Extraction, Classification, and Retrieval of Formulaic Expressions in Scientific Papers

（学術論文における定型表現の抽出，分類，検索に関する研究）

by

Kenichi Iwatsuki
岩月憲一

A Doctoral Thesis
博士論文

Submitted to
the Graduate School of Information Science and Technology,
the University of Tokyo
on 4 December 2020
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Information Science and Technology
in Computer Science

Thesis Supervisor: Prof. Akiko Aizawa　相澤彰子 教授

ABSTRACT

It is widely known that patterns of human linguistic representation are limited even though grammars and lexicons can generate infinite patterns; thus, languages are to some extent formulaic. Formulaic expressions are defined as continuous or discontinuous word sequences that are memorised and retrieved in the brain rather than composed according to grammars and lexicons. For second language speakers to use the language as native speakers do, formulaic expressions are important.

Formulaic expressions, such as '*in this paper, we propose*', appears frequently in scholarly articles. They convey communicative functions, such as *showing the aim of the paper*, which are closely connected to logical structures of scientific articles. Thus, formulaic expressions are indispensable to communicate easily because not only native speakers but also non-native speakers write and read research articles.

In order to make the most of formulaic expressions in scholarly papers, methodology to retrieve desirable formulaic expressions from a large amount of formulaic expressions is required. For the formulaic expression retrieval, keyword-matching has so far been a dominant method in existing studies. However, with the keyword-matching-based method, it is difficult to search for a variety of formulaic expressions, which is needed in sophisticated paper writing; e.g. to avoid repeating the same phrases or wordings.

In this thesis, we propose methodology to suggest diverse formulaic expressions according to users' purposes. In Chapter 1, we first describe the motivation of this thesis and obstacles to the effective use of formulaic expressions. We also propose a framework, where diverse formulaic expressions can be retrieved by using communicative functions of formulaic expressions as a query in addition to keywords. To realise this framework, a communicative-function-labelled formulaic expression database is indispensable, and to construct it, both communicative-function-based sentence classification and FE extraction from scholarly papers should be tackled. In Chapter 2, existing computer-based academic writing-assistance systems are introduced, and we argue that retrieving and suggesting formulaic expressions or phrases is common to them. We then illustrate how formulaic expressions and communicative functions in scholarly articles have been defined and analysed. We also describe existing computational methodology for the communicative-function-based sentence classification and formulaic expression extraction. In Chapter 3, we explain how the datasets that are used for the communicative-function-based sentence classification, the formulaic expression extraction, and the evaluation for them are constructed, and the corpora used to construct the dataset. In Chapter 4, we propose a method for the communicative-function-based sentence classification in a supervised learning manner. We also show that it still works even if the disciplines between the training and inference dataset are different. In Chapter 5, we propose a formulaic expression extraction method. We compare it to existing extraction methods, and show that the proposed method is more suitable than the others to extract communicative-function-oriented formulaic expressions. In Chapter 6, we analyse discipline- and communicative-function-specific formulaic expressions, using the proposed communicative-function-labelled formulaic expression database. Additionally, we show that the formulaic expression retrieval where a variety of formulaic expressions are suggested can be performed in the proposed framework. In Chapter 7, we discuss the granularity of communicative function sets and the communicative function units from the viewpoint of the suggestion of diverse formulaic expressions. In Chapter 8, we conclude our contributions made in this thesis, and indicate a future direction.

Our contributions made it possible to automatically and computationally construct the large communicative-function-labelled formulaic expression databases, which was almost impossible because of the expensive manual labour necessary to the communicative function label assignment. They also enabled the suggestion of a variety of formulaic expressions using communicative functions, which was difficult in the keyword-matching manner. These achievements brings a new approach to the important linguistic phenomena, formulaic expressions and communicative functions, to computational linguistics, and they are also promising in that the applications to the computer-based academic writing assistance and scholarly paper analyses are suggested.

論文要旨

自然言語による表現は，語彙・文法上可能である組合せと比べて，実際には相当に少ないパターンしか出現せず，定型性があることが知られている。定型表現は，連続または非連続の単語列で，都度構成されるのではなく，そのまま記憶され使用されるという特徴を持つ。特に第二言語においては，定型表現の使用がネイティブらしさの観点から重要である。

学術論文においては，‘*in this paper, we propose*’ のような，特有の定型表現が多用されている。こうした定型表現には，*showing the aim of the paper* のような伝達機能を具現する働きがあり，文章の論理構造と密接に結びついている。そのため，非英語母語話者も多く執筆し読むことになる学術論文においては，定型表現がスムーズな情報伝達に欠かせないものとなっている。

学術論文における定型表現の活用にあたっては，大量の定型表現の中から目的のものを検索する手法が必要である。これまでの研究では，定型表現の検索手法は，キーワードマッチングによるものが多数であった。しかし，キーワードに依存した検索では，多様な定型表現を検索できず，特定の表現を繰り返し使用することを避けたいといったより洗練された論文執筆というユーザの要求に応えることができないという課題がある。

本論文では，検索意図に添いつつも多様な定型表現を提示するために必要な技術について提案を行う。第 1 章では，まず本論文の背景及び定型表現の利活用における課題について述べる。更に，キーワードに加え定型表現の伝達機能をクエリとして用いることによって，多様な定型表現を検索するフレームワークを提案する。このフレームワークには，伝達機能ラベル付き定型表現データベースが必要であり，これを構築するためには，伝達機能に基づく文分類技術と，コーパスに対する定型表現抽出技術が必須であることを述べる。第 2 章では，まず既存の英語論文執筆支援システムを俯瞰し，実質的に定型表現あるいは何らかのフレーズを検索・提示することに集約されることを示す。次に，学術論文における定型表現及び伝達機能がどのように定義され，また分析されてきたかを述べる。更に，伝達機能に基づく文分類と定型表現抽出に対して，計算機を用いた既存手法を述べる。第 3 章では，伝達機能に基づく文分類と，定型表現抽出およびそれらの評価に必要なデータセットの構築手法と，そのために用いる論文コーパスについて述べる。第 4 章では，伝達機能に基づく文分類を教師あり学習を用いて行う手法を提案する。また，訓練データの学術論文の分野と推定データの分野が異なっていても機能することを示す。第 5 章では，定型表現の抽出手法を提案する。既存の定型表現抽出手法を比較し，提案手法が伝達機能に着目した定型表現を抽出するのに適していることを示す。第 6 章では，提案手法によって構築した伝達機能ラベル付き定型表現データベースを用い，分野及び伝達機能別の定型表現について分析する。更に，提案した定型表現検索フレームワークによって，実際に多様な定型表現が検索できることを示す。第 7 章では，多様な定型表現を検索するという観点から，伝達機能の粒度と単位について議論する。第 8 章では，本論文の貢献をまとめ，今後の課題について述べる。

以上の提案によって，これまで伝達機能に基づく分類に人手を要した故に困難であった大規模な伝達機能ラベル付き定型表現データベースを計算機を用いて自動的に構築することが可能になった。また，伝達機能を用いることで，キーワードマッチングによる検索では不可能だった多様な候補の提示が可能になった。これらの成果は，定型表現および伝達機能という重要な言語現象に対して新たな計算言語学的アプローチをもたらすものであり，ま

た計算機による論文解析や論文執筆支援への応用可能性が示されている点でも有望である。

Acknowledgements

The very beginning of this research was a conversation with my supervisor, Prof. Dr. Akiko Aizawa, on how to write English scientific papers more efficiently. Then I thought that this topic would satisfy my interest in computer science, English, linguistics, and convention in scholarly communities. In my early life in the college at Komaba, computer science did not intrigue me much; instead, I was fascinated by languages and linguistics. The turning point was a class taught by Prof. Dr. Osamu Sudo, an economist who introduced several projects he led that had a great impact on society by making the most of computer technology; I was impressed at the potential of computers and determined to go on to a PhD programme. I was working for a private cram school to teach English; thus, I was enthusiastic about English grammar and language learning. The extracurricular activities on scientific sports aroused my curiosity about scholarly papers. Thanks to the supervisor, my interests were successfully satisfied, and I enjoyed resources and environments for my research provided by National Institute of Informatics.

I am most grateful to Assoc. Prof. Dr. Florian Boudin for his great pieces of advice on my research, without which I could not have achieved anything. During his stay in Tokyo, we discussed my research again and again, and he completely understood what I was doing. Besides, he and the Atlanstic 2020 committee accepted my proposal to visit his laboratory at Nantes for five months with full financial support, but the pandemic of COVID-19 prevented me from carrying it out.

I would like to express my gratitude to Prof. Dr. Simone Teufel, who gave me advice on the direction of my research, and Lecturer Mr. Tatsuya Ishii, who gave me knowledge of research on phraseology and lexical bundles.

I thank members of my PhD thesis committee, Prof. Dr. Hidetsugu Nanba, Prof. Dr. Yoshimasa Tsuruoka, Prof. Dr. Seiya Imoto, Assoc. Prof. Yoshihide Yoshimoto, and Prof. Dr. Yusuke Miyao. They made beneficial comments on this thesis.

I express my appreciation to the members of the laboratory, especially to Dr. Saku Sugawara, who was the only member who always read and made insightful comments on my drafts of research papers and grant proposals.

I must mention financial support provided by Japan Society for the Promotion of Science for the last two years in my life at the university. Additionally, the University of Tokyo exempted me from paying tuition fees fully for seven and a half years and half for the rest. The support was definitely indispensable in pursuing my research.

Last but not least, I thank my family. They always encouraged me for nine years after I moved to Tokyo. The nine years were not short, many things happened; however, I had stayed calm and felt relieved by remembering the place I should return to.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

It is widely known that patterns of human linguistic representation are limited even though grammars and lexicons can generate infinite patterns of expressions (Wray & Perkins, 2000). In other words, human languages are to some degree *formulaic*.

In scientific papers, a host of formulaic expressions are used, such as '*in this paper, we propose*'. Past studies pointed out that the usage of academic English differs between native and non-native English speakers (Wu, Mauranen, & Lei, 2020) and between students and scholars (Zhao, 2017). The usage of formulaic expressions of non-natives is also different from that of natives (Chen & Baker, 2010), and moreover, learning formulaic expressions improves non-native speakers' writing (AlHassan & Wood, 2015; Pérez-Llantada, 2014; Peters & Pauwels, 2015). Based on the usefulness, computer systems that suggest formulaic expressions for academic writing assistance were proposed (Liu, Wang, Liu, & Wang, 2016; Mizumoto, Hamatani, & Imao, 2017).

To make the most of formulaic expressions when writing scientific papers, it is important to effectively search for formulaic expression candidates that are suitable for the writer's purpose. However, existing computer systems that suggest formulaic expressions (Liu et al., 2016; Mizumoto et al., 2017) or other kinds of phrasal expressions (Chang & Chang, 2015; Jeong, Nam, & Park, 2014; Yen, Wu, Chang, Boisson, & Chang, 2015) use keyword matching. The problem of the keyword matching is that only formulaic expressions that contain keywords specified by users are retrieved. For instance, when a user intends to write about the paucity of previous work and to find expressions other than '*there are few studies on*', the keyword matching will not find expressions such as '*little attention has been paid to*'. This is because the keyword matching only compares the overlapping of the two formulaic expressions. These two formulaic expressions do not overlap each other at all although both can be used to refer to the paucity of previous work. Therefore, the challenge lying in the formulaic expression retrieval is how to find alternative formulaic expressions that are different from the user's query and that satisfy the user's purpose.

In this thesis, we propose a new framework for the formulaic expression retrieval (Iwatsuki & Aizawa, 2018), which enables users to find a variety of formulaic expressions that can be used as alternatives to the query formulaic expression. Our framework uses *communicative functions* as a query in addition to keywords provided by users. A communicative function of a linguistic unit is a purpose of writing the unit, and communicative functions are based on the structure of documents. The communicative function structure of scientific papers have

been investigated and proposed by many studies (Cotos, Huffman, & Link, 2015; Maswana, Kanamaru, & Tajino, 2015; Swales, 1981, 1990, 2004). For example, Swales (2004) advocated that introduction sections can be split into three communicative functions: *establishing a territory*, *establishing a niche*, and *occupying the niche*.

The diversity of formulaic expressions are realised by their lexical and syntactic variety as long as the communicative functions of the formulaic expressions are the same. For instance, '*in this paper, we propose*' and '*little attention has been paid to*' have different lexicon and syntax, and the communicative functions are also different; these are not alternative to each other. On the other hand, in the above example, '*there are few studies on*' and '*little attention has been paid to*' have the same communicative function although the lexicon and syntax are different. Therefore, communicative function labels should be assigned to each formulaic expression so that a set of formulaic expressions that have the same communicative function as the query can be searched. In our framework, a database of communicative-function-labelled formulaic expressions is constructed in advance.

1.2 Details of Proposed Framework

Figure 1.1 illustrates the proposed framework of the formulaic expression retrieval. First, based on the query keywords, the query communicative function is determined (step 1 and 2). Subsequently, formulaic expressions that have the same communicative function label as the query are retrieved (step 3 and 4).

To realise this formulaic expression retrieval, the communicative-function-labelled formulaic expression database is indispensable. There are two approaches to the construction of the database: the top–down and bottom–up approaches (Biber, Connor, & Upton, 2007; Durrant & Mathews-Aydınlı, 2011). The top–down approach is that communicative function labels are first assigned to text and then formulaic expressions are extracted from the communicative-function-labelled text. The bottom–up approach is that formulaic expressions are first extracted from a corpus and then communicative functions are assigned to each formulaic expression. In either case, the construction is two-fold: the communicative function assignment and formulaic expression extraction.

There is no consensus as to what constitutes the minimal text span that realises a communicative function. For example, to convey the communicative function, *describing the limitations of current research*, some may regard '*beyond the scope*' as the minimal formulaic expression, while others may consider a larger span such as '*is beyond the scope of this paper*'. Here, we follow past research (Dayrell et al., 2012; Fiacco, Cotos, & Rosé, 2019; Hirohata, Okazaki, Ananiadou, & Ishizuka, 2008) and deal with this issue by regarding a whole sentence as the minimal unit of a communicative function. In other words, we assume that one sentence is to be assigned one communicative function label.

The communicative function label assignment is regarded as a classification problem; the top–down approach requires a sentence classification while the bottom–up approach does a formulaic expression classification. We select the top–down approach because of recent advancements in pre-trained models for sentences.

Based on the top–down approach, the formulaic expression extraction task is reduced to assignment of a formulaic or non-formulaic label to each word of a sentence. We assume that a sentence consists of formulaic and non-formulaic

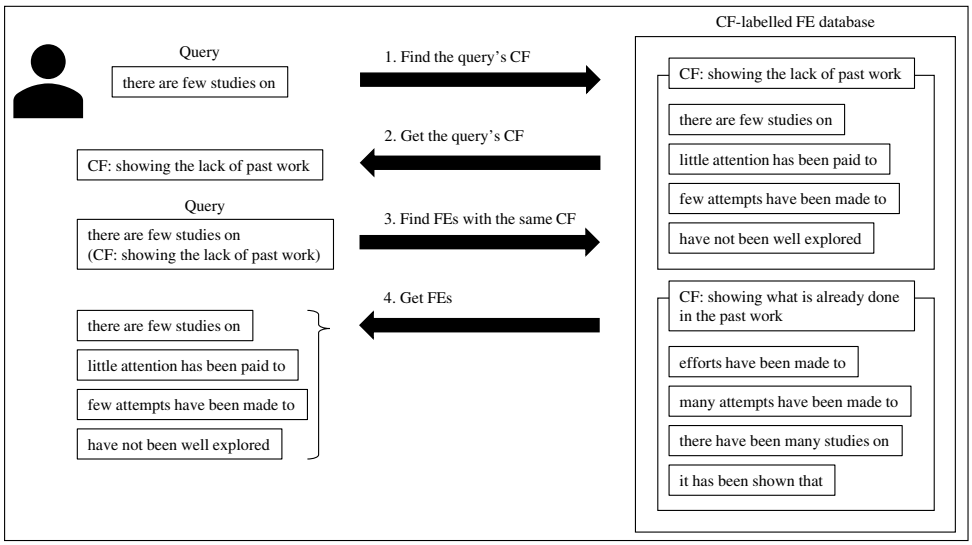

Figure 1.1: Proposed framework for communicative-function-based formulaic expression retrieval.

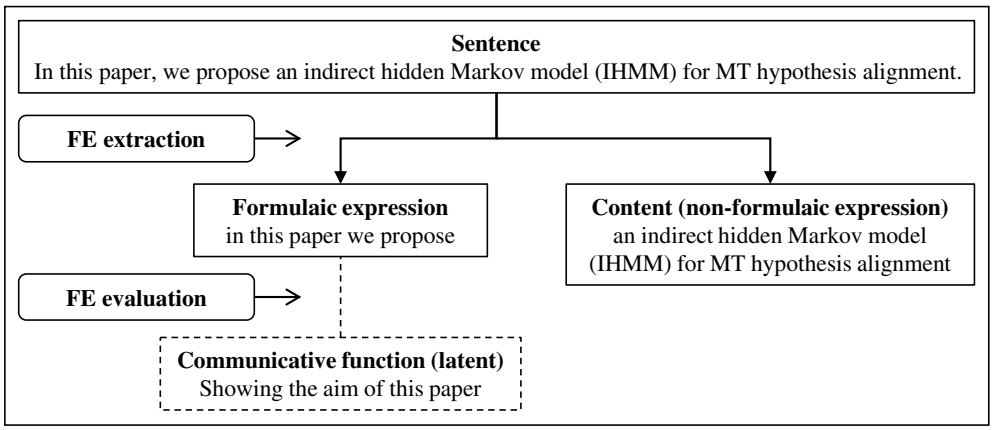

Figure 1.2: We assume that a sentence consists of a formulaic expression realising a communicative function of a sentence and content part. The formulaic expression conveys the communicative function of the sentence. The sentence is cited from He et al. (2008).

parts (Figure 1.2). The formulaic part conveys a communicative function of the sentence, while the non-formulaic part represents content of the sentence.

1.3 Challenges

The main purpose of this thesis is to construct the communicative-function-labelled formulaic expression database for the proposed formulaic expression retrieval. The construction consists of two parts: the communicative function label assignment and formulaic expression extraction. Obviously, both are difficult to perform manually. Thus, computational methodology is needed to automate the construction process.

Few studies have tackled the automated communicative function assignment (Dayrell et al., 2012; Hirohata et al., 2008; Soonklang, 2016), and they focused only on abstracts of scientific papers. The largest problem for the communicative function assignment is the paucity of the dataset of communicative-function-

labelled sentences to which supervised machine-learning can immediately be applied. Even if the dataset is available, it is still unclear whether a dataset for one discipline is enough to cover other disciplines. If not, the cost of creating the training data for many disciplines is too much.

So far, formulaic expression extraction methods have not been investigated intensely. In many studies, frequent word n-grams, which they referred to as *lexical bundles*, were extracted (Cortes, 2013; Esfandiari & Barbary, 2017; Jalali & Moini, 2014; Mizumoto et al., 2017; Pan, Reppen, & Biber, 2016), but no comparison between the frequent word n-grams and other methods was made. Moreover, whether extracted n-grams convey communicative functions has not been evaluated though there are several attempts at extracting formulaic expressions that are peculiar to a specific communicative function (Ädel, 2014; Cortes, 2013; Mizumoto et al., 2017).

To sum up, we tackle the two challenges to construct the communicative-function-labelled formulaic expression database:

1. the assignment of communicative function labels to sentences with supervised machine learning

2. the extraction of formulaic expressions that convey communicative functions of a sentence.

There are three tasks in the communicative function label assignment. First, a communicative-function-annotated sentence dataset is needed to train a classifier. Second, whether recent pre-trained models perform well on the communicative-function-based classification task should be evaluated. Third, the effect of disciplines, such as chemistry and psychology, on the classification performance is to be tested.

There are also two tasks in the formulaic expression extraction. First, how the formulaic expressions that realise sentential communicative functions can be extracted should be investigated. Second, evaluation ways for the formulaic expression extraction should be found out.

After tackling these five tasks, we constructed a communicative-function-labelled formulaic expression database and tested whether the proposed formulaic expression retrieval framework worked well in that it provided diverse alternative formulaic expressions.

1.4 Contributions

The contributions of this thesis are as follows.

1. We proposed a more effective framework for the formulaic expression retrieval (Chapter 1).

2. We created a communicative-function-annotated sentence dataset for training sentence pre-trained models that are used for the communicative function label assignment (Chapter 3).

3. We showed that the SciBERT classifier (Beltagy, Lo, & Cohan, 2019), which is one of the pre-trained models, performed well when trained on one discipline and applied to another discipline (Chapter 4).

4. We proposed a new formulaic expression extraction method (Chapter 5).

5. We proposed an automated evaluation method for the formulaic expression extraction methods and evaluated the proposed and existing formulaic expression extraction methods both manually and automatically (Chapter 5).

6. We constructed the communicative-function-labelled formulaic expression database and evaluated the overall performance of the proposed formulaic expression retrieval (Chapter 6).

1.5 Outline of Thesis

The remainder of this thesis is organised as follows. In Chapter 2, we provide the background of this research. First, we illustrate computer-based academic writing-assistance systems, and then, we explain how formulaic expressions and communicative functions have been investigated so far. In Chapter 3, we explain the corpora and how to create the datasets, which were used in both the communicative function label assignment and formulaic expression extraction. Finally, two datasets were presented: FECFeval dataset and the communicative-function-annotated sentence dataset. In Chapter 4, we describe how the communicative function label assignment was conducted using the communicative-function-annotated sentence dataset. We also present the communicative-function-labelled sentence dataset, which was created by the proposed communicative function assignment method. In Chapter 5, we illustrate the proposed formulaic expression extraction method, and compare it to existing formulaic expression extraction methods. In Chapter 6, using the communicative-function-labelled formulaic expression database, which was constructed with the proposed communicative function label assignment and formulaic expression extraction methods, we performed the formulaic expression retrieval, and manually evaluated whether the proposed framework worked well. In Chapter 7, we discuss the inherent difficulties in the construction of the database and indicate future direction of research in formulaic expression and communicative functions.

Chapter 2

Background

2.1 Genre of Scientific Papers in Natural Language Processing

2.1.1 Characteristics of Scientific Papers and Processing Scientific Papers

Scientific papers form a genre that has a peculiar writing style and structure of documents. One article has a title, names of authors and affiliations, and an abstract as bibliographic data in addition to the metadata: the name of journals or conference where the article is published, the year of publication, the number of pages, and several identifications including the uniform resource indicator and document object identifier. The main body of the article consists of not only text but also headings, figures, lists, and tables. The text is structured; sections and paragraphs are components of the article that generate the logical flow of the content. Thus, to process scientific papers, special attention should be paid to these characteristics unique to scientific papers.

Computationally processing scientific papers is an important task. A pile of scientific papers are knowledge of the world as such. Thus, to understand the state of the human knowledge, scientific papers should be searched. A single article may be a solution to some problems, but usually knowledge drawn from multiple papers that relate to each other provides more solutions. Since a growing number of scientific papers are published every year, it is difficult to manually connects one paper to another that may look unrelated.

Scientific papers are often read by researchers whose expertise is the same as the discipline of the papers. However, they are also read by other researchers, which is important to interdisciplinary research. People not in scientific communities sometimes have need to understand scientific papers, but it is difficult because of a lot of jargons and tacit knowledge in the field. Thus, summarisation or simplification is important approach to scientific papers.

Publishing a scientific paper is another perspective of scientific paper processing. Findings of research should be published as soon as possible to share the knowledge with humans, but writing scientific papers is not an easy task for researchers. Additionally, the quality of writing is also important to convey an accurate message to readers and reviewers. Assisting composition of scientific papers includes not only sentence-level perspective such as grammars but also document-level such as logical flow and rhetorical/discourse structures.

Most part of a scientific paper is text. Therefore, techniques of natural language processing are to be applied to the text for various purposes including information retrieval, information extraction, and constructing citation graphs. Research communities hold workshops on processing scientific papers collocated with conferences on natural language processing, digital libraries, and informa-

tion retrieval. The workshops on mining scientific publications have been held since 2012, the workshops on bibliometric-enhanced information retrieval have been held since 2014, and the workshops on scientific document analysis have been held since 2016. Two more related workshops have started: the workshop on scholarly document processing and workshop on natural language processing and data mining for scientific text in 2020 and the workshop on scientific document understanding in 2021. The research field of processing scientific papers has been growing as the need for it has become larger.

2.1.2 Document Analysis

Most of the scientific papers published decades ago were formatted in papers; thus, if they are digitalised, they are still just scanned documents. The scanned documents as such are not eligible for text processing because of lack of text data and bibliographic information. Most recent papers were formatted in the portable document format (PDF), which is also difficult to process directly. The PDF papers must be converted into a computer-readable format in pre-processing stage. Some journals provide HTML- or XML-formatted papers. These papers can be easily parsed by computers, but the usage of tags are not consistent across journals or platforms.

Mathematical formulae are peculiar to scientific papers and difficult to process. Detecting mathematical expressions is not an easy process because it often appear in narrative texts as *inline* mathematical expressions (Iwatsuki, Sagara, Hara, & Aizawa, 2017). Not only extracting the formulae but also understanding them is indispensable. For example, what variants such as x and y indicates is important information to understand the mathematical formulae. The explanation of the formulae are written in text; connecting the formulae to the text is a problematic task. Mathematical formulae often work as a summary of methodology; thus, mathematical-formula-based retrieval of scientific papers (math IR) is another important problem (Kristianto, Topić, & Aizawa, 2017; Schubotz et al., 2018). There happen two different mathematical expressions having the same meanings; e.g. $\sin 2x$ and $2 \sin x \cos x$ are mathematically the same.

Tables are very familiar to scientific papers, but these are also problematic components in scientific papers. A table often conveys a summarisation of characteristics of methods or data presented in a paper by comparison. It has a two-dimensional structure, but semantics of each row and column is not always clear. Sometimes it has more complicated structures.

2.1.3 Citation Analysis

Citing articles is convention unique to scientific communities. Science is succession to past work; citations reveal which work is based on which work. Thus, exploiting citations will enable us to measure the impact of research articles, to draw a big picture of one field and relations between other disciplines, and to search for related work.

Citation extraction is not an easy task. The citation and bibliography formats differ across journals. This makes it difficult to identify what part of a text is a citation. Also, identifying names of authors, names of journals, publication dates, and page numbers in a bibliographical information is a tough task. The name of author is not always written in the same form; the journal and conference titles are often abbreviated.

Recognition of citation intention is another important task related to citation

analysis (Cohan, Ammar, van Zuylen, & Cady, 2019; Jurgens, Kumar, Hoover, McFarland, & Jurafsky, 2018; Teufel, Siddharthan, & Tidhar, 2006). In one scientific paper, citations are used to provide background of research, indicate methods that authors use, compare authors' work to existing ones, and so on. Thus, recognising these intentions will be helpful in better understanding of relations between citing and cited articles.

2.2 Formulaic Expressions and Communicative Functions in Scientific Papers

2.2.1 Problems Lying in Academic Writing Assistance Systems

When writing a research article, authors are often faced with a situation where they are not able to think of a desirable phrase to explain something or they wish to determine whether their wording is grammatically and conventionally correct. In such cases, they try to find better phrases or wordings by consulting books on academic writing or they search the web for phrases that appear more frequently. Because this process takes much time and effort, some computer systems have been proposed to automate this process.

Existing writing assistance systems are classified into three types. First, the most direct approach for computer-based writing assistance is that in which user-input sentences are used to retrieve example sentences. Search results are shown with concordances (Wu, Chang, Liou, & Chang, 2006) or dependency structures (Kato, Matsubara, & Inagaki, 2006).

Another approach is similar to an input method in which users can input non-alphabetical languages. FLOW (Chen, Huang, Hsieh, Kao, & Chang, 2012) suggests an English translation from words written in another language. WINGS (Dai, Liu, Wang, & Liu, 2014) suggests full Chinese sentences and words from pinyin. Full sentences are suggested on the basis of searches for sentences that contain words that are the same as or similar to the input.

The third approach is combined with an authoring system. With this approach, candidate English expressions that follow user input are listed; then the users can choose one of them (Chang & Chang, 2015; Chang et al., 2015; Jeong et al., 2014; Liu et al., 2016; Mizumoto et al., 2017; Yen et al., 2015). Some systems allow users to specify the categories of formulaic expressions. Such categories include the introduction, methods, results and discussion (IMRaD) structure (Jeong et al., 2014), argumentative zone (Chang et al., 2015; Teufel, 1999), and move–step structure (Mizumoto et al., 2017; Swales, 1990). The drawback of these systems is that users must designate which category to use. Thus, users must know what kind of categories are prepared by the systems. AWSuM (Mizumoto et al., 2017) provides six sections (abstract, introduction, methods, results, discussion, and conclusion) and 25 communicative function categories. It is not easy for users to select one of them every time they write something.

In most cases, phrases or wordings are extracted from linguistic resources and recorded in a database in advance, and a system searches for one of them based on the users' writing. In order to extract frequently used word n-grams, Jeong et al. (2014) relied on PubMed structured abstracts as a resource, in which sentences are labelled with the following functions: introduction, methods, results and discussion. However, this convention of writing abstracts is specific to PubMed; thus, this work will not be applicable to other disciplines. Chang and Chang (2015) proposed WriteAhead2[1]. They extracted approximately 3,000

[1] `http://writeahead.nlpweb.org`

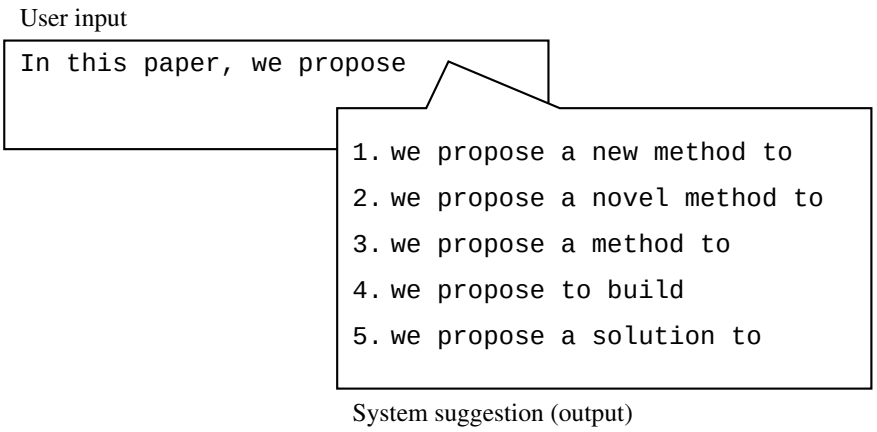

Figure 2.1: Image of keyword-matching-based formulaic expression retrieval. All the suggested formulaic expressions contain the query '*we propose*'.

part-of-speech (POS) patterns from an English dictionary. Subsequently, using 700 keywords, which were derived from Academic Keyword List and the POS patterns, they extracted phrasal patterns from CiteSeerX. Thus, the system suggests not fixed formulaic expressions but grammatical phrasal frames with POS-based placeholders and examples. The system they proposed is useful to find a correct usage of specific words. Liu et al. (2016) extracted frequent word n-grams from Elsevier's ScienceDirect and paraphrased them using WordNet synonyms to extend their database. AWSuM (Mizumoto et al., 2017) utilises a database where fixed length word n-grams were assigned communicative function labels; these labels were assigned manually to sentences in corpora they used. The corpora comprised originally 1,000 articles from 10 journals on applied linguistics, but to date the corpora has been updated; now the system covers computer science (approximately 300 articles), material science, and medicine (the sizes of the latter two are not provided).

Despite the differences in the methods used to create databases, the method of recommendation of phrases and wordings is similar among the systems mentioned here. When a user writes something, all systems show examples or phrases that follow the user's input. For example, if a user writes '*we propose*', the systems only show phrases that contain '*propose*' (Figure 2.1). WriteAhead2 uses the last word of a user input to search the database. It returns phrasal frames that begins with the last word. AWSuM uses the last few words of a user input; users can select the number of words the system use to search and the number of words of resulting word n-grams. Examples of WriteAhead2 and AWSuM are shown in Table 2.1. The input text is '*in this paper we*'. For AWSuM, the computer science corpus, the introduction section, the *presenting study* function, and four-word length were selected. In both results from the two systems, the suggested candidates are to follow the user input '*in this paper we*'. Thus, these systems assume that the user input is always correct and users always come up with the beginning part of a formulaic expression, which is clearly not the case. This is a limitation of keyword-based search in existing writing-assistance systems. In order to help users find phrases with different wordings, the use of communicative functions as queries, rather than keywords alone, can be beneficial.

