Applying an Information-theoretic Approach for Automatic Identification of German Multi-word Expressions

Sergei Bagdasarov
Saarland University, Germany
sergeiba@lst.uni-saarland.de

Elke Teich
Saarland University, Germany
e.teich@mx.uni-saarland.de

Abstract

In this study, we apply a largely information-theoretic approach to identify and extract German multi-word expressions (MWEs) from selected corpora. We analyze the register-specific use of MWEs, comparing medical scientific abstracts and newspaper articles. Our results show that the two registers vary in the use of MWEs both structurally and semantically. For instance, scientific texts exhibit more discourse-related MWEs and expressions that denote objects or phenomena, whereas in newspaper texts we observe more MWEs denoting persons or functioning as references to time or place. Furthermore, we interpret these differences in terms of the situational context, i.e. field (topic), tenor (attitude), and mode (discourse).

1 Introduction

Multi-word expressions (MWEs), formulaic sequences of at least two words that tend to co-occur together, make up a considerable proportion of language. MWEs vary greatly in their nature, ranging from highly opaque idiomatic expressions like *da liegt der Hund begraben*, complex prepositions (*in Bezug auf*) to light verb constructions (*zur Verfügung stellen*) and terminological patterns (*lokale Tumorkontrolle*). Despite their heterogeneity, MWEs seem to share a common property of being processed holistically as whole units or at least providing predictable transitions from one token to the other. This results in a processing advantage in comparison to non-formulaic sequences of comparable length (Tremblay and Baayen, 2010; Siyanova-Chanturia et al., 2017).

This processing advantage make MWEs an important factor influencing the formation of an efficient code of communication, especially within clearly distinguishable functional varieties of language, i.e., registers. However, the functional variation in the use of formulaic language, particularly in languages other than English, seems to be an under-researched topic in linguistics. Another yet unresolved question, despite more than two decades of active research, is the accurate identification, extraction, and classification of MWEs.

Our goal here is, therefore, twofold. First, we test a novel information-theoretic approach for MWE extraction (Gries, 2022a) that, to the best of our knowledge, has not been used on German data yet. Second, we aim to analyze the formulaic language in German scientific and newspaper texts and establish how the contextual differences between these two registers impact the use of MWEs.

The remainder of the paper is structured as follows. Section 2 offers an overview of related work on register variation in English and German, as well as different approaches to MWE identification. Section 3 presents our data and briefly describes the MWE identification procedure used in this study. In Sections 4 and 5 we present the MWEs extracted from our data and discuss the use of formulaic language in German scientific and newspaper texts. And finally, Section 6 summarizes the present work and outlines the potential for future research.

2 Related Work

2.1 Register Variation

Conceptually, our work is rooted in register theory, register being defined as a language variety that can be described in terms of its context of situation (field, tenor, and mode (Halliday and Hasan, 1985)). The field of discourse refers to the topic of a text as well as to the overall purpose of communication (e.g., transmit information, describe world phenomena, etc.). Thus, variation in field is especially evident at the lexical level including specific terminology in specialized texts (such as scientific texts). The tenor of discourse describes the relationship between the author and the audi-

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 295–305, 2025

ence (e.g., professional vs. layperson) as well as the author's attitude towards the subject matter of a text. Finally, the mode refers to the channel and medium of communication (spoken vs. written) as well as to the general discourse organization of a text.

Building on this theoretical framework, Biber and Gray (2010) and Biber and Conrad (2019) conducted extensive research on English scientific writing, comparing it to other registers. There are also numerous studies that adopt a diachronic perspective showing the evolution and formation of a register. For instance, Biber and Gray (2011) show how scientific English evolved towards a more nominal style of writing. Similarly, Degaetano-Ortlieb and Teich (2022) and Degaetano-Ortlieb (2021) apply information-theoretic methods to trace how scientific English became more conventionalized and informationally denser over a period of 300+ years.

While the research on register variation is mostly focused on English, the German language has been studied from this perspective as well. For example, Neumann (2014) carried out a comprehensive study of eight registers in German and English based on a broad selection of linguistic features; or Kunz and Lapshinova-Koltunski (2015) and Krielke (2021) explore discourse-related and syntactic aspects of register variation in German, also drawing a comparison with English.

However, register-based studies on MWEs seem to be rather limited. Conrad and Biber (2005) analyze lexical bundles in different English registers. Similarly, Breeze (2013) studies the same type of MWEs in several legal subregisters. Adopting a broader perspective on MWEs, Alves et al. (2024a) and Alves et al. (2024b) conducted diachronic analyses of different MWE types in English scientific writing. However, to our knowledge, no comparative register-based study of German MWEs has been carried out yet.

2.2 MWE Identification

In MWE research, the problem of identification and extraction of MWEs is of key importance. Despite diverse and persistent efforts invested into developing extraction solutions in recent decades, this problem still remains a "pain in the neck" for linguists and NLP scholars (Sag et al., 2002).

