types of cross connections.

The structures of MWEs and MW patterns, the distinctive characteristics of their usage, their interrelations, and the regularities of the underlying patterns are presented primarily on the basis of the arrangement and hierarchical ordering of KWIC lines. In this way one can learn about language use by studying numerous real-word usage examples with related characteristics. Filler tables like the ones cited in this article, collocation profiles and other data contribute to the understanding of the MWE and its place in networks and hierarchies.

We want to explain the main ideas of this concept using the example of the node [*aus Gründen* SUB-G] ('for reasons NP') (as mentioned before) and the MWE node *aus welchen Gründen auch immer* ('for whatever reasons'). Figure 8 shows a snippet of the pattern node [*aus Gründen* SUB-G] ('for reasons NP'):

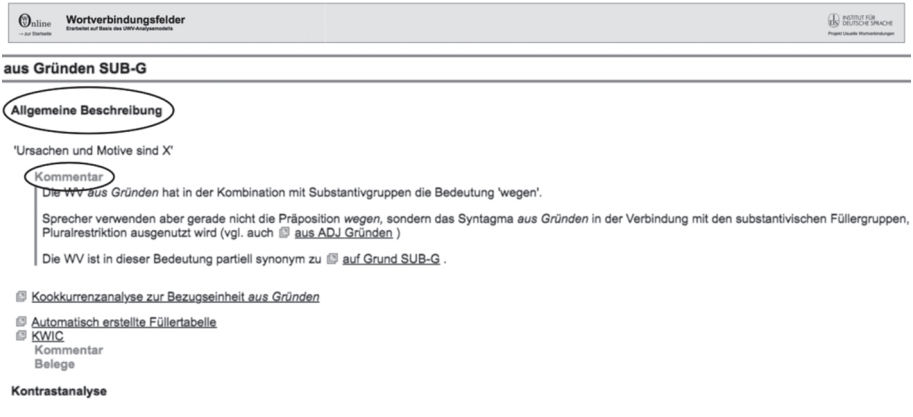


Figure 8. "Multiword fields" online: Node *aus Gründen* SUB-G ('for reasons NP') (snippet)

This node contains narrative comments, e.g. on the abstract meaning of the pattern *aus Gründen* SUB-G ('for reasons NP') (see "Allgemeine Beschreibung") and on pragmatic features (see "Kommentar"). Furthermore, there are links to the CA

profile of this pattern (see snippet in Figure 9) and the filler table of the NP slot. This table is complemented by a linguistic comment about pattern restriction based on the word building phenomena of the noun fillers.

In addition to this, the section "Kontrastanalyse" ('contrast analysis') provides information on a quasi-equivalent of the chunk *aus Gründen* ('for reasons), the prepositional single lexeme *wegen* ('because of') (pattern: [*wegen* NP] ['because of NP']) by showing its CA profile. Our online platform allows the user to compare those two profiles (Figure 9).

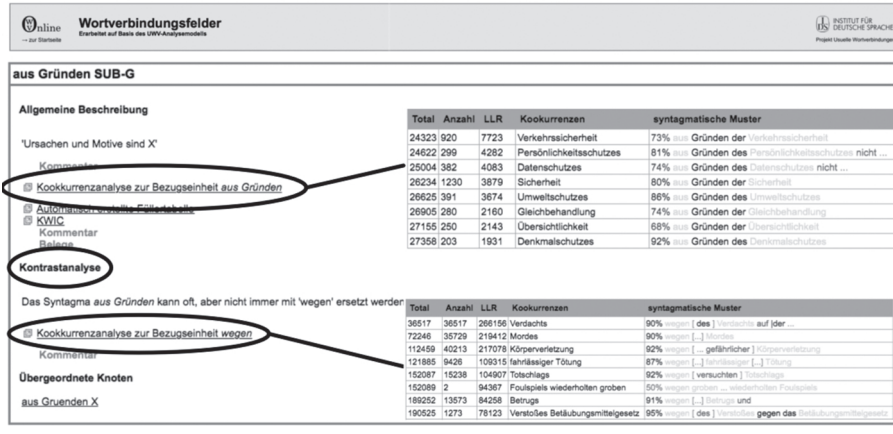
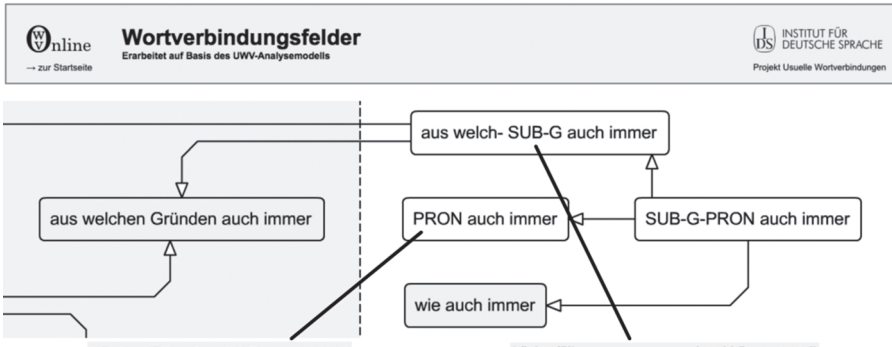


Figure 9. "Multiword fields" online: Contrasting CA profiles of *aus Gründen* und *wegen* (snippets)

In this way, the user gets an idea of the significant usage differences between *wegen* and *aus Gründen*.

The MWE node *aus welchen Gründen auch immer* 'for whatever reasons' is interesting because several levels of abstraction are at play: the word form *Gründen* can be substituted with different nouns in the MWP *aus welch-* SUB-G *auch immer*, e.g. *Motiv* ('motive') or *Richtung* ('direction'). But the lexical components *auch immer* are part of another, more abstract pattern as well: [PRON *auch immer*]. In the MWP, the place of PRON is taken by interrogative pronouns e.g. *was* ('what'), *wo* ('where'), *wer* ('who'), and *warum* ('why').

One of those pronoun fillers, *wie* ('how'), is much more frequent than the others, and this is why there is a separate MWE node *wie auch immer*. This MWE can be translated as 'howsoever' or 'to whatever extent'. The complicated abstractions and relations are visualised by the MWE network (Figure 10). By means of this network the user can understand the progressing abstraction from the lexical surface to the pattern meaning ('sth is not known or not comprehensible') and the



Lückenfüller	Anzahl	Prozentanteil
Wie	9611	10,08
wie	7389	7,75
was	5289	5,55
Was	1953	2,05
Wer	1825	1,91
warum	1529	1,60
wo	1486	1,56
wer	1446	1,52
wem	1292	1,35
Wo	556	0,58
Warum	312	0,33
wann	263	0,28
welche	186	0,20

Lückenfüller	Anzahl	Prozentanteil
Gründen	2986	72,02
Grund	577	13,92
Motiven	91	2,19
Grunde	35	0,84
gründen	26	0,63
Beweggründen	21	0,51
Quellen	13	0,31
Quelle	12	0,29
Gruenden	12	0,29
Richtung	10	0,24
Ecke	9	0,22
Motivation	9	0,22
grund	9	0,22
Motivation heraus	7	0,17
Land	7	0,17
Motiv	7	0,17
Anlass	6	0,14

Figure 10. "Multiword fields" online: Pattern hierarchy and filler tables (snippets)

communicative function that is stable on every level (Speakers express a certain degree of scepticism, doubt or criticism).

Currently, the central ideas and components of this lexicographical concept are implemented in the cooperation project *Preposition-Noun-Combinations in Context* (PREPCON).[20]

PREPCON explores preposition-noun-combinations (PNCs) with a recurrent zero gap as autonomous lexical items (e.g. *durch Zufall* ('by chance') or *über Jahre* ('over years')). PREPCON has three different forms of data representation: (i) a fully automated database of 80,000 German PNCs (autonomous and as part of other constructions) with KWICs and frequencies; (ii) a semi-automatic database of temporal PNCs including a collection of 150 autonomous temporal MWEs that also fulfil modal or discourse marker functions; (iii) a contrastive pilot study based on the UWV methodology for a trilingual contrastive description of lexical patterns.

20. The cooperation partners are: IDS UWV group (head: Kathrin Steyer, Mannheim), FRASESPAL (head: Carmen Mellado Blanco, Santiago de Compostela/Spain), WICOL (head: Peter Ďurčo, Trnava/Slovakia).