Table 2.1: Examples of existing writing assistance systems: WriteAhead2 and AWSuM. The input text was '*in this paper we*'. Only the top-five results are shown. The parentheses are examples shown in WriteAhead2.

WriteAhead2	AWSuM
we do (we present a) (we propose a)	in this paper we propose a novel
we do something (we present an algorithm for) (we present experimental results showing)	in this paper we propose two contributions
we did (we investigated the)	in this paper we study the learning
we did something (we developed a system) (we evaluated the method using)	in this paper we attempt to address
	in this paper we propose an approach

2.2.2 Formulaic Expressions in Scientific Papers

There has been no established definition of formulaic expressions, and more than forty terms have been used to refer to formulaicity or formulaic expressions (Wray & Perkins, 2000). Brooke, Šnajder, and Baldwin (2017) used the term *formulaic sequences* and considered them as a wider concept that overlaps multi-word expressions and constructions. Many studies used the term *lexical bundles* (Biber & Barbieri, 2007; Durrant, 2017; Hyland, 2008) or 'phraseology' (Simpson-Vlach & Ellis, 2010; Vincent, 2013) to refer to word n-grams that occur in a corpus more frequently than by chance. A survey of definitions of formulaic expressions shows that there are three ways of defining them (Durrant & Mathews-Aydınlı, 2011). The first definition is a *phraseological* approach. Using this approach, *formulaicity* is definable by non-compositionality of word sequences. However, this definition is not for formulaic expressions but for idioms because the semantics of formulaic expressions are often compositional. For example, '*have been explored by many researchers*' has a compositional meaning but it is nonetheless a formulaic expression. The second definition is a *frequency-based* approach. In this approach, frequently co-occurring word sequences are considered formulaic expressions. However, noisy phrases such as '*is one of the*' cannot be removed. Also, formulaic expressions do not always occur frequently (Simpson-Vlach & Ellis, 2010). The third definition is a *psychological* approach, which defines formulaic expressions as word sequences that are processed and remembered as a whole in the human brain. Wray and Perkins (2000) defined it as *a sequence, continuous or discontinous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.* According to Mizumoto et al. (2017), formulaic expressions are the term referring to their psychological features rather than lexical bundles. The psychological approach seems to work well, but computational formalisation is difficult. Considering all these discussion, we regard a formulaic expression as a continuous or discontinuous word sequence that conveys a communicative function of a sentence.

Biber, Conrad, and Cortes (2004) analysed the usage of lexical bundles in an academic context. They defined lexical bundles as *the most frequent recurring lexical sequences in a register*. Their results showed that lexical bundles are not always syntactically structured. In fact, they often contain some fragments such as '*is based on the*', '*I don't know if*' and '*a little bit of*'.

Along with lexical bundles, Gray and Biber (2013) specifically examined phrase frames (p-frames): discontinuous word sequences with a slot '*' that is filled by any word. The number of lexical bundles used in corpora is larger than that of phrase frames, but examining particularly those occurring in at least five texts, phrase frames are more numerous than lexical bundles. They classified phrase frames into three types: verb-based frames, frames with other content words, and function word frames.

The advantage in utilising formulaic expressions is that formulaic expressions are grammatically and conventionally correct so that they can be used without modification. Past studies (Ädel & Erman, 2012; Chen & Baker, 2010) showed that the usage of formulaic expressions between native and non-native English speakers was different. It is important to make the most of formulaic expressions in order to write scientific papers fluently (Conklin & Schmitt, 2008; Ellis, Simpson-vlach, & Maynard, 2008).

The usage of formulaic expressions also differs across disciplines (Hyland, 2008; Nekrasova-Beker, 2019) although there are multiple studies that worked on collocations or lexicons used in common (Ackermann & Chen, 2013; Coxhead, 2000). Discipline-specific studies on formulaic expressions, including mathematics (Cunningham, 2017), social sciences (Lu, Yoon, & Kisselev, 2018), medicine (Jalali & Moini, 2014), psychology (Esfandiari & Barbary, 2017), and applied linguistics (Qin, 2014), were conducted. Therefore, not only general-purpose formulaic expressions but also discipline-specific formulaic expressions should be collected for writing assistance.

Generally, multi-word expression is a different concept to formulaic expression but there is some overlap between the two concepts. Multi-word expressions do not always convey a communicative function. According to the survey by Constant et al. (2017), multi-word expressions can be categorised in several ways. For instance, '*kick the bucket*' is a typical multi-word expression and categorised into the *idiom* class and '*International Business Machines*' is categorised into the *multi-word named entity* class. However, both do not convey any specific communicative function in scientific papers.

PARSEME (Savary et al., 2017) is the most comprehensive dataset for multi-word expression identification. In this dataset, multi-word expressions are classified into three categories: general, quasi-general, and other; these categories are not based on communicative functions. Therefore, state-of-the-art models for identification of multi-word expressions trained on the dataset (Saied, Candito, & Constant, 2019; Waszczuk, Ehren, Stodden, & Kallmeyer, 2019) cannot be directly applied to the extraction of formulaic expressions.

2.2.3 Communicative Functions in Scientific Papers

Text of a scientific paper has its rhetorical structure to report research logically, and each component of the text plays its own role, such as providing background information, explaining methodology, and discussing experimental results. These roles are referred to as communicative functions, and communicative functions represent authors intentions of how each part of the text should be read by readers. Sections can be regarded as communicative functions. For example, the

Table 2.2: CARS model proposed by Swales (2004).

Introduction	
Move 1 Establishing a territory	
Step 1	Claiming centrality
Step 2	Making topic generalization(s)
Step 3	Reviewing items of previous research
Move 2 Establishing a niche	
Step 1A	Indicating a gap
Step 1B	Adding to what is known
Step 2	Presenting positive justification
Move 3 Occupying the niche	
Step 1	Announcing present research descriptively and/or purposively
Step 2	Presenting research questions or hypotheses
Step 3	Definitional clarifications
Step 4	Summarizing methods
Step 5	Announcing principal outcomes
Step 6	Stating the value of the present research
Step 7	Outlining the structure of the paper

section structures of the introduction, methods, results, and discussion (IMRaD) have communicative functions: introduction to research, explaining method, reporting results, and discussing findings.

The sections are a coarse set of communicative functions; finer-grained analyses were conducted by (Swales, 1981, 1990, 2004). He proposed the Creating-A-Research-Space model (CARS model), which explained the communicative function structures of the introduction sections of research articles. In the model, the introduction section consists of three *moves*, and each move consists of several *steps* (Table 2.2). A move was defined as '*a unit that relates both to the writer's purpose and to the content that s/he wishes to communicate*' by Dudley-Evans and John (1998).

Following his work, a host of studies extended the concept to all parts of a scientific paper. Most studies focused on very limited part of scientific papers; only the introduction (Ozturk, 2007), methods (Cotos, Huffman, & Link, 2017; Lim, 2006), results (Basturkmen, 2009; Lim, 2010), discussion sections (Basturkmen, 2012; Peacock, 2002), or abstracts (Darabad, 2016; Lorés, 2004; Rashidi & Meihami, 2018; Saboori & Hashemi, 2013). On the other hand, Kanoksilapatham (2005) proposed a communicative function structure of all the sections in biochemistry papers. Maswana et al. (2015) also presented communicative functions of a whole paper including an abstract in engineering disciplines. Cotos et al. (2015) used scholarly papers ranging from humanities to sciences to investigate communicative function structures of the four sections.

Table 2.3 lists the different communicative function structures of scholarly papers proposed by Cotos et al. (2015); Kanoksilapatham (2005); Maswana et al. (2015). The numbers of communicative functions are different, but communicative functions do not completely differ. For example, in the introduction sections, *stating purpose(s)*, *announcing present research purposefully*, and *reference to research purpose* are alike in that the communicative functions are related to referring to the purpose of research. Granularity of the communicative function sets is also different; e.g. *describing procedures* and *presenting findings* in Kanoksilapatham (2005) are integrated into *reference to main research procedure*

Table 2.3: Communicative function structures of scholarly articles. Moves are in bold.

Kanoksilapatham (2005)	Cotos et al. (2015)	Maswana et al. (2015)
Introduction	**Introduction**	**Introduction**
Announcing the importance of the field	**Establishing the territory**	**Presenting the background information**
Claiming the centrality of the topic	Claiming centrality	Reference to established knowledge in the field
Making topic generalizations	Providing general background	Reference to main research problems
Reviewing previous research	Reviewing previous research	**Reviewing related research**
Preparing for the present study	**Identifying a niche**	Reference to previous research
Indicating a gap	Indicating a gap	Reference to limitations of previous research
Raising a question	Highlighting a problem	**Presenting new research**
Introducing the present study	Raising general questions	Reference to research purpose
Stating purpose(s)	Proposing general hypotheses	Reference to main research procedure and outcome
Describing procedures	Presenting justification	
Presenting findings	**Addressing the niche**	
	Introducing present research descriptively	
	Announcing present research purposefully	
	Presenting research questions	
	Presenting research hypotheses	
	Clarifying definitions	
	Summarizing methods	
	Announcing principle outcomes	
	Stating the value of present research	
	Outlining the structure of the paper	
Methods	**Methods**	**Methods & Results**
Describing materials	**Contextualizing the study methods**	**Identifying source of data and method adopted in collecting them**
Listing materials	Referencing previous works	Indicating source of data
Detailing the source of the materials	Providing general information	Indicating data size
Providing the background of the materials	Identifying the methodological approach	Indicating criteria for data collection
Describing experimental procedures	Describing the setting	Indicating data collection procedure
Documenting established procedures	Introducing the subjects/participants	Providing background details of data
Detailing procedures	Rationalizing pre-experiment decisions	**Describing experimental procedures**
Providing the background of the procedures	**Describing the study**	Identifying main research apparatus
Detailing equipment	Acquiring the data	Recounting experimental process
Describing statistical procedures	Describing the data	Indicating criteria for success
	Identifying variables	**Describing data analysis procedures**
	Delineating experimental/study procedures	Defining terminologies
	Describing tools/instruments/materials/equipment	Indicating process of data classification
	Rationalizing experiment decisions	Identifying analytical instrument/procedure
	Reporting incrementals	Indicating modification to instrument/procedure
	Establishing credibility	**Reporting results**
	Preparing the data	
	Describing the data analysis	
	Rationalizing data	

(Continued)

	processing/analysis	Restating data analysis procedures
Results	**Results**	Restating research questions
Stating procedures	**Approaching the niche**	Stating general findings
Describing aims and purposes	Providing general orientation	Stating specific findings
Stating research questions	Restating study specifics	**Commenting on results**
Making hypotheses	Justifying study specifics	Interpreting results
Listing procedures or methodological techniques	**Occupying the niche**	Comparing results with previous studies
Justifying procedures or methodology	Reporting specific results	Evaluating results (or research)
Citing established knowledge of the procedure	Indicating alternative presentation of results	
Referring to previous research	**Construing the niche**	
Stating results	Comparing results	
Substantiating results	Accounting for results	
Invalidating results	Explicating results	
Stating comments on the results	Clarifying expectations	
Explaining the results	Acknowledging limitations	
Making generalizations or interpretations of the results	**Expanding the niche**	
Evaluating the current findings	Generalizing results	
Stating limitations	Claiming the value	
Summarizing	Noting implications	
	Proposing directions	
Discussion	**Discussion/Conclusion**	**Conclusion**
Contextualizing the study	**Re-establishing the territory**	**Highlighting overall results and their significance**
Describing established knowledge	Drawing on a/theoretical general background	**Explaining specific research outcomes**
Presenting generalizations, claims, deductions, or research gaps	Drawing on study-specific background	Stating a specific outcome
Consolidating results	Highlighting principal findings	Interpreting the outcome
Restating methodology (purposes, research questions, hypotheses restated, and procedures)	Previewing the discussion 'road map'	Indicating significance of the outcome
	Framing the new knowledge	Contrasting present and previous outcomes
Stating selected findings	Explicating results	Indicating limitations of outcomes
Referring to previous literature	Accounting for results	**Stating research conclusions**
Explaining differences in findings	Clarifying expectations	Indicating research implications
Making overt claims or generalizations	Addressing limitations	Promoting further research
Exemplifying	**Reshaping the territory**	
Stating limitations of the study	Supporting with evidence	
Limitations about the findings	Countering with evidence	
Limitations about the methodology	**Establishing additional territory**	
Limitations about the claims made	Generalizing results	
Suggesting further research	Claiming the value	
	Noting implications	
	Proposing directions	

and outcome in Maswana et al. (2015).

Preferences over communicative functions also vary across disciplines. Cotos et al. (2015) found that in the result sections, *comparing results* was preferred in chemical engineering papers, while *clarifying expectations* was used many times in psychology papers. The usage of communicative functions are conventionally established by the research community to make papers easily understandable.

In summary, there is no established communicative function set yet, and some communicative functions are not used or are frequently used in a specific discipline. Proposing a new communicative function set is beyond the scope of this thesis; however, we must select a set of communicative functions. We adopted the communicative function set used in Academic Phrasebank (Morley, n.d.) and modified them (explained afterwards). Specifically, we use the categorisation system that is adopted in Academic Phrasebank made by Morley (n.d.) because the categorisation of this resource is similar to move–step structures and many example expressions are listed in this resource. In Academic Phrasebank there are six sections: Introducing Work, Referring to Sources, Describing Methods, Reporting Results, Discussing Findings and Writing Conclusions and 77 categories such as *establishing the importance of the topic for the discipline* and *giving reasons why a method was adopted or rejected*, which roughly correspond to steps. This resource was made of 100 postgraduate theses of various disciplines.

Units where communicative functions are realised are flexible. Halliday and Matthiessen (2014) conducted broader analyses of functions in different levels of linguistic units ranging from multiple sentences to phrases. Several sentences sometimes realise one communicative function, while a clause may also do. However, it is difficult to detect the precise spans that corresponds to one communicative function. In previous work (Dayrell et al., 2012; Fiacco et al., 2019; Hirohata et al., 2008), a sentence was regarded as a unit of communicative function. We follow this manner; we assume that one sentence has a communicative function and thus one sentence has one formulaic expression that conveys the communicative function.

2.2.4 Communicative-Function-Based Classification

Regardless of communicative function units, the communicative-function-based classification was conducted manually in most of the past work (Ädel, 2014; Cortes, 2013; D. Liu, 2012; Mizumoto et al., 2017; Simpson-Vlach & Ellis, 2010). There exist several studies that tackled the automated communicative-function-based classification. Hirohata et al. (2008) adopted conditional random fields (Lafferty, McCallum, & Pereira, 2001), Dayrell et al. (2012) used a classifier chain with sequential minimum optimisation (Read, Pfahringer, Holmes, & Frank, 2009), and Rakel with the J48 algorithm (Tsoumakas & Vlahavas, 2007), Soonklang (2016) used a Bayes classifier and decision tree, and Hashimoto, Soonklang, and Hirokawa (2016) extracted feature words of each communicative function and applied support vector machines to them. However, these studies only focused on abstracts of scientific papers. Moreover, no communicative-function-labelled sentence corpus is available to the public.

2.2.5 Extraction of Formulaic Expressions

Two approaches are used for extracting formulaic expressions: corpus- and sentence-level approaches. Based on the intuition that formulaic expressions appear frequently or words composing formulaic expression are strongly associated,

Table 2.4: The length and frequency threshold of formulaic expressions (FEs) were different across past studies. Pmw means *per million words*.

Reference	FE Length	FE frequency
Simpson-Vlach and Ellis (2010)	3–5 words	-
Cortes (2013)	4 words	20 pmw
	5 words	10 pmw
	6 or 7 words	8 pmw
	longer	6 pmw
Mizumoto et al. (2017)	4 words	top 200
Jalilifar, Ghoreishi, and Roodband (2016)	3–5 words	10 pmw

most studies use the corpus-level approach, in which statistical metrics, such as frequency or mutual information, are applied to a whole corpus. To extract formulaic expressions, word n-grams were collected from a whole corpus by using the metrics (Biber et al., 2004; Kermes, 2012; Kermes & Teich, 2020; Mizumoto et al., 2017; Simpson-Vlach & Ellis, 2010). However, this approach results in the extraction of an explosive number of overlapping n-grams, thus causing a serious problem in the communicative-function-labelled formulaic expression database construction. For instance, suppose '*in this paper we propose*', '*this paper we propose a*', and '*in this paper we propose a new method*' are extracted, a criterion is needed to determine which of these are regarded as formulaic expressions; however, determination of such a criterion is difficult and different values were used (Table 2.4).

The n-gram lattice method (Brooke et al., 2017) is one approach to address this problem; here, scores of various aspects of *formulaicity* are first calculated for all word n-grams. Next, an objective function that contains all scores of the n-grams is maximised to determine which n-grams should be disregarded and which should remain. However, this method is still not focused on formulaic expressions conveying communicative functions but on general phrasal expressions, and is thus not suitable for our setting.

For extracting phrase frames, which have a slot where any suitable word can be inserted, different methods were proposed. Biber (2009) first extracted continuous word sequences according to frequency threshold, and then removed a word from them to collect p-frames. Gray and Biber (2013) directly collected p-frames from a corpus. Vincent (2013) decomposed a candidate phrase into the phrasal core and its collocates. The phrasal core is a continuous or discontinuous word sequence occurring with high frequency. Candidate phrases including the core were first identified in a corpus; then, the collocates were sought.

The sentence-level approach assumes that one formulaic expression occurs in one sentence. In this way, '*in this paper we propose a new method*' can be extracted but '*this paper we propose a*' will not be extracted from a sentence. This approach is also useful for extracting formulaic expressions with a slot like p-frames, such as '*however, * have not been reported*'. This setting is regarded as a sequence-labelling problem, in which each word of a sentence is labelled as either formulaic or non-formulaic. Liu et al. (2016) proposed removing topic-specific words as non-formulaic words, using latent Dirichlet allocation. They used a corpus consisting of papers from various disciplines, and tried to remove discipline-specific vocabulary. Thus, this is not suitable for extracting discipline-specific formulaic expressions.

The evaluation of formulaic expression extraction methods is another problem.

Table 2.5: Properties of existing and proposed methods for construction of communicative-function-labelled formulaic expression databases. The approach of Morley (n.d.) is unknown. For the communicative function label assignment, we adopted supervised machine-learning. The formulaic expression extraction was conducted manually using a corpus- or sentence-level method.

	Approach	CF assignment	FE extraction
Simpson-Vlach and Ellis (2010)	bottom–up	manual	corpus
Morley (n.d.)	-	manual	manual
Mizumoto et al. (2017)	top–down	manual	corpus
Lu et al. (2018)	bottom–up	manual	corpus
Ours	top–down	automated	sentence

Brooke et al. (2015) pointed out that the comparison of newly extracted formulaic expressions with existing reference was unreasonable because if a reference was on point, a new lexicon did not need to be created. Manual evaluation has been a common method of the formulaic expression evaluation. Simpson-Vlach and Ellis (2010) asked 20 experienced English-for-academic-purposes instructors and testers to rate the extracted word n-grams. The experts were divided into three groups, in which they checked phrases according to one of the three criteria: (1) formulaic or not, (2) having cohesive function or not, and (3) worth teaching or not. Brooke et al. (2015) asked three judges, who were native English speakers, to check whether the extracted formulaic expressions were canonical or not. The canonical formulaic expressions were defined as word sequences whose consisting words were considered to be formulaic.

Additionally, the flexibility of formulaic expressions also makes automated intrinsic evaluations difficult, where extracted formulaic expression candidates are evaluated by their properties, such as frequency and mutual information. For example, both '*beyond the scope*' and '*is beyond the scope of this paper*' are good formulaic expressions that convey the same communicative function, i.e. '*describing the limitations of current research*'. Therefore, even if manually annotated formulaic expressions are available, there are still other allowable formulaic expressions as long as they convey the same communicative function.

2.2.6 Communicative-Function-Labelled Formulaic Expression Databases

Databases comprising communicative-function-labelled formulaic expressions are required from a pedagogical perspective (Martinez & Schmitt, 2012), and a computer-based academic writing assistance system[2] that uses such communicative-function-labelled formulaic expressions has been proposed (Mizumoto et al., 2017). Several attempts have been made to extract formulaic expressions from scientific corpora and categorise them based on communicative functions (Ädel, 2014; Cortes, 2013; Lu et al., 2018; Mizumoto et al., 2017; Morley, n.d.; Simpson-Vlach & Ellis, 2010). A communicative-function-labelled formulaic expression database can be constructed using two main approaches: top–down and bottom–up approaches (Biber et al., 2007). By using the top–down approach, sentences are first assigned communicative function labels, and then formulaic expressions are extracted, while in the case of the bottom–up approach, formulaic

[2] `http://langtest.jp/awsum/`

Table 2.6: Statistics of existing formulaic expression (FE) databases and lists. Some studies did not disclose the number of documents or formulaic expressions. Either formulaic expressions specific to one discipline are extracted or formulaic expressions used in a corpus in which several disciplines are mixed are extracted. The number of documents used for extraction and the extracted formulaic expressions of the existing and presented database are shown. Morley (n.d.) constantly revises the database, and therefore the number of formulaic expressions is not fixed. CF stands for communicative function.

	Discipline	CFs	Documents	FEs
Simpson-Vlach and Ellis (2010)	mixed	15	-	200
Morley (n.d.)	mixed	146	100	$\simeq 2,000$
Mizumoto et al. (2017)	specific	52	1,000	-
Lu et al. (2018)	mixed	12	600	454
Ours	specific	32	61,728	285,183

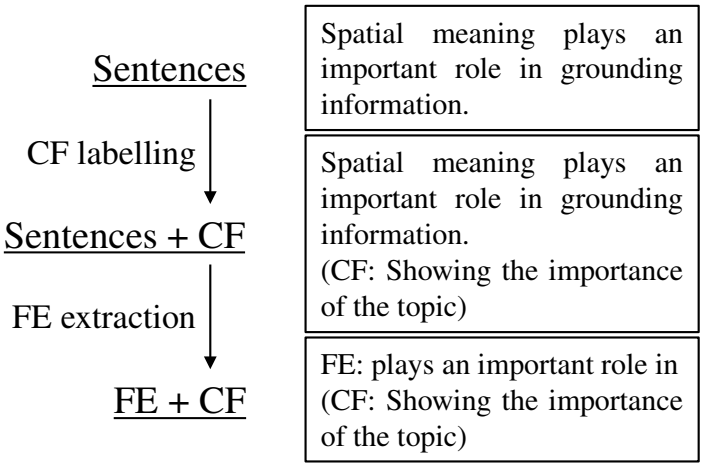

Figure 2.2: Process of creating formulaic expression database. The sentence is cited from Schulte im Walde et al. (2018).

expressions are first extracted and then assigned communicative function labels. So far, both the approaches have been adopted because the communicative function assignment is performed manually (Table 2.5). In this thesis, we propose a fully automated construction of the communicative-function-labelled formulaic expression database, where we consider that the top–down approach to be more beneficial (Figure 2.2). This is because the bottom–up approach requires the classification of formulaic expressions, which is difficult because a perfect formulaic-expression-extraction technique has not yet been realised and formulaic expression embeddings have not been investigated intensively. The top–down approach requires sentence classification, which has highly improved with the recent advancements in pre-trained models.

Table 2.6 describes the existing studies that tried to combine communicative functions and formulaic expressions. Except Academic Phrasebank (Morley, n.d.), the studies did not aim at presenting formulaic expression databases; thus, very few formulaic expressions were collected. Moreover, communicative functions were manually assigned in these past studies, which made it difficult to construct a large database of communicative-function-labelled formulaic expressions. Previous studies have shown that formulaic expressions are discipline-specific, and the resource of academic vocabulary should be presented for each discipline

(Hyland & Tse, 2007; D. Liu, 2012). Thus, the development of communicative-function-labelled formulaic expression databases for each discipline is important; however, many studies focused on *general* formulaic expressions, which were extracted from a mixed corpus consisting of scientific papers on multiple disciplines. Some studies adopted the discipline-specific approach; Mizumoto et al. (2017) considered only the journals on applied linguistics, while Lu et al. (2018) used only the introductions of social-science papers. Moreover, only a small number of documents were used because the existing resources require manual labour for assigning communicative function labels. Hence, we contend that the automated communicative-function-based classification is helpful for constructing a large, comprehensive communicative-function-labelled formulaic expression database.

Academic Phrasebank (Morley, n.d.) is a comparatively large database for academic writing. In this database, example expressions containing formulaic expressions were categorised based on their communicative functions and other writing purposes. The communicative functions are classified into section-based categories: *introducing work*, *referring to sources*, *describing methods*, *reporting results*, *discussing findings*, and *writing conclusions*. The other types of categories are called general language functions, which include *being cautious*, *being critical*, *classifying and listing*, *compare and contrast*, *defining terms*, *describing trends*, *describing quantities*, *explaining causality*, *giving examples*, *signaling transition*, and *writing about the past*. The total number of expressions is approximately 1,000 for the communicative-function-based categories and 1,000 for the general language functions. The expressions were collected from 100 PhD theses of the University of Manchester. The disciplines of the theses were not disclosed, but apparently they were not discipline-specific because technical terms were included in the expressions such as '*metabolism*', '*Aristotle*', and '*development economics*'. The expressions are fragments of sentences that contain formulaic expressions, placeholders, technical terms, and proper nouns; e.g. '*It has been demonstrated that a high intake of X results in damage to ⋯ (Smith, 1998; ⋯).*'. These non-formulaic parts may be useful for humans, but are noises for computers. Removing non-formulaic words to obtain a formulaic expression from these expressions is not easy because it is essentially extraction of a formulaic expression from a sentence. Therefore, this resource cannot be used as it is for construction of formulaic expression database in a computational manner.

2.3 Processing Phrase and Sentences

2.3.1 Word Association Measures and Extraction of Phrasal Expressions

There are many word association measures that have been proposed (Pecina, 2008). The word association measures indicate how strongly two words are connected. One of the most popular measures is point-wise mutual information (PMI) (Church & Hanks, 1990). PMI measures how often a pair of words co-occur; if the two words co-occur more frequently than expected to occur independently, the PMI is larger than 0. PMI performs well on collocation detection (Pecina, 2010).

However, the drawback of PMI is that excessively high scores are assigned to infrequent words. To alleviate this, local mutual information (LMI), normalized PMI (NPMI), positive PMI (PPMI), PMI^2 and PMI^3 were proposed and

formulated as follows:

$$\begin{aligned}
\mathrm{PMI}(a,b) &= \log \frac{p(a,b)}{p(a)p(b)} && (2.1)\\
\mathrm{LMI}(a,b) &= f(a,b)\cdot \mathrm{PMI}(a,b) && (2.2)\\
\mathrm{NPMI}(a,b) &= -\frac{\mathrm{PMI}(a,b)}{\log p(a,b)} && (2.3)\\
\mathrm{PPMI}(a,b) &= 2^{\mathrm{PMI}(a,b)+\log p(a,b)} && (2.4)\\
\mathrm{PMI}^2(a,b) &= \log \frac{p(a,b)^2}{p(a)p(b)} && (2.5)\\
\mathrm{PMI}^3(a,b) &= \log \frac{p(a,b)^3}{p(a)p(b)} && (2.6)
\end{aligned}$$

where a and b denote a word, a,b denotes the co-occurrence of the words, $p(a)$ is a probability of occurrence of a, and $f(a)$ is a frequency of a in a corpus (Role & Nadif, 2011).

These measures are very useful to detect collocation, a pair of two words, but cannot directly be applied to three-word or longer phrases (Constant et al., 2017). In our study, formulaic expressions with different length should be compared, but it is unclear whether the measures of phrases with different length can be compared as such.

2.3.2 Extraction of Informative Phrasal Expressions

In information retrieval and information extraction, phrasal expressions often play an important role. They are used as features that represents a longer text such as a whole document or paragraph.

Zhong, Li, and Wu (2012) tried to overcome the problem that in text mining, phrasal expressions that had more information than single words had not improved performance. They proposed an algorithm to distinguish effective patterns from ineffective ones.

Zhang, Marin, Hutchinson, and Ostendorf (2013) utilised phrases as a feature for text classification. They considered both lexical bundles and phrase frames. In addition to words, they used word class, the part-of-speech and polarity tags.

Marin, Holenstein, Sarikaya, and Ostendorf (2014) utilised a knowledge graph to construct phrasal patterns for text classification. From a corpus, they first created a graph structure in which each word was a node. Then, using the graph, clusters of words were formed. From the clusters, phrasal pattern were extracted.

J. Liu, Shang, Wang, Ren, and Han (2015) improved frequency-based phrase extraction. Generally, longer phrases occur less frequently than shorted ones. However, shorter phrases are sometimes fragments of longer phrases. Thus, the counts of shorter phrases include those of longer phrases. They proposed a method to adjust the frequency.

Bing et al. (2015) used phrases to generate a summary of a document. They first extracted feature words from a document as a pool of concepts and facts. Then, sentences were generated by choosing phrases.

Phrasal expressions as features are basically content-focused expressions. Formulaic expressions for academic writing are functional-focused expressions. Thus, generally used phrases such as '*in this paper*' are not useful to the content-focused text mining, while informative phrases such as '*support vector machine*' are not useful to academic writing assistance.

2.3.3 Sentence Representations

Since the sentence is one of the fundamental units of languages, vector representations of sentences have attracted much research attention. Following successful word embeddings such as word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014), unsupervised methods to acquire sentence embeddings, such as Skip-Thought Vectors (Kiros et al., 2015) have been proposed. Conneau, Kiela, Schwenk, Barrault, and Bordes (2017) found that even a supervised method trained on a dataset for natural language inference yielded universal sentence representations that performed well on various tasks. The current trend in the acquisition of sentence representations is the use of outputs from pre-trained language models such as BERT (Devlin, Chang, Lee, & Toutanova, 2019).

Skip-Thought model (Kiros et al., 2015) is the first neural model for acquiring sentence representations for general purposes not by combining word embeddings but by directly calculate them. The model was inspired by the Skip-gram model (Mikolov et al., 2013), where the input is a word and the output is surrounding words; in Skip-Thought, the input is a sentence and the output is surrounding sentences. Each sentence was encoded and decoded with recurrent neural networks including long short-term memories. This vectors were tested on various tasks, such as classification and semantic relations, and showed promising results.

InferSent (Conneau et al., 2017) is a different model for sentence representations. They constructed a bi-directional long short-term memory architecture to obtain sentence representations utilised for a natural language inference task, where whether two given sentences were entailed, contradicted, or not logically related was judged. After training the model, they tested the sentence representations for various tasks and showed it performed better than Skip-Thought vectors. This research had a great impact upon research on sentence representations because the approach was different from unsupervised ones for general purposes or supervised ones for specific tasks.

These past studies utilised neural architectures such as recurrent neural networks, long short-term memories, and convolutional neural networks, all of which were considered to be computationally time-consuming. Vaswani et al. (2017) proposed using only attentions instead of recurrent architectures. This is called Transformers, which consists of an encoder-decoder containing multiple attentions. The model achieved good performance although the architecture is quite simple.

In addition to transferring trained models and the Transformer model, Devlin et al. (2019) introduced fine-tuning to the usage of neural network models for solving natural language processing tasks. The BERT model consists of 12 transformer encoders. The input of the model is the summation of token, segment, and position embeddings. It can be a single sentence or two sentences; two sentences are split by a special token [SEP]. A special token [CLS] is put in the beginning of an input sequence. The segment embedding denotes the segment of the sentences. The position embedding indicates the position of each word (sub-word); Transformer originally has this embedding.

The pre-training was conducted with two tasks: prediction of masked words and prediction of next sentences. In the former setting, 15% of random sub-words were masked and the model was trained to predict the masked tokens. In the latter setting, a pair of sentences were given and the model judged whether the sentence pair was contiguous or not. After the pre-training, this model can be used as a supervised machine-learning model. When it is fine-tuned for classifi-

cation tasks, the output of the [CLS] token is fed into another layer, such as a linear layer, as a sentence representation. This model and setting were successful; it performed much better on many tasks than other existing models.