The proposed approaches vary greatly in their conceptual basis and implementation. Some scholars rely on frequency as their criterion for MWE identification, extracting all sequences of n words that are more frequent than a defined threshold (lexical bundles) (Conrad and Biber, 2005; Breeze, 2013). Frequency is sometimes combined with some type of association measure such as mutual information (Simpson-Vlach and Ellis, 2010). Building on the idea of co-occurrence and association, Brunner and Steyer (2015) designed a corpus-driven MWE extraction approach based on collocation profiles. Again with a focus on German, Weller and Heid (2010) leverage syntactic relations to extract verbal MWEs by combining syntactic dependencies and association measures.

A variety of more computational extraction methods can also be found in the literature. For instance, Klyueva et al. (2017) use neural networks to extract verbal MWEs in 15 languages. In turn, Cap et al. (2013) trained a Conditional Random Fields classifier to detect light verb constructions in German. Moreover, a number of works rely on word embeddings for this task (Scherbakov et al., 2016; Loukachevitch and Parkhomenko, 2018).

While all these approaches have their advantages, they are subject to certain limitations. Methods that prioritize lexical bundles overlook less frequent MWEs or shorter word sequences that, however, might still be worthy of analysis (e.g., some named entities, terminological patterns). Syntactic approaches rely on previous parsing of texts and, therefore, are language-sensitive. Moreover, they often focus only on one specific type of MWEs. Automatic approaches are limited by the data sets used for training. For instance, systems that participated in the PARSEME shared task (Ramisch et al., 2020) only cover verbal MWEs in different languages, while works presented in the DiMSUM shared task (Schneider et al., 2016) focus only on English MWEs.

To overcome these limitations, some scholars combine different state-of-the-art methods in their MWE research (Alves et al., 2024a,b), which improves recall. Alternatively, one could apply a statistically motivated and language-independent extraction procedure that is capable of retrieving MWEs of different types. Such a procedure was developed by Gries (2022a) and will be described in the next section.

3 Data and Methods

3.1 Data

For the analysis of MWEs in German newspaper texts we use the Tiger corpus (version 2.2) (Brants

et al., 2004). The Tiger corpus contains newspaper articles extracted from *Frankfurter Rundschau*, consisting of over 800,000 tokens.

To cover texts of the medical domain, we use the German part of the Springer medical corpus available through the MuchMore project.[1] The corpus was compiled using medical abstracts from 41 different domains ranging from arthroscopy to legal medicine, with the German part containing roughly 1,000,000 tokens. General corpus statistics are summarized in Table 1

Corpus	Texts	Sentences	Tokens
Tiger	2,163	50,474	843,232
Springer	7,808	54,839	1,137,843

Table 1: Corpus statistics for Tiger and Springer datasets.

3.2 Extraction Method

We use a largely information-theoretic method proposed by Gries (2022a) to automatically identify and extract MWEs. The core idea of the method boils down to splitting a corpus into bigrams, iterating through the bigrams, and selecting the best MWE candidates based on eight dimensions. The best candidates are then merged into one single unit, and the process is repeated N number of iterations.

The eight dimensions[2] involved in this method capture different aspects of formulaicity that characterize MWEs. Given a hypothetical bigram *ab*, these dimensions are as follows:

Dimension 1 (frequency): total bigram frequency; here, we apply a threshold of at least 10 occurrences.

Dimension 2 (dispersion): how well a bigram is distributed throughout the corpus, as measured by normalized Kullback-Leibler divergence (KLD):

$$\mathrm{KLD}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (1)$$

$$\mathrm{KLD}_{\mathrm{norm}} = 1 - e^{-\mathrm{KLD}} \quad (2)$$

Dimensions 3 and 4 (type frequency): how many unique token types occur after *a* and before *b*.

Dimensions 5 and 6 (normalized entropy): how much uncertainty there is about what follows the token *a* and precedes the token *b*:

$$H_{\mathrm{norm}} = -\frac{\sum_{i=1}^{n} p_i \times \log_2 p_i}{\log_2 n} \quad (3)$$

Dimensions 7 and 8 (association): how strongly *a* is attracted to *b* and vice versa, as measured by normalized KLD (see again Equation 2).

After calculating the dimensions, they are normalized to range from 0 to 1 (if not already), creating an eight-dimensional vector. Euclidean distance (ED) is then computed to measure the distance of each bigram from the center of the eight-dimensional space and select the most suitable MWE candidate.

We implemented this extraction algorithm in Python, adding a preprocessing step in which all numbers, special characters and punctuation marks except for hyphens are removed. No lemmatization was performed during the preprocessing. For now, we limit the number of iterations to 200 due to the high computational load, with each iteration yielding 5 MWEs.

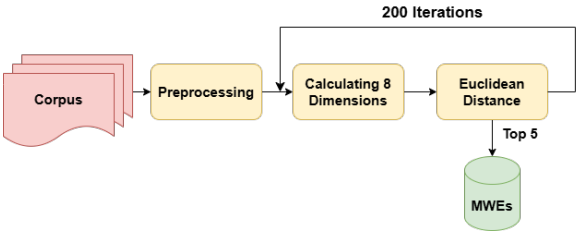

Figure 1: MWE extraction pipeline.