6. Conclusion

This paper discussed new pattern perspectives that are relevant not only to the field of phraseology, but also to usage-based linguistics and lexicography as a whole. We argued that large corpora and sophisticated automatic methods can only be fruitful for these fields if they are used in combination with qualitative interpretation. Qualitative interpretation requires deep reflection on the explanatory potential of each corpus linguistic method and also on its limitations. The UWV methodology described in this paper is an example of an integrative approach that is focused on discovering hidden structures of language use, especially frozen lexical building blocks and patterns and their functions in communication. Beside the contextualistic approach to multiword units based on collocation profiles, this methodology is focused on the detection and interpretation of patterns that have holistic meanings or functions. Using the UWV tool *lexpan* we illustrated several characteristics of frozenness: fixed lexical multiword expressions, multiword patterns with fixed lexical components and mandatory slots and extended multiword expressions with facultative, but also recurrent contextual extensions of the core. All these types have the potential for cognitive entrenchment. We showed the complex nature of filler tables, from the disproportional frequency of a few lexical units (indicating the prototypical realisation of the pattern) to a continuous range of frequencies. In the first case we can assume more than one entry in the mental lexicon: as an MWE and as a realisation of a pattern. Another interesting observation is that even fillers that occur only once (indicating occasional ad hoc realisations) follow the same predispositions that are apparent for the highest ranks. One of the main results of this research is the insight that there is logic and purpose behind the seemingly endless linguistic creativity. Nothing is arbitrary in language production, but it is not only grammar that guides language production and reception. The genesis of patterns and their functional restrictions cannot be deduced *a priori* based on rules but only *a posteriori* based on an inductive quantitative and qualitative reconstruction.

With regard to a new pattern-based lexicography several important questions emerge: How can we visualise these relationships? Which kind of representation is appropriate for which user group, for example for foreign language learners? The major challenge will therefore be to adapt the concept of the corpus-driven pattern dictionary didactically. We have to find an answer to the very complex question of how to guide a user through the universe of corpus data and networks.

References

Belica, C. (1995). *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode*. Mannheim: Institut für Deutsche Sprache.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Burger, H. (2015). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 5., neu bearb. Aufl. Berlin: Schmidt.

Burger, H., Dobrovol'skij, D., Kühn, P., & Norrick, N. R. (Eds.) (2007). *Phraseology. An International Handbook of Contemporary Research*. 2 vols. (HSK 28, 1/2). Berlin/New York: de Gruyter.

Dobrovol'skij, D. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch, & A. Ziem (Eds.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze* (pp. 111–130). Tübingen: Stauffenburg.

Feilke, H. (1996). *Sprache als soziale Gestalt: Ausdruck, Prägung und die Ordnung der sprachlichen Typik*. Frankfurt am Main: Suhrkamp.

Granger, S., & Meunier, F. (Eds.) (2008). *Phraseology. An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.139

Gries, S. Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger, & F. Meunier (Eds.), *Phraseology and linguistic theory: A brief survey* (pp. 3–25).

Hanks, P. (2013). *Lexical Analysis. Norms and Exploitations*. Cambridge, MA/London: MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Herbst, T., Faulhaber, S., & Uhrig, P. (Eds.) (2011). *The Phraseological View of Language. A Tribute to John Sinclair*. Berlin/Boston: de Gruyter. https://doi.org/10.1515/9783110257014

Hunston, S., & Gill, F. (2000). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.4

Langacker, R. W. (1987). *Foundations of Cognitive Grammar. Vol 1. Theoretical Prerequisites*. Stanford: Stanford University Press.

Mieder, W. (2009). *International Bibliography of Paremiology and Phraseology*. 2 vols. Berlin: de Gruyter. https://doi.org/10.1515/9783110215595

Moon, R. (1998). *Fixed Expressions and Idioms in English*. A Corpus-Based Approach. Oxford: Clarendon Press.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer, & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 128–143). London: Longman.

Schmid, H.-J. (2014). Lexico-grammatical patterns, pragmatic associations and discourse frequency. In T. Herbst, H.-J. Schmid, & S. Faulhaber (Eds.), *Constructions, Collocations, Patterns* (pp. 239–293). Berlin/Boston: Mouton de Gruyter.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stefanowitsch, A. (2013). In G. Trousdale, & T. Hoffmann (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 290–306). Oxford/New York: Oxford University Press.

Steyer, K. (2018). Lexikalisch geprägte Muster. Modell, Methoden und Formen der Onlinepräsentation. In K. Steyer (Ed.), *Sprachliche Verfestigung. Chunks, Muster, Phrasem-Konstruktionen* (pp. 227–264). Tübingen: Narr.

Steyer, K. (2016). Corpus-driven Description of Multi-word Patterns. In G. Corpas Pastor, J. Monti, V. Seretan, & R. Mitkov (Eds.), *Workshop Proceedings Multi-Word Units in Machine Translation and Translation Technologies (MUMTTT2015)* (pp. 13–18). Geneva: Editions Tradulex.

Steyer, K. (2015). Patterns. Phraseology in a state of flux. In D. Dobrovol'ski (Ed.), *Phraseology and Dictionaries (Special Issue). International Journal of Lexicography*, 28(3), 279–298. https://doi.org/10.1093/ijl/ecv021

Steyer, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

Steyer, K. (2009). Zwischen theoretischer Modellierung und praxisnaher Anwendung. Zur Korpusgesteuerten Beschreibung usueller Wortverbindungen. In C. Mellado Blanco (Ed.), *Theorie und Praxis idiomatischer Wörterbücher* (pp. 119–145). Niemeyer: Tübingen.

Steyer, K. (2000). Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und Lexikografische Möglichkeiten. *Deutsche Sprache*, 28(2), 101–125.

Steyer, K., & Brunner, A. (2014). Contexts, Patterns, Interrelations – New Ways of Presenting Multiword Expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)* (pp. 82–88). Gothenburg, Sweden.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Internet sources

(Accessed on 2 Juin 2018)

CII: COSMAS II. Korpusrecherche- und analysesystem. Institut für Deutsche Sprache. Mannheim. http://www.ids-mannheim.de/cosmas2/

Institut für Deutsche Sprache: Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2015-II (Release vom 28.09.2015). Mannheim: Institut für Deutsche Sprache. http://www1.ids-mannheim.de/kl/projekte/korpora/

lexpan: Lexical Pattern Analyzer. http://uwv.ids-mannheim.de/lexpan/

PDEV: Pattern Dictionary of English Verbs. http://pdev.org.uk/#about_cpa

PREPCON: Präposition-Nomen-Verbindungen im Kontext. http://uwv.ids-mannheim.de/prepcon/

SkE: Sketch Engine. http://www.sketchengine.co.uk/

Abbreviations

ADJ	adjective	N	noun
CA	IDS collocation analysis	NP	noun phrase
ECP	extended multiword pattern	P	preposition
LP	lexical pattern	PP	prepositional phrase
MWE	multiword expression	UWV	usuelle Wortverbindungen
MWP	multiword pattern		

How context determines meaning

Patrick Hanks
RIILP, University of Wolverhampton (WLV) and BCL, University of the West of England (UWE)

It is an extraordinary fact that, although most speakers and writers of the English language (or, we may presume, any other language) believe that they are capable of expressing any meaning that they want to with considerable precision, the behaviour of the words they use is highly variable, with much variation in phraseology as well as subtle semantic distinctions. Even more extraordinary is the fact that only some of the logically predictable variants of any given phrase are accepted by native speakers as idiomatic.

This chapter shows how meanings are associated with phraseological norms rather than with words in isolation. It also illustrates the phenomenon of alternation among phraseological norms and shows how phraseological norms are not merely conformed to, but also exploited creatively in ordinary language use. Underlying this paper is the proposition that words in isolation do not have a determinable meaning per se. Instead they have **meaning potential**, different facets of which are activated in different contexts.

By detailed corpus pattern analysis of the verb *blow*, which typically expresses the causation of movement, we explore the relationship between core meaning and a rich set of patterns of idiomatic phraseology – phrasal verbs, idioms, and proverbs.

Keywords: lexical sets, meaning potential, valency, collocation, corpus pattern analysis (CPA)

1. Patterns and valency

This chapter is based on an in-depth analysis of the phraseology associated with just one word, namely the verb *blow*, with some points being illustrated by other clauses and phrases.