The pre-training costs a lot of time, but a pre-trained model can be used for general purposes; thus, pre-trained models are publicly available. The original BERT was pre-trained on two datasets: BookCorpus and Wikipedia. BioBERT (Lee et al., 2019) is a BERT model pre-trained on abstracts in PubMed and research articles in PMC. SciBERT (Beltagy et al., 2019) is another BERT model pre-trained on AI conference papers and biomedical papers collected by Semantic Scholar.

The implementation of BERT is also provided by multiple organisations. Google published the BERT code and pre-trained models for TensorFlow; Hugging Face also made ones public for TensorFlow and PyTorch.

In any case, sentence representations for general purposes do not always contain every aspect of languages. Hence, it is important to investigate which linguistic aspects they contain, and comprehensive evaluation benchmarks have been proposed for this purpose (Conneau & Kiela, 2018; Wang et al., 2018). These benchmarks can well evaluate sentence representations in terms of semantic factors such as semantic relatedness, paraphrases and caption-image retrieval as well as logical factors such as entailment. Communicative functions, which the present thesis focuses on, are another perspective related to rhetorical structure. Basically, the discourse structure is realised in multiple sentences, but a sentence can play a role of a rhetorical unit to make discourse. Therefore, rhetorical information embedded in sentence representation is worth evaluating.

Chapter 3

Creating Datasets

3.1 Introduction

Formulaic expressions and their communicative functions have been investigated mainly in academic writing research to help people write papers more rapidly and fluently (Cortes, 2013; Mizumoto et al., 2017; Omidian, Shahriari, & Siyanova-Chanturia, 2018). There even exist some computer systems for academic-writing assistance[1][2] that rely on these communicative functions to improve the user's writing skills by suggesting commonly-used, alternative formulaic expressions. This is especially helpful for users whose native language is not English (AlHassan & Wood, 2015; Chen & Baker, 2010).

Writing-assistance systems use pre-compiled lists of formulaic expressions labelled with communicative functions for each discipline. There are two approaches to create such lists (Biber et al., 2007): 1) the top–down approach, in which communicative functions of sentences are first identified and formulaic expressions are subsequently extracted from the sentences, and 2) the bottom–up approach, in which formulaic expressions are first extracted from a corpus and their communicative functions are subsequently identified. With either approach, problems arise when computational methods are applied to create the lists. For the top–down approach, no evaluation dataset is publicly available for classifying sentences into communicative functions. Moreover, evaluation datasets are expensive and time-consuming to build. To alleviate this issue, only smaller portions of papers, such as the abstract (Dayrell et al., 2012; Hirohata et al., 2008; Wu et al., 2006) or introduction (Pendar & Cotos, 2008), were annotated, and a limited number of disciplines were used (Cortes, 2013; Mizumoto et al., 2017). The bottom–up approach is not much better, because there is no established evaluation dataset for detecting formulaic expressions. Previous work, therefore, relied on domain experts to manually assess the quality of extracted formulaic expressions (Brooke et al., 2015; Iwatsuki & Aizawa, 2018), which, in addition to being costly, hinders replicability. Overall, the unavailability of annotated resources for both communicative functions and formulaic expressions has hindered the development of automated methods for detecting communicative functions.

There are, nonetheless, closely related resources for academic writing, in which examples of phrases and wordings are collected and classified into communicative functions. Academic Phrasebank (Morley, n.d.) is one of them. However, the use of this resource as a ground-truth dataset is not straightforward, as it was made with the purpose of helping scholars write and organise scientific papers. Therefore, it contains mostly incomplete sentences as example expressions

[1] `http://langtest.jp/awsum/`
[2] `http://pep-rg.jp/abst/`

(see Figure 3.1) and lacks the contextual information needed to detect communicative functions. Another problem with Academic Phrasebank is that example expressions were retrieved from papers belonging to a wide variety of disciplines ranging from humanities to medicine. Since section structures (Thelwall, 2019), vocabulary, word usage and the use of communicative functions differ among disciplines (Hyland, 2008), it is not reasonable to evaluate classifiers of communicative functions on that resource if one hopes to draw meaningful conclusions.

The present study attempts to address the aforementioned problems by building a new evaluation dataset (Iwatsuki, Boudin, & Aizawa, 2020a). The proposed dataset contains unaltered, contextualised sentences collected from a discipline-specific corpus, that is, the ACL Anthology Sentence Corpus (AASC)[3]. Sentences are annotated with communicative functions (and minimal formulaic expressions) by using a set of labels derived from Academic Phrasebank.

For the communicative-function-based sentence classification, we created a communicative-function-annotated sentence dataset for supervised learning. The dataset consists of a small number of sentences that are assigned communicative function labels. We collected the sentences from scientific papers of multiple disciplines: computational linguistics, chemistry, oncology, and psychology. The collection was conducted by using the minimal formulaic expressions, and to ensure the quality of the dataset, we performed the evaluation on Amazon Mechanical Turk.

Introduction Section

Stating the purpose of the current research

- The specific objective of this study was to …
- An objective of this study was to investigate …
- This thesis will examine the way in which the …
- This study set out to investigate the usefulness of …

…

Describing the research design and the methods used

- Data for this study were collected using …
- Five works will be examined, all of which …
- This investigation takes the form of a case-study of the …
- This study was exploratory and interpretative in nature.

…

Figure 3.1: Example expressions from Academic Phrasebank that are classified into communicative functions (written in bold).

The contributions of this chapter are as follows:

- we presented the FECFeval dataset, where communicative-function-labelled CoreFEs and sentences were collected.
- we presented the communicative-function-annotated sentence dataset for the supervised communicative function label assignment.

3.2 Preparation

3.2.1 Overview

This section describes the process we followed for building our dataset, which consists of sentences labelled with CFs. Figure 3.2 presents an illustration of this

[3] https://github.com/KMCS-NII/AASC

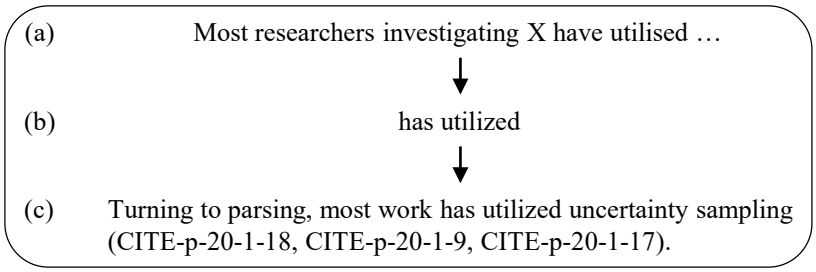

Figure 3.2: Sentences are collected in the following steps. (a) An example expression collected in Academic Phrasebank, which is not a complete sentence. Most of the expressions do not appear in a corpus. Even the formulaic expression in the example expression are not used in a corpus because they are too long. (b) We choose a core FE (core FE) by shortening an formulaic expression. (c) By using the core FE as a query, we retrieve several sentences from a corpus. The sentence (c) is cited from Osborne and Baldridge (2004).

process. Starting from the example expressions provided in Academic Phrasebank, we queried a collection of scientific papers for candidate sentences, each of which was assigned a communicative function label. As most of the example expressions are domain dependent or too specific, we also performed an intermediate manual shortening step to generalise expressions and retrieve more sentences.

3.2.2 Academic Phrasebank

In the first step, we used Academic Phrasebank (Morley, n.d.), which contained many example expressions labelled with communicative functions. An example is shown in Figure 3.1. Each example expression bore an formulaic expression, which was not explicitly marked. More than one thousand example expressions were collected and classified into 72 communicative functions (see Table 3.1). However, this resource has the two problems described in the introduction: incomplete sentences without context and expressions that are not domain specific. Therefore, it cannot be used as a ground-truth dataset.

Communicative functions were also modified because some were (1) based not on the rhetorical structure of a paper but rather on a grammatical perspective, (2) not distinguishable between each other or (3) not relevant for natural language processing (NLP), the discipline of the corpus we used. We present some examples here. Because of (1), '*Describing the process: infinitive of purpose*' and '*Describing the process: verbs used in the passive*' were integrated into one category named '*Describing the process*'. Because of (2), '*Reference to a previous investigation: researcher prominent*' and '*Reference to a previous investigation: investigation prominent*' were integrated. Because of (3), we removed the function '*Giving reasons for personal interest in the research*' as it was not common in the NLP community. After our modifications, the number of core FEs is 397, and the number of CFs is 39 (see Table 3.1). All the CFs of Academic Phrasebank and the modified CFs are shown in Table 3.2 (introduction section), 3.3 (background section), 3.4 (methods section), 3.5 (results section), and 3.6 (discussion section).

Table 3.1: Numbers of example expressions (EEs) and communicative functions (CFs) in Academic Phrasebank that we modified because many example expressions do not appear in the corpus and some communicative functions are not based on the rhetorical structure of scientific papers. We call the modified expressions the CoreFEs and only one CoreFE is annotated in each example expression.

	Original		Modified	
	EEs	CFs	EEs	CFs
Introduction	328	17	104	11
Background	232	15	92	7
Method	210	14	82	6
Results	173	14	58	6
Discussion	153	12	61	9
Total	1,096	72	397	39

3.2.3 CoreFEs

We retrieved sentences from the corpus by using formulaic expressions as queries. Formulaic expressions were extracted from the example expressions by hand, but because they were very specific or sometimes contained irrelevant content, some queries returned no results. Therefore, we simplified and shortened the formulaic expressions and obtained what we call the CoreFEs to retrieve more sentences. For example, '*by adapting the procedure used by*' is an formulaic expression recorded in Academic Phrasebank, but it was not used in our corpus. Thus, we modified it to the CoreFE '*by adapting*'. All the CoreFEs are listed in Table 3.7. CoreFEs are not only continuous word sequences: in the table, discontinuous sequences, such as '*the main disadvantage of * is*' (*showing the main problem in the field* in the introduction) and '*selected * on the basis of*' (*showing criteria for selection* in the methods), and single-word CoreFEs, such as '*understudied*' (*showing limitation or lack of past work* in the introduction) and '*historically*' (*history of the related topics* in the background). The usage of CoreFEs caused noisy results; thus, we manually selected sentences that had an intended communicative function after retrieving candidate sentences.

Table 3.7: List of CoreFEs in each communicative function (CF).

Section	CF	CoreFE
introduction	showing the importance of the topic	is fundamental to
		has a central role in
		is becoming a key
		plays a vital role in
		plays a critical role in
		is essential for
		play an important role in
		plays a crucial role in
		become a central issue
		is among the most important
		there is a growing body of
		is an important component in
		a key aspect of * is
		is a classic problem in
		is an important aspect of
		is at the heart of

(Continued)

Section	CF	CoreFE
		has been studied by many researchers
		has been the subject of
		has been instrumental in
		is an important area
		has received considerable attention
		recently there has been interest in
		in recent years, there has been an increasing interest in
		recent developments in
		decades have seen
		recent trends in
		the last decade has seen
		has been attracting
		recent developments in
		recent years have seen
	showing the main problem in the field	is a major problem in
		one of the main obstacles
		one of the greatest challenges
		a key issue is
		the main disadvantage of * is
		the main challenge is
		is a major problem
		there is an urgent need
	showing what is already done in the past work	recent evidence suggests that
		it has previously been observed that
		several attempts have been made to
		previous research has established that
		previous research has found
		there is a growing body of work
		theories have been proposed
		it is well established that
		have been explored in
		studies have provided
	showing controversy within the field	have been raised about
		has been challenged
		major issue concerns
		subject of debate
		subject to discussion
	showing limitation or lack of past work	have only focused on
		studies are limited to
		has tended to
		has been restricted to
		no work exists
		it is unclear if
		there is no agreement
		no previous study
		has not been investigated
		there has been little analysis
		little attention has been paid to
		understudied
		few studies have investigated
		little discussion

(Continued)

Section	CF	CoreFE
		has not been closely examined
		is still lacking
		there have been no studies
		has received little attention
		there are few studies
		studies have attempted to
		little understanding
		has not been established
		it is not known
		less is known about
		remains unclear
		very little is known
		there is uncertainty
		poorly understood
		little is known about
		it is not clear what
		not fully understood
		very little is known
		few studies have investigated
		has not been investigated
	showing the aim of the paper	in this paper we argue that
		this paper attempts to
		it will be argued
		in this paper, we attempt to
		the aim of this paper is
		the purpose of this paper is
		this paper argues that
		this paper gives
		this paper discusses
		this paper attempts to
		this paper provides
		this paper reviews
		this paper reports
		this paper explores
		this paper considers
		this paper examines
		this paper proposes
		this paper compares
		this paper investigates
		this paper describes
		the objective of this paper is to
		the objective of this work is to
		this paper aims to
	showing brief introduction to the methodology	this * takes the form of
		this work uses * approach
		data * is drawn from
		the approach to * is
		by employing
		is adopted to
		are used in this
		approach * taken in this
	showing the importance of the research	this is the first study
		this paper offers

(Continued)

Section	CF	CoreFE
		will help
		for the first time
		make contributions to
		it is hoped that
		is not * of this paper
		is beyond the scope of this
	showing the outline of the paper	section of this paper will
		this paper begins
		the remaining part of the paper is organized as follows
		addressed in this paper
		this paper is divided into
	showing explanation or definition of terms or notations	throughout this paper
		the term * has been used
		can be defined as follows
		adopt the definition
background	general introduction to past work	has highlighted
		exist in the literature
		there are relatively few
		a large body of literature
		there is a small body of literature
		has been published on
		previous findings
		has revealed
	history of the related topics	has a long history
		over the past decade
		in recent years
		early examples
		over the past two decades
		historically
		it is only since * that
		first articulated
		it was not until * that
	what is done in past work	has utilized
		using this approach
		have been undertaken
		several studies have investigated
		has focused on
		previous studies have explored
		have examined
		researchers have considered
		have attempted to
	what is found or suggested in past work	suggest that
		has established
		have shown that
		it has been argued that
		have been published
		have been found to
		have argued that
		there is consensus

(Continued)

Section	CF	CoreFE
		have identified
		it has been demonstrated that
		it has been suggested that
		it has been shown that
		several studies have used
		studies have found
		studies have reported
		studies have shown that
		studies have indicated that
		have suggested that
		have demonstrated that
		have confirmed
		have revealed
		have highlighted
	what is done in past work	cite- * compared
		cite- * measured
		cite- * used
		cite- * identified
		cite- * carried out
		cite- * studied
		cite- * analyzed
		cite- * performed
		cite- * reviewed
		cite- * conducted
		cite- * investigated
	what is done in past work	a recent study
		a study by
		a recent literature review
		preliminary work on
		was first carried out
		was presented by
		the study by
		in an analysis of * found
		in a recent study
		in a recent experiment
		was originally
		was first studied by
		was first reported
		was studied extensively
		cite- * provides
		cite- * examines
		cite- * identifies
		cite- * highlights
		cite- * uses
		cite- * mentions
		cite- * considers
		cite- * discusses
		cite- * defines
	what is found or suggested in past work	according to cite-
		as noted by cite-
		cite- * argues that
		cite- * offers
	comparison among past work	similarly cite-
		in the same vein cite-

(Continued)

Section	CF	CoreFE
		every paper
	comparison between the present and past work	unlike cite-
		in contrast to cite-
	comparison among past work	a broader perspective
		conversely cite-
		likewise, cite-
	summary of past work	taken together
		together these
		all of the work
		such studies
methods	showing methodology used in past work	the most well-known
		traditionally
		a number of techniques
		methods have been proposed
		in a variety of ways
		methods exist
		one of the most common
		a long tradition
		recent advances in
		there are a number of methods
		the most popular methods
		have been developed
		a well-established approach in
		have been used in the past
	showing reasons why a method was adopted or rejected	a major advantage of
		the benefit of this approach
		was selected for
		approach was used to
		this method is useful for
		was employed since
		was chosen because
		the advantages of
		one advantage of
		another advantage of
		have a number of advantages
		was used to
		was chosen to
		was adopted to
		the main disadvantage of
		there are problems
	using methods used in past work	according to the procedure
		using the same method as
		based on * proposed by
		by adapting
	showing the characteristics of samples or data	was divided into
		were recruited from
		were representative
		were recruited for
		over half

(Continued)

Section	CF	CoreFE
		met the criteria
		were included in
		were divided into
		were interviewed
	showing criteria for selection	criteria for selecting
		only included in
		was chosen for
		inclusion criteria
		selected * on the basis of
		was drawn from
	description of the process	in order to identify
		in order to understand
		in order to establish
		in order to measure
		in order to determine
		in order to rule out
		in order to control
		in order to assess
		to see if
		to enable
		to increase
		to compare
		to prevent
		in order to remove
		in an attempt to make
		were sent
		were normalised
		was obtained from
		were administered
		were generated
		were approved by
		were used
		were collected
		were run
		were completed
		were taken from
		was set at
		were performed
		were identified
		were gathered
		were coded
		were searched
		the first step was to
		prior to
		after training
		after collection
		after testing
		were asked to
		was carried out
		it was necessary to
		once * were completed
		were then
		was then
		and then
		finally

(Continued)

Section	CF	CoreFE
		the final stage
		was calculated using
results	restatement of the aim or method	aimed to
		the purpose of * was to
		was used to
		were compared
		was tested
		were used to
	reference to tables or figures	table * shows
		table * compares
		table * presents
		figure * provides
		the table * illustrates
		the top half of the table
		the bottom half of the table
		as shown in
		as can be seen from
		it can be seen from
		are summarized in table
		are presented in
		are shown in
		it is apparent from
		highlighted in table
		table * revealing that
		from this table
	description of the results	the mean score for * was
		further analysis showed that
		revealed that
		were shown to
		evidence was found
		significant at
		the results indicate that
		there was * correlation
		the difference * was significant
		there was a significant difference
		no * was detected
		no * was observed
		no * were found
		none of * statistically significant
		no * was found
		unaffected by
		only * were detected
		there was no evidence
		did not show
		did not affect
		found no
		did not increase
		a significant increase
		no significant difference
	describing interesting or surprising results	interestingly
		counterintuitive
		more surprising
		surprisingly

(Continued)

Section	CF	CoreFE
		the most surprising
		interesting because
		the most striking
	comparison of the results	a comparison of * results
		comparing * it can be seen that
	summary of the results	these results suggest that
		these results indicate that
		these results show that
		taken together these results
		the results in this section
discussion	showing background provided by past work	as mentioned in
		prior work that has
		previous studies
		has been reported
	restatement of the results	interesting finding is
		the most interesting finding is
		was found to
		the results of this study
		experiments did not
		the most important finding
		it is interesting to note that
	unexpected outcome	surprisingly
		what is surprising
		was unexpected
		it is somewhat surprising that
		contrary to expectations
	comparison of the results and past work	this study confirms
		also reported
		is consistent with
		comparison * confirms
		accords with
		corroborates
		these results corroborate
		in accordance with * cite-
		are consistent with * cite-
		are in line with
		in contrast to earlier
		than that of * previous
	explanation for findings	a possible explanation for
		may be explained by
		can be explained by
		there are several possible explanations
		may explain
		may be due to
		results are likely to
		could be attributed to
		it is difficult to explain
		cannot be ruled out
		it may be that
		the reason for this is
		might be explained
		may be limited
		it could be argued that
		need to be interpreted

(Continued)

Section	CF	CoreFE
		should be interpreted
	suggestion of hypothesis	these findings suggest that
		we can infer that
		support the hypothesis
		it can be hypothesized that
		suggest that * exists
		these results provide
	implications of the findings	it can therefore be assumed that
		an implication of this
		raises the possibility
		important implications
		results raise
	comments on the findings	disappointing
		encouraging
		was successful
		results are significant
	suggestion of future work	for future research
		there are still
		questions remain
		further work is required to
		for further progress
		a further study
		future studies
		additional studies

3.3 FECFeval Dataset

3.3.1 Sentence Selection

We used the ACL Anthology Sentence Corpus (AASC) as our main source of sentences for several reasons. First, this dataset covers a limited range of disciplines that are all related to NLP, thereby standardising the usage of communicative functions and allowing us, as NLP researchers, to do annotation work. Second, each sentence in AASC is labelled with one out of five section headers (introduction, background, methods, results and discussion), which can be used to narrow down the number of possible communicative functions. To prevent research-topic-sensitive effects, all the sentences were retrieved from different papers in the corpus. Figure 3.3 shows a few instances in the proposed dataset. Each sentence has a sentence ID that corresponds to the sentence ID in AASC. Therefore, the surrounding context of each sentence can be easily retrieved if a classifier needs it.

3.3.2 Quality Analysis of the Dataset

Method

In order to ensure that the sentence selection was correctly conducted and to assess the difficulty in detecting communicative functions, we performed manual evaluation for the dataset. Figure 3.4 shows the detailed design. Evaluators solved quizzes that were made from the dataset. In one quiz, three sentences were picked from a section in the dataset. One sentence was the targeted sentence and another sentence was the correct choice. Both had the same communicative function. The other sentence was the wrong choice (distractor) and had a different

Table 3.2: The communicative functions (CFs) in the introduction section of Academic Phrasebank are modified for three reasons: (2) because the CFs are not distinguishable between each other and (3) because the CFs are not relevant in scientific papers.

CFs of Academic Phrasebank	Modified CFs	Reason
Giving reasons for personal interest in the research	(removed)	(3)
Describing the research design and the methods used	Showing brief introduction to the methodology	
Identifying a controversy within the field of study	Showing controversy within the field	
Explaining key terms used in the current work	Showing explanation or definition of terms or notations	
Explaining the inadequacies of previous studies Identifying a knowledge gap in the field of study Identifying the paucity or lack of previous research	Showing limitation or lack of past work	(2)
Stating the focus, aim, or argument of a short paper Stating the purpose of the current research	Showing the aim of the paper	(2)
Explaining the significance of the current study	Showing the importance of the research	
Establishing the importance of the topic (time frame given) Establishing the importance of the topic for the discipline Establishing the importance of the topic for the world or society	Showing the importance of the topic	(2)
Describing the limitations of the current study	Showing the limitation of the research	
Establishing the importance of the topic as a problem to be addressed	Showing the main problem in the field	
Outlining the structure of the paper or dissertation	Showing the outline of the paper	
Referring to previous work to establish what is already known	Showing what is already done in the past work	

communicative function. The communicative function of the targeted sentences was given. Figure 3.5 shows an example of the quizzes.

Each evaluator was asked to guess the communicative function of the sentences and choose the one that seemed to have the same communicative function as the targeted sentence. Because sentences were retrieved from different papers, the contents could be unrelated to each other, but the targeted sentence and the correct choice should be alike in terms of communicative functions. If an evaluator did not decide the answer, we did not include them as an evaluator for the quiz when calculating the accuracy. Four evaluators were assigned to

Table 3.3: The communicative functions (CFs) in the background section of Academic Phrasebank are modified for three reasons: (1) because the CFs are not based on the rhetorical structure of a paper but on grammar, (2) because the CFs are not distinguishable between each other, and (4) because CoreFEs were not found.

CFs of Academic Phrasebank	Modified CFs	Reason
Some ways of introducing quotations	(removed)	(4)
Stating what is currently known about the topic	(removed)	(4)
Synthesising material: bringing sources together	Comparison among past work	
Emphasising the difference between the present study and past work	Comparison between the present and past work	
General comments on the relevant literature	General introduction to past work	
Summarising the review or parts of the review	Summary of past work	
Previous research: area investigated Previous research: methodological approaches taken Reference to what other writers do in their text	What is done in past work	(2)
Reference to a previous investigation: investigation prominent Reference to a previous investigation: researcher prominent Reference to a previous investigation: time prominent Reference to a previous investigation: topic prominent	What is done in past work	(1)
Previous research: a historical perspective	History of the related topics	
Previous research: what has been established or proposed Reference to another writer's idea or position	What is found or suggested in past work	(2)

introduction and background sections while five evaluators were assigned to the remaining sections (the different numbers of evaluators are coincidental).

After the evaluation, we calculated the accuracy and inter-evaluator agreement using Fleiss' κ. The accuracy indicates how likely evaluators were to choose the correct answers, while the agreement indicates the degree to which they made the same choice. Thus, if the sentence selection in the process of creating the dataset fails to make pairs of sentences with the same communicative functions, the accuracy will be low but the agreement will be high. In other words, a low accuracy and high agreement indicate that the dataset is of low quality. In addition, if the task of detecting communicative functions is very difficult, evaluators

Table 3.4: The communicative functions (CFs) in the methods section of Academic Phrasebank are modified for three reasons: (1) because the CFs are not based on the rhetorical structure of a paper but on grammar, (2) because the CFs are not distinguishable between each other, and (4) because CoreFEs were not found.

CFs of Academic Phrasebank	Modified CFs	Reason
Indicating methodological problems or limitations	(removed)	(4)
Describing the process: adverbs of manner Describing the process: expressing purpose with 'for' Describing the process: infinitive of purpose Describing the process: questionnaire design Describing the process: sequence words Describing the process: statistical procedures Describing the process: 'using' + instruments Describing the process: verbs used in the passive	Description of the process	(1)
Indicating criteria for selection or inclusion	Showing criteria for selection	
Describing previously used research methods Indicating the use of an established method	Showing methodology used in past work	(2)
Giving reasons why a method was adopted or rejected	Showing reasons why a method was adopted or rejected	
Describing the characteristics of the sample	Showing the characteristics of samples or data	

will become confused, resulting in both a low accuracy and low agreement.

Results and Discussion

Table 3.8 presents the statistics of the dataset and the results. The accuracy and agreement in the table are macro averages of the accuracy and agreement for each communicative function. The results show that all the sections except *methods* yielded high accuracy and agreement, which implies that the dataset is of sufficient quality and the task is not too difficult. Confusion matrices for each section are shown in Table 3.9, 3.10, 3.11, 3.12, and 3.13. Communicative functions for *introduction* yielded the highest scores even though the number of functions is higher than those of the others. Confusions rarely happened probably because the communicative function set was properly created and did not overlap with each other. The *background* section indicates a little confusion. This is because all the communicative functions in the section are to some degree related to past work, which caused the confusion. The *methods* section yielded a moderate ac-

Table 3.5: The communicative functions (CFs) in the results section of Academic Phrasebank are modified for two reasons: (2) because the CFs are not distinguishable between each other, and (3) because the CFs are not relevant in scientific papers.

CFs of Academic Phrasebank	Modified CFs	Reason
Surveys and interviews: Introducing excerpts Surveys and interviews: Reporting participants' views Surveys and interviews: Reporting proportions Surveys and interviews: Reporting response rates Surveys and interviews: Reporting themes	(removed)	(3)
Transition: moving to the next result	Comparison of the results	
Highlighting interesting or surprising results	Describing interesting or surprising results	
Reporting positive and negative reactions Stating a negative result Stating a positive result	Description of the results	(2)
Highlighting significant data in a table or chart Referring to data in a table or chart	Reference to tables or figures	(2)
Referring back to the research aims or procedures	Restatement of the aim or method	
Summarising the results section	Summary of the results	

curacy and low agreement, which implies that the task is more difficult than the four other sections. The communicative function, *description of the process* was found to be confused with others, probably because this communicative function is broader than the others. In other words, all sentences in *methods* could be labelled with that function. However, it is difficult to define communicative functions more finely for *methods* because methodology varies too widely among papers. In the *results* section, a similar problem occurred; *description of the results* was found to be confusing because this is also a broad communicative function. In the *discussion* section, *suggestion of hypothesis* seemed confusing.

Table 3.14 lists the number of quizzes at different accuracy thresholds. We note that 64.7% of the data showed 100% accuracy, and the accuracy for 84.4% of the data is greater than 75%, which implies that the majority of the quizzes are easy to answer. Thus, the task of detecting the CFs of sentences is not too difficult for humans. It can also be said that CFs are understandable regardless of the content of a sentence. The accuracy is recorded in the dataset so that other researchers can use specific part of the data such as only data with 100% accuracy. The dataset is available at `https://github.com/Alab-NII/FECFevalDataset`.

Table 3.6: The communicative functions (CFs) in the discussion section of Academic Phrasebank are modified for three reasons: (1) because the CFs are not based on the rhetorical structure of a paper but on grammar, (2) because the CFs are not distinguishable between each other, and (3) because the CFs are not relevant in scientific papers.

CFs of Academic Phrasebank	Modified CFs	Reason
Providing background information: reference to the question	(removed)	(4)
Commenting on the findings	Comments on the findings	
Comparing the result: contradicting previous findings Comparing the result: supporting previous findings	Comparison of the results and past work	(2)
Advising cautious interpretation of the findings Offering an explanation for the findings	Explanation for findings	(2)
Noting implications of the findings	Implications of the findings	
Restating the result or one of several results	Restatement of the results	
Providing background information: reference to the literature	Showing background provided by past work	
Suggesting general hypotheses	Suggestion of hypothesis	
Giving suggestions for future work	Suggestion of future work	
Indicating an unexpected outcome	Unexpected outcome	

3.4 Communicative-Function-Annotated Sentence Dataset

3.4.1 Corpora of Scientific Papers

In this study, we considered the corpora satisfying the following conditions. First, because we use full text of scientific papers and have made all the data public, papers must be open access. Second, to construct a comprehensive database, the size of corpora is important. Third, for cross-discipline analyses, a discipline-specific journal is preferred to a multidisciplinary journal. We selected a corpus containing at least 10,000 papers.

Under these three conditions and based on the diversity of the disciplines, we selected four corpora: ACL Anthology Sentence Corpus for computational linguistics (CL), Molecules[4] for chemistry (Chem), Oncotarget[5] for oncology (Onc), and Frontiers in Psychology[6] for psychology (Psy). Papers of the latter three journal are available at PMC[7]. Each corpus comprises more than 10,000 papers and is open access to full text (creative commons licence).

For pre-processing, we performed sentence splitting using ScipaCy (Neumann, King, Beltagy, & Ammar, 2019) and replaced citations and mathematical formu-

[4] https://www.mdpi.com/journal/molecules
[5] https://www.oncotarget.com/
[6] https://www.frontiersin.org/journals/psychology
[7] https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/

Section: Introduction
Function: Limitation or lack of past work
Core FE: has not been investigated
Sentence: Also the extent to which inclusions pose a problem to existing NLP methods has not been investigated.
Sentence ID: D07-1016_s-2-1-0-3

Section: Background
Function: Comparison between the present and past work
Core FE: in contrast to cite-
Sentence: Also, in contrast to CITE-p-14-1-21, we respect the consistency constraint discussed in Section 1.
Sentence ID: E14-1009_s-3-1-2-6

Section: Methods
Function: Criteria for selection
Core FE: selected * on the basis of
Sentence: The verbs were selected from Levin's classes on the basis of our intuitive judgment that they are likely to be used with sufficient frequency to be found in the corpus we had available.
Sentence ID: E99-1007_s-8-1-4-0

Section: Results
Function: Reference to tables or figures
Core FE: figure * provides
Sentence: Figure 5 provides a more detailed characterization of LNQ's performance.
Sentence ID: P18-1029_s-12-6-1-0

Section: Discussion
Function: Suggestion of future work
Core FE: further work is required to
Sentence: Further work is required to reconcile our results with prior work on topic differences and audience size (CITE-p-12-3-2).
Sentence ID: N18-2022_s-10-1-2-1

Figure 3.3: Examples recorded in the proposed dataset (FECFeval). Information on a section, communicative function, and CoreFE is provided. The original sentences are cited from Alex et al. (2007); Pavlopoulos and Androutsopoulos (2014); Srivastava et al. (2018); Stevenson and Merlo (1999); Stewart et al. (2018).

lae with a special token. By using a simple rule-based method, section labels were normalised into five classes: introduction, methods, results, discussion, and other. Each sentence was assigned a section label; we did not use sentences belonging to the 'other' class. The numbers of sentences and documents are listed in Table 3.15.

3.4.2 Communicative Function Set and CoreFEs

We used a set of communicative functions proposed in Section 3.2.2. Table 3.16 describes the numbers of communicative functions in each section. Unlike the FECFeval dataset, we used the only four section labels: introduction, methods, results, and discussion, because the background sections are unfamiliar to the corpora: Chem, Onc, and Psy. We used CoreFEs to create the communicative-function-labelled sentence dataset.