This extraction logic is visualized in Figure 1 and can be further illustrated with the example of the MWE *in der laufenden Periode*. First, an MWE candidate composed of tokens *laufenden* and *Periode* was ranked second by ED in iteration 29 and merged into one unit *laufenden Periode*. Then, an MWE candidate consisting of elements *der* and *laufenden Periode* was ranked third by ED in iteration 32 and merged into the unit *der laufenden Periode*. And finally, in iteration 151, an MWE candidate consisting of elements *in* and *der laufenden Periode* was ranked second by ED and merged into the final MWE *in der laufenden Periode*.

For the follow-up analysis of the extracted MWEs, we rely on the association measure again (Equations 1 and 2). However, instead of calculating the strength of the forward and backward

[1] https://muchmore.dfki.de/resources1.htm

[2] Here, we just give a general overview of these dimensions. For more information on calculations and reasons for selecting the dimensions, see Gries (2022a), Gries (2020) and Gries (2022b).

association, we obtain one measure for the whole MWE. For instance, if an MWE consists of two elements, we calculate the association *a–>b* and *b–>a* and take the mean of both values. If an MWE has more than two elements, we calculate the association for each MWE part as described above and then take the mean of the resulting values. For example, the association strength of the MWE *in der laufenden Periode* would be the mean of the association scores of *in* <–> *der*, *in der* <–> *laufenden*, and *in der laufenden* <–> *Periode*.

4 Results

4.1 Extraction Results

The extraction procedure applied as described above yielded 1,000 potential MWEs for each corpus. However, thorough filtering was still required to reduce noise. First, we automatically removed all MWE candidates that were included in longer sequences. For example, *jeden Fall* was rejected because it also occurs in *auf jeden Fall*. This led to the exclusion of 194 MWE candidates from the Springer corpus and 176 MWE candidates from the Tiger corpus.

Subsequently, a human annotator with a solid linguistic background was tasked with the manual evaluation of the remaining MWE candidates. The annotator labeled the potential MWEs as successful and unsuccessful with 1s and 0s respectively according to the following criteria:

- A successful MWE consists of at least two words.
- A successful MWE constitutes a semantic unit but not necessarily a non-compositional one.
- A successful MWE does not consist only of two tokens that are grammatically or syntactically inherently bound to each other.
- A successful MWE can be allocated to one of the 10 predefined categories.[3]

The categories used for the classification of the extracted MWEs are:

Object/Phenomenon: material objects, abstract concepts, world phenomena (*der koronaren Herzkrankheit, moderne Wirtschaftspolitik*);

Person: persons' names or MWEs referring to individuals or groups and/or their function or status (*Bundestagspräsidentin Rita Süssmuth, ein Sprecher des*);

Organization: political/economic entities (*der Bundesrepublik Deutschland, der Europäischen Union*);

Action/Process/Condition: verbal or nominal MWEs denoting actions and processes or describing persons or objects (*scharf kritisiert, zur Verfügung gestellt, die statistische Auswertung*);

Discourse organizer: MWEs used to create text cohesion (*Ziel dieser Arbeit, aus diesem Grund, vor allem*);

Place: geographic locations, directions, references to places with varying degree of specificity (*in den USA, an unserer Klinik, in der Nähe*);

Time: time references with varying degree of specificity (*im nächsten Jahr, unmittelbar vor, nach Herztransplantation*);

Quantity: MWEs indicating the amount of measurable objects (*Milliarden Mark, ein paar, eine Vielzahl von*);

Attribute: characteristics or features of persons or objects (*im Alter von, Kilometer langen, statistisch signifikanten*);

Mode of Action: MWEs describing an action or process (*in vitro, immer wieder, aus Protest gegen*).

After the filtering according to these criteria, we retained 609 MWEs extracted from the Tiger corpus and 553 MWEs extracted from the Springer corpus, resulting in a precision of 60.9% and 55.3% respectively. Youssef (2024) reported a comparable precision of 48% for English data. Note that the higher precision values in our results are likely due to the exclusion of all bigrams containing symbols or punctuation during preprocessing, reducing the number of unsuccessful MWE candidates. Recall could not be calculated, as the total number of MWEs present in the corpora is unknown.

We grouped the successful MWEs into several broad syntactic patterns based on the Universal Dependencies POS tags (de Marneffe et al.) of their components (see Table 2). We found more nominal and adjectival MWEs and in the Springer corpus, structures that reflect terminology formation patterns typical of scientific language (*einer retrospektiven Studie*). Verbal MWEs are also more prominent in the Springer corpus. These are mostly expressions used for topic introduction or method description (*retrospektiv analysiert, wir untersuchten*).