The normal, conventional patterns of use of the verb *blow* are freely available in the *Pattern Dictionary of English Verbs* (http://pdev.org.uk). For PDEV, we analysed a sample from the British National Corpus (BNC) of just over 1500 uses of this verb.

https://doi.org/10.1075/ivitra.24.15han

PDEV records no less than 77 patterns for *blow* as a verb, of which 33 are classified as phrasal verbs and 22 as idioms. It is entirely possible – indeed, probable – that some additional patterns, not found in our BNC sample, remained to be recorded, but if they exist at all, they are either extremely rare or recent developments, or both.

Our analysis shows how each phraseological pattern of the verb is associated with a meaning. This association is redolent of both precision and flexibility. Users of English can convey their meanings with great precision, but at the same time the language itself obliges them to make innumerable phraseological choices in everything they say, and these choices activate differences of meaning and implicature that can be very broad and general but can also be of the greatest delicacy.

PDEV currently distinguishes 77 phraseological patterns for the verb *blow*. Of these, 28 patterns are associated with 9 phrasal verbs (*blow about, blow apart, blow away, blow down, blow in, blow off, blow out, blow over, blow up*) and 22 are idioms such as 'blow the whistle on someone or something'.

Speakers and writers of a language exercise these choices with varying degrees of skill, depending on their background and their ability as users of the language. A natural language accommodates all its users, but of course not all of them are equally skilled. At the same time, language events (especially in spoken language) do not tolerate much agonising. Here, too, natural languages are accommodating. Fuzziness, vagueness, and flexibility are design features of natural language (not deficiencies, as was once thought.) At the same time, some features of natural language allow users to take short cuts, for example by not always obliging them to say what is common knowledge or has been previously established.

Examples of usage in this chapter are authentic and are taken from the British National Corpus (BNC). BNC was compiled in the early 1990s. Language change is a very slow process. Great attention is paid by the media and others to new coinages (which may be ephemeral) and occasional sudden changes in meaning of a lexical item, but this kind of journalistic excitement tends to obscure the essential stability of phraseological patterns of English (and other languages, too), most elements of which persist for decades and even for centuries.

The phraseological patterns mentioned in this chapter are numbered with capital letters, to distinguish them from examples of actual usage, which are numbered with Arabic numerals. The patterns are taken from the *Pattern Dictionary of English Verbs*, PDEV: http://pdev.org.uk, which is work in progress. In PDEV, each lexicogrammatical pattern is associated with a meaning. Only a pattern as a whole can be regarded as meaningful. Strictly speaking, a verb in isolation must be regarded as having **meaning potential**, rather than meaning *per se*. The range of choices available to a skilled language user in order to express a particular meaning is often large – sometimes astonishingly large, as in the case of the verb *blow*, for which PDEV distinguishes nearly 80 distinct patterns. The patterns range from the

frequent and commonplace to the rare and colourful – but the range cannot be said to be infinite. Moreover, most patterns can be exploited creatively or rhetorically. As explained in Hanks (2013, especially Chapter 8), exploitation of normal usage is itself a rule-governed linguistic process, quite different from the rules governing clause construction.

To take a simple example, the meaning 'destroy something with explosives' is most normally expressed by a phrasal verb formed with the base verb *blow* + an adverbial particle, typically *up*. Example (1) below is commonplace: it is an example of the most normal and least colourful way of expressing this particular meaning. Examples (2), (3), and (4) are less normal but still perfectly idiomatic. Example (5), which has no adverbial, is, arguably, also an idiomatic way of expressing the same meaning, but it seems to be normal only in military parlance.

(1) During their retreat, the Germans had added to the destruction by *blowing up* all ten of the bridges over the river.

(2) Many critics … believe that the aircraft was *blown apart* by a terrorist bomb.

(3) Gandhi walked towards a platform to begin his speech and was *blown to pieces* by a bomb, along with 14 other people.

(4) The truck was *blown to kingdom come*.

(5) Because the Germans had *blown* the bridges over the Po, there were only temporary bridges.

Example (5) prompts mention of the important distinction between unusual usage and a mistake. In the absence of a satisfactory theory of linguistic usage, this distinction has not yet been properly studied. Considerable skill is required to select appropriate and consistent levels of generalisations when compiling entries for a phraseological dictionary such as PDEV – a fact that can lead to interminable, often unproductive agonising by linguists, lexicographers, and students of language.

2. The verb is the pivot of the clause

Since Lucien Tesnière's valency theory was published in 1959 (*Éléments de syntaxe structurale*), linguists have recognised, not only that clauses are the basic building blocks of discourse (written and spoken), but also that the verb is the pivot of the clause and that clauses consist of structured selections of interdependent pattern elements. Normally, each pattern consists of a verb plus between one and three **arguments** (**subject**, **object**, and **adverbial**). To this must be added **subvalency features** (otherwise known as **subargumental cues**). For example, in order to determine unambiguously the meaning of a sentence containing the verb *take* with

the noun *place* in the direct object slot, it is necessary also to note the presence or absence of a subvalency feature, namely a determiner governing the noun and, if a determiner is present at all, it must be subclassified as a possessive determiner,[1] a reflexive determiner, or an ordinal number. *Events take place*; *people take their place at a table; you can take someone else's place*; while a competitor may hope to *take first place*.

Despite such complexities, we must not lose sight of the fact that normal clause structure is relatively straightforward and that disambiguation – or at any rate, assignment of a meaning – is entirely possible for most if not quite all of the clauses uttered in English and (presumably) any other language. Ordinary language users do it all the time quickly and instinctively (though, in speech, often imperfectly), with little or no agonising or uncertainty. But this chapter asks how they do it, which is a very different kind of question. It is one thing to know what someone means, but a very different thing to know how you know. Getting a satisfactory answer to the latter you question will require a willingness to reappraise our assumptions – sometimes cherished assumptions – about how language works.

3. Collocations and lexical sets

According to Corpus Pattern Analysis (Hanks 2004), each clause role in a pattern other than the verb is populated by a lexical set of nouns and noun phrases. Writers and speakers choose lexical items from a lexical set in order to express a meaning. Lexical items are grouped into sets according to their semantic type. Lexical sets are given names and are conventionally printed in double square brackets. Lexical sets are open-ended – i.e., in principle there can be any number of lexical items in a lexical set. A hierarchical inventory of the semantic types used by PDEV is available in an on-line ontology. Each lexical set has a name, but it would be a mistake to regard these names as defining set membership. Nothing should be read into the terminology of semantic types; their function is no more than mnemonic. The real value of each lexical set will be found in its *extensional* definition, i.e. the list of words that have been identified as set members according to contexts in corpora. So, for example, the nouns *round* and *salvo* cannot reasonably be defined as kinds of projectile, but in pattern A of the verb *fire*, these nouns constitute a lexical set. This lexical set contrasts with another lexical set, namely firearms. These lexical sets are important because they facilitate the attribution of contrasting implicatures (entailments) to what is said. If you fire a gun or other firearm, it stays where it is.

1. Possessive phrases, as in *She took the place of David Cameron*, are classified as sub-possessive determiners, interchangeable grammatically with *She took his place*.

If you fire a bullet or other projectile, it moves. You can fire all sorts of projectiles, but you cannot fire one that does not move. The English language does not allow it. If you fire something that stays where it is after being fired, you are into a different lexical set and, arguably, a different meaning of *fire*.

CPA reveals that the lexical set [[Projectile]] includes at least the following canonical lexical items in relation to the verb *fire*:

bullet, shell, missile, rocket, grenade, torpedo, tear gas, very light, shot, salvo, blank ….

This lexical set is already different from WordNet's synset for *bullet*, which consists only of *bullet* itself and *slug*. If we look at the hyponyms in WordNet, we can consider augmenting these two items with a few hyponyms, which are:

dumdum, dumdum bullet, full metal jacket, rifle ball, rubber bullet.

If TNE (the Theory of Norms and Exploitations, as explained in Hanks 2013) is right that meaning depends, at least in part, on context – i.e. collocations – a lot more work needs to be done on lexical sets. An example of the sort of problem encountered in grouping collocations into lexical sets is that some canonical set members don't fit well into the set semantically, e.g. *round* (as in 6):

(6) The police … fired more than 20,000 *rounds* a day.

It is only in the context of firearms that the noun *round* denotes a projectile. A projectile is a physical object, whereas a round is more typically an event. Nevertheless, the phraseology of 6 is perfectly normal: it is not metaphorical or unusual in any way.

A second kind of phraseological problem concerns collocates that are unusual in some way; they cannot be regarded as canonical members of a lexical set. Consider Examples (7) and (8).

(7) He strings his own bow while hunting, takes the *arrows*, sets and fires them.

(8) He got angry and fired *a bottle of whiskey* at a shelf.