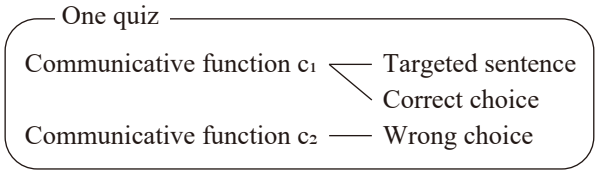

Figure 3.4: Design of the quizzes made from the dataset. The quiz consists of three sentences: a targeted sentence, correct choice and wrong choice. The targeted sentence and correct choice have the same communicative function (c_1), while the wrong choice has a different communicative function (c_2), which is not shown to evaluators.

Q: The purpose of this paper is to outline the main aspects of our ongoing and future work.
Function: The aim of the paper

(1) The aim of this paper is to deal with the first of these steps, i.e. question analysis module.

(2) This work uses a Maximum Entropy Markov Model (MEMM) based approach, which allows to combine different features.

Figure 3.5: Example of a quiz made from the dataset. The targeted sentence is denoted as Q. The communicative function of the targeted sentence is also shown. Evaluators are asked to choose a sentence that they think has the same communicative function out of (1) and (2). In this example, the answer is (1). The sentences are cited from Batista et al. (2008); Makkonen (2003); Przybyła (2013).

3.4.3 Communicative Function Label Annotation

For the communicative-function-based classification, we created a sentence dataset by using the aforementioned corpora. To effectively collect labelled sentences, we took the following procedures (Figure 3.6). First, the CoreFEs were used as queries to retrieve sentences from the corpora. Although the CoreFEs have communicative function labels, the retrieved sentences may not always have the same communicative functions.

Next, we used Amazon Mechanical Turk (AMT) to check if each sentence was assigned correct labels; this process was three-fold. First, a *correct* set of sentences was prepared. Two experts were asked whether the sentences in the correct set were correctly labelled, and sentences whose labels were judged incorrect by at least one expert were removed. Another set of sentences, called the *incorrect* set, was prepared, in which the same sentences were randomly assigned incorrect labels. Second, by using these sets, a pilot test was conducted on AMT. Five annotators were recruited and asked to check whether the labels were correct or not. The annotators satisfied all the following qualifications: the number of ever approved tasks was 1,000 or more, the approval rate of the tasks was 0.98 or more, and an annotator lived in the UK or US. The reward was 0.15 USD for each sentence. Based on this pilot test, we determined the threshold to cut off sentences. Finally, a larger set of sentences was prepared, which was different from the set used in the pilot test. Another five annotators were asked to perform the same task on the larger set. The final dataset comprises the sentences

Table 3.8: Numbers of sentences and communicative functions (CFs). The numbers of sentences and communicative functions are not balanced because the dataset is created based on Academic Phrasebank, which bears imbalance. The accuracy of annotators' choice and their agreement (κ, computed as Fleiss' Kappa) are also listed.

Section	CFs	Sentences	Accuracy	κ
Introduction	11	104	97.9	93.0
Background	7	92	87.7	62.5
Method	6	82	78.4	40.7
Result	6	58	84.4	60.0
Discussion	9	61	85.2	60.7

Table 3.9: Confusion matrix of communicative function annotation in introduction section. The communicative functions are denoted as follows: (1): Showing the outline of the paper, (2): Showing brief introduction to the methodology, (3): Showing the importance of the topic, (4): Showing the limitation of the research, (5): Showing what is already done in the past work, (6): Showing the main problem in the field, (7): Showing the aim of the paper, (8): Showing controversy within the field, (9): Showing limitation or lack of past work, (10): Showing the importance of the research, and (11): Showing explanation or definition of terms or notations.

		Annotations										
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Answer	(1)	11						1				
	(2)		28									
	(3)			86				1			1	
	(4)				8							
	(5)					28						
	(6)						27		1			
	(7)				1			87				
	(8)								4			
	(9)						1			103		
	(10)										16	
	(11)											16

satisfying the threshold.

The correct and incorrect sets consist of 55 sentences. The results of the pilot test are shown in Table 3.17. Accordingly, we set the threshold to 5/5 because high precision was important for creating the formulaic expression database rather than recall, and the strictest threshold did not significantly reduce the sentences. Table 3.18 lists the total number of sentences.

3.5 Conclusion

In this chapter, we presented the FECFeval dataset and task, which we showed could be used to evaluate the formulaic expression extraction methods. We also presented the communicative-function-annotated sentence dataset for the supervised communicative-function-label assignment. The sentence dataset is available at `https://iwa2ki.com/FE/`.

Table 3.10: Confusion matrix of communicative function annotation in background section. The communicative functions are denoted as follows: (1): History of the related topics, (2): Comparison between the present and past work, (3): What is found or suggested in past work, (4): What is done in past work, (5): General introduction to past work, (6): Comparison among past work, and (7): Summary of past work.

		Annotations						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Answer	(1)	32				3	1	
	(2)		8					
	(3)	6		80	3	5	2	
	(4)	2	1	6	145	3	6	1
	(5)					27	1	
	(6)	3	2		1		18	
	(7)		1					11

Table 3.11: Confusion matrix of communicative function annotation in methods section. The communicative functions are denoted as follows: (1): Showing methodology used in past work, (2): Showing reasons why a method was adopted or rejected, (3): Using methods used in past work, (4): Showing the characteristics of samples or data, (5): Showing criteria for selection, and (6): Description of the process.

		Annotations					
		(1)	(2)	(3)	(4)	(5)	(6)
Answer	(1)	57		1	1	1	
	(2)		55		1	5	4
	(3)	3		17			
	(4)			1	25	4	5
	(5)	1	1		2	20	1
	(6)	9	11	20	18	7	140

Table 3.12: Confusion matrix of communicative function annotation in results section. The communicative functions are denoted as follows: (1): Reference to tables or figures, (2): Describing interesting or surprising results, (3): Restatement of the aim or method, (4): Summary of the results, (5): Description of the results, and (6): Comparison of the results.

		Annotations					
		(1)	(2)	(3)	(4)	(5)	(6)
Answer	(1)	82					3
	(2)		35				
	(3)	1		26	3		
	(4)	1			19		
	(5)	11	7	7	8	82	5
	(6)						5

Table 3.13: Confusion matrix of communicative function annotation in discussion section. The communicative functions are denoted as follows: (1): Comments on the findings, (2): Comparison of the results and past work, (3): Unexpected outcome, (4): Restatement of the results, (5): Suggestion of hypothesis, (6): Implications of the findings, (7): Explanation for findings, (8): Suggestion of future work, and (9): Showing background provided by past work.

		Annotations								
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Answer	(1)	12		2	3	1		2		
	(2)	1	51		1					2
	(3)			24		1				
	(4)	3	1		22				3	1
	(5)					25	1	1		3
	(6)				2	4	14			
	(7)	1		4		5		58	1	1
	(8)			1					39	
	(9)					2				13

Table 3.14: Distribution of the quizzes in terms of the accuracy. 64.7% of the dataset showed 100% accuracy.

Accuracy (%)	100	≥ 75	≥ 50
Introduction	98(94%)	104(100%)	104(100%)
Background	61(66%)	78(85%)	90(98%)
Method	30(37%)	57(70%)	77(94%)
Result	33(57%)	45(78%)	53(91%)
Discussion	35(57%)	51(84%)	57(93%)
All	257(65%)	335(84%)	381(96%)

Table 3.15: Number of documents, sentences, and words in each corpus.

Corpus	Documents	Sentences	Words
CL	13,921	1,612,921	32,698,072
Chem	15,949	1,703,902	39,303,460
Onc	19,541	3,029,285	68,719,634
Psy	12,317	1,948,082	49,329,526

Table 3.16: Numbers of communicative functions for each section.

Section	Communicative functions
Introduction	11
Methods	6
Results	6
Discussion	9

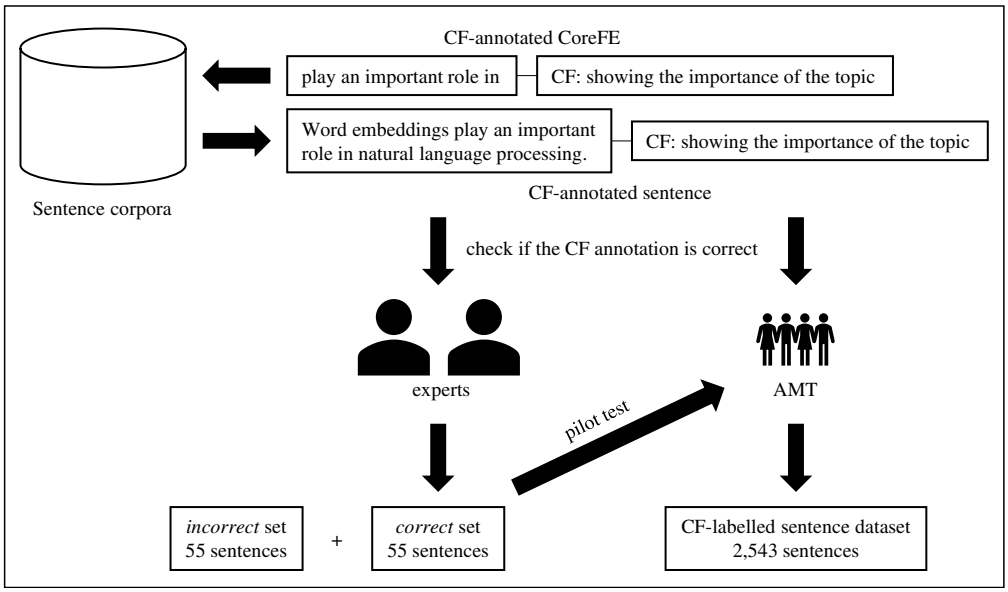

Figure 3.6: We first collected sentences using the CoreFEs. Next, we asked experts to check if the sentences was assigned the correct labels. Using the checked sentences, we conducted the pilot test on Amazon Mechanical Turk (AMT). Finally, we used AMT to check if the labels were correct ot not.

Table 3.17: Threshold indicates the number of annotators (out of five) who judged pairs of the sentence and CF label as correct.

Threshold	Precision	Recall
5/5	0.94	0.80
4/5	0.79	0.98
3/5	0.62	1.00
2/5	0.54	1.00
1/5	0.50	1.00

Table 3.18: Numbers of sentences in the final dataset for training (communicative-function-annotated sentence dataset).

Discipline	Sentence
CL	612
Chem	644
Onc	600
Psy	687

Chapter 4

Assignment of Communicative Function Labels

4.1 Introduction

The first step of the top–down approach to constructing the communicative-function-labelled formulaic expression database is assigning labels of communicative functions to sentences. Communicative functions of a sentence are different from semantics of a sentence. Sentences playing the same communicative functions can contain the information about the methodology or results, which may differ depending on topics of papers.

Recent advancement of pre-trained language models have been reported to achieve much better performance on various tasks of natural language processing than previous methodology. However, it is not evident that the models are able to capture communicative functions since evaluation of the models was conducted from semantic and logical perspective. This is because no dataset that contains sentences labelled with communicative functions has been available.

In this chapter, we address the assignment of communicative function labels to sentences. The assignment of communicative function labels is regarded as a problem of sentence classification. We adopted a supervised machine-learning approach, using SciBERT classifier (Beltagy et al., 2019).

We used the communicative-function-annotated dataset we presented in Section 3.4 for training and evaluating the classifiers. The dataset consists of a small number of sentences that are assigned communicative function labels. We collected the sentences from scientific papers of multiple disciplines. By using this dataset, we fine-tuned SciBERT classifier.

The SciBERT model was reported to be effective in various scientific paper processing tasks, but it is still unclear whether it can detect communicative functions of sentences. We show that the BERT-based models can be used for the communicative-function-based sentence classification.

We carefully considered multidisciplinary problems in the classification. Although the development of a training dataset for every discipline in the world is obviously impossible, demonstrating a successful classification using a single disciplinary dataset is not sufficient for practical use. In this study, we determined whether a model trained on a corpus of one discipline can be applied to that of another discipline. Moreover, the effects of a pre-training dataset were examined by comparing SciBERT and BERT (Devlin et al., 2019). The experimental results show that the SciBERT and BERT classifier performed fairly well in terms of both in-discipline and cross-discipline data.

Finally, we constructed communicative-function-labelled sentence dataset by applying the SciBERT classifier to the whole corpus. Because there are prefer-

ences for communicative function usage depending on disciplines and as preparation and covering of all communicative functions of every discipline are difficult, sentences to which any prepared communicative function label should not be assigned may appear in a corpus (no-CF sentences). These sentences have a bad effect on the classification, which deteriorates the classification performance. Thus, based on the recent work on out-of-distribution detection in natural language processing (Hendrycks & Gimpel, 2017; Hendrycks et al., 2020), we used the maximum value of the softmax layer as the threshold to filter no-CF sentences in order to alleviate the effects of no-CF sentences.

The contributions of our study are as follows:

- we showed that a simple SciBERT-based neural classifier performed reasonably well for the communicative-function labelling problem,
- we showed that the SciBERT classifier can be used even though the discipline of the training data is different from the inferred one, and
- we constructed the communicative-function-labelled sentence dataset.

4.2 Methods

4.2.1 Corpora and Datasets

Dataset for Training and Evaluation

To apply the supervised machine-learning methodology to the classification, a dataset that contains labels of communicative functions is indispensable. We used the communicative-function-annotated sentence dataset (Section 3.4) for training, parameter-tuning, and evaluating the classifier.

The dataset consists of sentences of four disciplines: computational linguistics (CL), chemistry (Chem), oncology (Onc), and psychology (Psy). Each sentence was assigned a communicative function label.

The set of communicative functions we used is the same as the communicative-function-annotated sentence dataset. The list of the communicative functions is in Table 3.2, 3.4, 3.5, and 3.6. The numbers of communicative functions in each section are as follows: 11 in the introduction, 6 in the methods, 6 in the results, and 9 in the discussion. We did not use the background section because the section was only used in the CL corpus.

The dataset was split into training/development and evaluation datasets. The evaluation dataset was created by randomly selecting four sentences for each communicative function because the number of sentences for each communicative function is imbalanced but for evaluating the classification, it is important to make sure that the classifier performs well for every communicative function class. The rest of the sentences in the dataset were used as the training/development dataset. The number of sentences is listed in Table 4.1.

Corpora of Scientific Papers

To create the communicative-function-labelled sentence dataset, the classifier should be applied to corpora of scientific papers. In this study, we used the corpora prepared in Section 3.4.1.

The corpora were made of scientific papers of four disciplines: computational linguistics (CL), chemistry (Chem), Oncology (Onc), and Psychology (Psy). They consist of sentences, which can be used directly as input to the classifiers.

Table 4.1: Number of sentences in communicative-function-annotated sentence dataset for training and evaluation. This dataset was split into training/development and evaluation datasets.

Discipline	Introduction	Methods	Results	Discussion
CL	166	124	137	185
Chem	174	127	137	206
Onc	207	89	112	192
Psy	226	128	135	198

4.2.2 Sentence Classification

Classifiers

The task of the communicative function assignment was regarded as a sentence-classification problem based on communicative functions. Thus, any sentence classifier can be applied to this task.

In this study, we used SciBERT (Beltagy et al., 2019), whose model is the same as BERT (Devlin et al., 2019) but pre-training datasets are different. BERT is a language model that utilises Transformers. To use BERT, two-step learning is needed: pre-training and fine-tuning. In the pre-training step, the model is trained on two tasks: masked language model and nexe sentence prediction. The first task requires the model to predict some masked sub-words. The second task is a binary classification task where two sentences are given to the model and the model answers whether the one sentence follows the other sentence. In the fine-tuning step, like any other supervised machine-learning model, the model is trained on labelled data for a specific task.

SciBERT was reported to perform well on various tasks related to scientific papers (Beltagy et al., 2019). The tasks were named entity recognition, sequence labelling, dependency parsing, and text classification. The text classification task was citation intention prediction, but it is still not clear if the classifier can detect communicative functions of a sentence.

In our experiment, we added a linear layer to the output of the SciBERT for the classification. The output of the linear layer was fed to a softmax layer. The loss function of the classifier was a cross-entropy function. For implementation we used the Huggingface's Transformers library with PyTorch[1]. The pre-trained model of SciBERT (scibert_scivocab_uncased) and BERT (bert-base-uncased) were automatically called in the Transformers library. We used the Trainer class and AutoModelForSequenceClassification class in the library, which did batch processing and model loading.

We also used BERT (Devlin et al., 2019) in addition to SciBERT to see if the difference of the pre-training datasets between the two models had any effect on the performance. The setting was the same as the SciBERT classifier described above.

To test the performance of the classifiers, we first fine-tuned them and then, we evaluated the classification accuracy. The fine-tuning was conducted as follows. First, we split the training dataset into five subsets, out of which one subset was used as a development dataset; the rest of the subsets were used for the training. Second, we trained the classifiers using different parameters. According to Devlin et al. (2019), although there are a host of parameters in the models adjusting the learning rate, batch size, and number of epochs is enough to acquire good

[1] `https://github.com/huggingface/transformers`

performance. Thus, we changed the three parameters to find out the best ones. Third, we calculated the classification accuracy using the development subset. We repeated these steps five times using different subset as a development subset (five-fold cross validation) and finally determined the parameters based on the average accuracy. The parameters were set for each discipline and section.

After the fine-tuning, we evaluated the classifiers using the evaluation dataset. The parameters were set to the results of the parameter-tuning. We calculated the classification accuracy for each discipline and section.

Multidisciplinary Perspectives

Since the usage of formulaic expressions differ across disciplines, databases of formulaic expressions should be constructed for each discipline. To create the multidisciplinary database, the classification must be applied to texts of various disciplinary. As it is costly to manually create a training dataset for each discipline, we tested whether the classifiers trained on a dataset of one discipline could be immediately applied to datasets of other disciplines.

There are two types of datasets used in BERT-based classifiers. The first one is the pre-training dataset. The BERT-based models are trained on a very large corpus in advance and then, the fine-tune is conducted with a smaller dataset for a certain task. The SciBERT was pre-trained on scientific papers from the Semantic Scholar[2] (Beltagy et al., 2019), while the BERT was pre-trained on the book corpus and Wikipedia (Devlin et al., 2019). By comparing the SciBERT and BERT, we show how the difference of the pre-training data have an effect on the classification. The corpora used in this study are open access and were also included in Semantic Scholar. Thus, it may be the case that the cross-disciplinary adaptation is successful because the sentences are (partly) contained in the pre-training dataset.

The second one is the fine-tuning dataset. Although the set of communicative functions does not vary to a great extent across disciplines, preferences for communicative functions are different; some communicative functions are frequently used in one discipline but less in another one. Thus, it is not evident whether the training data made of text of one discipline can be used for different disciplines. To test this, we conducted the training and evaluation using datasets of different disciplines. Our dataset was made of four disciplines; thus, we tested 16 combinations of training and evaluation data.

4.2.3 Creating Communicative-Function-Labelled Sentence Dataset

Using the SciBERT classifier, which was fine-tuned on each disciplinary dataset, we constructed the communicative-function-labelled sentence dataset. In the dataset, every sentence was assigned a communicative function label. The SciBERT classifier was applied to each corpus we prepared in Section 3.4.1.

Because sets of communicative functions in scientific papers have not been established, the communicative function set we used cannot satisfactorily cover all sentences written in papers. Additionally, pre-processing errors, such as sentence splitting, sometimes result in no-CF sentences. These no-CF sentences may have a bad effect on the classification and the performance with the corpora will be worse than that with the training data which do not contain no-CF sentences.

It is not easy to detect no-CF sentences because the no-CF class is not contained in the training dataset; thus, this problem is regarded as the out-of-

[2] `https://www.semanticscholar.org/`

distribution detection problem. Although the maximum value of the softmax layer is not a perfect metrics for out-of-distribution detection, pre-trained Transformers, such as BERT and RoBERTa, with a softmax layer are good detectors of out-of-distribution data (Hendrycks & Gimpel, 2017; Hendrycks et al., 2020).

To remove the no-CF sentences from the resulting dataset, we used the maximum softmax value of the classifier and verified its performance. The verification was performed in the same manner as the creation of the communicative-function-labelled sentence dataset. We set six ranges of the maximum softmax value: [0.00, 0.60], (0.60, 0.70], (0.70, 0.80], (0.80, 0.90], (0.90, 0.99], and (0.99, 1.00]. Most sentences in the corpora were assigned the score higher than 0.99 and thus we set the (0.99, 1.00] range.

Next, we applied the fine-tuned SciBERT classifier to each corpus: 16 corpora (the combinations of four disciplines and four sections). In addition to the communicative function label, the value of the softmax layer was collected from the classifier.

To see the relationship between the softmax range and the classification accuracy, we randomly picked out 100 sentences from each range. As we did in the creation of the communicative-function-annotated sentence dataset, we asked five Amazon Mechanical Turk (AMT) annotators whether the output label was correct. The threshold was also the same: a label was considered correct when all the five annotators labelled it correct. The qualifications for annotation were also the same: the number of ever approved tasks was 1,000 or more, the approval rate of the tasks was 0.98 or more, and an annotator lived in the UK or US. The reward was 0.15 USD for each sentence.

4.3 Results

4.3.1 Sentence Classification with SciBERT

We calculated the classification accuracy on the evaluation dataset and the results of the classification are shown in Table 4.2. Except the methods section in the oncology corpus, the accuracies were very high. The average accuracies were more than 80% for all the disciplines.

We integrated the datasets into one dataset for training and evaluation. The result shows that the performance was better. The sentences that contained CoreFEs account for 4.41% (343,579/7,784,317) in the sentence dataset we finally created in this chapter.

The accuracy of each communicative function is listed in Table 4.3. The communicative functions of *showing criteria for selection* and *Description of the process* in the methods were found to be confusing communicative functions.

The parameters, the batch size and number of epochs, we tuned are listed in Table 4.4. We tested larger number of batch sizes such as 16 and 32 and smaller number of epochs, which were reported in Devlin et al. (2019), but the accuracy was quite low. The learning rate was 5e-5.

4.3.2 Effects of Disciplines of Training Datasets

We verified how the difference in pre-training datasets affected the classification and how the difference between training and evaluation dataset for the fine-tuning did. We conducted the classification with the training and evaluation dataset using the BERT classifier in addition to the SciBERT classifier. We also tested 16 combinations of training and evaluation datasets (4×4 disciplines).

Table 4.2: Accuracy scores of each section in each discipline that were obtained by SciBERT classifier. The average indicates the macro average. In this table, the training and evaluation datasets are from the same disciplines, but the All denotes the integrated training dataset and respective evaluation dataset.

Discipline	Introduction	Methods	Results	Discussion	Average
CL	0.83	0.83	1.00	0.91	0.90
Chem	0.95	0.79	0.88	0.89	0.89
Onc	0.92	0.63	0.92	0.92	0.88
Psy	0.93	0.88	0.96	0.81	0.84
All	0.97	0.92	0.98	0.94	0.95

The results of the BERT classifier are shown in Table 4.5. Compared to the SciBERT results (Table 4.2), the average accuracies of SciBERT are slightly better for the CL and Onc corpora but worse for the Psy corpus. The observed difference was so small that it can be concluded no clear difference was found between the two models.

Next, we tested whether SciBERT and BERT trained on one discipline can be applied to different disciplines. The results are shown in Table 4.6 and 4.7. Except the Onc-CL pair, the classification accuracies were more than 80%; the performance did not deteriorate even though the disciplines of the training and evaluation datasets were different.

4.3.3 Communicative-Function-Labelled Sentence Dataset

Filtering no-CF Sentences

The accuracies of each range are listed in Table 4.8. The table also shows the ratio of sentences belonging to each range in the whole corpora.

The accuracy becomes much lower when the maximum softmax value is 0.80 or lower. Thus, for database construction, we removed the sentences with a score of 0.80 or lower to improve the overall accuracy in the resulting dataset. As a result, approximately 8% of the sentences in the corpora were removed.

Statistics of Sentence Dataset

The statistics of the resulting dataset are listed in Table 4.9. Each disciplinary subset contains more than one million labelled sentences. This dataset is much larger than existing ones, where communicative function labels were assigned manually.

4.4 Discussion

4.4.1 BERT-Based Classifiers for Communicative-Function-Based Sentence Classification

The classification accuracy was quite high and thus the results can be a good baseline for communicative-function-based sentence classification task. Thus, it can be inferred that the BERT-based classifiers can learn the sentential communicative functions.

The number of sentences that contain CoreFEs in the final dataset is 343,579, which accounts for only 4.4% of the dataset. All sentences in the training dataset contained CoreFEs; thus, the classifiers might learn automatically the CoreFEs

Table 4.3: Accuracies in each communicative functions. The accuracies were calculated for each discipline and section and then averaged.

Communicative Function	Accuracy
Introduction	
Showing the importance of the topic	0.88
Showing the main problem in the field	1.00
Showing what is already done in the past work	0.69
Showing controversy within the field	0.79
Showing limitation or lack of past work	0.94
Showing the aim of the paper	1.00
Showing brief introduction to the methodology	0.94
Showing the importance of the research	0.94
Showing the limitation of the research	0.92
Showing the outline of the paper	0.88
Showing explanation or definition of terms or notations	0.81
Methods	
Showing methodology used in past work	0.94
Showing reasons why a method was adopted or rejected	0.81
Using methods used in past work	0.94
Showing the characteristics of samples or data	0.88
Showing criteria for selection	0.63
Description of the process	0.69
Results	
Restatement of the aim or method	1.00
Reference to tables or figures	1.00
Description of the results	0.88
Describing interesting or surprising results	1.00
Comparison of the results	0.75
Summary of the results	1.00
Discussion	
Showing background provided by past work	0.75
Restatement of the results	0.75
Unexpected outcome	1.00
Comparison of the results and past work	0.81
Explanation for findings	0.94
Suggestion of hypothesis	0.94
Implications of the findings	0.94
Comments on the findings	0.88
Suggestion of future work	1.00

as a clue to sentential communicative functions. However, from Table 4.8, the classifiers assigned correct communicative function labels to most of the sentences that did not contain the CoreFEs. In other words, communicative-function-based learning might be used to find a formulaic part that realises a sentential communicative function, into which further investigation should be conducted.

The accuracies obtained with the training and evaluation datasets (Table 4.2) were higher than those obtained with the corpora (Table 4.8). The difference between the training and evaluation datasets and the corpora might explain the difference of the accuracies. The training and evaluation datasets were so created that most of the communicative function labels were correct. However, the

Table 4.4: Parameters we tuned in SciBERT. We tuned the batch size and number of epochs (formatted in batch/epoch).

Discipline	Introduction	Methods	Results	Discussion
CL	1/20	1/10	2/15	1/20
Chem	1/15	1/15	1/20	1/10
Onc	4/10	3/15	2/15	1/15
Psy	1/15	1/20	2/15	3/10
All	1/5	1/15	4/20	1/10

Table 4.5: (BERT) Accuracy scores of each section in each discipline. The average indicates the macro average.

Discipline	Introduction	Methods	Results	Discussion	Average
CL	0.90	0.84	0.96	0.93	0.88
Chem	0.93	0.87	0.93	0.93	0.89
Onc	0.92	0.66	0.94	0.95	0.86
Psy	0.92	0.88	0.95	0.89	0.92

corpora contain no-CF sentences, which decreased the accuracies. Therefore, we estimate that approximately 10% (the difference in the accuracies) of the sentences in the corpora were no-CF sentences.

The no-CF detection worked fairly. From Table 4.8 it can be said that the maximum value is often too high; 30% of the communicative function labels assigned scores higher than 0.99 were incorrect. However, much lower (≤ 0.80) scores tended to cause lower accuracy. Thus, this approach is useful to improve overall precision, which is more important to construct a communicative-function-labelled formulaic expression database than recall.

4.4.2 Problems in Multidisciplinary Data

We raised two questions: Can the classifier trained on one discipline be applied to other disciplines? Do the pre-training data affect the classification performance?

The results of the sentence classification imply that the SciBERT classifier trained on a dataset of one discipline can be applied to datasets of other disciplines. This mitigates the labour of creating a training dataset for all other disciplines. Therefore, we argue that to create another communicative-function-labelled sentence dataset of another discipline, the CF-labelled sentence dataset we created can be used as a training dataset.

The comparison of SciBERT (Table 4.6) and BERT (Table 4.7) denies that the cross-discipline adaptation worked as long as the discipline was included in pre-training data. Thus, the ability of disciplinary adaptation does not come from the pre-training dataset, which implies that the classifier could be used whether a discipline is covered by the pre-training dataset or not.

In this study, we used four disciplinary corpora. We did not use interdisciplinary journals because formulaic expressions differ across disciplines and we intended to test the effects of disciplinary difference. Thus, each corpus is single-disciplinary dataset, but the coverages of each one is different.

The ACL Anthology Sentence Corpus (AASC) is the corpus made of papers collected in the ACL Anthology, a repository for computational linguistics papers. Computational linguistics is a smaller discipline than computer science or linguistics. Still, the papers can be divided into far smaller fields; e.g. sentiment

Table 4.6: Average accuracy scores by SciBERT. The training and evaluation datasets comprise different discipline.

		Evaluation			
		CL	Chem	Onc	Psy
Training	CL	0.90	0.88	0.86	0.84
	Chem	0.84	0.89	0.91	0.84
	Onc	0.75	0.89	0.88	0.82
	Psy	0.88	0.89	0.88	0.84

Table 4.7: Average accuracy scores by BERT.

		Evaluation			
		CL	Chem	Onc	Psy
Training	CL	0.88	0.87	0.82	0.85
	Chem	0.85	0.89	0.91	0.86
	Onc	0.74	0.91	0.86	0.82
	Psy	0.87	0.92	0.88	0.92

analysis, summarisation, or machine translation. Oncotarget is a journal of oncology, which is part of medicine. Oncology can also be divided into wet and dry or medical, surgical, and radiation oncology. Molecules is a journal of chemistry, whose focus is wider than the other journals. Frontiers in Psychology is a journal of psychology including clinical psychology and cognitive psychology.

The usage of communicative functions and formulaic expressions can vary across these finer fields. In our settings, it can be said that the classifiers covered these differences within one broader discipline because of the training dataset containing various smaller fields.

Table 4.10 shows examples of sentences classified with SciBERT and BERT. The first sentence clearly has a communicative function of *showing the outline of the paper* and the second sentence also clearly conveys *showing the aim of the paper*. However, BERT failed to classify these sentences into the correct categories although SciBERT worked well. Both models worked well on the training/evaluation dataset, which was constructed using CoreFEs, but the most of the sentences in the corpus did not contain CoreFEs. Thus, it may be true that BERT paid more attention to CoreFEs, while SciBERT learned communicative functions of sentences better.

4.5 Conclusion

In this chapter, we addressed the assignment of communicative function labels to sentences automatically, using the SciBERT classifiers. In addition to the fact that the SciBERT achieves good results on various NLP tasks including named entity recognition and dependency parsing, we showed that the model has the ability to recognise communicative functions of sentences. We also showed that the classifier can be applied to disciplines that are different from training dataset. Moreover, we showed that the difference in the pre-training data of BERT-based models does not have much effect on the communicative-function-based sentence classification task.

Using the fine-tuned SciBERT classifier, we constructed the communicative-function-labelled sentence dataset, which was used to extract formulaic expressions afterwards. In order to alleviate the effect of the sentences that should not

Table 4.8: Accuracy scores of each range of the maximum value of the softmax layer, and the proportion of sentences in the corpora.

Range	Accuracy	Proportion
(0.99, 1.00]	0.69	76.1%
(0.90, 0.99]	0.67	12.4%
(0.80, 0.90]	0.74	3.7%
(0.70, 0.80]	0.51	2.4%
(0.60, 0.70]	0.51	2.1%
(0.00, 0.60]	0.43	3.3%

Table 4.9: Number of sentences in communicative-function-labelled sentence dataset. The no-CF sentences were removed from this dataset.

Corpus	Introduction	Methods	Results	Discussion	Total
CL	266,904	362,477	507,592	111,052	1,248,025
Chem	285,810	376,583	721,960	175,266	1,559,619
Onc	441,141	976,205	1,069,044	834,641	3,321,031
Psy	484,615	429,155	288,754	453,118	1,655,642
Total	1,478,470	2,144,420	2,587,350	1,574,077	7,784,317

be assigned any prepared communicative function label, we utilised the maximum value of the softmax layer. As consistent with the previous work, it worked well to remove a set of sentences with lower accuracies.

