A methodology for the extraction of information about the usage of formulaic expressions in scientific texts

Hannah Kermes

Universität des Saarlandes
Universität Campus A2.2, 66123 Saarbrücken, Germany
h.kermes@mx.uni-saarland.de

Abstract

In this paper, we present a methodology for the extraction of formulaic expressions, which goes beyond the mere extraction of candidate patterns. Using a pipeline we are able to extract information about the usage of formulaic expressions automatically from text corpora. According to Biber and Barbieri (2007) formulaic expressions are "important building blocks of discourse in spoken and written registers". The automatic extraction procedure can help to investigate the usage and function of these recurrent patterns in different registers and domains.

Keywords: tools, extraction, register analysis

1. Introduction

Formulaic expressions are commonplace not only in everyday language but also in scientific writing. Patterns such as *in this paper (we), the (total) number of, on the basis of* are often used by scientists to convey research interests, the theoretical basis of their studies, results of experiments, scientific findings as well as conclusions and are used as discourse organizers. For Hyland (2008) they help to "shape meanings in specific context and contribute to our sense of coherence in a text". Findings about formulaic expressions are relevant for text and discourse analysis, language pedagogy, translation studies, and NLP tasks (information extraction, text segmentation).

We are interested in: (i) which and what (structural) type of formulaic expressions are used in scientific texts? (ii) the distribution of formulaic expression across different scientific disciplines, (iii) where do formulaic expressions occur within a text?

In section 2 we will describe the linguistic background and point to previous studies on formulaic expressions. In section 3 we will then describe the methodology and the tools we used. In section 4 we will present example results before we will give a conclusion of our study in section 5.

2. Background and Previous Studies

2.1. Formulaic expressions

Formulaic expressions are often referred to as extended collocations, lexical bundles, routines, fixed expressions, prefabricated patterns or simply as formula. We define them as follows: conventionalized/recurrent patterns of three or more tokens that have a statistical tendency to co-occur. Formulaic expressions are usually compositional, i.e. semantically and syntactically transparent, and may be structurally incomplete, crossing phrase boundaries (e.g. *the number of*). As *building blocks of discourse* (Biber and Barbieri, 2007) they help to structure language and to convey the content of a text.

2.2. Previous studies

Previous studies on formulaic expressions in scientific texts have focused on the extraction of (domain specific) lexical bundles (Biber et al., 1999) and a manual classification of the extracted patterns based on keyword-in-context information. The classification follows two dimensions: functional and structural.
(Biber et al., 2004) differentiates between three major groups of discourse functions:

1. stance expressions, which convey a writer's evaluation (*the fact that the, is assumed to be, be equal to the*),
2. discourse organizers, which structure the text (*in this paper (we), on the other hand, due to the*) and
3. referential expressions, which identify entities or specific parts of entities (*the presence of a, is based on, the number of (the)*).

(Hyland, 2008) modifies these classes slightly using the following classes: (i) participant oriented, (ii) text oriented, and (iii) research oriented. With respect to the structure, we find the following classes:

1. NP-based bundles (*the size of the, the fact that the*)
2. PP-based bundles (*with respect to the, on the basis of*)
3. VP-based bundles (*can be used to, be the set of, shown in table CARD*)

On this basis a number of studies have investigated the use of formulaic expressions with respect to their functional and structural distribution. Biber and Conrad (1999), Conrad and Biber (2004) and Biber et al. (2003) compare the use of lexical bundles in conversation and academic prose, Biber et al. (2004), DeCarrico and Nattinger (1988) and Nattinger and DeCarrico (1992) focus on differences between classroom teaching and textbooks, Cortes (2002) and Cortes (2004) investigates the use of formulaic expression by university students of history and biology, Cortes (2008) performs a multilingual study, Chen and Baker (2010) compare L1 and L2 academic writing, and Simpson-Vlach and

Ellis (2010) extract a list of academic formula for language teaching.