The problem with 7 is that an arrow, while having the normal semantics of a projectile when it is flying through the air towards a target, does not share all the semantic features of other projectiles. It is not normally fired from a gun, nor does it involve gunpowder. Thus, it may be considered as a slightly defective member of the lexical set of projectiles. There is nothing metaphorical about the phrase *firing an arrow*.

By contrast, 8 can clearly be classified as a metaphorical use of the verb fire, because obviously, a bottle of whiskey is not a canonical member of the set [[Projectile]], although in this context, where 'a bottle of whiskey' functions as the direct object of the verb *fire*, it must be interpreted as such. Classifying this

clause as a metaphor is not entirely satisfactory, because the meaning is literally that a projectile was directed at a target (although, presumably, no gunpowder was used). The direct object in this clause is an anomalous argument. This is a point of some theoretical importance, because it goes a long way towards explaining the phenomenon of linguistic creativity. It is also relevant to the theory of metaphor. Many so-called metaphors in the literature on the subject are no more than secondary conventions. They are called 'metaphors' by reason of their semantic resonance with another meaning of the same word or phrase, namely the literal meaning, not by reason of their originality. In discussions of meaning and metaphor, it is useful to distinguish the normal, conventional use of words from abnormal, creative uses.

4. Core meaning

The verb *blow* has an unusually rich range of phraseological pattern. However, the patterns are not separate and free-standing. There are interesting relationships among patterns. In the first place, as already noted, almost half of the patterns are phrasal verbs. In the second place, there are some regular alternations, for example between causative (transitive) uses and inchoative (intransitive) uses. For example, it is equally idiomatic to speak of a particular event as *the wind blew the fence down* and as *the fence blew down*. In PDEV these are treated as two different patterns, because the implicatures are different and may indeed be represented by different translations into some languages. Likewise, *an explosion blew the windows out* and *the windows blew out*.

Since publication of Lakoff and Johnson (1980) there has been widespread discussion of the phenomenon of metaphor. However, the evidence for much of this discussion is not empirically well founded. By contrast, TNE is a theory that is based firmly in a procedure called CPA (Corpus Pattern Analysis). This means that it aims to establish a theory of text meaning that is based on observed patterns of linguistic behaviour. Among other things, TNE argues that conventional metaphors are nothing more than secondary norms. They are not exploitations, but they do have resonance: part of the meaning of a word or phrase used metaphorically is that it resonates with some other meaning of the same word or phrase, which is the literal meaning or core meaning.

Most verbs have at least one core meaning. Core meaning is a rather vague and variable concept and is difficult if not impossible to tie down precisely. The core meaning of the verb *blow* has something to do with the movement of a current of air, created either by atmospheric conditions (9), (10) or by a human exhaling (11), (12).

(9) The next morning the wind still *blew* from the west.

(10) A south-westerly breeze *blew* into the faces of the anxious group of watchers above Whitsand Bay.

(11) It's no use *blowing* a whistle asking for help, if you know the policeman is a couple of miles away.

(12) [He wondered] whether he could risk *blowing* his nose.

Lexical items that realise the concept of an atmospheric condition that creates wind include: *wind, breeze, gale, storm, hurricane, typhoon, tornado*, and several other words. These lexical items constitute a lexical set, unified by a shared semantic concept. Lexical sets are of the greatest importance for understanding how meaning works. Many English verbs have more than one sense, but neither monolingual English dictionaries nor bilingual dictionaries make. Any one member of this lexical set of nouns, when used in the clause role of subject, serves to select the appropriate meaning of the verb *blow*. The verb is usually intransitive in this sense, but occasionally we find a pleonastic 'it' in the subject role and the wind type masquerading as a pseudo-direct object, as in (13).

(13) It was blowing a gale.

More frequent – and much more important from the point of view of phraseology and meaning analysis – is the fact that clauses with blow as the verb often take an adverbial of direction, as in Example (1) above.

Is 'explosive force' part of the core meaning of *blow*? Since the invention of gunpowder, many uses of the verb *blow* contain an implication of explosive force, as in the phrasal verb *blow up* and several other phrasal verbs and idiomatic expressions. However, it not clear that explosive force should be regarded as an element of the core meaning of *blow* itself. We shall return to this question when discussing the phraseology of phrasal verbs.

5. Phrasal verbs

The essential fact here is that CPA incontrovertibly points to the conclusion that meanings are associated with patterns of word use rather than with words in isolation. But this does not tell the whole story. Some dictionaries acknowledge that certain patterns of word use may themselves be ambiguous, but they do not tell us how the ambiguity is to be resolved. For example, Collins English Dictionary lists the phrasal verb **blow up** with six different senses:

> **blow up** *vb*. (*adv*.)**1**. to explode or cause to explode. **2**. (*tr*.) to increase the importance of (something): *they blew the whole affair up*. **3**. (*intr*.) *Informal*. to come into consideration: *we lived well enough before this thing blew up*. **4**. (*intr*.) to come into existence with sudden force: *a storm had blown up*. **5**. *Informal*. to lose one's temper (with a person). **6**. (*tr*.) to enlarge the size or detail of (a photograph).

Each of these senses can be supported by plentiful examples from corpus evidence, as in Examples (14)–(19), (21) below. Sense 1 requires two Examples (14), (15) to illustrate the causative/inchoative alternation.

(14) A worker who *blew up* a 19th century listed Methodist chapel…

(15) A listed Wesleyan chapel facade was *blown* up with gelignite in an abortive attempt to allow a hard-pressed property developer to build flats in its place.

(16) A van, loaded with explosives, *blew up* outside the Northern Bank.

(17) The degree of sensitivity on this issue is now such that the emigration of a single Serbian individual or family is *blown* up to a national crisis. [2]

(18) A major row *blew up* last night. [3]

(19) While they are at sea, a storm *blows up*. [4]

(20) The typical pattern with aggressiveness is that people attack, over-react, *blow up* and cause resentment. [5]

(21) He pointed to a large X-ray photograph *blown up* on a screen in front of us.

Additionally, PDEV records a pattern (or sense) of this phrasal verb that is not covered by Collins English Dictionary, namely 'to inflate (a balloon) by blowing air into it).

(22) There are two ways of *blowing up* a balloon.

Blow up is the most frequent phrasal verb formed with *blow*; it is also the most cognitively salient, though it is by no means the only one. Others include *blow about, blow apart, blow away, blow down, blow in, blow off, and blow over*.

The adverbial particles used with *blow* to form phrasal verbs all have a basic sense denoting movement in a particular direction. It is not always obvious whether a verb+particle construction should be regarded as a compositional construction or as a phrasal verb (where, according to one widely accepted criterion, a meaning cannot be assigned separately to the individual components). Thus, the verb + particle combination in 23 is pretty obviously compositional (i.e. not a phrasal verb), while in 24 it is clearly a phrasal verb. But what about 25? Here, there is a prepositional object, which for some scholars would rule out classification as a phrasal verb – but the meaning is almost identical two that in 24, which is clearly a phrasal verb.

(23) A breeze *blowing across* the top of the chimney

(24) A customs officer … woke to find the roof of his bedroom had *blown off*.

(25) They *blew* the doors *off* the van.

6. Exploiting established phraseology

The phrase *winds of change* has established itself as an intertextual classic. It is now used in many different contexts and the phraseology is exploited in many different and sometimes unusual ways. The relevant history is as follows:

In early 1960, the then British Prime Minister Harold Macmillan made a speech, first in Ghana, then in South Africa, clearly indicating the British government's commitment to ending colonialism. He said:

(26) The wind of change is blowing through this continent. Whether we like it or not, this growth of national consciousness is a political fact.

Macmillan was probably not the originator of this felicitous phrase, but he was surely its most significant exponent. Let us look at a few of the innumerable re-uses and exploitations of this felicitous metaphor as recorded in BNC, ranging from the sublime to the bathetic. The main point to note is the wording is not strictly constrained, but can be highly variable, being adapted to relevant aspects of the context in which the phrase is being used. This is a regular feature of idioms. Speakers and writers do not simply trot out the canonical wording of an idiom like automata. People like to take ownership of the phrases they use. If fact, it is sometimes difficult to decide what is the canonical wording of idiom.