In most studies, communicative function labels were assigned manually, which resulted in the small number of sentences in sentence data. Our contributions including the training dataset that is freely available make it possible to automatically construct a large collection of sentences with communicative function labels. The dataset is available at `https://iwa2ki.com/FE/`.

Table 4.10: Examples of sentences classified with SciBERT and BERT. Sentences and communicative functions assigned by the two models are shown. The sentences are cited from Pal et al. (2017), Dunietz et al. (2013), Yih (2009), and Chung (2004).

Sentence	Section 3 describes the experimental setup and presents the evaluation results.
SciBERT	*Showing the outline of the paper*
BERT	*Showing explanation or definition of terms or notations*
Sentence	In this paper, we present DAVID (Detector of Arguments of Verbs with Incompatible Denotations), a resource-based system for detecting preference violations.
SciBERT	*Showing the aim of the paper*
BERT	*Showing the outline of the paper*
Sentence	The choice of loss function for training model parameters depends on the true objective in the target application.
SciBERT	*Showing criteria for selection*
BERT	*Description of the process*
Sentence	Increased flexibility for customizing the model of the dialog is needed to enable the software to be applied to the development of other kinds of dialog systems.
SciBERT	*Suggestion of future work*
BERT	*Restatement of the results*

Chapter 5

Extraction of Formulaic Expressions

5.1 Introduction

In scientific papers, the authors often use several fixed phrasal patterns that are specific to the genre, such as '*in this paper we propose*'. These patterns are called *formulaic expressions* or *formulaic sequences*. Formulaic expressions convey the intentions of the authors to the readers, i.e. the manner in which a sentence should be understood. This characteristic of the formulaic expression is called *communicative function*. For example, the phrase '*in this paper we propose*' conveys the communicative function of the sentence meaning *showing the aim of the paper*. Formulaic expressions are useful for understanding the composition of a scientific paper and are helpful in writing the paper.

A few studies have been reported on addressing the extraction of formulaic expressions and subsequent assignment of communicative function labels to them (Cortes, 2013; Mizumoto et al., 2017). However, these works have not rigorously investigated whether the extracted formulaic expressions convey the communicative functions of a sentence. Extracting word n-grams with frequency thresholds has been reported in several studies, although frequent formulaic expressions do not always convey the sentential communicative functions. According to Swales (2019), they are *of little meaning or use to English language teachers and learners*. Machine-learning approaches have hitherto been scarcely adopted because of the dearth of sufficient formulaic-expression-annotated resources.

Evaluating extracted formulaic expressions is another problem. In tasks of extracting phrasal expressions such as named entity recognition and multi-word expression extraction, evaluation is conducted by comparing results to reference data that are created as ground truth in advance. However, as far as formulaic expressions are concerned, it is quite difficult to determine answer formulaic expressions. This is because formulaic expressions may have several acceptable word sequences. For example, both '*in this paper we propose*' and '*in this paper we propose a new method*' are acceptable formulaic expressions. There can be multiple answers in one sentence, it is not easy to annotate them and compare extracted formulaic expressions to them.

In most existing studies, evaluation of extracted formulaic expressions was conducted by some experts checking the quality of them. However, the standards of the judgement are not consistent. Brooke et al. (2015); Simpson-Vlach and Ellis (2010) asked evaluators whether extracted phrases are formulaic although formulaicity is an ambiguous concept.

In this chapter, we propose a new sentence-level formulaic expression extraction method and compare it to several existing methods. We assume that a single formulaic expression is extracted from each sentence because it conveys the en-

tirety of the communicative function of that sentence. The proposed method consists of two steps. First, the named and scientific entities are removed from the sentence. Second, two types of n-grams are extracted from the sentence.

Then, the extracted formulaic expressions were evaluated based on whether they conveyed the sentential communicative functions. The results of manual evaluations show that the proposed method can extract more formulaic expressions representing the communicative functions of sentences than existing methods.

Considering the compilation of a list of formulaic expressions, which will be a possible application of the formulaic expression extraction, removing noisy formulaic expressions and enhancing precision are important. Thus, we tested how effective filtering formulaic expressions based on the number of occurrence of a formulaic expression was, and show that it improved precision much.

For evaluation, we measured how much a formulaic expression conveys a communicative function. We conducted automated and manual evaluations from the viewpoint of communicative functions.

As the automated evaluation method, we propose a sentence retrieval task. This is an extrinsic task based on the idea that because a formulaic expression conveys a sentential communicative function, similarity of well-extracted formulaic expressions can be regarded as similarity of communicative functions. In this task, a query sentence is given and sentences that have the same communicative function as the query should be retrieved. Sentences are converted into vector representations and ranked according to their similarity with the query. To examine how much the formulaic part of a sentence conveys a communicative function, we created sentence vectors by assigning different weights to formulaic words and non-formulaic words in a sentence.

In order to show the proposed task can be used to evaluate formulaic expression extraction methods, we compare the CoreFEs to randomly extracted phrases. We show that the retrieval performance differ between the CoreFEs and random phrases, which infers that this task can be used to evaluate the extraction methods (Iwatsuki, Boudin, & Aizawa, 2020b).

Using the proposed method, we evaluated the proposed formulaic expression extraction method and existing methods. The results show that the every method work better than randomly extracting phrases.

As the manual evaluation method, we asked annotators to check whether each extracted formulaic expression had the same communicative function as a sentence where the formulaic expression was extracted and whether the formulaic expression was reusable for writing scientific papers. We compared the proposed extraction method to existing methods and show that the proposed method is more suitable for the formulaic expression extraction.

5.2 Extraction Methods

5.2.1 Pre-Processing

We used the communicative-function-labelled sentence dataset, which was created in the previous chapter (Chapter 4). Before extracting formulaic expressions, we cleaned the sentences. All the sentences were lowercased, and punctuation (except hyphens and underscores) were removed.

Because formulaic expressions are assumed to be used as they are, we did not lemmatise words of the sentences. Formulaicity sometimes does not allow the replacement of a word of an formulaic expression with another word or flection.

For example, tenses can be section-specific (present or past): '*in this paper we proposed*' rarely occurs in the introduction sections. Formulaicity also avoids grammatical errors such as '*little researches have been done*'. Many previous studies did not lemmatise formulaic expressions (Esfandiari & Barbary, 2017; Mizumoto et al., 2017; Pan et al., 2016; Simpson-Vlach & Ellis, 2010).

5.2.2 Two Approaches in Formulaic Expression Extraction

Since no sentence dataset in which annotations of formulaic expressions are given is available, supervised machine-learning approaches are not applicable to the extraction of formulaic expressions.

Our definition of the formulaic expression requires formulaic expressions to convey sentential communicative functions. Thus, it is natural to use information on communicative functions to extract formulaic expressions. However, communicative function labels are not available in general settings; past studies did not use the labels. In our settings, we can now use the communicative functions labels, but the classification accuracy is not 100% and thus the errors will be propagated to the extraction stage in the pipeline of the top–down approach. Moreover, it is not evident how the labels should be used for the extraction; extracting different word sequences depending on a communicative function without a supervised dataset is difficult. In this chapter, we therefore discuss formulaic expression extraction methods without using communicative function labels.

Two main approaches were considered here for extracting the formulaic expressions: corpus- and sentence-level approaches. In the corpus-level approach, the formulaic expressions are extracted from the entire corpus. A bunch of word n-grams are first extracted and then, based on statistical metrics, formulaic expressions are selected. As the metrics, frequency, mutual information, and word association measures including point-wise mutual information can be used. The corpus-level approach may cause problems with deciding the formulaic expression size and overlap between formulaic expressions (span problem) (Iwatsuki & Aizawa, 2018). For example, when 4-grams are extracted in the experiments, the phrases '*paper we propose a*' and '*we propose a method*' were both extracted, but it is difficult to determine which of these is the better formulaic expression.

In the sentence-level approach, a single formulaic expression is extracted from each sentence (Figure 5.1). This approach can be regarded as a sequence labelling problem, in which each word of a sentence is assigned either formulaic or non-formulaic label; then, only formulaic words are extracted. The sentence-level approach is free of the span problem because it does not have a fixed length for the n-gram. Since a single formulaic expression is extracted from each sentence, only '*in this paper we propose a method*' is extracted. Additionally, this approach is suitable for phrase frames that have a slot, where any word can be inserted. Therefore, we adopted the sentence-level approach in the remaining experiments.

In this chapter, we compared two corpus-level and two sentence-level methods with the proposed method. As the corpus-level approaches, we tested a frequency-based method and Lattice FS (Brooke et al., 2017). As the sentence-level approaches, we tested a frequency-based method and latent Dirichlet allocation (LDA)-based method (Liu et al., 2016).

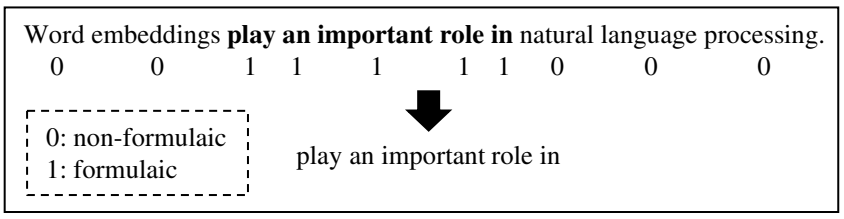

Figure 5.1: Sentence-level formulaic expression extraction. This is regarded as a sequence labelling problem. Each word of a sentence is assigned either a formulaic or non-formulaic label, after which only formulaic words are extracted as an formulaic expression.

Word *n*-grams	Frequency	
in this paper we propose	1,000 / 1,000,000	formulaic expressions
in this paper we propose a new	200 / 1,000,000	
in this paper we present	40 / 1,000,000	
this paper we propose an algorithm	7 / 1,000,000	→ discarded

Figure 5.2: Frequency-based corpus-level formulaic expression extraction method. After extracting word n-grams, they are filtered based on their frequency. In this figure, the frequencies are not actual numbers.

5.2.3 Corpus-Level Extraction

Frequent N-grams

This method regards frequent word n-grams as formulaic expressions. The extraction was conducted as follows. First, for each sentence in all documents, continuous word n-grams were extracted. The size of n-grams was three words or longer; thus, the longest n-gram was a whole sentence. It should be noted that in the corpora we used all the sentences were lowercased and punctuations were removed. Second, the n-grams were filtered based on their frequency in the corpus (Figure 5.2). Although various studies have used different lengths and frequency thresholds for the n-grams, we extracted formulaic expressions whose lengths were three words or greater, and followed the method used by Cortes (2013) for the frequency thresholds: 20 per million words (pmw) for four-word or shorter n-grams, 10 pmw for five-word phrases, 8 for six- and seven-word phrases, and 6 pmw for phrases longer than seven words. Only n-grams satisfying the thresholds remained as formulaic expressions.

Lattice FS (N-gram Lattice)

This approach was originally proposed by Brooke et al. (2015, 2017); Brooke, Tsang, Hirst, and Shein (2014), where n-grams were first extracted and later selected based on the concepts of *covering*, *clearing*, and *overlap*. Covering indicates that if the number of instances of '*we propose*' is almost the same as those of '*we propose a new*', the longer formulaic expression would explain the presence of the shorter formulaic expression. Clearing indicates the opposite idea to covering. Overlap indicates that the expressions '*in this paper we*' and '*this paper we proposed*' should not be accepted as formulaic expressions at the same time. These three concepts are expressed in mathematical form, and the formulaic expressions are optimised computationally. We used an implementation available

on the web[1].

5.2.4 Sentence-Level Extraction

Frequency-Based Filtering

Unlike the corpus-level frequency-based method, the sentence-level method extracts one formulaic expression from one sentence. The extraction was conducted as follows. First, for all the sections and corpora, the numbers of appearance of every word were counted to make the frequency lists. Second, for each sentence, words whose frequency did not satisfy thresholds were replaced with a slot '*'. Continuous slots were converted into a single slot; slots in the beginning and end of a sentence were removed. Words including the slots remaining were a resulting formulaic expression.

The frequency threshold has not been established. However, it is intuitive that formulaic words are more frequent than non-formulaic words. Thus, we removed infrequent words. We used two frequency thresholds, namely 1/50,000 words and 1/100,000 words.

LDA-Based Filtering

Instead of frequency, Liu et al. (2016) proposed utilising LDA (Blei, Ng, & Jordan, 2003). LDA is a topic model that assigns each word a probability of composing a document on a specific topic. A document on a topic is considered to be a set of the words, which occur probabilistically.

From a different perspective, words that occur specifically in a certain topic, the probability of the word is high in the topic. Thus, LDA can be used to extract topic-specific words that represents content of a topic.

The LDA-based formulaic expression extraction utilises the probability assigned to each word. Based on the idea that formulaic expressions are used regardless of topics, topic-specific words are regarded as non-formulaic words.

The extraction was conducted as follows. Each word of a sentence was judged as either topic-specific or topic-independent based on the following criterion:

$$\mathrm{P}(w) = 1 - \frac{\max p_w(i)}{\sum p_w(i)}, \tag{5.1}$$

where $p_w(i)$ is the probability of the word w in a topic i. If $P(w)$ is greater than the threshold, w is formulaic. We used $P(w) > 0.65$ and 10 topics, which they reported optimal.

Proposed Method

The proposed method comprises two steps: (1) removing named and scientific entities and (2) extracting longest word n-grams (Figure 5.3). The first step was based on the idea that the named and scientific entities, including places, organisations, materials, and methods, such as 'Helsinki' and 'word embeddings', do not constitute formulaic expressions. In the second step, dependency parsing was applied to the sentences to determine their roots. After removing the named and scientific entities, two types of word n-grams were labelled as formulaic:

1. the longest word n-gram satisfying a frequency threshold;

[1] `https://github.com/julianbrooke/LatticeFS`

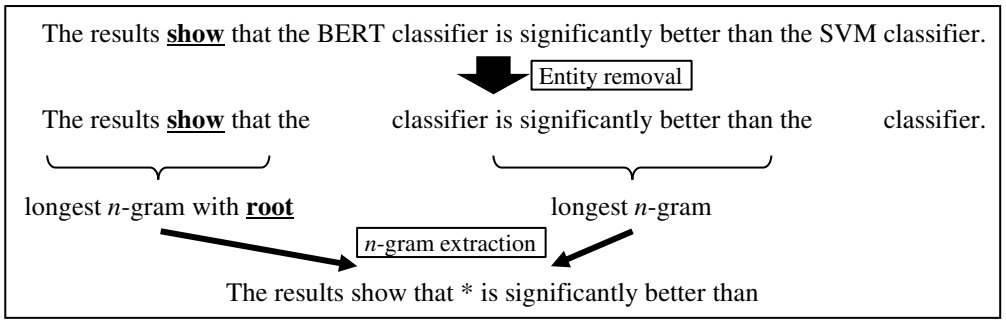

Figure 5.3: The proposed formulaic expression extraction method. The sentential root is in bold.

2. the longest word n-gram that contains a root of the sentence and satisfies the frequency threshold.

If multiple formulaic expressions of the same lengths were found, the most frequent one was prioritised.

We focused on the longest word sequences because Cortes (2013) observed that lengthy formulaic expressions, such as '*the rest of the paper is organized as follows*' existed. Additionally, we assumed that in several cases, sentential communicative functions were realised around the root of the sentence, so that two types of n-grams should be extracted. Specifically, n-grams whose lengths were less than three words were ignored because such formulaic expressions would be too short. The remaining words in the sentence after n-gram extraction were removed. The frequency threshold was thus set to 3 to collect the maximum number of formulaic expressions.

The dependency parsing and entity removal were conducted with ScispaCy (en_core_sci_sm model)[2] (Neumann et al., 2019). ScispaCy is a model for spaCy trained on scientific papers.

In the example in Figure 5.3, the root word is '*show*'. The longest n-gram satisfying the threshold and containing the root would thus be '*the results show that*', while '*is significantly better than*' would be another n-gram that does not contain the root. There could also be cases where these two types of formulaic expressions overlap or be the same.

5.2.5 Filtering Formulaic Expressions

For compiling a list of formulaic expressions, which is one of the applications of the formulaic expression extraction, it is not always necessary to use all these formulaic expressions extracted from every sentence. It is more important to discard non-formulaic expressions. Because the word sequences that occur only once or twice are not formulaic, filtering formulaic expressions based on the number of the occurrence is effective. Therefore, we set several thresholds of the number of formulaic expression occurrence in the dataset, and removed formulaic expressions not satisfying the thresholds.

[2] `https://allenai.github.io/scispacy/`

5.3 Evaluation Methods

5.3.1 Automated Evaluation

Sentence Representations

As mentioned in the introduction, we assume that a communicative function is conveyed by a formulaic expression and thus, the extraction can be evaluated by the strength of connection between a formulaic expression and a communicative function. Therefore, we created sentence vectors by assigning different weights to the formulaic and non-formulaic parts. It is a common way to average word embeddings of each word of a sentence to create a sentence vector. Unlike the ordinary method, we assigned different weights to word vectors of formulaic and non-formulaic parts when averaging them, which can be formalised as follows:

$$\mathrm{s}(W) = \frac{1}{|W|}\left\{\alpha \cdot \sum_{w_i \in \mathrm{FE}} \mathrm{v}(w_i) + (1-\alpha) \cdot \sum_{w_j \in \mathrm{nonFE}} \mathrm{v}(w_j)\right\}, \tag{5.2}$$

where $\mathrm{s}(\cdot)$ is a vector of a sentence, W is a sequence of words in the sentence, which consists of FE (formulaic expression) and nonFE (the remaining words in the sentence), $\mathrm{v}(w)$ is a function that returns a vector representation of w and $\alpha (0 \leq \alpha \leq 1)$ is a parameter determining the weights of the formulaic and non-formulaic parts. When $\alpha = 0.5$, the sentence vector is simply the average of each word embedding. When $\alpha = 1.0$, it consists of only the formulaic part.

Unlike standard sentence representations, where α was fixed to 0.5, we varied α. In our experimental setting, we used skip-gram models (Mikolov et al., 2013) for $\mathrm{v}(w)$ trained on AASC. The parameters of the skip-model models are as follows: the dimension was 200 and the window size was 2. To cover all words, we set the minimum count to 0. The other parameters were default value of an implementation we used[3]: learning rate was 0.025 and the number of iterations was 5. It should be noted that our experiments did not rely on specific word embedding models or parameters.

Sentence Retrieval Task

Instead of directly evaluating extracted formulaic expressions, we propose an extrinsic evaluation method that utilises communicative functions conveyed by formulaic expressions. We adopted the sentence retrieval task to measure the strength of connection between extracted formulaic expressions and communicative functions. In this task, a query sentence is given and then a retrieval system should return an ordered list of sentences ranked according to the similarities of communicative functions between the query and other sentences. Then, the top-N sentences in the list are selected and for evaluation, it is checked how many sentences have the same communicative function as the query (Figure 5.4).

In the system, sentences are converted into vector representation, as described above. Then, sentence vectors are ranked according to the cosine similarity with the query. Mean average precision (MAP) is used for evaluation of the retrieval task (Manning & Schütze, 1999), which is formulated as follows:

$$\mathrm{MAP}(S^i) = \frac{1}{|S^i|} \sum_{s_j \in S^i} \frac{1}{n_{s_j}} \sum_{k=1}^{|R_j^i|} \begin{cases} 0 & (\mathrm{CF}(r_k) \neq \mathrm{CF}(s_j)) \\ \mathrm{P}_j^i(k) & (\mathrm{CF}(r_k) = \mathrm{CF}(s_j)) \end{cases},$$

[3]We used an implementation available at `https://github.com/dav/word2vec`.

Target: Although CG is a radically lexicalist grammatical theory, little attention has been paid to the structure of the lexicon.

#	Sentences	Cosine	Correct
[1]	Recently there has been interest in the development of a general computational treatment of the comparative.	0.9046	
[2]	Dependency parsing is a basic technology for processing Japanese and has been the subject of much research.	0.8974	
[3]	Although it has been suggested that head-driven parsing has benefits for lexicalist grammars, this has not been established in practice.	0.8955	✓
[4]	While it has been observed informally that the internal sentence representations of such models can reflect semantic intuitions (CITE-p-15-4-3), it is not known which architectures or objectives yield the 'best' or most useful representations.	0.8820	✓
[5]	Below, it will be argued that these semantic representations are indeed too weak, but not only from the point of view of Natural Language Processing.	0.8801	

Figure 5.4: Illustration of ranking task. Cosine similarities between a targeted sentence and all the other sentences in its section are calculated, and sentences are ranked by the similarity score. The sentences that have the same communicative function as the targeted sentence are marked correct. In this example, sentences 3 and 4 have the same CF. The sentences are cited from Abekawa and Okumura (2006); Bouma and van Noord (1993); Dorrepaal (1993); Friedman (1989); Hill et al. (2016); van der Linden (1992).

where S^i is a set of sentences in section i, n_{s_j} is the number of correct answers when the query sentence is s_j, R^i_j is an ordered list of the sentence retrieval result, $\mathrm{P}^i_j(k)$ is the precision at position k-th in the list and $\mathrm{CF}(r_k)$ is a communicative function of the k-th ranked sentence $r_k \in R^i_j$.

Validity of the Evaluation Method

In the FECFeval dataset, the CoreFEs are labelled for each sentence. We used the CoreFEs as the result of manual extraction to compare other methods of extraction.

For comparison purposes, we prepared three other types of expressions: NonFE, OneWordCoreFE and NonFE+CoreFE. Figure 5.5 shows the examples of the four patterns. NonFE represents words that are randomly extracted from a sentence in which a CoreFE is removed. The length of NonFE expressions is the same as that of the corresponding CoreFE. These are regarded as bad formulaic expressions. OneWordCoreFE represents one word randomly picked from a CoreFE for each sentence. NonFE+CoreFE represents combinations of NonFE and CoreFE.

OneWordCoreFE simulates an extraction method that misses most parts of formulaic expressions. Putting more weight on OneWordCoreFE means applying less weight to most parts of formulaic expressions. Thus, the performance should start to deteriorate at some point. NonFE+CoreFE simulates an extraction method that extracts the same number of formulaic and non-formulaic words. This should cause lower performance than CoreFE because non-formulaic words are heavily weighted.

Sentence:	When comparing the two online learning models, it can be seen that MIRA outperforms the averaged perceptron method.
CoreFE:	comparing it can be seen that
NonFE:	MIRA outperforms the averaged perceptron method
OneWord:	can
Core+NonFE:	comparing it can be seen that MIRA outperforms the averaged perceptron method

Figure 5.5: Examples of four methods: CoreFE, NonFE, OneWordCoreFE (OneWord) and CoreFE+NonFE (Core+NonFE), all of which are extracted from the sentence. The original sentence is cited from McDonald et al. (2005).

Table 5.1: In the case where sentences whose accuracy was less than 100% were removed, MAP score of CoreFE did not change much and that of NonNE increased.

	All	= 100%
CoreFE	56.2%	56.2%
NonFE	26.9%	31.7%

We also tested the MAP scores of this retrieval task when the dataset was filtered according to the accuracy (Table 3.14). We removed sentences whose accuracy of the human annotation was less than 100%.

Results

In Figure 5.6 the MAP scores of CoreFE, NonFE, CoreFE+NonFE and OneWordCoreFE are shown. Comparing the performances between CoreFE and NonFE extraction, it can be said that good extraction methods improve the sentence retrieval performance as α increases while bad methods deteriorate the performance as α increases. Therefore, the MAP score at $\alpha = 1.0$ (Table 5.2) can be used as an indicator of effectiveness of extraction methods.

We conducted further analysis of the transitions of the performances according to α. As for CoreFEs, i.e. good formulaic expressions, MAP increases monotonically as α increases. Conversely, for NonFE, MAP decreases monotonically. MAP of CoreFE+NonFE is located between the two. The performance increases as well as CoreFEs, but due to non-formulaic words, it is not as good as CoreFEs.

However, for OneWordCoreFE, the peak is at, $\alpha = 0.8$, and MAP decreases after that. This phenomenon can be explained as follows. As α increases from 0.5 to 0.8, heavier weight on the one-word formulaic expressions has a good effect on the performance. In other words, less weight is put on the remaining formulaic expressions. This smaller weight on the remaining formulaic expressions deteriorates the performance with higher α.

From these observations, we argue that the sentence retrieval task is valid to evaluate extraction methods. Basically, comparing MAP scores at $\alpha = 1.0$ is a good indicator. The change of MAP score gives additional insight. If it increases monotonically, most formulaic words are extracted from a sentence. If there is a peak between $\alpha = 0.5$ and 1.0, the method seems to fail to extract a significant part of a formulaic expression.

Table 5.1 shows the MAP scores of the two cases, where all the sentences in the dataset were used and where the sentences whose accuracy was 100% were used. Hereafter, to alleviate the effects of sentences with low accuracy, we use the only sentences with 100% accuracy.

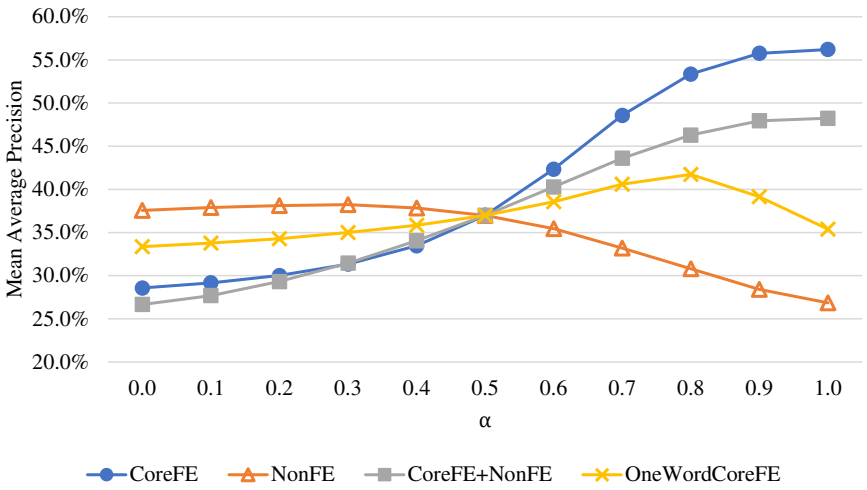

Figure 5.6: Relationships between MAP and α. MAP of CoreFE monotonically increases, while that of NonFE behaves inversely. CoreFE+NonFE shows that lower performance is attributed to extraction of unnecessary words. OneWordCoreFE shows that by missing indispensable words the peak of MAP appears between $\alpha = 0.5$ and 1.0.

Table 5.2: MAP scores of each prepared formulaic expression. CoreFE is the highest, NonFE is the lowest, and the other two are in between.

	CoreFE	NonFE	OneWordCoreFE	CoreFE+NonFE
MAP	56.2%	26.9%	35.4%	48.2%

5.3.2 Manual Evaluation

Basically, manual evaluation is conducted by asking annotators whether extracted expressions are formulaic or not. The problem is what the annotators should be based on in order to judge them.

In our experiment, we asked annotators to judge from two perspectives. First, formulaic expressions must have the same communicative function as a sentence from which the formulaic expression is extracted. This is based on our definition of formulaic expressions. Second, formulaic expressions must be reusable for writing other scientific papers. Considering applications of formulaic expression database including academic writing assistance, formulaic expression must be reusable. Only extracted expressions satisfying both were labelled as correct by annotators.

5.4 Results

5.4.1 Automated Evaluation

We evaluated the sentence-level methods using the FECFeval dataset and the retrieval task. The results are shown in Table 5.3. In addition to MAP scores, we added the ratio of sentences from which no formulaic expression was extracted to the table. With the sentence-level methods, if no formulaic expression is found in a sentence, nothing is extracted, which was considered wrong in this retrieval task.

There are three baselines: Full sentence, CoreFE and NonFE. The full sentence is that every word in a sentence is extracted as an formulaic expression.

Table 5.3: Results of formulaic expression extraction (FECFeval).

Method	MAP	Ratio of no-FE
Full sentence	0.41	0.00
CoreFE	0.56	0.00
NonFE	0.32	0.00
Frequency-based (1/50,000)	0.41	0.00
Frequency-based (1/100,000)	0.40	0.00
LDA-based	0.43	0.00
Proposed (step 1)	0.41	0.00
Proposed (step 2)	0.42	0.01
Proposed (step 1+2)	0.36	0.12

Since a sentence consists of an formulaic expression and content part, and the full sentence contains both, CoreFE was better than the full sentence. CoreFE is manually extracted formulaic expression fragments; thus, the MAP score of CoreFE is considered as maximum value in this automated evaluation. NonFE is a randomly extracted n-grams that do not contain CoreFE. In other words, NonFE is wrong formulaic expressions; thus, it is expected that good extraction methods should achieves better performance than NonFE.

The results show that the all tested methods performed better than NonFE, which implies that they extracted at least some part of formulaic expressions. However, it is difficult to compare each method because not much difference was observed. It can be said that there are still much room to improve the methods to achieve CoreFE scores.

Looking at the results of the proposed method, the MAP score of the combination of step 1 (the entity removal) and step 2 (the n-gram extraction) is worse than that of each step. This is because the ratio of no-FE of the combination was higher than that of each step. The proposed method only extracts formulaic expressions of three words or longer. However, the entity removal (step1) sometimes eliminates formulaic words (discussed in the discussion section below), which causes many short n-grams. For instance, '*Research in building factoid QA systems has a long history.*' became '*in * has a*' after the removal was performed. In this case, the second step (the n-gram extraction) was not able to extract meaningful n-grams longer than two words.

5.4.2 Manual Evaluation

We randomly chose 100 sentences from the sentence dataset to evaluate the formulaic expression extraction. For the sentence-level methods, a single formulaic expression was extracted from each sentence. For the corpus-level methods, the formulaic expressions and sentences were not clearly connected. Thus, we randomly selected a single formulaic expression from the set of extracted formulaic expressions for each sentence.

The evaluations were then conducted manually. Three annotators were asked to check if the formulaic expressions extracted with each method had the same communicative functions as the sentences from which they were extracted and if these were reusable when writing scientific papers. The formulaic expressions were presented to the annotators simultaneously, and the method that was applied to the formulaic expression was not disclosed. A total of 100 combinations of sentences and formulaic expressions were randomly selected for evaluations.

The results of the evaluations are shown in Table 5.4, and the proposed

Table 5.4: Ratios of formulaic expressions that two or three out of the three (≥2/3) and all three (3/3) annotators labelled as correct.

Method	≥ 2/3	3/3	Fleiss's κ
Frequent n-grams	0.30	0.09	0.36
Lattice FS	0.07	0.03	0.30
Frequency-based (1/50,000)	0.04	0.02	-0.36
Frequency-based (1/100,000)	0.05	0.02	-0.39
LDA-based	0.08	0.03	-0.20
Proposed (Step 1)	0.13	0.05	-0.27
Proposed (Step 2)	0.54	0.28	0.23
Proposed (Step 1+2)	0.58	0.39	0.44

Table 5.5: Ratios of formulaic expressions whose scores were 3/3 and filtering thresholds.

Occurrence	≥ 1	≥ 3	≥ 5	≥ 7
Ratio of 3/3	0.28	0.45	0.55	0.53
#	39/100	24/53	21/47	21/46

method is observed to show clear advantage over other baselines in the formulaic expression extraction. Each step of the proposed method had a good effect on the overall performance.

Table 5.5 shows the thresholds of the number of occurrence of formulaic expressions and scores. From the table, it can be seen that if formulaic expressions occurring less than three times in a corpus are ignored, the precision would change from 0.39 (39/100) to 0.45 (24/53). It should be noted that the recall cannot be calculated because there are no available formulaic-expression-annotated resources.

5.5 Discussion

5.5.1 Automated Versus Manual Evaluation

As discussed in Section 2.2.5, the majority of the evaluation way of the formulaic expression extraction has been manual evaluation. It is obvious that the manual evaluation is too costly to compare many extraction methods and parameters.

The manual evaluation showed that the proposed method was quite different from the LDA- or frequency-based sentence-level methods. However, the automated evaluation showed not much difference between them. It still showed the difference between the NonFE and the other methods. Therefore, the proposed automated evaluation method can be used to check if the methods are better than random results. Manual evaluation should be also conducted for the methods that achieved better scores than NonFE.