In contrast, the most common MWEs extracted from the Tiger corpus are prepositional phrases,

[3]These criteria are an adapted version of the annotation guidelines used in Youssef (2024)

probably due to the need for very specific time and place references in newspaper texts (*im Ausland, am heutigen Dienstag*).

Pattern	Tiger	Springer
NP	257	279
PP	255	117
AdjP	3	23
VP	50	101
AdvP	24	15
Misc	20	18

Table 2: MWEs POS patterns. NP: noun phrase, PP: prepositional phrase, AdjP: adjectival phrase, VP: verb phrase, AdvP: adverbal phrase

Among unsuccessful MWEs, the most commonly encountered POS patterns are: determiner + noun (*der Kunde, des Parteitags*),[4] determiner + adjective (*ein neues, ein wichtiges*), reflexive pronoun + verb (*einigten sich*), and particle + infitive (*zu finanzieren*). These four patters account for 87% and 84% of rejected MWEs in Tiger and Springer corpora respectively (see Table 3 for a more detailed overview).

Pattern	Tiger	Springer
DET NOUN	102	89
DET ADJ	39	42
PART VERB	33	49
VERB PRON	14	33
Misc	27	40

Table 3: POS patterns of unsuccessful MWEs. Category "Misc" includes POS patters that correspond to less than 10 unique MWEs.

In successful MWEs, the first component tends to be more strongly connected to the second component than vice versa. The opposite trend is observed in unsuccessful MWEs. This pattern is reflected in both forward and backward association measures, as well as in the distribution of items following the first component and preceding the second component (see Figures 2 and 3). These differences are in line with the POS patterns of the excluded MWEs most of which are determiner+noun or determiner+adjective sequences. No differences in frequency and dispersion between successful and unsuccessful MWEs were observed.

[4]However, we opted to keep some cases where determiner and noun/adjective do form a fossilized construction (*ein Viertel, dieses Jahres*).

4.2 MWE Categories

The most common MWE categories in the Tiger corpus are **Place**, **Time**, **Person**, **Discourse organizer** and **Action/Process/Condition** (see Figure 4). The latter two are also prominently attested in the Springer corpus. However, **Object/Phenomenon** is the most common MWE type in scientific texts. In the following paragraphs, we will have a closer look at the most relevant MWE categories in both corpora.

Object/Phenomenon. In the Springer corpus, this MWE type is mostly represented by terminology referring to (a) diseases (*koronare Herzkrankheit*), (b) body parts (*langen Röhrenknochen, unteren Extremitäten*), (c) substances (*Vitamin E, mit Antikörpern gegen*), or (d) abstract concepts (*therapeutische Konsequenzen, ein breites Spektrum*). Many of the Springer MWEs of this type are Latin or English terms (*Ramus interventricularis anterior, Injury Severity Score*). In the Tiger corpus, this MWE type is much less frequent and consists primarily of MWEs that denote abstract concepts from the field of politics, economy or sociology (*die absolute Mehrheit, rote Zahlen, gute Chancen*).

Discourse organizers. Springer data features MWEs that (a) describe authors' scientific activity (*wir beschreiben, wir untersuchten, berichten wir über*), (b) refer to the authors' paper or study (*Ziel dieser Arbeit, der vorliegende Artikel*), or (c) present or interpret results (*legen nahe daß, sprechen dafür daß, fanden wir*). In contrast, Tiger discourse organizing MWEs contain expressions that (a) indicate the source of information (*nach offiziellen Angaben, einem Bericht der*), (b) present overall context (*angesichts des*, *vor diesem Hintergrund*), or (c) describe personal attitude (*er hoffe, ich denke, nach Ansicht*). Apart from that, both corpora share a number of common cohesive devices (*in der Regel, in diesem Zusammenhang, vor allem, darüber hinaus*).

Action/Process/Conditions. As mentioned in Section 4.1, MWEs of this type comprise both verbal and nominal patterns. In the Springer texts, they often refer to diagnostic and analytical methods or treatments employed in the medical field (*retrospektiv analysiert, akustisch evoziert, Entfernung des Tumors*). In newspaper articles, these MWEs describe actions commonly seen in political or economic settings (*die Abschaffung der, im Kampf gegen*). Both sets of MWEs contain light verb