(27) Through the corridors of Scotland Yard, a wind of change is *blowing*.
The Guardian, November 1989

(28) The last National Serviceman left the Army in May 1963. By then, warnings were already *blowing in the post-imperial winds of change* that the all-regular Army might soon be overstretched trying to maintain stability in the final stages of Britain's withdrawal from Empire.
General Sir William Jackson (1990): *Britain's defence dilemma: An inside view (rethinking British defence policy in the post-imperial era)*

(29) The winds of neo-liberalism are blowing a gale through Prague.
Marxism Today, 1985–93

(30) The political winds had also been *blowing* from a different and more conservative direction.
D. Waddington (1992), *Contemporary issues in public disorder*

(31) A wind of change had already begun to *blow* at Newport Pagnell.
Management Today, 1991

Finally, it is interesting to note, in Example (32), an allusion to this well-established phrase in a work that discusses phraseological coinages and the creative use of language, a topic that is close to the very subject of this chapter!

(32) The emergence of new phrases is continuous; this is an attribute of a vigorous and developing language. Many original phrases, coined by keen minds for a specific occasion, have made strong impressions because of their freshness and force. Speakers of less originality, and writers too, pick up these phrases at second hand and use them until their first fire and punch are worn away. How many have tried to *blow* Harold Macmillan's winds of change when the context did not merit the phrase? How many have taken the edge off the phrase at the grass roots by applying it to inappropriate circumstances?
W. Fisher Cassie, *A Student's Guide to Success*.

A similarly well-established phrase is 'to know which way the wind is blowing', typically used to refer to someone's political or industrial awareness of facts that may be to his or her advantage or disadvantage.

(33) I suspect that people who work in the industry are under no illusions. They know the way that the wind is *blowing*, and would be only too pleased to be redeployed into another trade.

(34) Mr Dostam keeps a token force of fighters in Kabul, sometimes supporting Mr Masoud, sometimes opposing him, depending which way the wind is *blowing*.

6.1 Phraseology that is both literal and figurative

Some phraseology can be both literal and figurative, depending on the context in which it is. For example, *you can blow a hole in a wall*, but you can also *blow a hole in a proposition or plan*.

Another example is the phrase "to be blown off course', which is likewise used both figuratively and literally as Examples (35) and (36) demonstrate.

(35) Short-term prospects can be *blown off course* by inflation.

(36) Several ships were either wrecked or *blown off course*.

This idiom harks back to the days of sailing ships. Although it is less common now than it was in the 18th and 19th centuries, we still find it used literally to designate sailing boats that lose control over the direction in which their skippers wish to sail because of high winds. In exceptional circumstances (such as persistent gale-force winds and/or engine failure), even great ships It is also used occasionally of migrating birds that lose direction, for the same reason. But the idiomatic sense is more frequent nowadays.

By contrast, some idioms are used only figuratively. It is hard to imagine any circumstances in which the phrase *to blow someone's socks off* would be used literally.

(37) Gus Dudgeon [a music producer of the 1970s] popped down to my office, stuck this demo [by David Bowie] on the turntable and it just *blew my socks off*.

This idiom is quite rare in the corpus data that we looked at, but it still appears to be conventional, whereas in Example (38) it appears to be being exploited by someone more interested in teeth the socks.

(38) [The Caterham Seven (a type of lightweight sports car aimed at a narrow market of enthusiasts)] 'is likely to *blow* your teeth out at much over 80 mph'.

It is highly unlikely that the writer of this fragment (published in a 1992 issue of the journal *Accountancy*, thus reinforcing the view that accountants tend to have more money than sense) had any serious message in mind about the relationship between the car and the teeth of the driver. Rather, it seems to be nothing more precise than a colourful way of saying that the car is uncomfortable and impractical, but fun. The point is reinforced by the continuation (39), which cites another writer as saying about the same car that:

(39) your smile `will need surgically removing' after driving one.

This piece of hyperbole, like most hyperboles and many similes, cannot possibly have been intended to be taken literally. Evidently, it is a colourful way of saying that the experience of driving such a car will give you great pleasure.

7. Exploiting a proverb

There is a proverb in English. *It's an ill wind that blows nobody any good*, which means (roughly): 'Even very bad events may be of some benefit to someone somewhere'. In the following quotation, the proverb is being exploited to refer to an economic recession, which is usually considered to be a bad event.

(40) *It's an ill recession that blows nobody any good*. The financial markets may be in retreat and the pound may be on a slippery slope, but it's not just the insolvency practitioners and bailiffs that are doing well; some companies in sectors as diverse as retailing, restaurants and medical equipment are **also** doing more than just make ends meet. Many of these are young companies, debt-free and so not hampered by high interest rates. They are nimble enough and liquid enough to take advantage of the recession. They can get good deals in the property market.

8. Other IDIOMS with 'blow'

In conclusion, we list just ten of the 22 idiomatic phrases used with the verb 'blow', by way of illustrating the rich phraseological complexity of the expressions that have this verb as their pivot:

- *To blow the whistle on someone or something* [= expose wrongdoing]
- *To blow the cobwebs away* [= get rid of useless old ideas]
- *To blow one's own trumpet* [= boast]
- *To blow hot and cold* [=equivocate]
- *To be blown off course* [= lose control]
- *He blow his top* [= lost his temper]
- *He blew a lot of his money on gambling* [= spent]
- *It blew up in my face* [= went badly wrong]
- *Lawrence blew my cover* [= revealed that I was a spy]
- *He blew his brains out* [= killed himself by shooting himself in the head]

9. Conclusion

In this chapter we have offered a discussion of the proposition that "context determines meaning". We looked at some of the problems that lie in the way of an excessively literal interpretation of this idea. Corpus data supports the proposition for almost all normal uses although words and phrases we have been looking at, but anomalous uses must be recognised for what they are. Data can be sorted into lexical sets according to word use, but some uses turn out to be anomalous. Anomalies are found not only in freshly created linguistic metaphors, but also occasionally in regular phraseology that is semantically anomalous, for example 'firing a round', where a 'round' does not share the semantic attributes of projectiles. A *round* is stereotypically an event, whereas a projectile is stereotypically a physical object.

The most important conclusion of this study is that whereas words are typically multiply ambiguous, phraseological patterns, if described following the rules and procedures of CPA, are almost invariably unambiguous. Of course, much depends on the skill of the lexical analyst in getting just the right level of generalisation for this purpose. For this reason and others, our conclusion is tentative at the time of writing. Lexical analyst is a human being, not an automaton. But then, meanings are also aspects of human behaviour. The tentative conclusions of this paper need to be tested in detail by analysis of a great deal more phraseology and many more clauses showing the usage of many other verbs and indeed other predicators.

References

Hanks, P. (2004). Corpus pattern analysis. In G. Williams, & S. Vessier (Eds.), *Euralex Proceedings*. Lorient, France: Université de Bretagne-Sud.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Tesnière, L. (1959). *Éléments de syntaxe structurale* [Elements of Structural Syntax]. Paris, France: Klincksieck.

Detecting semantic difference

A new model based on knowledge and collocational association

Shiva Taslimipoor[1], Gloria Corpas Pastor[1,2] and Omid Rohanian[1]
[1]Research Group in Computational Linguistics, University of Wolverhampton / [2]University of Malaga

Semantic discrimination among concepts is a daily exercise for humans when using natural languages. For example, given the words, *airplane* and *car*, the word *flying* can easily be thought and used as an attribute to differentiate them. In this study, we propose a novel automatic approach to detect whether an attribute word represents the difference between two given words. We exploit a combination of knowledge-based and co-occurrence features (collocations) to capture the semantic difference between two words in relation to an attribute. The features are scores that are defined for each pair of words and an attribute, based on association measures, n-gram counts, word similarity, and Concept-Net relations. Based on these features we designed a system that run several experiments on a SemEval-2018 dataset. The experimental results indicate that the proposed model performs better, or at least comparable with, other systems evaluated on the same data for this task.

Keywords: semantic difference, collocation, association measures, n-gram counts, word2vec, Concept-Net relations, semantic modelling

1. Introduction

Semantic modelling in natural language processing requires attending to both semantic similarity and difference. While similarity is well-researched in the community (Mihalcea & Hassan, 2017), the ability of systems in discriminating between words is an under-explored area (Krebs et al., 2018). A semantic model is more robust if it becomes sensitive to differences alongside similarities in meaning. For example, the concepts rain and snow are very similar, however the attribute liquid is challenging for a system whose aim is to understand natural languages. Such

https://doi.org/10.1075/ivitra.24.16tas

attributes (that might be very easy for humans to grasp) have been also researched as a kind of commonsense knowledge reasoning (Davis, 1990).