5.5.2 Errors in Proposed Method

Errors in Entity Recognition

We analysed the errors (formulaic expressions that 1/3 or less annotators judged as correct) in the proposed method. The errors in the entity recognition (step 1) accounts for approximately 60% of all the errors. They can be classified into

Table 5.6: Examples of errors in named and scientific entity recognition. The sentences are cited from Kim et al. (2018); Xie et al. (2015). CF stands for communicative function.

CF	Full sentence	Sentence without entities
Reference to tables or figures	From this table, we observe that the topics learned by our method are better in coherence than those learned from the baseline methods, which again demonstrates the effectiveness of our model.	from this * we observe that the topics learned by our * are better in * than those learned from the * which again demonstrates the * of our
Showing limitation or lack of past work	Although the cellular uptake efficiency could be improved by adjusting the size and the sequence of DNPs in the previous study, it has not been investigated whether the DNPs can also be used in the in vivo environment rich in nucleases.	although the * could be improved by adjusting the * and the * of * in the previous * it has not been * whether the * can also be used in the * rich in

two types: (1) entities are not removed and (2) formulaic words are removed as entities though they are not entities. Most of the errors were the type (2).

Table5.6 lists the examples of this error. From this table, it can be seen that formulaic words such as '*table*' and '*investigated*', which are indispensable for representing the communicative functions, were removed. When formulaic words are removed at this stage, meaningful n-grams are not to be extracted in the step 2. This results infer that entity recognition is crucial to the proposed method, and the recognition should be improved much.

Errors in N-grams

Another type of errors is the errors in the n-gram extraction (step 2). In the proposed method, we extracted two different n-grams: the longest n-gram containing the sentential root and the longest n-gram that does not necessarily contain the root, both of which satisfied the threshold of the number of occurrence in the corpora.

The majority of this error is that the extracted two n-grams are the same but do not contain communicative-function-realising part. Table 5.7 lists the examples of this error. The span error occurred in the second example. Since '*both plasma and urine*' is content part, the formulaic expression should not include '*both*'. The other examples missed the communicative-function-realising part. In the first example, '*a common approach*' is important to the introduction to the methodology. In the third example, detail numbers were extracted. It should be noted that the numbers sometimes constitute a formulaic expression because in some disciplines, there exist very fixed numbers, such as '*a p value*

Table 5.7: Examples of errors in n-gram extraction. The sentences are cited from Guo et al. (2017); Phan et al. (2016); Sarkar (1998); Vivas et al. (2019); Xia et al. (2016). CF stands for communicative function; FE stands for formulaic expression.

CF	Sentence	FE
Showing brief introduction to the methodology	A common approach used to assign structure to language is to use a probabilistic grammar where each elementary rule or production is associated with a probability.	is to use a
Restatement of the results	For example, shared specific genomic aberrations were observed in both plasma and urine cfDNAs at loci of PTEN, TMPRSS2 and AR (Figure 1 and [CITATION]).	were observed in both
Description of the results	Rs679620 was also associated with increased OA risk in dominant (" TC-TT " , OR = 2.03, 95% CI: 1.03-4.01, P = 0.038) and over-dominant model analyses (" TC " , OR = 2.04, 95% CI: 1.05-3.96, P = 0.033).	p 0038 and
Using methods used in past work	The smoothness value used for the AlphaSim calculation was based on the smoothness of the residual image of the statistical analysis as proposed by [CITATION] .	was based on the
Showing controversy within the field	However, it should be noted that the biological involvement of many of these targets in HBD-3 activities has been challenged in recent years [[CITATION]].	however it should be noted that the

less than 0.05 was considered significant'. In the fourth example, the formulaic expression missed '*as proposed by*' to show the method was used in past work. In the last example, the controversy is represented by '*has been challenged*', which was not extracted.

The last example also shows that n-grams that contain the sentential root do not always convey the communicative function. It is true that the *that* clause conveys the communicative function *showing controversy within the field*, but the phrase in the main clause '*it should be noted that*' may have a different communicative function. This is a limitation when a sentence is regarded as a unit of a single communicative function because a long sentence may have more than one communicative function. However, it is difficult to determine the length that constitutes the minimum unit of a communicative function.

Table 5.8: Average number of formulaic expressions with 3/3 accuracy for all communicative functions (CFs).

CF	Avg. Acc.
Showing limitation or lack of past work	0.00
Comments on the findings	0.00
Showing explanation or definition of terms or notations	0.00
Unexpected outcome	0.00
Describing interesting or surprising results	0.00
Summary of the results	0.00
Comparison of the results	0.00
Showing the limitation of the research	0.00
Showing the characteristics of samples or data	0.00
Showing reasons why a method was adopted or rejected	0.00
Showing brief introduction to the methodology	0.20
Restatement of the aim or method	0.22
Showing background provided by past work	0.25
Showing controversy within the field	0.33
Reference to tables or figures	0.33
Restatement of the results	0.33
Showing what is already done in the past work	0.33
Description of the process	0.43
Description of the results	0.44
Showing the importance of the topic	0.60
Using methods used in past work	0.67
Showing the importance of the research	0.67
Comparison of the results and past work	0.67
Showing methodology used in past work	1.00
Suggestion of hypothesis	1.00
Showing the outline of the paper	1.00
Showing the aim of the paper	1.00
Suggestion of future work	1.00
Explanation for findings	1.00
Showing criteria for selection	1.00
Showing the main problem in the field	1.00

Table 5.8 shows the average number of formulaic expressions with 3/3 accuracy in each communicative function. It can be said that the difficulty in the formulaic expression extraction differs depending on the communicative functions. The communicative functions such as *describing interesting or surprising results* and *unexpected outcome* are often realised by an adverb or adjective, which is difficult to extract using the proposed method.

5.5.3 Error Analyses in Existing Methods

The existing formulaic expression extraction methods have different drawbacks. Table 5.9 lists the number of formulaic expressions extracted with the sentence-level methods after removing infrequent formulaic expressions occurring less than three times in the corpus. Compared to the proposed method, these methods extracted smaller numbers of formulaic expressions because most of these formulaic expressions rarely occur in the corpus. An example of sentence-level extraction is illustrated in Figure 5.7 and 5.8. The existing methods did not remove the non-

Original sentence	In order to avoid over fitting, PA with PCA was chosen for this study.
Frequency (1/50,000)	in order to avoid over fitting pa with * was chosen for this study
Frequency (1/100,000)	in order to avoid over fitting pa with pca was chosen for this study
LDA-based	in order to avoid over fitting * with * chosen for this study
Proposed	in order to avoid * was chosen for this

Figure 5.7: Example of formulaic expression extraction. The second step of the proposed method extracted two different n-grams. The original sentence is cited from An et al. (2018).

formulaic words sufficiently here because the focus is only on a single word, and words such as '*in*' or '*results*' do not always constitute the formulaic expression.

The corpus-level methods are different in this regard. The numbers of extracted formulaic expressions are 23,847 (frequent n-gram) and 2,480,935 (Lattice FS). The frequent n-gram method extracts a smaller number of formulaic expressions because of the frequency thresholds. Further, it achieved a relatively good quality score, which was still lower than that of the proposed method (Table 5.4). The Lattice FS extracts too many formulaic expressions, which can deteriorate the quality of the formulaic expressions.

Table 5.9: Number of formulaic expressions (FEs) that were extracted using the different methods and occurred at least three times in the dataset.

Method	FEs
Frequency-based (1/50,000)	13,722
Frequency-based (1/100,000)	12,840
LDA-based	18,033
Proposed	285,193

5.6 Conclusion

In this chapter, we proposed a new sentence-level formulaic expression extraction method to realise communicative-function-oriented analysis. We manually compared the proposed method to four existing methods, and our manual evaluations showed that the proposed method extracted communicative-function-realising formulaic expressions better than these other methods. We also used the automated evaluation and showed the limitation of the evaluation method. Although formulaic expression extraction has not been discussed in detail thus far in reported literature, we showed the existence of a more robust method than just extracting frequent n-grams, as adopted in the past studies.

Original sentence	There is an urgent need for the development and innovation of monitoring systems, which should be sensitive, quick, specific, inexpensive and convenient for users to monitor the quality of treated wastewater effluents as well as the natural water sources.
Frequency (1/50,000)	there is an * need for the development and * of monitoring systems which should be sensitive * specific * and convenient for * to monitor the quality of treated * as well as the natural water sources
Frequency (1/100,000)	there is an urgent need for the development and * of monitoring systems which should be sensitive quick specific inexpensive and convenient for * to monitor the quality of treated wastewater * as well as the natural water sources
LDA-based	there is an urgent need for the development and innovation of * systems which should be sensitive quick * inexpensive and * for users to * the * of treated * as well as the
Proposed	there is an urgent need for the

Figure 5.8: Example of formulaic expression extraction. The second step of the proposed method extracted the same n-gram. The original sentence is cited from Zhang et al. (2018).

Chapter 6

Construction of Communicative-Function-Labelled Formulaic Expression Database and Retrieval of Formulaic Expressions

6.1 Introduction

The existing writing assistance systems that suggest formulaic expressions or useful phrases use keyword-matching to search for formulaic expression candidates. The limitation of the keyword-matching is that the only formulaic expressions that are the same or lexically similar are retrieved. For example, if the query is '*little attention has been paid to*', one of the results of the keyword-matching method will be '*relatively little attention has been paid to*' (Figure 6.1). Of course, this result will be useful for learning collocations of the query, but the keyword-matching method is not useful to find alternative formulaic expressions.

To suggest diverse formulaic expressions, we propose the communicative-function-based formulaic expression retrieval. The communicative-function-based formulaic expression retrieval uses the communicative function labels in addition to the query. For instance, the communicative function of '*little attention has been paid to*' is *showing the lack of past work*, and thus the formulaic expressions that have the same communicative function are suggested such as '*only few studies have investigated*' (Figure 6.1). The suggested formulaic expressions can be lexically and syntactically different.

For this communicative-function-based formulaic expression retrieval, the communicative-function-labelled formulaic expression database is required. Using the communicative-function-labelled sentence dataset (Chapter 4) and the proposed formulaic expression extraction method (Chapter 5), we constructed the communicative-function-labelled formulaic expression database. Our manual evaluation shows that the 65% of the formulaic expressions in the database were useful and correct.

Additionally, we analyse the extracted formulaic expressions. We show the discipline- and communicative-function-specific formulaic expressions to reconfirm that formulaic expressions vary across disciplines and communicative functions.

We also conducted the communicative-function-based formulaic expression retrieval. To suggest diverse formulaic expressions, we used Jaccard index, which indicated how much vocabularies of two formulaic expressions overlapped with each other. We changed the maximum of Jaccard index to decrease vocabulary overlapping, and evaluated the results of the communicative-function-based and

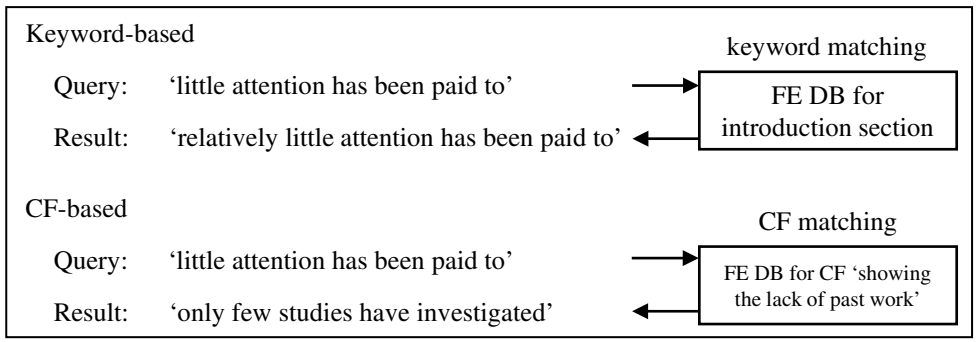

Figure 6.1: Keyword-matching-based and communicative-function-based formulaic expression retrieval.

keyword-matching-based formulaic expression retrieval in that each formulaic expression was assigned a correct communicative function label. The results show that the communicative-function-based retrieval successfully suggest diverse formulaic expressions that has the same function as the query.

Theoretically, all FEs suggested with the proposed method have the same communicative functions as the queries, but in our evaluation, not all formulaic expressions were judged so. We argue that this gap comes from not only the errors made by the formulaic expression extraction and communicative function label assignment, but also the granularity of communicative functions.

The contributions of this chapter are as follows:

- we constructed the communicative-function-labelled formulaic expression database,
- we showed the discipline- and communicative-function-specific formulaic expressions, and
- we proposed the communicative-function-based formulaic expression retrieval.

6.2 Methods

6.2.1 Database Construction

We created the communicative-function-labelled formulaic expression database in the following steps. Step 1: communicative function labels were assigned to each sentence in a corpus and no-CF sentences were removed. Step 2: formulaic expressions were extracted from each sentence. Step 3: Noisy formulaic expressions were filtered out. If an formulaic expression was assigned multiple communicative function labels, only one communicative function was selected by majority voting. If none of the communicative functions took the majority, the formulaic expression was removed. Any communicative-function-labelled formulaic expression occurring less than three times was also removed.

We evaluated the final database from two perspectives: whether a sentence was assigned a correct label and whether an formulaic expression was useful for writing a scientific paper.

The evaluation was conducted on the Amazon Mechanical Turk. A sentence and its communicative function label were shown to evaluators, and an formulaic expression was highlighted in the sentence (see Figure 6.2). The evaluators were

CF: Suggestion of future work
Sentence:
In the future, we plan to explore how to combine more features such as part-of-speech tags into our model.

Figure 6.2: Example of the database evaluation. A formulaic expression is underlined in the sentence, which has been retrieved from Cao et al. (2014).

asked whether the sentence conveyed the communicative function and whether the formulaic expression was useful. Each formulaic expression was annotated by five evaluators, and if it was not evaluated by all as correct or useful, it was regarded as incorrect or useless.

6.2.2 Communicative-Function-Based Formulaic Expression Retrieval

Overview

To compare the proposed formulaic expression retrieval framework to the existing framework, we performed formulaic expression retrieval using the database we created. We conducted the keyword-matching-based retrieval without the communicative function labels and the communicative-function-based retrieval.

In our settings, the similarity scores between the query formulaic expression and other candidate formulaic expressions were used to rank the candidates. This similarity score depended on the similarity of the surface expressions, i.e. the lexical overlap between them. However, to suggest diverse FEs, the similarity should be low, although lower similarity causes incorrect results.

Therefore, we measured how many resulting formulaic expressions had the same communicative functions as the query when the lexical similarity was lower, which meant more diverse. Without communicative function labels, only lexical overlap is used for the retrieval; thus, the lower similarity results in formulaic expressions with different communicative functions. With communicative function labels, the formulaic expressions with the same label are searched; thus, theoretically, all resulting formulaic expressions have the same labels.

Additionally, we utilised vector representations for formulaic expressions generated by SciBERT. Considering the similarity of the vectors as the similarity of communicative functions, we ranked formulaic expressions. This settings indicate the degree to which communicative functions are represented in the vectors generated in an unsupervised manner.

Query Selection

In our experiment, the query must be assigned a correct communicative function label. We randomly picked the queries out of the database we constructed. The formulaic expressions in the database are not always assigned correct labels because the communicative function label assignment was not perfect. Thus, we used coreFEs for selecting queries.

We prepared the queries in the following way. Firstly, we randomly chose coreFEs whose length was three words or longer for each communicative function. Secondly, formulaic expressions satisfying the following conditions were randomly picked out as the queries.

1. containing one of the coreFEs,
2. five-word or longer, and
3. occurring at least ten times in the corpus.

Retrieval and Evaluation

To assure the diversity of resulting formulaic expressions, we used Jaccard index, which is formulated as follows:

$$\mathrm{J}(x, y) = \frac{|\mathrm{V}(x) \cap \mathrm{V}(y)|}{|\mathrm{V}(x) \cup \mathrm{V}(y)|}, \tag{6.1}$$

where x and y are formulaic expressions, and $\mathrm{V}(x)$ is a set of vocabulary of x. We set the maximum value of the Jaccard index to assure the lexical diversity of formulaic expressions. We used three thresholds: 1.0, 0.5, and 0.1. If the Jaccard value is 1.0, it means that the vocabularies of two formulaic expressions are the same.

For the keyword-matching method, formulaic expressions in the dataset whose section label was the same as the query were ranked according to the Jaccard index. Formulaic expressions whose Jaccard index was higher than the threshold were ignored. Finally, top-five formulaic expressions were selected.

The communicative-function-based retrieval consisted of three steps. First, the communicative function label of the query was determined by searching for the same formulaic expression in the database. Second, formulaic expressions with the same communicative function label were ranked according to the Jaccard index. Finally, top-five formulaic expressions were selected.

For the SciBERT-based retrieval, we first removed formulaic expression candidates dissatisfying the Jaccard threshold. Second, we created vector representations of every formulaic expression. The input of SciBERT was a formulaic expression instead of a sentence. The output we used was the [CLS] vector. Subsequently, cosine similarities between the vector of the query and vectors of the candidates were calculated. The candidates were ranked according to the similarity scores and top-five formulaic expressions were selected.

The evaluation was conducted on Amazon Mechanical Turk. Three annotators were recruited for each query. The annotators satisfied all the following qualifications: the number of ever approved tasks was 1,000 or more, the approval rate of the tasks was 0.98 or more, and an annotator lives in the UK or US. The reward was 1.80 USD for each query. They were asked to check if each resulting formulaic expression had the same function as the query formulaic expression. Queries were randomly selected and we prepared 161 queries. We calculated the proportion of the number of correct labels to the total number of the queries.

6.3 Results and Discussion

6.3.1 Communicative-Function-Based Formulaic Expression Database

The communicative-function-labelled formulaic expression database was evaluated by sampling 200 formulaic expressions. The results are shown in Table 6.1.

The incorrect sentence-communicative function pairs were obtained because the classifier made errors and some sentences were not a complete sentence. An example of an incomplete sentence is '*of three independent experiments.*'; this was

Table 6.1: Results of the evaluation of the constructed communicative-function-labelled formulaic expression (FE) database.

		Sentence		
		Correct	Incorrect	Total
FE	Useful	130	12	142
	Useless	34	24	58
	Total	164	36	200

produced because of the error of sentence splitting. Examples of useful formulaic expressions are '*plays a crucial role in*' (communicative function: *showing the importance of the topic*) and '*no significant differences were detected in*' (communicative function: *description of the results*), while '*et al demonstrated that*' (communicative function: *showing background provided by past work*) and '*is to use a*' (communicative function: *showing brief introduction to the methodology*) were judged useless. The statistics of the database are shown in Table 6.2.

To show *general* formulaic expressions, which occurred in all the four corpora, we calculated average rank of each formulaic expression. Table 6.3 lists the top-10 general formulaic expressions ranked according to the average rank. It can be seen that several communicative functions have less than 10 formulaic expressions because no more general formulaic expressions were found. This implies that the number of the general formulaic expressions is smaller than the specific formulaic expressions. In this study we used only four disciplinary corpora, but if more corpora is applied, the number will be decreased.

On the other hand, to show discipline-specific formulaic expressions, we calculated average of odds ratio for each communicative function of each discipline. The odds ratio is formulated as follows:

$$\mathrm{Spec}(f, i) = \frac{1}{n-1} \sum_{j \neq i} \frac{\frac{p_i(f)}{1-p_i(f)}}{\frac{p_j(f)}{1-p_j(f)}}, \tag{6.2}$$

where i, j is a discipline, f is an formulaic expression, n is the number of corpora, and $p_x(f)$ is a probability of f in the discipline x. Table 6.4 illustrates the top-10 highest odds ratio formulaic expressions in each communicative function in each section in each corpus. In *description of the process* in the methods, formulaic expressions whose subject '*we*' appear in the CL corpus, which implies that the research community of computational linguistics prefers to use active voice. In the same communicative function, '*was carried out in accordance with the*' and '*were approved by the*' occur in the Psy corpus. These formulaic expressions are used to indicate that the research conforms to ethical criteria, which is important in the psychological community. Syntactical and lexical difference show that the conventions of how to write papers and conduct research, and formulaic expressions will be useful to fit the style into acceptable expression.

The differences between disciplines are relative, and these results might change if another corpus of a different discipline is added; however, preference for formulaic expressions still exists across disciplines. This reinforces the previous claim that formulaic expressions are discipline-specific (Durrant, 2017; Hyland, 2008; Hyland & Tse, 2007; Jalilifar et al., 2016).

Table 6.2: Number of formulaic expressions in communicative-function-labelled formulaic expression database.

Corpus	Section	CF	FEs
CL	introduction	Showing brief introduction to the methodology	9,153
		Showing controversy within the field	37
		Showing explanation or definition of terms or notations	168
		Showing limitation or lack of past work	346
		Showing the aim of the paper	458
		Showing the importance of the research	351
		Showing the importance of the topic	596
		Showing the limitation of the research	18
		Showing the main problem in the field	167
		Showing the outline of the paper	631
		Showing what is already done in the past work	601
	methods	Description of the process	3,782
		Showing criteria for selection	3,892
		Showing methodology used in past work	449
		Showing reasons why a method was adopted or rejected	799
		Showing the characteristics of samples or data	254
		Using methods used in past work	1,669
	results	Comparison of the results	108
		Describing interesting or surprising results	2,282
		Description of the results	3,292
		Reference to tables or figures	2,211
		Restatement of the aim or method	10,551
		Summary of the results	102
	discussion	Comments on the findings	24
		Comparison of the results and past work	29
		Explanation for findings	159
		Implications of the findings	65
		Restatement of the results	393
		Showing background provided by past work	966
		Suggestion of future work	876
		Suggestion of hypothesis	235
		Unexpected outcome	1,276
Chem	introduction	Showing brief introduction to the methodology	1,570
		Showing controversy within the field	66
		Showing explanation or definition of terms or notations	699

(Continued)

Corpus	Section	CF	FEs
		Showing limitation or lack of past work	642
		Showing the aim of the paper	453
		Showing the importance of the research	240
		Showing the importance of the topic	6,053
		Showing the limitation of the research	26
		Showing the main problem in the field	271
		Showing what is already done in the past work	2,147
	methods	Description of the process	13,203
		Showing criteria for selection	316
		Showing methodology used in past work	233
		Showing reasons why a method was adopted or rejected	1,195
		Showing the characteristics of samples or data	368
		Using methods used in past work	1,014
	results	Comparison of the results	19
		Describing interesting or surprising results	129
		Description of the results	18,007
		Reference to tables or figures	5,312
		Restatement of the aim or method	4,420
		Summary of the results	425
	discussion	Comments on the findings	296
		Comparison of the results and past work	709
		Explanation for findings	309
		Implications of the findings	27
		Restatement of the results	3,435
		Showing background provided by past work	894
		Suggestion of future work	698
		Suggestion of hypothesis	321
		Unexpected outcome	17
Onc	introduction	Showing brief introduction to the methodology	1,898
		Showing controversy within the field	97
		Showing explanation or definition of terms or notations	38
		Showing limitation or lack of past work	1,329
		Showing the aim of the paper	186
		Showing the importance of the research	709
		Showing the importance of the topic	12,090
		Showing the main problem in the field	387
		Showing what is already done in the past work	1,535
	methods	Description of the process	25,230
		Showing criteria for selection	1,190

(Continued)

Corpus	Section	CF	FEs
		Showing methodology used in past work	319
		Showing reasons why a method was adopted or rejected	1,574
		Showing the characteristics of samples or data	2,122
		Using methods used in past work	3,093
	results	Comparison of the results	18
		Describing interesting or surprising results	577
		Description of the results	24,217
		Reference to tables or figures	1,021
		Restatement of the aim or method	10,706
		Summary of the results	1,036
	discussion	Comments on the findings	117
		Comparison of the results and past work	1,272
		Explanation for findings	2,488
		Implications of the findings	127
		Restatement of the results	16,275
		Showing background provided by past work	5,993
		Suggestion of future work	2,075
		Suggestion of hypothesis	1,959
		Unexpected outcome	112
Psy	introduction	Showing brief introduction to the methodology	1,525
		Showing controversy within the field	77
		Showing explanation or definition of terms or notations	215
		Showing limitation or lack of past work	1,765
		Showing the aim of the paper	280
		Showing the importance of the research	937
		Showing the importance of the topic	2,209
		Showing the limitation of the research	14
		Showing the main problem in the field	317
		Showing the outline of the paper	211
		Showing what is already done in the past work	8,291
	methods	Description of the process	12,808
		Showing criteria for selection	245
		Showing methodology used in past work	474
		Showing reasons why a method was adopted or rejected	1,067
		Showing the characteristics of samples or data	2,013
		Using methods used in past work	434
	results	Comparison of the results	71

(Continued)

Corpus	Section	CF	FEs
		Describing interesting or surprising results	939
		Description of the results	6,588
		Reference to tables or figures	675
		Restatement of the aim or method	2,511
		Summary of the results	243
	discussion	Comments on the findings	1,459
		Comparison of the results and past work	942
		Explanation for findings	2,943
		Implications of the findings	451
		Restatement of the results	2,018
		Showing background provided by past work	1,856
		Suggestion of future work	1,312
		Suggestion of hypothesis	904
		Unexpected outcome	145

Table 6.3: List of top-10 general formulaic expressions (FEs) for each communicative function (CF), which occurs in all the corpus. These are ranked according to the average rank in the corpus.

Section	CF	FE
introduction	Showing the importance of the topic	is an important
		plays an important role in
		is a key
		play an important role in
		is one of the most important
		plays an important role in the
		is important for
		is essential for
		plays a key role in
		plays a crucial role in
	Showing brief introduction to the methodology	in addition we
		by using the
		by using a
		we used the
		in the current
		to this end we
		we were able to
		was used to
		we used a
		we developed a
	Showing what is already done in the past work	have shown that
		it has been shown that
		has shown that
		have shown that the
		it is well known that

(Continued)

Section	CF	FE
		have demonstrated that
		it has been demonstrated that
		it has been suggested that
		it was shown that
		have found that
	Showing the aim of the paper	in this paper we
		the aim of this
		in this paper we present
		in this paper we use
		the aim of the present
		here we present
		finally we discuss
		the purpose of this paper is to
	Showing explanation or definition of terms or notations	is defined as the
		is defined as
		refers to the
		are referred to as
		is referred to as
		refers to a
		is often referred to as
		is referred to as a
		is also referred to as
		will be referred to as
	Showing the importance of the research	this is the first
		for the first time
		is the first to
		was the first to
		were the first to
		for the first time in
		for the first time the
		we are the first to
		for the first time we
		this is the first time that
	Showing limitation or lack of past work	has not been
		there is no
		have not been
		has not yet been
		there are no
		however there is no
		there are few
		little is known about the
		there has been no
		have not yet been
	Showing the main problem in the field	there is a need to
		there is a need for
		is the lack of
		is a serious
		therefore there is a need to
		is a challenging

(Continued)

Section	CF	FE
		there is a clear need for
		there is an urgent need for
		there is a need for a
methods	Showing reasons why a method was adopted or rejected	is used to
		can be used to
		was applied to
		was also used to
		was designed to
		was used to identify
		was used in order to
		is designed to
		was used to provide
		can be applied to
	Description of the process	were added to the
		was carried out using
		was performed using
		was carried out by
		were allowed to
		was performed using the
		were determined by
		at the same
		were collected from the
		and then the
	Using methods used in past work	as described in
		as described by
		as described below
		as described in the previous
		as described earlier
		we followed the
		described in the previous
		as in the previous
	Showing criteria for selection	were selected from the
		were selected based on
		were chosen based on
	Showing methodology used in past work	is a widely used
		have been shown to
		has been widely used in
		is widely used in
		is a commonly used
		have been shown to be
		is a common
		have been used to
		have been reported
		is one of the most
	Showing the characteristics of samples or data	were excluded from the
		was divided into
		were included in the

(Continued)

Section	CF	FE
		were divided into
		were not included in the
		participated in the
		included in this
		were randomly divided into
		divided into two
		is divided into two
results	Reference to tables or figures	are shown in
		as shown in
		are presented in
		is shown in
		are summarized in
		are reported in
		are listed in
		are given in
		is presented in
		can be found in
	Restatement of the aim or method	in order to
		was used to
		we used the
		were used to
		we used a
		were used as
		we performed a
		to test the
		to determine the
		was used for
	Description of the results	compared to the
		showed that the
		we found that
		we found that the
		none of the
		was found to be
		were found to be
		most of the
		there was a
		there was no
	Summary of the results	this suggests that
		suggest that the
		this suggests that the
		this indicates that
		this indicates that the
		this shows that
		we conclude that
		this shows that the
		this suggests that a
		this confirms that
	Describing interesting or surprising results	it is not surprising that
		it is interesting that
		it is remarkable that

(Continued)

Section	CF	FE
		more importantly the
discussion	Suggestion of hypothesis	suggest that the
		this suggests that
		indicate that the
		this suggests that the
		suggests that the
		this indicates that
		we conclude that
		this indicates that the
		we conclude that the
		we suggest that the
	Restatement of the results	showed that the
		was found to be
		revealed that the
		were found to be
		was found to
		it was found that
		was shown to be
		it is interesting to note that
		it is important to note that
		we observed a
	Comparison of the results and past work	this is in
		are in line with
		is similar to the
		are in line with the
		is in line with the
		is in line with
		are consistent with the
		this is in contrast to the
		in line with the
		this is similar to the
	Showing background provided by past work	it is well known that
		it is known that
		have focused on the
		has not been
		it has been
		have focused on
		is known to
		in the previous
		are known to be
		it is well-known that
	Explanation for findings	may be due to the
		can be explained by the
		this may be due to the
		could be due to
		may be due to
		might be due to the
		can be explained by
		could be attributed to the

(Continued)

Section	CF	FE
		could be related to the
		can be attributed to
	Suggestion of future work	is needed to
		needs to be
		need to be
		are needed to
		we are currently
		remains to be
		will be needed to
		in the future
		further work is needed to
		needs to be further
	Unexpected outcome	as expected the
		it is not surprising that the
		it is not surprising that

Table 6.4: List of top-10 formulaic expressions (FEs) specific to each communicative function (CF), section, and discipline. These are ranked according to the odds ratio across the corpora.