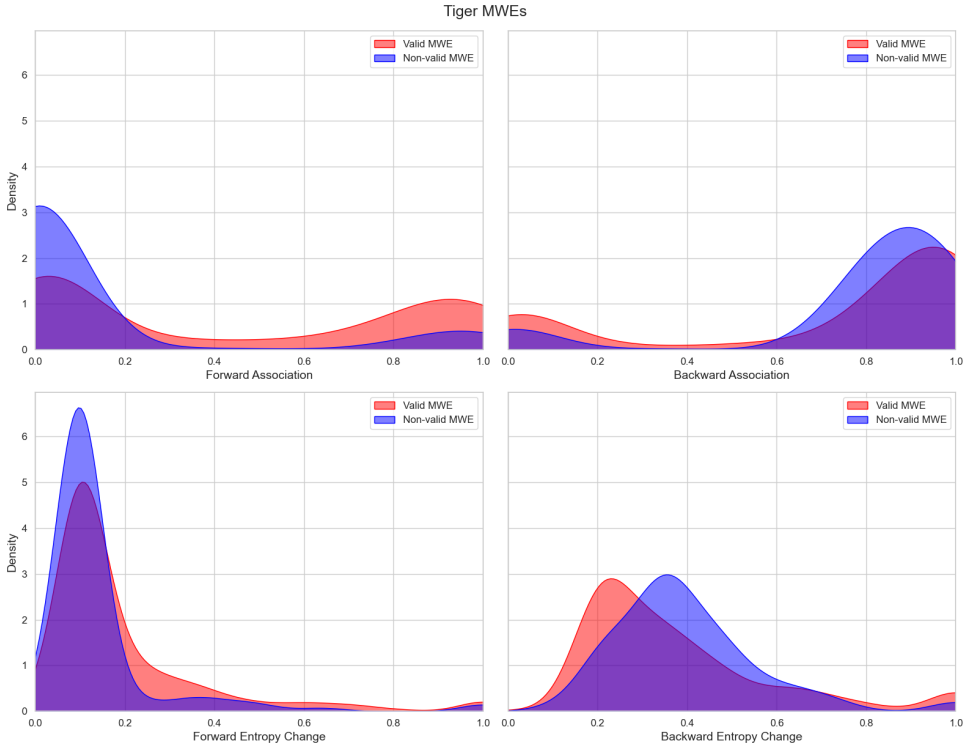

Figure 2: Comparison of successful and unsuccessful MWEs in the Tiger corpus.

constructions. However, those from the Springer corpus indicate cognitive and analytical processes (*in Betracht gezogen, in Erwägung gezogen*), while those from the Tiger corpus are more varied in their semantics (*in Frage gestellt*, *über die Bühne* [gehen], *ums Leben gekommen*, *ins Auge* [fassen]).

Place. With only seven unique expressions, MWEs of this type are not well represented in the Springer corpus. However, in the Tiger corpus this MWE type shows a wide range of realizations, with the most common being preposition + city/country/region (*in Genf, in Indien, in Afrika*). Additionally, we also found more generic and relative place reference (*im Ausland, nördlich von, im Süden*).

Time. Time MWEs in scientific texts mostly indicate the sequence of processes or events (*nach Beendigung der, nach dem Unfall, nach Abschluss der*). We also found some fixed time-related Latin expressions (*bis dato, post mortem*) and references to time periods (*in diesem Zeitraum, einen Zeitraum von*). Time MWEs in newspaper texts are especially used for comparing or presenting economic and financial statistics in different periods (*zum Jahresende, als im Vorjahr, in der laufenden Periode*). Specific time references for event contextualization are also richly attested in the Tiger corpus (*am späten, am Donnerstag Abend, der Nacht zum Freitag*). Both registers share some common MWEs used for generic time references (*in der Vergangenheit, seit langem, vor kurzem*).

Person. This type of MWEs is relatively sparsely attested in medical texts, describing mainly different groups of patients (*Patienten bei denen, Patienten mit koronarer Herzerkrankung*). In newspaper text, however, this category is primarily represented by proper names (*Bundespräsident Roman Herzog, Bundeskanzler Helmut Kohl*). References to people's social or political functions and different groups of people are also attested in our data (*ausländische Investoren, bosnische Serben, stellvertretende Vorsitzende*).

4.3 Association

In terms of association, MWEs in scientific texts tend to show stronger internal ties between the elements. Figure 5 shows that the Springer corpus contains a larger proportion of MWEs with an association score between 0.5 and 1, which shows a moderate to high association strength. In contrast, newspaper texts are more characterized by MWEs with a score between 0.2 and 0.5, indicating a rather weak association strength. Interestingly, the tendency towards higher MWE-internal ties in scientific texts also holds for 35 out of 49 MWEs that are attested in both corpora (see Table 4).