Distributional semantics, which is a common research area in semantic representation from the early times (cf. Firth, 1957 [1968]), is mainly exploited for finding similarities between words (Turney & Pantel, 2010). The main idea behind distributional semantics states that linguistic items with similar distributions have similar meanings (Blevins, 2016). The idea is further developed in state-of-the- art word representation models such as Mikolov et al. (2013). The effectiveness of a word representation model can be more rigorously evaluated by quantifying its strength in finding differences between words. Santus et al. (2018) state that the task can also be a useful addition for the creation of ontologies and other types of lexical resources.

For this study, semantic difference is operationalised as follows. Given two semantically related words and a discriminative feature, the feature word should only characterise the first one. An example is the triplet *apple*, *banana*, *red*, in which *red* can only be an attribute for *apple* (i.e. the attribute *red* can discriminate *apple* from *banana*). In this sense, discriminative attributes are properties that people tend to find important for a given concept. The idea is that one can express semantic differences between concepts by referring to attributes of the concepts. This practice is defined by Krebs and Paperno (2016) as an evaluation set that captures differences between concepts. The so-called non-trivial semantic task was proposed by Krebs et al. (2018) as a competition in the Semantic Evaluation (SemEval 2018) conference. In such a competition, known as SemEval shared task, participants are provided with a shared annotated dataset and they are asked to design systems that automatically predict the annotation labels. After that, they are provided with a test dataset. All systems are then evaluated on the common dataset and compared to each other.

We propose two automatic approaches to capture discriminative attributes. One is a supervised support-vector machine (SVM) model (Cortes & Vapnik, 1995) and the other is a K-means clustering method (MacQueen, 1967). The features we design for both methods are scores computed for word pairs and triples with the aim of capturing different semantic relations. The first category of scores that we propose comes from co-occurrence statistics of the words. The motivation behind this is that the attribute discriminates a word, if its co-occurrence with the word is more salient compared to its co-occurrence with the second word. This lies at the heart of collocations (Cf. Smadja & McKeown, 1990; Hausmann, 2007). In this sense, an attribute is discriminative of a word with which it collocates. Two common features to extract this property are n-gram features and association measures which are further explained in Section 3.

Another related category of scores that we use comes from distributional similarity hypothesis. We expect that the attribute word should have a significantly higher similarity to the word that it discriminates compared to the other word. For this we use the recent word embedding methodology (Mikolov et al., 2013) which is widely adopted by state-of-the-art natural language processing systems.

The third category of scores is related to the hypothesis that discriminative attributes are common sense knowledge about a word. One promising resource to extract these knowledge-based features are semantic networks (Sowa, 1991) and we exploit ConceptNet (Speer & Havasi, 2013), in particular. In Section 3, we describe a formula that we propose to compute numerical features for each attribute word corresponding to an ordered pair of words. Our classification and clustering methodologies based on a knowledge-based ontology and co-occurrence counts are further evaluated and the results are reported and compared with other systems designed for the SemEval shared task. Our system ranked the fourth among the systems applied to the dataset of SemEval 2018 (Krebs et al., 2018). This study further elaborates on the advantages of the applied lexical features and discusses similarities and differences of the system with other systems that participated in the competition.

2. Related work

In the task of capturing discriminative attributes for words, different features have been used. These include collocational or co-occurrence-based features (Santus et al., 2018; Taslimipoor et al., 2018), word similarity features (Shiue et al., 2018), word embeddings (Santus et al., 2018), and finally the features extracted from taxonomy relations such as hypernymy (Is-A) or meronomy (Has-A) (Lai et al., 2018).

The term collocation was introduced by Firth (1957 [1968], 1968) to mean a mode of semantic analysis (meaning by collocation) and a stylistic means to characterise restricted languages. Later on collocation was equated with usual or habitual co-occurrence. Halliday's redefinition of collocation in probabilistic terms marks the beginning of the distributional or statistical approach to collocation: "the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a*, *b*, *c* …" (Halliday, 1966). The traditional lexico-semantic approach to collocation presupposes certain sense relations between the constituents of a collocation. Thus, collocations exhibit a bipartite structure, conventionally restricted, in which both collocates have a different semantic status: for example, in *commit suicide*, the base is the semantically autonomous word (*suicide*) and the

verb *to commit* is the collocate, that is, the semantically dependent component (cf. Hausmann, 2007).

Word similarity and embeddings can all be grouped as distributional similarity features. The main idea of distributional similarity is that words that occur in the same contexts tend to have similar meanings. Distributional similarity can be approximated by different similarity measures between word vectors, including cosine, Jaccard coefficient, Euclidean distance, etc. (Lee, 1999). In this way, semantic difference can be modelled as the reverse order of similarity or can be judged based on the distributional similarity with a third word (Attia et al., 2018). However not all semantic differences can be adequately captured using this method. There are many cases where the difference between two words originates from the absence or the presence of a feature that cannot be directly mapped to the vector difference between two related words. One such example is *dolphin* and *narwhal* that only differ in having a *horn* (Krebs & Paperno, 2016). Such attribute is more visual and rarely occurs in text. Therefore, combining linguistic and conceptual information would potentially strengthen a semantic model in capturing the meaning of a word.

To tackle this issue, some studies rely on human annotated list of different attributes related to a concept which are called feature norms (McRae et al., 2005). Despite their strength in encoding semantic knowledge, feature norms have not been widely used in practice because they are usually small in size and require a lot of work to assemble (Fagarasan et al., 2015). Lazaridou et al. (2016) is an earlier attempt at identification of discriminative features which focuses on visual attributes.

The need for conceptual information also exists for systems that have to cope with commonsense reasoning such as question answering (Mcskimin, 1977) and word sense disambiguation (Sussna, 1993). This information can be obtained from manually or automatically created semantic networks such as BabelNet (Navigli & Ponzetto, 2012), ConceptNet (Speer & Havasi, 2013), etc. A semantic network is usually a directed or undirected graph structure consisting of nodes of concepts and edges which represent semantic relations between concepts. ConceptNet is one such knowledge base including but not limited to relations such as *RelatedTo*, *IsA*, *HasA*, *PartOf*, *UsedFor* and *HasProperty*. Extracting any of these relations between a word and an attribute results in informative features to capture whether the attribute is discriminative of the word (Speer & Lowry-Duda, 2018). The representations learned on ConceptNet have also been proven successful in capturing discriminative attributes (Vinayan et al., 2019).

3. Methodology

Our goal is to define a simple interpretable metric that can be used to gauge semantic difference and to identify discriminative attributes. We hypothesise that for a triplet in this task, a stronger relation between the first word and the attribute (in comparison with the second word and the attribute)[1] is indicative of the attribute word being discriminative between the two words.

For each triple we define a discriminative score *Disc Score* (*w*1, *w*2, *attr*) as follows:

$$Disc\_Score\ (w1, w2, attr) = Score\ (w1, attr) - Score\ (w2, attr) \quad (1)$$

where *w*1, *w*2 and *attr* are the first, second, and third word respectively. *Score* is a variable function of relation between two words that can be any of the scores explained in Sections 3.1, 3.2, 3.3, and 3.4.

3.1 Association-based score

Statistical association measures have a long history in language processing. With the availability of huge corpora, these measures can be even more effective than before in finding collocations and associations between words. Meaning by collocation is essentially a corpus-driven/corpus-based and distributional model of linguistic analysis which strives to statistically uncover significant word co-occurrences. This model also presupposes an underlying extraction method based on the analysis of discontinuous co-occurrences and word distance, where the units thus retrieved are also termed collocations or collocates (cf. Stubbs, 2002).

Collocational behaviour between two words is a strong signal that suggests that one of the words can identify the other. As an example, in the triplet (*hair, body, curly*), the association score in (*hair, curly*) is much higher than (*body, curly*), suggesting that *curly* is a discriminative attribute between the other two words. For each triplet in this task, collocational behaviour of the attribute word with the first two words is measured to see whether the first word can be a better collocate than the other. To this end, we use several different association measures[2] to compute the outputs of the *Score* function in Equation 1.

---

1. This stronger relation corresponds to more common semantic context and/or higher co-occurrence probability.

2. Statistical methods for the automatic extraction of collocations require large corpora and the use of an association measure or a combination of association measures: mutual information (MI), chi-square ($\chi 2$), phi-square ($\Phi 2$), log-likelihood (LR), etc. For a comprehensive list of association measures see Evert (2005).