3. Methods and Tools

3.1. Corpus

As corpus basis we use the SciTex corpus (see (Teich and Holtz, 2009; Teich and Fankhauser, 2010; Degaetano-Ortlieb et al., 2012)), which contains texts from the four 'contact' disciplines under investigation (computational linguistics, bioinformatics, digital construction, microelectronics; B subcorpora) and the five 'seed' disciplines (computer science (A subcorpus), linguistics, biology, mechanical engineering, electrical engineering (C subcorpora)). The corpus contains approx. 34M tokens and is physically divided into two separate corpora from two different time periods: the DaSciTex, which covers the early 2000s and the SaSciTex (mid 70s). Each corpus is organized in subcorpora (one for each discipline, cf. Figure 1) and is tagged for tokens, lemma, part-of-speech and sentence boundaries using the TreeTagger (Schmid, 1994). Moreover, it includes document structure annotation (text parts, sections, paragraphs, headlines) and meta-information about the included papers: academic discipline, author, title, year of publication, journal.

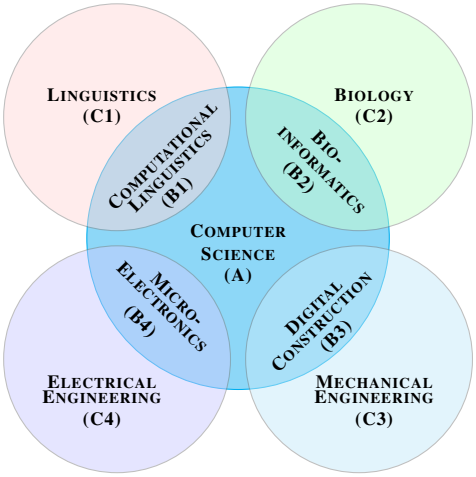

Figure 1: Composition of the SciTex corpus

3.2. The extraction pipeline

In order to automate the investigation of the usage of formulaic expressions, we developed an extraction pipeline. The idea is to be able to extract as much information as possible automatically making the process reproducible and applicable on other corpora and to other domains. We can apply several queries subsequently extracting frequency distribution across subcorpora, e.g., academic discipline, text part and year of publication. The pipeline is kept as modular as possible by using parameter files and program options. The query sequence and parameters for the extraction are stored in a parameter file. Possible parameters are: pattern range, token type, subcorpora (academic discipline, text part, year of publication), frequency cut-off, reference query. It is possible to give different parameters for each query. The results of the first query in the parameter file may be treated as reference for the the following queries. In this case, the results of the other queries are filtered against this reference set, taking all results included in the reference set without any frequency cut-off. Although in this paper we focus on the extraction of formulaic expression, the pipeline can be used for other extraction processes as well. Basically, it can be used for any process aiming at the extraction of distributional information. For the extraction queries we use the Corpus Workbench (CWB, 2010) which allows to use Perl-Scripts to apply and post-process the queries. The queries can be stored in macro files. This is especially helpful, if the queries are complex or to test the queries.

In a first step, we extract candidate ngrams, i.e., any four subsequent tokens, etc. We use a frequency cut-off of 10 per million words for 4grams, and 40 per million for 3grams. To avoid idiosyncrasies, the candidate ngrams have to occur in at least three different texts. Both frequency cut-off and restriction on text number are parameters, and may be changed. Additional filters exclude patterns with more than one number, containing punctuation marks or non-word items. The resulting list serves as reference for the subsequent queries.

Second, we apply a sequence of queries in order to collect information about the usage of the extracted formulaic expression. We use the following features:

- textual distribution
  - beginning/end of section/paragraph/sentence
  - first paragraph/sentence in a section
  - first sentence in a paragraph
  - last paragraph/sentence in a section
  - last sentence in a paragraph
  - occurrence in certain text parts (Abstract, Introduction, Main Part, Conclusion)
- grammatical distribution
  - following nouns/prepositions/(past tense/finite) verbs
  - preceding nouns/(past tense/finite) verbs, subordinate clauses

We select features which can be extracted relatively easy and reliably from a corpus lacking syntactic annotation. However, we use the document structure annotation present in the SciTex corpus. According to (Nattinger and DeCarrico, 1992) formulaic expressions are fixed expressions with a clear pragmatic function. As such, we expect them to occur frequently in certain exposed positions of a text, e.g. at the beginning or end of sections/paragraphs/sentences. We expect certain formula to show preferences for certain text parts, e.g., the introduction of a text. [1]

Third, the results are sorted and filtered. We calculate the frequency distributions of the extracted formulaic expressions in the whole corpus as well as across the subcorpora

[1] For corpora lacking document structure mark-up, the process can easily be adapted applying only those queries that build on basic token annotation (word, lemma, pos, sentences markers).

specified in the parameter files (e.g. academic disciplines, text parts, year of publication). Figures for each feature are stored in separate files. Besides, we combine the results of all queries for a corpus or subcorpus in a summary file displaying the frequency distribution across the single features. We also extract and calculate corpus and subcorpus sizes as well as the number of matches for each feature as reference for the normalization of frequencies. In order to calculate frequency distribution for other pattern types (lemma, structural type), we map the word patterns to lemma, pos and structural type patterns. We use two different structural types, the basic types described in (Biber et al., 2004) and an extended set of structural types derived from these basic types using rules (cf. Table 2).