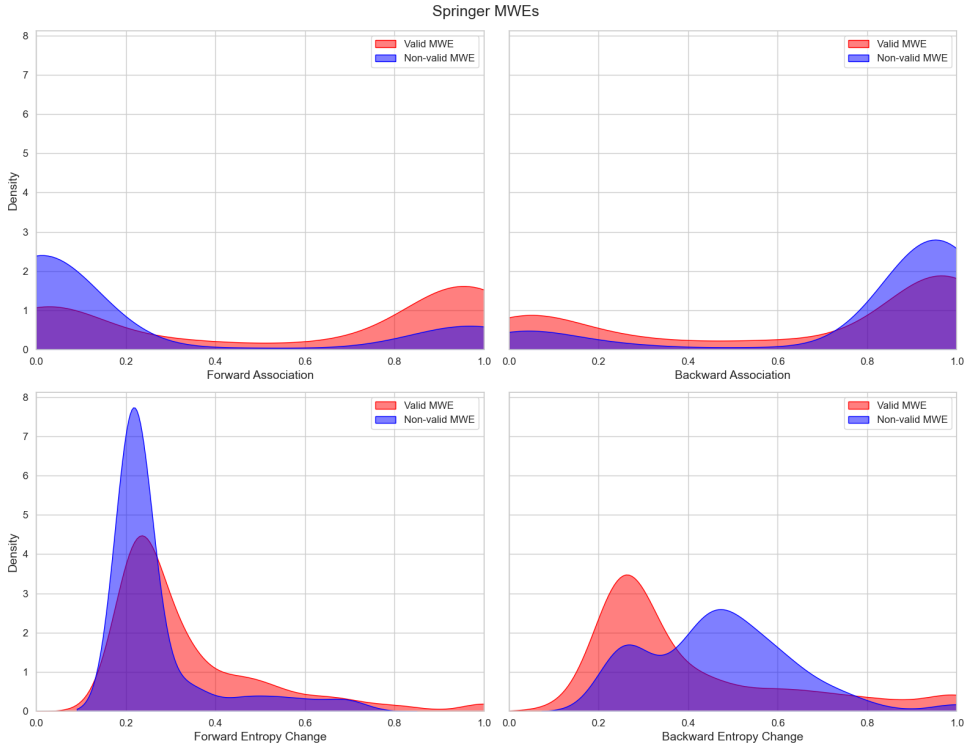

Figure 3: Comparison of successful and unsuccessful MWEs in the Springer corpus.

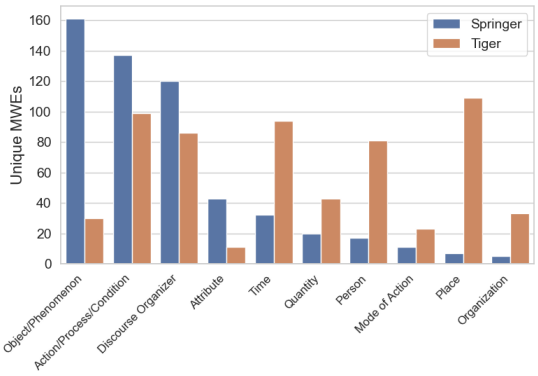

Figure 4: Comparison of MWE types in both corpora.

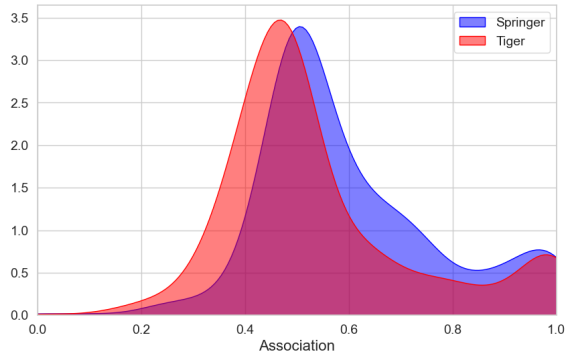

Figure 5: Comparison of mean association scores in Tiger and Springer MWEs.

MWE	Springer	Tiger
darüber hinaus	0.99	0.95
in erster Linie	0.73	0.68
doppelt so	0.59	0.52
Hinweis auf	0.52	0.46
die Tatsache daß	0.46	0.41
in der Regel	0.45	0.33

Table 4: Association scores for selected common MWEs.

Figure 6 offers a more fine-grained comparison of MWEs in different categories. **Object/Phenomenon**, **Person** and **Organization** are the categories showing the highest association strength of MWEs, while all other types show a rather moderate association strength among the elements of MWEs. The MWEs of scientific texts have a stronger association in almost all categories. The difference is by far most noticeable in the category **Object/Phenomenon**. In contrast, MWEs in the categories **Person** and **Organization** exhibit stronger word-to-word connections in newspaper texts.

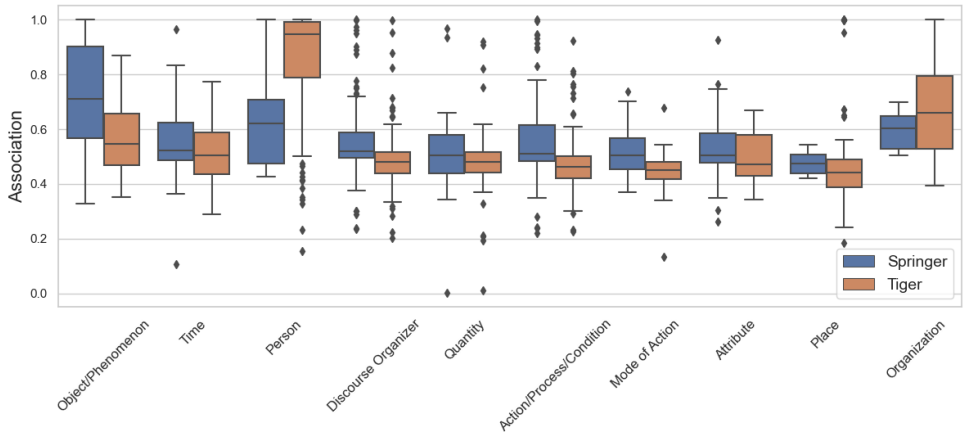

Figure 6: Strength of association across categories.