We measure the association of two words based on their co-occurrence within a 5-word span. We use SketchEngine (Kilgarriff et al., 2004, 2014) to extract these statistics from the huge enTenTen corpus (Jakubíček, Kilgarriff, Kovář, Rychlý & Suchomel, 2013). Specifically, for each pair of words, we extract PMI (Church & Hanks, 1990) (known as MI in SketcEngine), MI3 (Oakes, 1998), log-likelihood (Dunning, 1993), T-score (Krenn & Evert, 2001), log-Dice (Dice, 1945), and Salience (Kilgarriff et al., 2014) all as defined in SketchEngine.

3.2 Google N-Grams

A second quantitative method to extract collocations is based on n-gram frequency analysis (continuous co-ocurrences) and it also requires very large data. In this case, the units of analysis are continuous sequences of two or more words which are retrieved from corpora according to a specified frequency threshold, regardless of their meanings (compositional or non-compositional) and their structural status (Stubbs, 2002). Unlike collocations, which can be discontinuous, n-grams are always a set of continuous co-occurring words. For instance, *excruciating pain* is an Adj. + N. collocation, but it does not constitute one of the 3 bigrams that can be extracted from the sentence "The pain was excruciating" (1. *the pain*, 2. *pain was*, 3. *was excruciating*). This is a fundamental difference for their automatic extraction, as they require different techniques and procedures.

N-grams are frequently used in computational linguistics for a variety of purposes including language modelling and association measures based on lexical co-occurrence. A well-known collection of n-grams is Google Books Ngram Dataset.[3] This Dataset is a collection of phrases (between 1 and 5 words long) extracted from over 8 million books printed between 1500 and 2008.

We use PhraseFinder (Trenkmann, 2016), a free web API that makes it possible to look up words or phrases from this dataset using a wildcard-supporting query language. Using this resource, we derive two different features. In the first one, we only consider bigrams, and in the other, we consider up to 5-grams. In both cases, we count the number of times that words occur near one another within a given span, regardless of order. We follow the same formula as defined in Equation 1. In order to eliminate the bias of high/low frequency words we divide *Disc_Score* by *Score* (*w*1, *attr*) + *Score* (*w*2, *attr*) that we compute from n-gram co-occurrence counts.

3. https://books.google.com/ngrams.

3.3 Word embedding-based score

Word embedding is a type of word representation that allows words with similar meaning to be understood by machine learning algorithms: words are mapped into vectors of real numbers using a neural network.[4] The basic assumption is that this model can create vectors that categorise similar words together and place them far away from vector representation of different words. Thus, words that have the same meaning have a similar representation. For instance, word embedding will create the vector representation of drinks (*water*, *coffee*, *tea*, *juice*, *milk*, *wine*, etc.) as clearly separated from the vector of furniture (*table*, *cupboard*, *chair*, *bed*, *sofa*, *chest of drawers*, etc.).

In distributional semantics, word embeddings are used to induce meaning representations for words. These methods are inspired by neural network language modelling and have become a basic building block for most applications in computational linguistics. The most popular word embedding method is word2vec (with the skip-gram architecture) which learns dense vector representations for words using an unsupervised model. Word2vec's training objective is based on DH, defined so that the model can learn word vectors that are good at predicting nearby words (Mikolov et al., 2013). Another popular embedding technique is GloVe which, like word2vec, preserves semantic analogies in the vector space. One major difference between the two models is that GloVe uses corpus statistics by training on global co-occurrence counts rather than local context windows (Pennington, Socher, & Manning, 2014).

In our system we use a concatenation of two sets of pre-trained embeddings. The first is trained on English Wikipedia using a variation of word2vec (Bojanowski, Grave, Joulin, & Mikolov, 2016). The other called ConceptNet Numberbatch (Speer & Lowry-Duda, 2017), is an ensemble of pre-trained Glove and word2vec vectors whose values are readjusted using a technique called retrofitting (Faruqui et al., 2014). In retrofitting, the values of the embeddings are updated using a training function that considers relational knowledge.

Using each word embedding, we compute cosine similarity between each word in a triplet and the attribute word to account for the statistics *Score* (*w*1, *attr*) and *Score* (*w*2, *attr*) in Equation 1.

4. Word embedding is also termed distributed semantic model, distributed represented model or (semantic) vector space model.

3.4 ConceptNet score

Co-occurrence based measures are not sufficient to account for all the various semantic relations that can exist between two words. Knowledge-based ontologies (e.g. ConceptNet, BabelNet etc) encode information about words and their relations in a structured way. This additional source of semantic information can be used to determine whether or not an attribute is discriminative. Because of its free web interface and ease of use, we use ConceptNet to empower our system with relational knowledge (Speer & Havasi, 2013). ConceptNet provides a large semantic graph to be used by computer applications. It represents general background knowledge and the way it is expressed in natural language (words and common phrases).

For any given (*w*1, *w*2, *attr*) triplet, using ConceptNet's REST API we query *w*1, limiting the number of search results to 1,000. The output is a JSON file that contains all relations between the queried word and other concepts. We traverse all the relations and count the number of times *attr* is linked to *w*1 to compute *score* (*w*1, *attr*). We repeat the procedure for *w*2 and compute *score* (*w*2, *attr*) and substitute them in Equation 1.

4. Experiments

4.1 Data

The dataset provided by Krebs et al. (2018) as part of the shared task on "capturing discriminative attributes" (as explained in Section 1) contains manually verified triplets of the form <word1, word2, attribute>. The attribute characterises the first word only and hence based on this definition, semantic difference in this dataset is asymmetric (Krebs et al., 2018). The data includes both positive and negative examples. Positive examples are like <*tray*, *pan*, *rectangular*> and negative examples can range from the one that the attribute can refer to: none of the words (e.g. <*squirrel*, *leopard*, *fur*>) or both words (e.g. <*dresser, cupboard, large*>).

The triplets are divided into three sets: one set for training, a second set for validation and a third set for testing. The test set would be kept blind and the models are trained on training set and hyper-parameters are optimised on the validation set. In order to ensure that models do not rely on attribute memorisation, the division is done so that no attribute in the test set or the validation set is also present in the training set. The statistics about the dataset are represented in Table 1 from (Krebs et al., 2018).

Table 1. The distribution of data into train, validation and test

	train	validation	test
positive	6591	1364	1047
negative	11191	1358	1293
total	17782	2722	2340

4.2 Experimental setup

The final feature set is the collection of Disc-Score measures based on the set of proposed scores. As a result we have six association-based scores, two google n-gram based scores, two embedding based scores, and one ConceptNet score. In total, we have eleven scores as our features.

In ConceptNet, reliability of each relation is given by a weight score. We decided to ignore this information and opted for raw counts because it did not help performance. Furthermore, binarising the scores based on raw counts (with 0 as a threshold) slightly improved the results. We use the features in both a supervised scenario (using SVM) and an unsupervised scenario (using KMeans). In both cases, all eleven features are exploited.

4.3 Evaluation metrics

The evaluation in this shared task is in terms of the average of positive and negative F1-scores which are standard in binary classification tasks. In this chapter, we report the precision, recall and F1-score for both positive and negative labels separately, along with the average F1-score.

The baseline system adopted by Krebs et al. (2018) is a simple unsupervised method that classifies a triplet as positive if the similarity of the attribute and the first word is greater than its similarity to the second word. The performance of the baseline is reported in Section 5. They also calculate the upper bound performance by human on the dataset which is F1-score of 0.9.

5. Results and discussion

Table 2 shows the results on both validation and test sets. The validation set is available to the system at the time of training and we perform hyperparameter optimisation on that. The test set however is blind to the system. We report the results

of both our supervised (SVM) and unsupervised (KMeans) models and compare them with the baseline and the top system (Lai et al., 2018) applied to this dataset.

Table 2. Results on Validation and TEST sets

			Precision	Recall	F1-score	Average F1-score
Validation	SVM	pos	0.7679	0.5652	0.6512	0.6913
		neg	0.6548	0.8284	0.7315	
	KMeans	pos	0.7039	0.6833	0.6935	**0.6972**
		neg	0.6910	0.7113	0.7010	
TEST	baseline		–	–	–	0.607
	SVM	pos	0.7299	0.6065	0.6625	**0.7142**
		neg	0.7197	0.8183	0.7658	
	KMeans	pos	0.6464	0.7001	0.6722	0.6930
		neg	0.7396	0.6899	0.7139	
	Top System 1	–	–	–	–	0.75

According to Table 2, our systems significantly outperform the baseline and underperform the Top System 1 by lower F1-score of less than 0.04. It is surprising that the unsupervised model (KMeans) can cluster the validation data as well as or even better than the supervised classification approach (SVM). Unsupervised models do not require training data. These models only use the validation data for hyperparameter optimisation.