Finally, we implement a number of functions for basic statistic analyzes of the results including: normalization, type-token-ratio, significance tests, comparisons between subcorpora. These analysis are performed automatically for each token type, and for the different types of subcorpora. Again, the process is modular: we can specify which statistical analysis is to be performed for which token type or subcorpus. We can specify the types of tokens and subcorpora. The analysis is performed for all result files of the extraction process. As a result we get a series of tables (e.g. frequency percentage of formulaic expressions, frequency per million tokens, type-token-ratio) and figures (e.g. barplots, parallel coordinate plots) which allow display the usage of the extracted formulaic expressions and allow to look at the results from different perspectives.

4. Results

In this section, we will give examples results of the extraction process described above. We will focus on 4grams extracted from the DaSciTex Corpus.

Table 1 shows the type-token distribution of formulaic expressions across academic disciplines. Figures for tokens are given in frequency per million.

	types	tokens	ttr
CompSci (A)	224	5382	0.0416
CompLing (B1)	228	4013	0.0568
BioInf (B2)	229	4515	0.0507
DigConstr (B3)	224	4680	0.0479
MicroElec (B4)	231	6994	0.033
Ling (C1)	216	3294	0.0656
Bio (C2)	205	2789	0.0735
MechEng (C3)	222	5307	0.0418
ElectroEng (C4)	228	5034	0.0453

Table 1: Type-token-ratio of 4gram formulaic expressions across academic disciplines

We can observe that *Linguistics* and *Biology* differ from the other disciplines having a rather low token number and a high type-token-ratio (*Computational Linguistics* being in between). *Micro-Electronics* uses more than twice as many formulaic expression as *Linguistics* and *Biology*, the type-token-ratio being considerably lower. This indicates a stronger trend to formalize language in *Micro-Electronics*. The usage of formulaic expressions seems to reflect the type of content of the articles, which is likely to be more technical in *Micro-Electronics* than in *Biology* or *Linguistics*. In general, we can conclude that *Linguistics* and *Biology* - and to a certain extend also *Computational Linguistics* - differ from the other academic discipline with respect to formalization of language.

Figure 2 shows the distribution of structural types across text parts in decreasing order of frequency as a parallel coordinate plot.

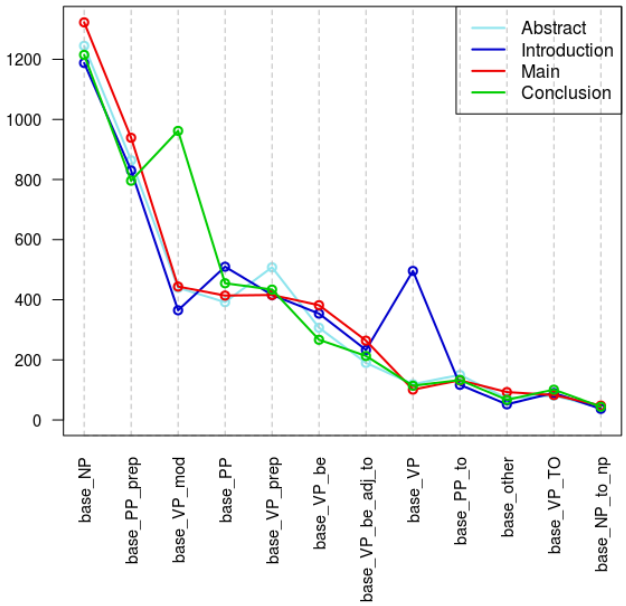

Figure 2: 4gram distribution of structural types across text parts (frequency per million)

The structural types are derived from the three classes (NP-based, PP-based and VP-abed) mentioned in (Biber et al., 2004). In this case, we use an extended the classification scheme with 12 classes (cf. Table 2) including one mixed class. As mentioned above, the classification of the language formula into structural types is performed automatically on the basis of rules.