Section	CF	CL	Chem	Onc	Psy
introduction	Showing the importance of the research	to the best of our	was the first	it is important to	allows us to
		we aim to	is the first	for the first time that	this allowed us to
		this is the first work to	this is the first report on the	we demonstrate	allowed us to
		in this paper we will	this is the first report on	for the first time that	this would suggest that
		in this work we aim to	it is expected that the	we aimed to	is expected to
		this is the first work that	may be a promising	therefore it is important to	should be able to
		we would like to thank	in the first step	are required to	would suggest that
		would like to thank	was one of the first	we show for the first time that	it should be possible to
		is to build a	the first step in the	in the present *	should be more
		we will discuss the	for the first time by	we aimed to	would be the
	Showing limitation or lack of past work	it is difficult to	it is known that	provide new insights into the	to the best of our
		it is hard to	it is known that the	are needed to	can not be
		it is not clear how to	has been extensively studied	however the role of	is not a
		is it possible to	have not been reported	has not been fully elucidated	may not be
		there has been little work on	are known for their	remains to be elucidated	is not limited to
		it is not trivial to	has been studied	remain to be elucidated	is not an
		are not suitable for	has been paid to the	remains to be determined	this is not to say that
		is not able to	has been paid to	have not been fully elucidated	are not necessarily
		there is a large	however there are only a few	however the precise	none of the
		it is not	has been extensively	have focused on the	has been paid to
	Showing the importance of the topic	it is important to	the use of	has been shown to	it is important to
		contributed equally to this work	as well as	we found that	it is important to note that
		is useful for	can be used as	have been shown to	it is assumed that
		is important for many	due to the	was shown to	the importance of
		it is crucial to	it is important to	has been shown to be	it is not surprising that
		there has been a growing interest in	there has been an	has been reported to	it is assumed that the
		is crucial to	due to their	as well as	is the ability to

(Continued)

Section	CF	CL	Chem	Onc	Psy
		can be useful for	belonging to the	in addition the	there is a growing
		is useful for many	it is necessary to	was found to	it is necessary to
		is more important than	in the last few	as well as the	therefore it is important to
	Showing controversy within the field	it is important to note that	by the fact that	this has led to the	have questioned the
		it is not surprising that	it should be pointed out that the	has been paid to	was introduced by
		this is in contrast to	it is not surprising that	has been focused on	was inspired by the
		it should be noted that	therefore it is not surprising that	has been limited by the	has been challenged by
		are those of the * and do not necessarily reflect the	however it should be noted that	has been focused on the	have been raised
		are not necessarily endorsed by the	the need for new	has been paid to the	this raises the
		this is in contrast with	it is worth mentioning that the	has been challenging	was inspired by
		it should be noted that the	it is not surprising that the	has been hampered by the	it has been debated whether
		are those of the	has been a hot topic	this has led to	was also supported by
		this is especially true for	has prompted the	has been controversial	there is an ongoing debate regarding the
	Showing what is already done in the past work	are widely used in	have been developed for the	we previously reported that	cite-
		in the past	have been reported	we previously demonstrated that	et al cite-
		have been used for	have been reported to possess	we and others have shown that	eg cite-
		have been used to	et al reported that the	it is believed that	as well as
		previous work has focused on	showed that the	recently it has been reported that	according to the
		have been proposed to address the	have been reported to	it is now clear that	such as the
		recent work on	has been reported to have	it is known that	for example the
		have been successfully applied to	have been reported to exhibit	et al reported that the	for example cite- found that
		recent work has focused on	it was reported that the	it should be noted that	are more likely to
		are common in	has been shown to	it is important to note that	on the other hand
	Showing explanation or definition of terms or notations	we refer to this	have been used as	is defined as a	refer to the
		we call this	have been used to	are defined as * longer than 200	to the ability to
		we use the term	has been used in	to describe the	refers to the ability to
		is defined as follows	has been used as a	are generally defined as	we use the term
		to denote the	have been used in	they are referred to as	has been referred to as
		we refer to	has been used to	has been termed	we refer to
		is said to be	have been used for	are defined as * more than 200	refers to an
		we refer to such	has been used for	are defined as	it refers to the
		we will refer to the	are used in	has been referred to as the	we will refer to this
		we denote by	have been used for the	has been referred to as	we will use the term
	Showing the main problem in the field	one of the main	are urgently needed	are urgently needed	there is a lack of
		is that it	one of the main	therefore it is	it is difficult to
		there are two major	therefore it is necessary to	is urgently needed	need to be
		one of the major	is highly desirable	remains a major	the need for
		is that they are	is one of the most serious	it is necessary to	we need to
		is that they	is urgently needed	therefore there is an urgent need to	makes it difficult to
		with this approach is that	therefore there is an urgent need to develop	therefore it is necessary to identify	make it difficult to

(Continued)

Section	CF	CL	Chem	Onc	Psy
		is one of the main	thus it is necessary to develop	thus there is an	need to be able to
		a key challenge in	therefore it is of great	thus it is	needs to be
		is a very challenging	is still needed	has become a major	making it difficult to
	Showing the aim of the paper	in this paper we propose a	the aim of this work was to	the purpose of this	the aim of the current
		in this paper we propose a novel	herein we report the	of the present	the purpose of this
		in this paper we address the	herein we describe the	the aim of our	the purpose of the present
		in this paper we focus on	the purpose of this	in this work we	the aim of the present
		in this paper we present a	the aim of the present work was to	the aims of this	in this paper
		in this paper we present an	of the present	therefore the aim of this	aims to explore the
		in this paper we propose	we describe the	we describe the	the second aim was to
		this paper describes a	focuses on the	was to determine whether	in the current paper we
		we present a	therefore the aim of this	the aim of this	the aim of the
		in this paper we propose an	in the present work we	we discuss the	the main aim of the present
	Showing brief introduction to the methodology	for example the	in our previous	we hypothesized that	were presented with
		as well as	in order to	we demonstrated that	we aimed to
		in terms of	et al developed a	we examined the	were asked to
		show that our	et al studied the	we explored the	to examine the
		such as the	were characterized by	have been used to	cite- used a
		the number of	et al reported the	therefore in this	we examined the
		on the other hand	et al used	here we demonstrate that	were required to
		according to the	prompted us to	to explore the	therefore the present
		are added by the	were determined by	we demonstrated that the	were presented with a
		show that the	led to the	in our previous	we examined whether the
	Showing the outline of the paper	are presented in	is shown in		
		of this paper are as follows	were as follows		
		we conclude in	we describe the		
		of this paper are	the first is the		
		finally we conclude in	in the following we will		
		for future work	is illustrated in		
		the related work	the first is		
		related work in	in what follows we		
		are summarized as follows	in the following we		
		4 presents the	this is followed by an		
	Showing the limitation of the research	is not a trivial	is referred to	it is beyond the	
		is still an	is referred to the	of this paper	
		is not an easy	can be found elsewhere	is beyond the scope of this paper	
		is not trivial	are mainly focused on	which is the focus of the present	
		is still an open	is provided in	is the focus of the current	
		is an area of	are not included in this	of the current paper	
		has been the focus of	only focus on the	this is the focus of the present	
		has been the topic of	we focus here on the	is the focus of the present	
		is out of the	which are the focus of this	not the main focus of the	
		has been the focus of much	are discussed below	is the topic of the present	
methods	Showing reasons why a method was adopted or rejected	is used for	was used as a	was defined as the	was used to
		are used to	was used for	were selected for	was set to
		are used for	was used to	was defined as the time from	we decided to
		is used as the	was used as the	was performed to identify	we chose to

(Continued)

Section	CF	CL	Chem	Onc	Psy
		is used as a	were used for	was considered as	was used to assess
		is that it can	was used as	was considered as a	it is possible to
		are used in	was used for the	were used to identify	it is a
		was used for	were used to	were used to determine the	allowed us to
		are used in the	were used as	was conducted to	was used to analyze the
		are used as	were used for the	is defined as the	was found to be
	Showing criteria for selection	for example the	were approved by the	p 005 was considered	was defined as the
		is the number of	was approved by the	005 were considered	was defined as
		is the set of	were as follows	less than 005 were considered	were defined as
		is a set of	was defined as the lowest	005 was considered	is defined as
		can be found in	were selected for	p005 was considered	was defined as a
		note that the	was selected as the	were as follows 1	were selected for
		1 is the	was defined as the	a p value 005 was considered	is defined as the
		be the set of	was defined as the amount of	of p 005 were considered	was defined by the
		the set of	and approved by the	p 005 was considered to be	were defined as the
		for example in	were selected as the	when p 005	was defined as an
	Description of the process	we compute the	were performed in	was used to	was approved by the
		we need to	was purified by	was used for	was carried out in accordance with the
		this allows us to	m h found	were used for	in accordance with the
		it is possible to	were conducted in	were as follows	were approved by the
		in order to	was washed with	at 4 c	in the present
		we calculate the	c n and	were used to	were presented on a
		we would like to	70 ev mz	supplemented with 10	gave written informed * in accordance with the
		we train a	was dried over	were stained with	was conducted in accordance with the
		we create a	was extracted with	were washed with	the order of
		it is necessary to	were dissolved in	was used as a	were presented in a
	Using methods used in past work	based on the	according to the	according to the	is shown in
		we use a	using the following	was approved by the	can be found in
		is based on the	according to the following	were approved by the	was based on the
		is shown in	was calculated using the following	as previously described	as shown in
		is based on	is in accordance with that reported in	as described previously	according to the
		is given by	was prepared from	was performed as previously described	is based on the
		we propose a	by the following	as previously described cite-	is presented in
		is as follows	11 40 ml was reacted according to	was performed as described previously	was used in this
		is defined as	was performed according to the	was performed according to the	was developed by
		is defined as follows	as previously described	were kindly provided by	adapted from the
	Showing methodology used in past work	we consider two	it is possible to	is based on the	eg cite-
		have been proposed in the	a number of	is based on	has been used in previous
		is to use	the most common	is defined as	has been shown to have good
		is closely related to the	a wide range of	have been described	cite- is a
		previous work on	is the most	have been described cite-	has been found to be
		rely on the	can also be	has been described	has been used to
		in two ways	there are two	is based on a	have shown that the
		there are many ways to	more and more	is directly proportional to the number of	has been found to

(Continued)

Section	CF	CL	Chem	Onc	Psy
		there are two main	some of these	it is a	has been shown to have
		there are a number of	a series of	are referred to as	has been validated in
	Showing the characteristics of samples or data	in total there are	are listed in	as the mean	took part in the
		we split the	were used in this	are presented as the mean	a total of
		is divided into	were considered as	were presented as mean	were recruited through
		are included in the	005 were considered	were repeated at least three	the majority of
		are more likely to	were listed in	were classified as	were recruited via
		included in the	of p 005 were considered	was repeated three	included in the
		is split into	used in this * are listed in	were repeated three	most of the
		are split into	served as a	was repeated at least three	the majority of the
		there are a total of	were randomly divided into four	were randomly divided into four	half of the
		are divided into	are described in	were performed at least three	at the time of
results	Describing interesting or surprising results	on the other hand	it is interesting to note that	the most common	note that the
		in contrast the	it is interesting to note that the	of note the	a number of
		on the other hand the	it was interesting that	interestingly we found that	the importance of
		this is because the	interestingly in the	interestingly we observed that	the most common
		can not be	it is worth mentioning that	interestingly we found that the	on the other hand
		this is because	it is worth mentioning that the	interestingly we observed a	seemed to be
		this is due to the	it was notable that the	in line with this	as expected the
		it is worth	interesting to note that	interestingly we observed	for example in
		in general the	it is interesting that the	interestingly we found	this is the
		what is the	is the presence of the	similarly in the	it should be noted that
	Description of the results	show that the	was obtained as a	it has been reported that	there was a main
		we observe that the	in the present	as well as	revealed a significant main
		achieves the best	was confirmed by	fig cite- and	revealed a main
		we see that the	in the presence of	the number of	revealed a significant main effect of
		indicates that the	was determined to be	respectively table cite-	there was no main
		we find that the	was identified as	is known to	showed a significant main
		is able to	were confirmed by	we have previously shown that	there was no significant main
		indicate that the	was confirmed by the	has been reported to	showed a main
		we note that the	due to the	suggested that the	revealed a main effect of
		is significantly better than	was obtained as	it has been shown that	there was also a main
	Comparison of the results	we compare our	it could be seen that the	it can be seen that	it can be seen that
		table 3 compares the	cite- compares the	it can be seen that the	it can be seen that the
		we compare the	one can see that the	we can see that the	we can see that
		table 1 compares the	it could be seen that	an example of the * is shown in	we now turn to
		in table 5 we	one can see that	a search of the	we see that the
		table 3 compares our	and this included	a search for	we can see that the
		with previous work	for the 50 different	comparison of the mean of each * multiple comparisons test indicated that	point showed that at time 3 f1 3
		it has been shown that	it was possible to observe that the	revealed that the * was significantly lower in the	we report the
		comparison on the	it can be seen that there is no	indicating there were no substantial	we now turn to the
		our approach with two	it is possible to notice that	an example of a * is shown in	we will focus on the
	Restatement of the aim or method	- 2 -	were characterized by	was confirmed by	were entered as

(Continued)

Section	CF	CL	Chem	Onc	Psy
		we use the	were subjected to	the role of	we predicted that
		we use a	was subjected to	we next examined the	we conducted a 2
		we use the same	were prepared by	were confirmed by	in addition to
		according to the	were prepared by the	and found that	we ran a
		as well as	was determined as	with or without	were conducted for each
		is the number of	were prepared according to the	were subjected to	by subtracting the
		we follow the	was suspended in	was further confirmed by	based on the
		as well as the	were selected for further	we first examined the	were coded as
		we use two	was prepared by	to determine if	and the two
	Summary of the results	this indicates that our	this result indicated that	taken together these	this means that
		this suggests that our	this result indicated that the	taken together our	in sum the
		this shows that our	based on these	indicate that the	this means that the
		this demonstrates that our	it seems that the	strongly suggest that	this indicated that the
		this shows that a	are in agreement with previous	taken together these * suggest that the	this indicated that
		this suggests that for	show that the	show that the	it appears that
		we conclude that our	indicate that the	may contribute to the	are in line with the
		this confirms our	are in accordance with the	taken together these * indicate that the	it shows that the
		in summary we can conclude that	this means that	may contribute to	provide partial support for
		this indicates that when	this suggested that the	all together these	this would suggest that
	Reference to tables or figures	table 3 shows the	is an important	as shown in fig	see table cite-
		table 1 shows the	it has been reported that	cite- shows the	cite- shows the
		table 4 shows the	it is known that	cite- shows that the	cite- presents the
		are shown in table 1	it is well known that	cite- shows that the	cite- shows that the
		table 5 shows the	it should be noted that the	were obtained in	cite- displays the
		we can see that the	shows that the	cite- shows a	cite- shows a
		6 shows the	cite- a shows the	were summarized in	cite- provides the
		results on the	is based on the	were listed in	we present the
		are shown in table 4	it should be noted that	as shown in *	was observed in cite- shows that
		4 shows the	can be attributed to the	cite- showed the	cite- for the
discussion	Restatement of the results	show that our	in summary we have	this is the first	indicated that the
		show that the	involved in the	in the present	also showed that
		show that the proposed	in conclusion the	in addition the	for example cite- found that
		table 3 shows the	show that the	as well as	also found that
		table 4 shows the	we have shown that	et al showed that	were found for
		table 5 shows the	depending on the	to the best of our	also found that the
		table 1 shows the	shows that the	as well as the	were found for the
		show that our proposed	in the present	we showed that	more specifically the
		showed that the proposed	in conclusion we	was reported to	were more likely to
		as can be seen in	responsible for the	et al demonstrated that	was stronger for
	Comparison of the results and past work	is based upon work supported by the	this is the first	also demonstrated that	in contrast the
		is based upon work supported in part by the	et al reported that	similar to the	is supported by the
		is based in part on	was confirmed by	similar to our	according to this
		with previous work	also showed that	in contrast to	is supported by
		is based upon work supported by	confirmed that the	et al also reported that	with the idea that
		in line with previous work	were confirmed by	in contrast to the	by contrast the

(Continued)

Section	CF	CL	Chem	Onc	Psy
		this corresponds to the fact that	than that of	it was also reported that	in contrast to the
		is supported by the	reported that the	in line with this	it is reasonable to
		is supported by the fact that	et al showed that	also reported that	in contrast to
		is confirmed by	than that of the	also found that	in contrast in the
	Explanation for findings	are due to	due to the presence of	therefore it is possible that	is that the
		it may be possible to	has been attributed to	we can not exclude that	it is also possible that
		is due to	attributed to the	we can not exclude the	it is also possible that the
		this is mainly due to the fact that	was attributed to the	it is possible that	it should be noted that the
		we attribute this to the	mainly due to the	we can not rule out	it is possible that
		we attribute this to the fact that	has been attributed to the	may be more	it seems that the
		we believe this is because the	were attributed to the	should be considered	it seems that
		it may be better to	can be attributed to the presence of	can not be ruled out	it should be noted that
		can be handled by	is caused by	may not be	it is possible that the
		this can be done by	could be attributed to its	therefore it is likely that	not be ruled out
	Suggestion of future work	in future work we	it is likely that	however the role of	it would be interesting to
		we would like to	it is necessary to	is still unknown	we suggest that future
		as future work we	and will be reported in due	are required to	should examine the
		we would also like to	it is likely that the	remains to be determined	it would be
		in future work we will	it is possible that	remains to be elucidated	it would also be interesting to
		for future work we	is required to	there are some	we recommend that future
		we are also	therefore it is necessary to	are needed to confirm our	it would be important to
		in future work we would like to	are currently in	remains largely unknown	should address this
		we will also	it is expected that	there were some	it is necessary to
		there are a number of	will be useful for	however the underlying	it would be useful to
	Comments on the findings	this is an encouraging	it is clear that	are currently in	we were able to
		we are encouraged by the	was successfully applied to	was well tolerated	were able to
		the most successful	it is clear that the	is currently in	can be used to
		is effective for	it is suggested that	has shown promising	we have shown that
		are very promising	it is believed that	have shown promising results in	it is possible to
		it is our hope that	was successfully applied to the	is a promising strategy for	we were not able to
		is promising as it	it was suggested that	we successfully established a	could be used to
		are very encouraging	it was suggested that the	have shown promising	allowed us to
		is a promising	it is believed that the	has emerged as a promising	in this way
		seems to be promising to	was achieved by	represents a promising	we believe that
	Suggestion of hypothesis	can be used to	suggesting that the	in conclusion our	this is the first
		we can see that	suggested that the	we show that	in sum the present
		we can see that the	may be a potential	here we show that	taken together the
		can be used for	which indicates that the	we demonstrate that	is the first to
		this allows us to	is a potential	in summary our	this is the first study to
		it is clear that	may be a promising	suggested that the	in sum the
		can be used as a	taken together the	show that the	in summary the
		indicates that the	may be involved in	here we demonstrate that	highlight the importance of
		it is clear that the	basis for the	we speculate that	this supports the
		can be used as	it may be concluded that	in conclusion we have	it can be
	Implications of the findings	is an important	the possibility of	the possibility of	it is important to
		is useful for	there is a possibility that	the possibility that	contributes to the

(Continued)

Section	CF	CL	Chem	Onc	Psy
		can be applied to other	have the potential to be used as	raising the possibility that	it is important that
		has the potential to	this is of	the possibility of a	it is therefore possible that
		may be useful for	this could lead to	there is a possibility that	this is an important
		is an important step towards	the need for further	may have significant	adds to the
		may be useful in	have the potential to	suggest the possibility that	it can be assumed that
		will be useful for	may find applications in	highlight the need to	it is also important to
		it is important to	is of crucial importance	support the possibility that	highlights the importance of
		is also useful for	this does not exclude the	raise the possibility that	it is important to consider the
	Showing background provided by past work	in this paper we	as shown in	et al reported that	most of the
		we proposed a	was reported to	have shown that	however in the
		in this paper we presented a	to the best of our	has been shown to	as described in the
		we presented a	plays an important role in	we have shown that	see cite- for a
		in this paper we proposed a	plays an important role in the	it has been reported that	is known to be
		in this paper we have	it has been reported that	have demonstrated that	have argued that
		in this paper we propose a	have shown that	has been reported to	have shown that * et al cite-
		in this paper we have presented a	is known to be	we have demonstrated that	it has been argued that
		we propose a	was reported to be	it has been shown that	have shown that
		in this paper we present a	this is the first report on the	can lead to	cite- suggested that
	Unexpected outcome	for example the	on the contrary	surprisingly we found that	this was not the
		we have shown that	more importantly the	would be expected to	it was expected that
		the number of	interestingly we found that the	therefore it is not surprising that	as expected we found that
		we found that	was prevented by	as expected we found that	we expected to find
		on the other hand	interestingly we found	would be predicted to	we expected that
		we show that	was observed only in the	it is therefore not surprising that	it is perhaps not surprising that
		we showed that	most importantly the	thus it is not surprising that	is not surprising
		we find that	this is not surprising since the	it is not surprising that	thus it is not surprising that
		we find that the	it is thus not	therefore it is not surprising that the	therefore it is not surprising that
		on the other hand the	as it was expected	it is expected that	is not surprising given the

6.3.2 Formulaic Expression Retrieval

The results of the evaluation are illustrated in Table 6.5. When the Jaccard index threshold is 1.0 or 0.5, the proportion of the correct labels is not much different, but when it is 0.1 (most diverse), the communicative-function-based method is better than the keyword-matching-based method. Overall, the communicative-function-based method works better even though Jaccard index is small.

Theoretically, the score of the communicative-function-based retrieval must be 1.00 because all resulting formulaic expressions had the same communicative function labels as the query. This gap is attributed to two problems: the quality of the database and the communicative function set. The assignment of the communicative function labels and the formulaic expression extraction are not perfect; thus, some formulaic expressions are not assigned correct labels or are not extracted correctly.

Another problem lies in the set of communicative functions we used. The performance differs across communicative functions. Table 6.6 lists the top-five communicative functions whose proportions of the correct labels are high, and Table 6.7 lists the worst five communicative functions. There is a large gap be-

Table 6.5: Results of evaluation for formulaic expression retrieval. Lower Jaccard index means more diverse formulaic expressions.

Jaccard index	Method	Correct label ratio
1.0	keyword-matching-based	0.77
1.0	communicative-function-based	0.76
1.0	SciBERT-based	0.76
0.5	keyword-matching-based	0.53
0.5	communicative-function-based	0.59
0.5	SciBERT-based	0.63
0.1	keyword-matching-based	0.35
0.1	communicative-function-based	0.40
0.1	SciBERT-based	0.43

tween the best one and worst one. Table 6.8 shows the results of the retrieval with the query '*very little is known about*' in *showing limitation or lack of past work*. When the Jaccard limitation was 1.0 or 0.5, the results were almost the same, and formulaic expressions were very similar in that they used many of the same words. However, in the case where the Jaccard limitation was 0.1, formulaic expressions suggested by the keyword-matching-based method changed too much to represent the same communicative function. The formulaic expressions retrieved by the communicative-function-based method still conveyed the same formulaic expressions though formulaic expressions were diversified. The diversity of the formulaic expressions was not only lexical but also syntactic; e.g. '*however there is a lack of*' was syntactically different from the query formulaic expression.

On the other hand, there are some cases where the diversity does not work effectively. Table 6.9 shows the results of the query '*there were no significant differences in*' in *description of the results*. The formulaic expressions extracted by the communicative-function-based method with $J \leq 0.1$ describe the results in a sense, but '*were found to contain the*' seems quite different from the query. The query formulaic expression is used when comparing some numbers as a result of some experiments, but the resulting formulaic expression is used to explain some ingredients. This difference seemed large to the annotators. Indeed, considering the situation where a user is looking for alternative formulaic expressions to the query, '*were found to contain the*' is not useful. Probably, formulaic expressions that can be used to show the statistical significance are more helpful.

This problem is reduced to the granularity of the communicative function set. In other words, the communicative function *description of the results* is too broad. Communicative functions regarding methodology and results of research should be finer-grained, while communicative functions such as suggestion of future work and showing the limitation seem appropriate.

The results also imply that the SciBERT-based vector representation does not improve the performance without communicative function labels. In other words, without any further labelled dataset or tuning, the vectors do not represent communicative functions sufficiently. Further investigation is needed into communicative-function-aware formulaic expression representations.

6.4 Conclusion

In this chapter, we constructed the communicative-function-labelled formulaic expression database and the evaluation showed that the 65% of the formu-

Table 6.6: Top-five highly scored communicative functions (CFs).

CF	Correct label ratio
suggestion of future work	0.60
showing the main problem in the field	0.58
showing the aim of the paper	0.50
implications of the findings	0.50
showing the importance of the topic	0.47

Table 6.7: Five worst communicative functions (CFs) in retrieval.

CF	Correct label ratio
showing the characteristics of samples or data	0.33
comments on the findings	0.33
showing the importance of the research	0.33
restatement of the results	0.36
comparison of the results and past work	0.36

laic expressions in the DB was correct and useful. The DB is available at `https://iwa2ki.com/FE/`. We also reconfirmed that formulaic expressions were discipline-specific by showing formulaic expressions specific to each communicative function, section, and discipline, ranked by the average odds ratio. We presented the communicative-function-based formulaic expression retrieval and the evaluation showed that the proposed method can suggest diverse formulaic expressions better than the keyword-matching-based method.

Table 6.8: The proposed method suggests formulaic expressions (FEs) whose funtions are the same as the query communicative function (CF).

Query: very little is known about		
CF: showing limitation or lack of past work		
Method	Jaccard	FE
CF	0.1	however it is still unknown whether
	0.1	however there is a lack of
	0.1	however there is no direct evidence
	0.1	so far there is no research
	0.1	limited information is available on the
Keyword	0.1	is a direct target gene of
	0.1	is a member of the family of
	0.1	is a tumor suppressor gene that
	0.1	is one of the most frequently
	0.1	is a member of the tripartite
CF	0.5	little is known about whether and how
	0.5	little is known about the impact of
	0.5	little is known about the effects of
	0.5	little is known about the contribution of
	0.5	however to date little is known about
Keyword	0.5	little is known about whether and how
	0.5	little is known about the impact of
	0.5	little is known about the effects of
	0.5	little is known about the contribution of
	0.5	however to date little is known about
CF	1.0	very little is known about
	1.0	very little is known about the
	1.0	little is known about
	1.0	so far very little is known about
	1.0	little is known about the
Keyword	1.0	very little is known about
	1.0	very little is known about the
	1.0	little is known about
	1.0	so far very little is known about
	1.0	little is known about the

Table 6.9: The proposed method fails to suggest formulaic expressions (FEs) with the same communicative function (CF) as the query when lower Jaccard index is set.

Query: there were no significant differences in		
CF: description of the results		
Method	Jaccard	FE
CF	0.1	are in excellent agreement with
	0.1	showed the highest activity in
	0.1	caused a decrease in the
	0.1	were found to contain the
	0.1	were similar to those of
Keyword	0.1	are shown in cite- b
	0.1	are shown in cite- a
	0.1	are reported in the cite-
	0.1	as shown in cite- a
	0.1	it is known that in
CF	0.5	there is no significant difference in
	0.5	significant differences in
	0.5	there was no significant change in
	0.5	no significant differences were observed between
	0.5	there were significant differences among the
Keyword	0.5	there is no significant difference in
	0.5	significant differences in
	0.5	there was no significant change in
	0.5	no significant differences were observed between
	0.5	there were significant differences among the
CF	1.0	there were no significant differences in
	1.0	there were no statistically significant differences in
	1.0	there were no significant differences in the
	1.0	however there were no significant differences in
	1.0	there were no significant differences in leaf
Keyword	1.0	there were no significant differences in
	1.0	there were no statistically significant differences in
	1.0	there were no significant differences in the
	1.0	however there were no significant differences in
	1.0	there were no significant differences in leaf

Chapter 7

Discussion

7.1 Granularity of Communicative Function Set

Considering useful applications including the communicative-function-based formulaic expression retrieval, the granularity of communicative functions is an important issue. Especially, communicative functions related to describing methodology or results, which differ to a great extent depending on research topics, should be divided into minimum purposes of writing a linguistic unit (e.g. a sentence).

The question is what the minimum granularity of purposes in scientific papers is. It is difficult to answer to the question directly. Still, from the viewpoint of academic writing assistance, the number of linguistic units associated with the same communicative function is a reasonable clue. What is important in the communicative-function-based retrieval is that the number of candidate formulaic expressions can be reduced using the communicative function label. Coarse-grained communicative functions mean that the linguistic units belonging to the same communicative function appear many times in a single document. Accordingly, the total number of formulaic expressions extracted from such a communicative function becomes considerable large.

For instance, in the CL corpus, the number of formulaic expressions in *showing the limitation of the research*, in which the communicative-function-based formulaic expression retrieval performed well, was 81, while the number of formulaic expressions in *description of the process*, which resulted in a bad score, was 22,980. It is impossible that more than 20,000 formulaic expressions are all useful for writing about a specific process of research.

However, finer-grained communicative function sets will probably be discipline-specific. The communicative functions in methods section in management research articles investigated by Lim (2006) are listed in Table 7.1. For example, the table contains the communicative function, *describing the location of the sample*, which is common in the management research, but rarely appears in computational linguistics papers.

If a communicative function set should be composed for each discipline, the training dataset for the communicative function label assignment should also be constructed manually for each discipline, which is very costly. Hence, automated construction of a communicative function set should be considered. The proposed formulaic expression extraction method does not depend on communicative function labels; thus, the formulaic expressions could be used to represent communicative functions in documents instead of using a full sentence to create a communicative-function-based vector space, which is virtually the bottom-up approach.

Table 7.1: Part of communicative functions in methods section in management research articles presented by Lim (2006).

Move 1: Describing data collection procedure(s)
Step 1: Describing the sample
(a) Describing the location of the sample
(b) Describing the size of the sample/population
(c) Describing the characteristics of the sample
(d) Describing the sampling technique or criterion

7.2 Unit of Communicative Function

Another problem related to communicative functions is the determination of units representing a communicative function. As discussed in Section 2.2.3, we used a sentence as a unit of a communicative function, following the past work. However, in the formulaic expression extraction, this caused problems.

In early work on communicative function analysis (Swales, 1981), communicative function labels were not assigned to each sentence, but a whole section was split into several communicative functions; thus, each unit might correspond to more than one sentence.

In our observations, there were also several cases where one sentence had multiple communicative functions. Thus, it might be a better approach to consider the communicative function label assignment as a sequence labelling problem, where a whole section is an input, and each word, clause, or sentence is assigned a communicative function label. Hirohata et al. (2008) adopted conditional random fields and regarded one sentence as a unit of a sequence to assign communicative function labels. However, the work was only focused on the abstract of scholarly papers. The whole paper is much longer than the abstracts, and communicative functions are more complex. Moreover, there is no large resource in which communicative function labels are assigned to scientific papers available.

Therefore, the future direction will be dataset annotation; a large, multidisciplinary dataset in which communicative function labels are assigned to a smaller unit than a sentence should be constructed.

Chapter 8

Conclusion

In this thesis, we proposed the framework for the communicative-function-based formulaic expression retrieval, which is able to suggest more diverse formulaic expressions than the existing keyword-matching-based method. The primary aim of this study was to construct the communicative-function-labelled formulaic expression database, in which formulaic expressions were assigned communicative function labels, to realise the formulaic expression retrieval. The construction of the database consisted of two parts: the communicative function label assignment to each sentence in the corpora, and the formulaic expression extraction from the communicative-function-labelled sentence dataset. After constructing the database, we conducted the communicative-function-based formulaic expression retrieval, and showed that the proposed method was better at suggesting diverse alternative formulaic expressions than the keyword-matching-based retrieval.

In Chapter 2, we described the existing writing assistance systems based on keyword-matching formulaic expression retrieval. We also introduced the existing studies regarding formulaic expressions and communicative functions in scholarly articles, and showed that the formulaic expression extraction had not been investigated extensively and little work had been conducted on the communicative function label assignment.

In Chapter 3, we presented how to collect communicative-function-annotated sentences from scientific corpora using the CoreFEs, which were manually created by shortening example expressions in Academic Phrasebank. We also presented the FECFeval dataset. Additionally, we presented the communicative-function-annotated sentence dataset used to train the SciBERT and BERT classifier for the communicative function label assignment.

In Chapter 4, we conducted the communicative function label assignment in a supervised machine-learning manner. We showed that the SciBERT classifier worked well, even though the disciplines of the training data and inference data were different in both pre-training and fine-tuning. We also showed that the maximum value of the softmax layer of the classifier was useful in filtering no-CF sentences

In Chapter 5, we proposed a new formulaic expression extraction method, which utilised the named and scientific entity removal and longest n-gram extraction. We manually and computationally evaluated the proposed and existing formulaic expression extraction methods.

In Chapter 6, using the methods and dataset proposed in Chapter 4 and 5, we constructed the communicative-function-labelled formulaic expression database. We showed the general and communicative-function-specific formulaic expressions. We conducted the formulaic expression retrieval, and compared it to the keyword-matching-based formulaic expression retrieval. The results showed that

the proposed method suggested diverse formulaic expressions whose functions are the same as the query.

In Chapter 7, we discussed what should be done for better performance of the formulaic expression retrieval from the viewpoint of communicative functions. We argued that the communicative function sets should be fine-grained so that the number of formulaic expressions will not be too large, and to do so, the automated communicative function set construction is an urgent task. We also argued that the dataset with finer-grained communicative function labels should be created in order to solve the communicative function unit problem.

Future work should explore the two problems described above. Additionally, this work is focused on the genre of English for Academic Purposes, but formulaic expressions are a common linguistic phenomenon in any other genre. Thus, we hope this work will accelerate research on formulaic expressions and communicative functions in order to make human communication and language learning easier and more efficient.

References

Abekawa, T., & Okumura, M. (2006). Japanese dependency parsing using co-occurrence information and a combination of case elements. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 833–840). doi: 10.3115/1220175.1220280

Ackermann, K., & Chen, Y.-H. (2013). Developing the academic collocation list (ACL) — a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. doi: 10.1016/j.jeap.2013.08.002

Ädel, A. (2014). Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves. *Journal of English for Academic Purposes*, *16*, 68–80. doi: 10.1016/j.jeap.2014.09.001

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. doi: 10.1016/j.esp.2011.08.004

Alex, B., Dubey, A., & Keller, F. (2007). Using foreign inclusion detection to improve parsing performance. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 151–160).