This can be explained by the fact that the features we employ for this task are all computed using a formula that is specifically defined to represent semantic difference, and that finding whether a feature is discriminative between two words closely correlates with the semantic difference between them. Another reason could be that the training dataset is very noisy (cf. Krebs et al., 2018). The best performing system (Top System 1), in fact, got the best result by being trained directly on the validation data, otherwise by training on both train and validation data, their performance was reported to be 0.721 (Lai et al., 2018). This system is similar to our system in the sense that they are using SVM and word similarities as one of their feature types. One difference is that they rely on taxonomy relations from Probase, which can be considered a limitation when such taxonomies are not available.

We can see from the results that our features are well generalised as they lead to even better performance on the held-out test data. In order to see the effectiveness of the scores we obtained from ConceptNet, we retrained the model excluding the ConceptNet based measure and also the vectors derived from Numberbatch embedding. As a result, the validation performance dropped to 0.6857 and the test result decreased to 0.6969 in terms of average F1-score.

One advantage of our system is its ability to capture genealogical and kinship relations, as in (*grandson, brother, male*). Some train and test triplets require hierarchical reasoning, as in (*invertebrate, insect, shell*). Our model captures these kinds of relations very well, as it has access to information from a knowledge base. It is worth noting that a large part of the test triplets requires the knowledge to understand whether something is a constituent of another entity, as in (*beer, wine, foam*). It appears that these relations are well captured using co-occurrence-based metrics (collocations) alone since deleting knowledge-based features leaves the results for these triplets for the most part unchanged.

6. Conclusion

Semantic similarity is a well-represented research topic in Computational Linguistics. There are plenty of procedures and metrics to compute semantic similarity among words or even texts that use statistics from corpora. In this paper we have described an alternative procedure from the opposite perspective: computing semantic difference. Our model provides a simple metric in order to discriminate among words in relation to an attribute. The approach is based on a combination of knowledge-based and co-occurrence features (collocations, n-grams and word embeddings). Simple and robust, our method can be successfully used as an addition to semantic modelling, as it computes the difference among words optimally.

References

Attia, M., Samih, Y., Faruqui, M., & Maier, W. (2018). GHH at SemEval-2018 task 10: Discovering discriminative attributes in distributional semantics. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 947–952). New Orleans, LA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S18-1155. https://doi.org/10.18653/v1/S18-1155

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199593545.001.0001

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Davis, E. (1990). *Representations of commonsense knowledge*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26, 297–302. https://doi.org/10.2307/1932409

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Evert, S. (2005). *The statistics of word co-occurrences: word pairs and collocations*. (PhD Thesis, University of Stuttgart, Stuttgart). Retrieved from http://elib.uni-stuttgart.de/opus/volltexte/2005/2371.

Fagarasan, L., Vecchi, E. M., & Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th international conference on computational semantics* (pp. 52–57).

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*. https://doi.org/10.3115/v1/N15-1184

Firth, J. (1957 [1968]). A synopsis of linguistic theory, 1930–1955. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth, (1952–59)* (pp. 168–205). London: Longmans.

Firth, J. (1968). Linguistic analysis as a study of meaning. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth, (1952–59)* (pp. 12–26). London: Longmans.

Halliday, M. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In memory of John Firth* (pp. 148–162). London: Longman.

Hausmann, F. (2007). Die Kollokationen im Rahmen der Phraseologie – Systematische und historische Darstellung. *Zeitschrift für Anglistik und Amerikanistik*, 55, 217–234.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., & Suchomel, V. (2013). The tenten corpus family. In *7th international corpus linguistics conferencecl* (p. 125–127).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V. (2014). The Sketch engine: ten years on. *Lexicography*, 1(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Rychlỳ, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams, & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–116). Lorient: Université de Bretagne-Sud.

Krebs, A., Lenci, A., & Paperno, D. (2018). SemEval-2018 task 10: Capturing discriminative attributes. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 732–740). New Orleans, LA: Association for Computational Linguistics.

Krebs, A., & Paperno, D. (2016). Capturing discriminative attributes in a distributional space: Task proposal. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP* (pp. 51–54). https://doi.org/10.18653/v1/W16-2509

Krenn, B., & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations* (pp. 39–46).

Lai, S., Leung, K. S., & Leung, Y. (2018). SUNNYNLP at SemEval-2018 task 10: A support-vector-machine-based method for detecting semantic difference using taxonomy and word embedding features. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 741–746). New Orleans, LA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1118

Lazaridou, A., Baroni, M., et al. (2016). The red one!: On learning to refer to things based on discriminative properties. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 213–218). https://doi.org/10.18653/v1/P16-2035

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (pp. 25–32). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1034678.1034693

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1: Statistics* (pp. 281–297). Berkeley, CA: University of California Press. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200512992

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4), 547–559. https://doi.org/10.3758/BF03192726

Mcskimin, J. R. (1977). The use of a semantic network in a deductive question- answering system. In *Proc. IJCAI* 5 (pp. 50–58).

Mihalcea, R., & Hassan, S. (2017). Similarity. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217–250. https://doi.org/10.1016/j.artint.2012.07.001

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-116

Santus, E., Biemann, C., & Chersoni, E. (2018). BomJi at SemEval-2018 task 10: Combining vector-, pattern- and graph-based information to identify discriminative attributes. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 990–994). New Orleans, LA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1163

Shiue, Y.-T., Huang, H.-H., & Chen, H.-H. (2018). NTU NLP lab system at SemEval-2018 task 10: Verifying semantic differences by integrating dis- tributional information and expert knowledge. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1027–1033). New Orleans, LA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1171

Smadja, F. A., & McKeown, K. R. (1990). Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th annual meeting of the association for computational linguistics* (pp. 252–259). Pittsburgh, PA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P90-1032 https://doi.org/10.3115/981823.981855

Smadja, F. A., & McKeown, K. R. (1990). Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th annual meeting of the association for computational linguistics* (pp. 252–259). Pittsburgh, PA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P90-1032 https://doi.org/10.3115/981823.981855

Sowa, J. F. (Ed.) 1991. *Principles of Semantic Networks. Explorations in the Representation of Knowledge*. San Mateo, California: Morgan Kaufmann.

Speer, R., & Havasi, C. (2013). Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web meets NLP* (pp. 161–176). Springer. https://doi.org/10.1007/978-3-642-35085-6_6

Speer, R., & Lowry-Duda, J. (2017). Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th international workshop on semantic evaluation (semeval- 2017)* (pp. 85–89). https://doi.org/10.18653/v1/S17-2008

Speer, R., & Lowry-Duda, J. (2018). Luminoso at SemEval-2018 task 10: Distinguishing attributes using text corpora and relational knowledge. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 985–989). New Orleans, LA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1162

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7 (2), 215–244. https://doi.org/10.1075/ijcl.7.2.04stu

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international conference on information and knowledge management* (pp. 67–74). New York, NY: ACM.

Taslimipoor, S., Rohanian, O., Ha, L. A., Corpas Pastor, G., & Mitkov, R. (2018). Wolves at SemEval-2018 task 10: Semantic discrimination based on knowledge and association. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 972–976). New Orleans, LA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1160

Trenkmann, M. (2016). *PhraseFinder – Search millions of books for language use*. http://phrasefinder.io/ (Accessed: 2018-01-30).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141–188. Retrieved from http://dl.acm.org/citation.cfm?id=1861751.1861756

Vinayan, V., Anand Kumar, M., & Soman, K. P. (2019). Capturing discriminative attributes using convolution neural network over conceptnet number-batch embedding. In V. Sridhar, M. Padma, & K. R. Rao (Eds.), *Emerging research in electronics, computer science and technology* (pp. 793–802). Singapore: Springer Singapore.

Index

Whether you wish to *deliver on a promise*, *take a walk down memory lane* or even *on the wild side*, phraseological units (also often referred to as phrasemes or multiword expressions) are present in most communicative situations and in all world's languages. *Phraseology*, the study of phraseological units, has therefore become a rare unifying theme across linguistic theories.

In recent years, an increasing number of studies have been concerned with the computational treatment of multiword expressions: these pertain among others to their automatic identification, extraction or translation, and to the role they play in various Natural Language Processing applications. Computational Phraseology is a comparatively new field where better understanding and more advances are urgently needed. This book aims to address this pressing need, by bringing together contributions focusing on different perspectives of this promising interdisciplinary field.

ISBN 978 90 272 0535 3

John Benjamins Publishing Company