We can observe that most of the structural types spread more or less evenly throughout the different text parts. The structural type *base_VP_mod* shows a clear tendency to occur in the conclusion. The modals *can, could* are used to reflect on the presented research, and the modal *would* is used to express acknowledgments and to point to future work.

We can also observe a rather high peak for *base_VP* in the Introduction. In order to explain this, we have to look more closely at the lexical fillers. There are only 11 different formula in this class. Three of these formula occur almost exclusively in the Introduction (around 90% of the occurrences) and are among the top ten of formula in this text part but not among the top 50 overall: *the paper is organized, paper is organized as, is organized as follows*.

The picture gets more obvious, if we look at the ranking of the 10 most frequent formulaic expressions occurring in the Introduction (cf. Table 3). We can see that the ranks of these formulaic expressions are extremely low for all other text parts (the lowest rank for the Abstracts is 54, for the

NP-based	base_NP	*a large number of, the results of the*
	base_NP_to_np	*to the fact that, to the number of*
VP-based	base_VP	*in order to obtain, it is well known*
	base_VP_mod	*can be used to, we would like to*
	base_VP_prep	*is based on the, are shown in table*
	base_VP_be	*is the number of, is the same as*
	base_VP_be_adj_to	*it is possible to, to be able to*
	base_VP_TO	*is related to the, was found to be*
PP-based	base_PP	*in this paper we, on the other hand*
	base_PP_prep	*in the case of, in terms of the*
	base_PP_to	*with respect to the, in addition to the*

Table 2: Extended structural types

	Abs	Intro	Main	Concl
on the other hand	1	1	1	4
is organized as follows	43	2	48	39
paper is organized as	47	3	48	39
in the case of	2	4	2	5
of this paper is	38	5	46	32
can be used to	3	6	8	6
in the presence of	15	7	11	17
the paper is organized	49	8	49	39
in the context of	10	9	30	18
as well as the	8	10	12	7

Table 3: Ranking of the 10 most frequent formulaic expressions occurring in the Introduction

Main Part 49, and for the Conclusion 39). All of these formulaic expressions are discourse markers introducing a specific content: how the paper is structured. A content typical for the Introduction of a paper.

Figure 3 displays the distribution of the top 20 formulaic expression across academic disciplines.

At a first glance, we can see three major peaks: *if and only if* occurs predominantly in Computer Science, *in the presence of* in Biology and *as a function of* in Mechanical Engineering. Less evident are the peaks for *in the case of, on the basis of* and *the fact that the* for Linguistics, and *the total number of* and *the size of the* for Micro-electronics. In general, the picture looks rather inhomogeneous, which strengthens the hypothesis that most language formula are register specific.

5. Conclusion

In this paper, we have presented an extraction pipeline for the acquisition of information about the usage of formulaic expressions. Our tool can extract a variety of different features automatically from text corpora. It groups and sorts the results according to subcorpora and features, and provides a number of basic statistical analysis, which allow to investigate the results from different perspectives. Thus, it can be used as a basis for the study of formulaic expressions in different registers, genres and domains.

We want to extend the tool further including multi-corpus analysis to enable diachronic studies and comparisons with other text genres and types.

Besides, we want to add an automatic classification of the formulaic expressions based on the extracted features and a clustering algorithm. Preliminary studies show promising results.

We will also add new features to investigate formula density more closely both with respect to academic disciplines as well as with respect to text parts and paragraphs. We are especially interested in the role that formulaic expressions play with respect to information density.

6. References

Douglas Biber and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3):263–286.

Douglas Biber and Susan Conrad. 1999. Lexical Bundles in Conversations and Academic Prose. In H. Hasslegard and S. Oksefijell, editors, *Out of corpora: studies in honour of Stig Johansson*, pages 181–190. Rodopi.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman.

Douglas Biber, Susan Conrad, and Vivian Cortes. 2003. Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson, and T. McEnery, editors, *orpus linguistics by the Lune: a festschrift for Geoffrey Leech*, pages 71–93. Peter Lang, Frankfurt.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25(3):371–405.

Yu-Hua Chen and Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2).

Susan Conrad and Douglas Biber. 2004. The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica*, pages 56–71.

Vivian Cortes. 2002. Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurice, and D. Biber, editors, *Using corpora to explore linguistic variation*, pages 131–145. John Benjamins, Amsterdam.

Viviana Cortes. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23:397–423.

Viviana Cortes. 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3:43–57, May.