5 Discussion

As shown in Section 4.1, the output of the statistical method for MWE identification applied in this study is not perfect and still requires post-hoc filtering (see also the evaluation of the method in the original paper (Gries, 2022a, p. 14)).

Moreover, this extraction procedure, sequential in nature, sometimes struggles to cope with the variation in German syntax and morphology. For instance, some of the prepositional MWEs we encountered in the corpora are actually part of light verb constructions (e.g. *zu Buche* [schlagen], *auf der Kippe* [stehen], *zur Last* [legen], etc.). The verbs, however, were not extracted as part of the MWEs because in some of the MWE realizations they were separated from the prepositional phrase by other tokens. However, this problem could potentially be solved by increasing the number of iterations.

The flexible German syntax also led to variations of the same expression being retrieved as different MWEs, for instance, *es handelt sich* and *handelt es sich* in the Springer corpus. While this might seem redundant, we do believe that it is useful to retrieve both versions, since it allows us to further investigate the usage of MWEs. So, *handelt es sich* is more frequent and shows a slightly higher association strength than *es handelt sich*, suggesting that this expression tends to be used more in sentences with a pragmatically marked word order.

Despite its limitations, the applied extraction method has proved to be a versatile tool for MWE identification since it captures both high-frequency MWEs such as discourse organizers (*vor allem, wir berichten über, in der Regel*) and relatively rare word combinations that, however, do function as single units, for instance, proper nouns, locations, names of organizations and scientific terminology (*rheumatoider Arthritis, Bundesinnenminister Manfred Kanther, der Frankfurter Rundschau*).

The analysis of MWE types clearly reflects how the differences in the use of formulaic language are related to the contextual dimensions of register variation, i.e. field, tenor and mode.

In terms of the **field of discourse**, the main objective of scientific texts is to present new knowledge. This inevitably involves an extensive use of terminology. Consequently, we observe more nominal and adjectival MWE patters in the Springer corpus, and the MWE category Object/Phenomenon is by far more common in scientific texts compared to newspaper articles. Moreover, terminology in the scientific domain becomes more fossilized due to its repetitive use in similar contexts, indicated by a higher internal association strength between MWE components. In contrast, newspaper articles are more event-oriented than descriptive. For this reason, we see more MWEs providing local and temporal references, which are essential for establishing the context of an event, as well as MWEs referring to people and organizations who are protagonists in events.

The **tenor of discourse** can potentially explain the presence of foreign-language MWEs in scientific texts. Scientific texts are usually written by professionals in a specific field for their peers who share the same background knowledge. This shared knowledge allows them to use linguistic conventions that would not be understandable outside their community (e.g., Latin terminology). Newspaper

articles, in contrast, are written for a broad audience and tend to use expressions from "general language". Even if newspaper texts touch upon topics from specialized domains such as economics or politics, they rely on general terms that can be understood by readers with no in-depth knowledge of these domains. Similarly, a scientific text intended for a broader audience (e.g., a blog entry or a plain-language summary of a scientific paper) tends to avoid specialized terminology (especially Latin terms).

In terms of the **mode of discourse**, we can see similarities and differences between scientific and newspaper texts. With the two registers using the written channel of communication, both show a high proportion of cohesive devices, especially those used for topic elaboration or cause-effect relations. However, unlike the Tiger corpus, the Springer corpus contains a considerable fraction of MWEs that include references to the authors themselves or to their texts. Such self-references play an important role in topic introduction and guide the reader's attention. Scientific texts also exhibit a number of MWEs that are used as cohesive devices for presentation and interpretation of results, which are key elements in scientific studies.

6 Conclusion and Future Work

In this study, we used a statistically motivated and language-independent method for MWE identification proposed by Gries (2022a) to extract MWEs from German texts. The method leverages both frequency-based and information-theoretic measures to better capture the essence of formulaic word sequences and enables the retrieval of different MWE types. Although the extraction output requires subsequent filtering by human experts, this method is a promising approach for MWE identification, especially in languages other than English.

Our analysis showed that the variation in the use of formulaic language in German scientific and newspaper texts can be explained in terms of field, tenor, and mode of discourse, which is in line with previous research on functional variation (see Section 2 for a brief overview).

The main focus in our future work will be on enhancing the efficiency of our Python implementation of the extraction method and increasing the number of iterations to improve the MWE recall. Additionally, we would like to explore whether the addition of further dimensions can improve the extraction precision. In this respect, surprisal should be a suitable measure to favor the retrieval of those word sequences that exhibit the most predictable transition from one token to another. Apart from that, we also plan to add more registers to our analysis.

Limitations

The Springer corpus contains only texts from medical journals, albeit covering a broad variety of medical fields. Therefore, it might fail to capture some features of language use characteristic of other scientific domains. Moreover, it is only composed of paper abstracts, which tend to be written in a more condensed way than the body of the papers. Similarly, the Tiger corpus only comprises texts coming from one newspaper – *Frankfurter Rundschau* – and may therefore be biased in terms of style.

Our research design only allowed us to evaluate the precision of the extraction method. Since it is not known how many MWEs are present in the corpora, it is not possible to calculate the recall. An additional evaluation on a dataset with gold-standard MWE annotation would provide valuable insights into the method's performance in terms of recall.

Although the annotator performed the classification of the extracted MWEs to the best of their abilities, we are aware that alternative judgments might also be possible in some cases.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt, and Elke Teich. 2024a. Diachronic analysis of multi-word expression functional categories in scientific English. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 81–87, Torino, Italia. ELRA and ICCL.

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024b. Multi-word expressions in English scientific writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL*

2024), pages 67–76, St. Julians, Malta. Association for Computational Linguistics.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2nd edition. Cambridge University Press, Cambridge.

Douglas Biber and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9:2–20.

Douglas Biber and Bethany Gray. 2011. The historical shift of scientific academic prose in English towards less explicit styles of expression: Writing without verbs. In Vijay Bathia, Purificación Sánchez, and Pascual Pérez-Paredes, editors, *Researching specialized languages*, pages 11–24. John Benjamins, Amsterdam.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Journal of Language and Computation*, 2:597–620.

Ruth Breeze. 2013. Lexical bundles across four legal genres. *International Journal of Corpus Linguistics*, 18(2):229–253.

Annelen Brunner and Kathrin Steyer. 2015. Corpus-driven study of multi-word expressions based on collocations from a very large corpus. In *Proceedings of the 4th Corpus Linguistics conference, Birmingham*, page 12.

Fabienne Cap, Marion Weller, and Ulrich Heid. 2013. Using a rich feature set for the identification of German MWEs. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technologies*, Nice, France.

Susan Conrad and Douglas Biber. 2005. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, 20(2004):56–71.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. 47(2):255–308.

Stefania Degaetano-Ortlieb. 2021. Measuring informativity: The rise of compounds as informationally dense structures in 20th century scientific English. In Elena Soave and Douglas Biber, editors, *Corpus approaches to register variation*, chapter 11, pages 291–312. John Benjamins Publishing Company.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

Stefan Th. Gries. 2020. Analyzing dispersion. In Magali Paquot and Stefan Th. Gries, editors, *A Practical Handbook of Corpus Linguistics*, pages 99–118. Springer, Berlin and New York.

Stefan Th. Gries. 2022a. Multi-word units (and tokenization more generally): A multi-dimensional and largely information-theoretic approach. *Lexis*, (19). Online since 26 March 2022.

Stefan Th. Gries. 2022b. What do (some of) our association measures measure (most)? association? *Journal of Second Language Studies*, 5(1):1–33.

M.A.K. Halliday and Ruqaiya Hasan. 1985. *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford University Press, Oxford.

Natalia Klyueva, Antoine Doucet, and Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 60–65, Valencia, Spain. Association for Computational Linguistics.

Marie-Pauline Krielke. 2021. Relativizers as markers of grammatical complexity: A diachronic, cross-register study of english and german. *Bergen Language and Linguistics Studies*, 11(1):91–120.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14(1):258–288.

Natalia Loukachevitch and Evgeny Parkhomenko. 2018. Recognition of multiword expressions using word embeddings. In Sergey Kuznetsov, Gennady Osipov, and Vladimir Stefanuk, editors, *Artificial Intelligence. RCAI 2018*, volume 934 of *Communications in Computer and Information Science*. Springer, Cham.

Stella Neumann. 2014. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. De Gruyter Mouton, Berlin, Boston.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Ivan A. Sag, Tim Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Berlin. Springer.

Andreas Scherbakov, Ekaterina Vylomova, Fei Liu, and Timothy Baldwin. 2016. Vectorweavers at semeval-2016 task 10: From incremental meaning to semantic unit (phrase by phrase). In *International Workshop on Semantic Evaluation*.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4):487–512.

Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter J.B. van Heuven. 2017. Representation and processing of multi-word expressions in the brain. *Brain and Language*, 175:111–122.

Antoine Tremblay and Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood, editor, *Perspectives on Formulaic Language: Acquisition and Communication*, page 151–173. Continuum, London.

Marion Weller and Ulrich Heid. 2010. Extraction of German multiword expressions from parsed corpora using context features. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Chadia Ben Youssef. 2024. *mMERGE: A Corpus Driven Multiword Expressions Discovery Algorithm*. Ph.D. thesis, University of California, Santa Barbara. ProQuest ID: BenYoussef_ucsb_0035D_16720; Merritt ID: ark:/13030/m5457324.