2010. The IMS Open Corpus Workbench. http://www.cwb.sourceforge.net.

J. DeCarrico and J. Nattinger. 1988. Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes*, 7:91–102.

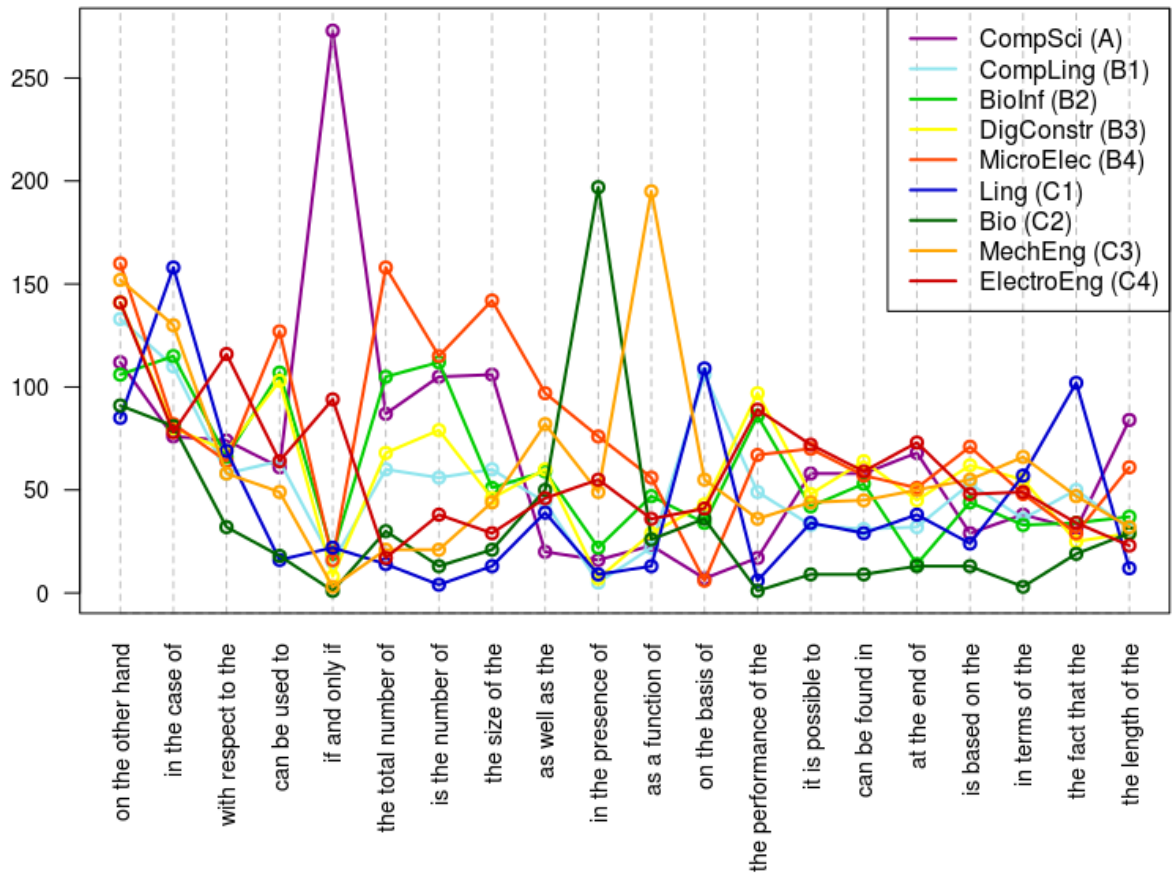

Figure 3: Frequency distribution of formulaic expression across academic disciplines (fpm)

Stefania Degaetano-Ortlieb, Kermes Hannah, Ekaterina Lapshinova-Koltunski, and Teich Elke. 2012. SciTex - A Diachronic Corpus for Analyzing the Development of Scientific Registers. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, volume 2 of *Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP*. Narr, Tübingen. to appear.

K. Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

J. Nattinger and J. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford University Press, Oxford.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

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

Elke Teich and Peter Fankhauser. 2010. Exploring a corpus of scientific text using data mining. In Stefan Th. Grief, S. Wulf, and M. Davies, editors, *Corpus-linguistic applications: Current studies, new directions*. Rodopi.

Elke Teich and Mônica Holtz. 2009. Scientific registers in contact. An exploration of the lexicogrammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics*, 14(4):524–548.