AlHassan, L., & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, *17*, 51–62. doi: 10.1016/j.jeap.2015.02.001

An, M., Zhang, X., & Zhang, X. (2018). Identifying the validity and reliability of a self-report motivation instrument for health-promoting lifestyles among emerging adults. *Frontiers in psychology*, *9*, 1222. doi: 10.3389/fpsyg.2018.01222

Basturkmen, H. (2009). Commenting on results in published research articles and masters dissertations in language teaching. *Journal of English for Academic Purposes*, *8*, 241–251. doi: 10.1016/j.jeap.2009.07.00

Basturkmen, H. (2012). A genre-based investigation of discussion sections of research articles in dentistry and disciplinary variation. *Journal of English for Academic Purposes*, *11*, 134–144. doi: 10.1016/j.jeap.2011.10.004

Batista, F., Mamede, N., & Trancoso, I. (2008). Language dynamics and capitalization using maximum entropy. In *Proceedings of ACL-08: HLT, short papers* (pp. 1–4).

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3615–3620). doi: 10.18653/v1/D19-1371

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, *14*(3), 275–311. doi:

10.1075/ijcl.14.3.08bib

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. doi: 10.1016/j.esp.2006.08.003

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structures.* John Benjamins Publishing.

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

Bing, L., Li, P., Liao, Y., Lam, W., Guo, W., & Passonneau, R. J. (2015). Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (pp. 1587–1597). doi: 10.3115/v1/P15-1153

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Bouma, G., & van Noord, G. (1993). Head-driven parsing for lexicalist grammars: Experimental results. In *Sixth conference of the European chapter of the association for computational linguistics.*

Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G., & Shein, F. (2015). Building a lexicon of formulaic language for language learners. In *Proceedings of the 11th workshop on multiword expressions* (pp. 96–104).

Brooke, J., Šnajder, J., & Baldwin, T. (2017). Unsupervised acquisition of comprehensive multiword lexicons using competition in an n-gram lattice. *Transactions of the Association for Computational Linguistics*, *5*, 455–470. doi: 10.1162/tacl_a_00073

Brooke, J., Tsang, V., Hirst, G., & Shein, F. (2014). Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of the 25th international conference on computational linguistics: Technical papers* (pp. 753–761).

Cao, Z., Li, S., & Ji, H. (2014). Joint learning of Chinese words, terms and keywords. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1774–1778). doi: 10.3115/v1/D14-1186

Chang, J., & Chang, J. (2015). WriteAhead2: Mining lexical grammar patterns for assisted writing. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: Demonstrations* (pp. 106–110). doi: 10.3115/v1/N15-3022

Chang, J., Hsu, H.-L., Boisson, J., Peng, H.-C., Wu, Y.-H., & Chang, J. S. (2015). Learning sentential patterns of various rhetoric moves for assisted academic writing. In *Proceedings of the 29th Pacific Asia conference on language, information and computation: Posters* (pp. 37–45).

Chen, M., Huang, S., Hsieh, H., Kao, T., & Chang, J. S. (2012). FLOW: A first-language-oriented writing assistant system. In *Proceedings of the ACL 2012 system demonstrations* (pp. 157–162).

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30–49. doi: 10125/44213

Chung, G. (2004). Developing a flexible spoken dialog system using simulation. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)* (pp. 63–70). doi: 10.3115/1218955.1218964

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information,

and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3586–3596).

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72–89. doi: 10.1093/applin/amm022

Conneau, A., & Kiela, D. (2018). SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 670–680). doi: 10.18653/v1/D17-1070

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 837–892. doi: 10.1162/COLI_a_00302

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, *12*(1), 33–43. doi: 10.1016/j.jeap.2012.11.002

Cotos, E., Huffman, S., & Link, S. (2015). Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes*, *19*, 52–72. doi: 10.1016/j.jeap.2015.05.004

Cotos, E., Huffman, S., & Link, S. (2017). A move/step model for methods sections: Demonstrating rigour and credibility. *English for Specific Purposes*, *46*, 90–106. doi: 10.1016/j.esp.2017.01.001

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. doi: 10.2307/3587951

Cunningham, K. J. (2017). A phraseological exploration of recent mathematics research articles through key phrase frames. *Journal of English for Academic Purposes*, *25*, 71–83. doi: 10.1016/j.jeap.2016.11.005

Dai, X., Liu, Y., Wang, X., & Liu, B. (2014). WINGS: Writing with intelligent guidance and suggestions. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 25–30). doi: 10.3115/v1/P14-5005

Darabad, A. M. (2016). Move analysis of research article abstracts: A cross-disciplinary study. *International Journal of Linguistics*, *8*(2), 125–140. doi: 10.5296/ijl.v8i2.9379

Dayrell, C., Candido Jr., A., Lima, G., Machado Jr., D., Copestake, A., Feltrim, V., ... Aluisio, S. (2012). Rhetorical move detection in English abstracts: Multi-label sentence classifiers and their annotated corpora. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 1604–1609).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). doi: 10.18653/v1/N19-1423

Dorrepaal, J. (1993). On the notion of uniqueness. In *Sixth conference of the*

European chapter of the association for computational linguistics.

Dudley-Evans, T., & John, M. J. S. (1998). *Developments in English for specific purposes.* Cambridge University Press.

Dunietz, J., Levin, L., & Carbonell, J. (2013). The effects of lexical resource quality on preference violation detection. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 765–770).

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, *38*(2), 165–193. doi: 10.1093/applin/amv011

Durrant, P., & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, *30*(1), 58–72. doi: 10.1016/j.esp.2010.05.002

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. doi: 10.1002/j.1545-7249.2008.tb00137.x

Esfandiari, R., & Barbary, F. (2017). A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes*, *29*, 21–42. doi: 10.1016/j.jeap.2017.09.002

Fiacco, J., Cotos, E., & Rosé, C. (2019). Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 310–319). doi: 10.1145/3303772.3303808

Friedman, C. (1989). A general computational treatment of the comparative. In *27th annual meeting of the association for computational linguistics* (pp. 161–168). doi: 10.3115/981623.981643

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, *18*(1), 109–136.

Guo, W., Xu, P., Jin, T., Wang, J., Fan, D., Hao, Z., ... Wang, J. (2017). MMP-3 gene polymorphisms are associated with increased risk of osteoarthritis in chinese men. *Oncotarget*, *8*(45), 79491–79497. doi: 10.18632/oncotarget.18493

Halliday, M. A. K., & Matthiessen, C. M. (2014). *Halliday's introduction to functional grammar.* Routledge.

Hashimoto, K., Soonklang, T., & Hirokawa, S. (2016). Feature words of moves in scientific abstracts. In *2016 5th IIAI international congress on advanced applied informatics* (pp. 144–149).

He, X., Yang, M., Gao, J., Nguyen, P., & Moore, R. (2008). Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 98–107).

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *The fifth international conference on learning representations.*

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., & Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2744–2751). doi: 10.18653/v1/2020.acl-main.244

Hill, F., Cho, K., & Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 conference*

of the north American chapter of the association for computational linguistics: Human language technologies (pp. 1367–1377). doi: 10.18653/v1/N16-1162

Hirohata, K., Okazaki, N., Ananiadou, S., & Ishizuka, M. (2008). Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the third international joint conference on natural language processing: Volume-I* (pp. 381–388).

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21. doi: 10.1016/j.esp.2007.06.001

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary?". *TESOL Quarterly*, *41*(2), 235–253. doi: 10.1002/j.1545-7249.2007.tb00058.x

Iwatsuki, K., & Aizawa, A. (2018). Using formulaic expressions in writing assistance systems. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2678–2689).

Iwatsuki, K., Boudin, F., & Aizawa, A. (2020a). An evaluation dataset for identifying communicative functions of sentences in English scholarly papers. In *Proceedings of the 12th language resources and evaluation conference* (pp. 1705–1713).

Iwatsuki, K., Boudin, F., & Aizawa, A. (2020b). *Extraction and evaluation of formulaic expressions used in scholarly papers.* arXiv:2006.10334.

Iwatsuki, K., Sagara, T., Hara, T., & Aizawa, A. (2017). Detecting inline mathematical expressions in scientific documents. In *Proceedings of the 2017 ACM symposium on document engineering* (pp. 141–144). doi: 10.1145/3103010.3121041

Jalali, Z. S., & Moini, M. R. (2014). Structure of lexical bundles in introduction section of medical research articles. *Procedia - Social and Behavioral Sciences*, *98*, 719–726. doi: 10.1016/j.sbspro.2014.03.473

Jalilifar, A., Ghoreishi, S. M., & Roodband, S. A. E. (2016). Developing an inventory of core lexical bundles in English research articles: a cross-disciplinary corpus-based study. *Journal of World Languages*, *3*(3), 184–203. doi: 10.1080/21698252.2017.1301279

Jeong, S., Nam, S., & Park, H.-Y. (2014). An ontology-based biomedical research paper authoring support tool. *Science Editing*, *1*(1), 37–42. doi: 10.6087/kcse.2014.1.37

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, *6*, 391–406. doi: 10.1162/tacl_a_00028

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, *24*, 269–292. doi: 10.1016/j.esp.2004.08.003

Kato, Y., Matsubara, S., & Inagaki, Y. (2006). A corpus search system utilizing lexical dependency structure. In *Proceedings of the fifth international conference on language resources and evaluation.*

Kermes, H. (2012). A methodology for the extraction of information about the usage of formulaic expressions in scientific texts. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 2064–2068).

Kermes, H., & Teich, E. (2020). Formulaic expressions in scientific texts: Corpus design, extraction and exploration. *Lexicographica*, *28*(1), 99–120. doi: 10.1515/lexi.2012-0007

Kim, K.-R., Röthlisberger, P., Kang, S. J., Nam, K., Lee, S., Hollenstein, M., &

Ahn, D.-R. (2018). Shaping rolling circle amplification products into DNA nanoparticles by incorporation of modified nucleotides and their application to in vitro and in vivo delivery of a photosensitizer. *Molecules*, *23*(7). doi: 10.3390/molecules23071833

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 3294–3302). Curran Associates, Inc.

Kristianto, G. Y., Topić, G., & Aizawa, A. (2017). Utilizing dependency relationships between math expressions in math IR. *Information Retrieval Journal*, *20*(2), 132–167. doi: 10.1007/s10791-017-9296-8

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning* (pp. 282–289).

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. doi: 10.1093/bioinformatics/btz682

Lim, J. M. H. (2006). Method sections of management research articles: A pedagogically motivated qualitative study. *English for Specific Purposes*, *25*, 282–309. doi: 10.1016/j.esp.2005.07.001

Lim, J. M. H. (2010). Commenting on research results in applied linguistics and education: A comparative genre-based investigation. *Journal of English for Academic Purposes*, *9*, 280–294. doi: 10.1016/j.jeap.2010.10.001

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, *31*, 25–35. doi: 10.1016/j.esp.2011.07.002

Liu, J., Shang, J., Wang, C., Ren, X., & Han, J. (2015). Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 1729–1744). doi: 10.1145/2723372.2751523

Liu, Y., Wang, X., Liu, M., & Wang, X. (2016). Write-righter: An academic writing assistant system. In *Thirtieth AAAI conference on artificial intelligence* (pp. 4373–4374).

Lorés, R. (2004). On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes*, *23*(3), 280–302. doi: 10.1016/j.esp.2003.06.001

Lu, X., Yoon, J., & Kisselev, O. (2018). A phrase-frame list for social science research article introductions. *Journal of English for Academic Purposes*, *36*, 76–85. doi: 10.1016/j.jeap.2018.09.004

Makkonen, J. (2003). Investigations on event evolution on TDT. In *Proceedings of the HLT-NAACL 2003 student research workshop* (pp. 43–48).

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. In (p. 535). MIT Press.

Marin, A., Holenstein, R., Sarikaya, R., & Ostendorf, M. (2014). Learning phrase patterns for text classification using a knowledge graph and unlabeled data. In *The 15th annual conference of the international speech communication association* (pp. 253–257).

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320. doi: 10.1093/applin/ams010

Maswana, S., Kanamaru, T., & Tajino, A. (2015). Move analysis of research

articles across five engineering fields: What they share and what they do not. *Ampersand*, *2*, 1–11. doi: 10.1016/j.amper.2014.12.002

McDonald, R., Crammer, K., & Pereira, F. (2005). Online large-margin training of dependency parsers. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)* (pp. 91–98). doi: 10.3115/1219840.1219852

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* (Vol. 26, pp. 3111–3119).

Mizumoto, A., Hamatani, S., & Imao, Y. (2017). Applying the bundle–move connection approach to the development of an online writing support tool for research articles. *Language Learning*, *67*(4), 885–921. doi: 10.1111/lang.12250

Morley, J. (n.d.). *Academic Phrasebank.* `http://www.phrasebank.manchester.ac.uk/`.

Nekrasova-Beker, T. M. (2019). Discipline-specific use of language patterns in engineering: A comparison of published pedagogical materials. *Journal of English for Academic Purposes*, *41*, 1–12. doi: 10.1016/j.jeap.2019.100774

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th bionlp workshop and shared task* (pp. 319–327). doi: 10.18653/v1/W19-5034

Omidian, T., Shahriari, H., & Siyanova-Chanturia, A. (2018). A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts. *Journal of English for Academic Purposes*, *36*, 1–14. doi: 10.1016/j.jeap.2018.08.002

Osborne, M., & Baldridge, J. (2004). Ensemble-based active learning for parse selection. In *Proceedings of the human language technology conference of the north American chapter of the association for computational linguistics: HLT-NAACL 2004* (pp. 89–96).

Ozturk, I. (2007). The textual organisation of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes*, *26*(1), 25–38. doi: 10.1016/j.esp.2005.12.003

Pal, S., Naskar, S. K., Vela, M., Liu, Q., & van Genabith, J. (2017). Neural automatic post-editing using prior alignment and reranking. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers* (pp. 349–355).

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in telecommunications research journals. *Journal of English for Academic Purposes*, *21*, 60–71. doi: 10.1016/j.jeap.2015.11.003

Pavlopoulos, J., & Androutsopoulos, I. (2014). Multi-granular aspect aggregation in aspect-based sentiment analysis. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 78–87). doi: 10.3115/v1/E14-1009

Peacock, M. (2002). Communicative moves in the discussion section of research articles. *System*, *30*(4), 479–497. doi: 10.1016/S0346-251X(02)00050-7

Pecina, P. (2008). *Lexical association measures* (Doctoral dissertation). Charles University.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation*, *44*(1–2), 137–158. doi: 10.1007/s10579-

009-9101-4

Pendar, N., & Cotos, E. (2008). Automatic identification of discourse moves in scientific article introductions. In *Proceedings of the third workshop on innovative use of NLP for building educational applications* (pp. 62–70).

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* (pp. 1532–1543). doi: 10.3115/v1/D14-1162

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, *14*, 84–94. doi: 10.1016/j.jeap.2014.01.002

Peters, E., & Pauwels, P. (2015). Learning academic formulaic sequences. *Journal of English for Academic Purposes*, *20*, 28–39. doi: 10.1016/j.jeap.2015.04.002

Phan, T. K., Lay, F. T., Poon, I. K., Hinds, M. G., Kvansakul, M., & Hulett, M. D. (2016). Human β-defensin 3 contains an oncolytic motif that binds PI(4,5)P2 to mediate tumour cell permeabilisation. *Oncotarget*, *7*(2), 2054–2069. doi: 10.18632/oncotarget.6520

Przybyła, P. (2013). Question analysis for Polish question answering. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop* (pp. 96–102).

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System*, *42*, 220–231. doi: 10.1016/j.system.2013.12.003

Rashidi, N., & Meihami, H. (2018). Informetrics of scientometrics abstracts: a rhetorical move analysis of the research abstracts published in scientometrics journal. *Scientometrics*, *116*, 1975–1994. doi: 10.1007/s11192-018-2795-6

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier chains for multi-label classification. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine learning and knowledge discovery in databases* (pp. 254–269).

Role, F., & Nadif, M. (2011). Handling the impact of low frequency events on co-occurrence based measures of word similarity. In *Proceedings of the international conference on knowledge discovery and information retrieval (KDIR-2011)* (pp. 218–223). doi: 10.5220/0003655102260231

Saboori, F., & Hashemi, M. R. (2013). A cross-disciplinary move analysis of research article abstracts. *International Journal of Language Learning and Applied Linguistics World*, *4*(4), 483–496.

Saied, H. A., Candito, M., & Constant, M. (2019). Comparing linear and neural models for competitive MWE identification. In *Proceedings of the 22nd Nordic conference on computational linguistics* (pp. 86–96).

Sarkar, A. (1998). Conditions on consistency of probabilistic tree adjoining grammars. In *36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, volume 2* (pp. 1164–1170). doi: 10.3115/980691.980759

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., ... 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). doi: 10.18653/v1/W17-1704

Schubotz, M., Greiner-Petter, A., Scharpf, P., Meuschke, N., Cohl, H. S., & Gipp, B. (2018). Improving the representation and conversion of mathematical

formulae by considering their textual context. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries* (pp. 233–242). doi: 10.1145/3197026.3197058

Schulte im Walde, S., Köper, M., & Springorum, S. (2018). Assessing meaning components in German complex verbs: A collection of source-target domains and directionality. In *Proceedings of the seventh joint conference on lexical and computational semantics* (pp. 22–32). doi: 10.18653/v1/S18-2003

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, *31*(4), 487–512. doi: 10.1093/applin/amp058

Soonklang, T. (2016). Move classification in scientific abstracts using linguistic features. In *The eleventh international symposium on natural language processing.*

Srivastava, S., Labutov, I., & Mitchell, T. (2018). Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 306–316). doi: 10.18653/v1/P18-1029

Stevenson, S., & Merlo, P. (1999). Automatic verb classification using distributions of grammatical features. In *Ninth conference of the European chapter of the association for computational linguistics* (pp. 45–52).

Stewart, I., Pinter, Y., & Eisenstein, J. (2018). Si o no, que penses? Catalonian independence and linguistic identity on social media. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 136–141). doi: 10.18653/v1/N18-2022

Swales, J. M. (1981). *Aspects of article introductions.* The University of Michigan Press.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications.* Cambridge University Press.

Swales, J. M. (2019). The futures of eap genre studies: A personal viewpoint. *Journal of English for Academic Purposes*, *38*, 75–82. doi: 10.1016/j.jeap.2019.01.003

Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text* (Doctoral dissertation, University of Edinburgh). doi: 1842/11456

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 103–110).

Thelwall, M. (2019). The rhetorical structure of science? a multidisciplinary analysis of article headings. *Journal of Informetrics*, *13*(2), 555–563. doi: 10.1016/j.joi.2019.03.002

Tsoumakas, G., & Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Machine learning: Ecml 2007* (pp. 406–417). Springer Berlin Heidelberg.

van der Linden, E.-J. (1992). Incremental processing and the hierarchical lexicon. *Computational Linguistics*, *18*(2), 219–238.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008).

Vincent, B. (2013). Investigating academic phraseology through combinations of very frequent words: A methodological exploration. *Journal of English for Academic Purposes*, *12*(1), 44–56. doi: 10.1016/j.jeap.2012.11.007

Vivas, A. B., Paraskevopoulos, E., Castillo, A., & Fuentes, L. J. (2019). Neurophysiological activations of predictive and non-predictive exogenous cues: A cue-elicited eeg study on the generation of inhibition of return. *Frontiers in Psychology*, *10*, 227. doi: 10.3389/fpsyg.2019.00227

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 353–355). doi: 10.18653/v1/W18-5446

Waszczuk, J., Ehren, R., Stodden, R., & Kallmeyer, L. (2019). A neural graph-based approach to verbal MWE identification. In *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019)* (pp. 114–124). doi: 10.18653/v1/W19-5113

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: an integrated model. *Language & Communication*, *20*(1), 1–28. doi: doi.org/10.1016/S0271-5309(99)00015-4

Wu, J.-C., Chang, Y.-C., Liou, H.-C., & Chang, J. S. (2006). Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions* (pp. 41–44). doi: 10.3115/1225403.1225414

Wu, X., Mauranen, A., & Lei, L. (2020). Syntactic complexity in English as a lingua franca academic writing. *Journal of English for Academic Purposes*, *43*. doi: 10.1016/j.jeap.2019.100798

Xia, Y., Huang, C.-C., Dittmar, R., Du, M., Wang, Y., Liu, H., ... Kohli, M. (2016). Copy number variations in urine cell free dna as biomarkers in advanced prostate cancer. *Oncotarget*, *7*(24), 35818–35831. doi: 10.18632/oncotarget.9027

Xie, P., Yang, D., & Xing, E. (2015). Incorporating word correlation knowledge into topic modeling. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 725–734). doi: 10.3115/v1/N15-1074

Yen, T.-H., Wu, J.-C., Chang, J., Boisson, J., & Chang, J. (2015). Writeahead: Mining grammar patterns in corpora for assisted writing. In *Proceedings of ACL-IJCNLP 2015 system demonstrations* (pp. 139–144). doi: 10.3115/v1/P15-4024

Yih, W.-t. (2009). Learning term-weighting functions for similarity measures. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 793–802).

Zhang, B., Marin, A., Hutchinson, B., & Ostendorf, M. (2013). Learning phrase patterns for text classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(6), 1180–1189. doi: 10.1109/TASL.2013.2245651

Zhang, W., Liu, Q. X., Guo, Z. H., & Lin, J. S. (2018). Practical application of aptamer-based biosensors in detection of low molecular weight pollutants in water sources. *Molecules*, *23*(2). doi: 10.3390/molecules23020344

Zhao, J. (2017). Native speaker advantage in academic writing? conjunctive realizations in EAP writing by four groups of writers. *Ampersand*, *4*, 47–57. doi: 10.1016/j.amper.2017.07.001

Zhong, N., Li, Y., & Wu, S. (2012). Effective pattern discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*, *24*(1), 30–44. doi:

10.1109/TKDE.2010.211

Appendix

Table 1: Top-50 frequent formulaic expressions in each communicative function (CF) in each discipline in the communicative-function-labelled formulaic expression database we presented.

CL		Chem		Onc		Psy	
Section: introduction							
CF: Showing the importance of the topic							
is an important	163	as well as	477	as well as	1069	is an important	191
it is important to	61	is an important	346	is the most common	1041	it is important to	168
plays an important role in	39	due to the	324	has been shown to	854	is one of the	111
play an important role in	38	is one of the most	238	is one of the most common	661	it is important to note that	94
is crucial for	36	due to their	238	also known as	496	plays an important role in	92
contributed equally to this work	35	is one of the	224	is one of the most	477	it is assumed that	90
is important for	33	belonging to the	212	is the most	461	is one of the most	82
is one of the most important	31	belongs to the	204	we found that	404	is important for	70
is useful for	25	such as the	178	is an important	404	is one of the most important	66
it is crucial to	22	the use of	176	plays an important role in	373	the importance of	63
plays a crucial role in	21	is the most	167	have been shown to	369	is the ability to	63
is a key	21	it is a	164	is one of the	365	play an important role in	63
there has been a growing interest in	20	is one of the most important	162	is a member of the	347	is considered a	62
is crucial to	18	also known as	151	in a variety of	327	is a common	62
has become a	18	due to its	151	such as the	322	one of the most	59
one of the most important	17	plays an important role in	131	was shown to	315	is a key	54
is essential for	17	on the other hand	125	is a common	315	it is a	53
is important for many	17	play an important role in	123	in addition to	315	it is not surprising that	50
can be useful for	15	is the most common	118	is involved in	311	it is necessary to	49
is essential to	15	are the main	115	has been shown to be	302	is considered to be a	48
are important for	14	on the other	113	has been reported to	300	therefore it is important to	48
plays a key role in	14	depending on the	109	is characterized by	277	it is assumed that the	47
plays a critical role in	14	are the most	107	the majority of	259	is crucial for	46
is at the	13	leads to the	106	due to the	253	there is a growing	42
this is an important	13	is a common	105	are involved in	249	plays a key role in	39
is an important step in	13	the most common	100	is a key	247	is considered as a	38
play a crucial role in	13	because of the	100	acts as a	246	plays an important role in the	38
is closely related to	13	is the main	100	is the leading cause of	240	is a form of	37
is critical for	13	one of the	99	leads to the	240	one of the most important	37
is an important aspect of	13	the number of	98	leading to the	228	plays a crucial role in	37
is part of the	12	leading to the	98	is a major	225	is the most	36
has been a	12	involved in the	97	involved in the	224	it is therefore important to	36
is important to	12	because of their	97	the most common	221	this is because	35
therefore it is important to	12	is characterized by	96	is involved in the	216	it is important to understand the	35
is a fundamental	12	for example the	94	belongs to the	213	it is also important to	33
it is an important	12	plays an important role in the	93	is required for	212	it is crucial to	33
is important in	12	belong to the	91	in addition the	211	it is the	32
are useful for	12	in addition to	88	is the second most common	210	it is important to note that the	31
has received a	11	play an important role in the	88	was found to	209	thus it is important to	30
have become an important	11	are involved in the	84	as well as the	205	play a crucial role in	29
play a key role in	11	which is a	84	was found to be	205	plays a role in	29

(Continued)

CL		Chem		Onc		Psy	
are crucial for	11	are the major	84	resulting in the	203	is the most common	29
as an important	11	resulting in the	82	which in turn	203	is an important aspect of	28
is useful for many	11	is a major	82	at the time of	201	as an important	28
interest in the	11	can lead to	82	and the second leading cause of	196	play a role in	27
is necessary for	11	are responsible for the	82	cite- and	190	play an important role in the	27
is more important than	10	are able to	81	plays a critical role in	187	is critical for	26
plays an important role in the	10	on the other hand the	80	has been implicated in	187	is essential for	26
is an important part of	10	is one of the most common	79	are involved in the	185	is assumed to	26
has become an important	10	is responsible for the	78	can lead to	185	is a crucial	26
CF: Showing brief introduction to the methodology							
for example the	378	in the present	217	in the present	339	in the present	374
based on the	309	in order to	177	based on the	143	in the current	189
as well as	290	based on the	154	we hypothesized that	143	were asked to	156
we propose a	215	is based on the	148	we show that	140	we used a	126
in terms of	207	was used to	108	we examined the	135	was used to	96
in order to	202	led to the	83	we demonstrated that	125	we used the	78
we show that	187	by means of	79	in addition we	121	to this end we	77
show that our	176	were used to	76	we demonstrate that	113	was designed to	77
such as the	174	in our previous	73	we performed a	112	to examine the	72
a set of	169	is based on	72	here we show that	85	were presented with	65
the number of	166	a series of	54	was used to	84	we aimed to	59
on the other hand	166	et al developed a	50	we explored the	83	we examined the	58
we use the	158	to determine the	49	we conducted a	83	by using a	56
is based on	157	was applied to	49	can be used to	80	we conducted a	52
is based on the	155	in addition we	48	and found that	78	to address this	51
according to the	154	et al studied the	47	in the current	75	were used to	50
can be used to	149	by using the	42	we found that the	74	cite- used a	49
we use a	149	are based on the	41	is based on the	72	are asked to	45
are added by the	147	were characterized by	40	to determine the	71	to test this	44
show that the	140	in this work	40	we focused on the	69	are used to	44
we show that the	134	we decided to	38	we used the	67	to test the	42
the use of	129	et al used	36	we developed a	67	were required to	41
in the form of	124	on the basis of	36	is based on	62	to explore the	41
for example in	124	based on a	36	we used a	61	we focused on the	40
is used to	120	was determined by	35	were used to	57	in addition we	37
as well as the	119	are based on	35	have been used to	57	therefore the present	37
we show that our	115	can be obtained by	34	to explore the	56	were presented with a	36
due to the	114	can be achieved by	34	has been used to	55	we examined whether the	35
attribution 40 international licence	111	were determined by	34	therefore in this	55	we focused on	34
are based on	109	et al reported the	32	to test this	50	by means of a	33
for example in the	106	focused on the	31	to this end we	50	we explored the	32
in this work we	106	was performed to	31	we tested the	47	with and without	32
can not be	104	was also studied	30	here we demonstrate that	45	we set out to	31
in which the	103	in the current	30	is used to	44	we conducted two	29
we focus on	102	prompted us to	30	in our previous	43	cite- used the	27
as shown in	101	we focused on the	30	we observed that	42	and asked them to	25
for example a	101	were obtained by	30	we demonstrated that the	39	we examine the	25
with respect to the	101	were applied to	30	we were able to	38	is used to	25
is available at	95	by using a	29	we hypothesized that the	38	aimed to examine the	24
in addition to	94	in this work the	28	we compared the	37	using the same	24
of the same	94	was based on the	28	we determined the	37	and the other	24
are used to	93	was employed to	28	therefore we performed a	37	we sought to	23
we find that	92	was carried out	28	are used to	36	was conducted by	23
in the same	92	was obtained by	28	we focused on	36	in the present * we aimed to	22
for a given	88	was selected as the	27	we have developed a	35	to do so we	22
is based on a	85	was developed for the	27	to identify the	34	therefore in the present	22
is a set of	82	were designed and	27	based on these	34	they were asked to	22
the set of	82	is based on a	27	we show that the	34	by using the	22
is described in	82	were performed to	26	and found that the	33	we use a	22
in terms of the	82	can be easily	26	therefore we conducted a	33	by examining the	21
CF: Showing what is already done in the past work							
have been proposed to	94	have shown that	491	have shown that	1040	cite-	2345
have been proposed	81	have been reported	279	have demonstrated that	484	et al cite-	1762
have been proposed for	73	it has been reported that	250	it has been reported that	469	eg cite-	778

(Continued)

CL		Chem		Onc		Psy	
have shown that	54	have demonstrated that	206	it has been shown that	249	cite- found that	629
have been applied to	47	has been shown to	186	have suggested that	216	as well as	552
it has been shown that	46	showed that the	166	have shown that the	213	have shown that	520
has shown that	44	have been shown to	161	have reported that	212	has shown that	337
have been developed to	43	it is well known that	160	have indicated that	208	according to the	319
have been shown to	42	have shown that the	157	has shown that	181	showed that the	272
have been used to	40	have been reported to	138	it is well known that	179	such as the	267
have been used for	39	it has been shown that	127	it was reported that	148	they found that	255
have been developed for	36	have focused on the	126	et al reported that	138	has been shown to	242
have been shown to be	28	have been developed to	117	have revealed that	136	for example the	239
have been proposed in the	28	et al reported that	109	it has been suggested that	134	for example cite- found that	217
have been used in	27	has been reported to	106	it is known that	133	have been shown to	206
there have been several	26	have reported that	98	it has been demonstrated that	132	it has been shown that	204
have focused on	26	have also been	97	we have previously shown that	126	are more likely to	200
have been made to	25	it was reported that	96	we previously reported that	106	it has been suggested that	179
has focused on	25	have been developed	94	have demonstrated the	103	have found that	172
have been widely used in	25	have indicated that	94	it is reported that	99	found that the	166
it is well known that	23	have reported the	91	it has been	97	on the other hand	164
are widely used in	23	it has been demonstrated that	88	have demonstrated that the	92	and found that	163
has been shown to	23	it was found that	86	has demonstrated that	89	suggest that the	147
have been developed	21	it is well known that the	86	it has been reported that the	84	have suggested that	146
has been shown to be	20	have revealed that	84	have found that	84	have shown that the	144
previous work has shown that	19	it was found that the	77	it is well established that	83	have demonstrated that	138
it is clear that	19	has shown that	76	it has been proposed that	79	have been found to	136
in the past	19	have been developed for	76	reported that the	75	has been found to	136
there have been many	19	it has been reported that the	74	we have shown that	72	a number of	134
there have been a number of	18	have focused on	73	have also shown that	64	cite- found that the	132
recent work has shown that	17	have led to the	73	we previously demonstrated that	64	has been found to be	130
it is well-known that	17	have been developed for the	71	we have previously demonstrated that	60	see cite- for a	126
have also been proposed	17	have demonstrated that the	71	we and others have shown that	58	was found to be	125
have focused on the	16	it was shown that	71	suggests that the	58	suggests that the	123
have also been	15	revealed that the	66	we have previously reported that	58	we predicted that	123
was proposed by	15	have suggested that	66	it has also been reported that	57	has been shown to be	122
have shown that the	15	have demonstrated the	66	it is believed that	56	is related to	122
has been successfully applied to	14	have been reported in the	64	it is estimated that	56	as compared to	121
have been successfully applied to	14	have been reported for	64	have confirmed that	52	it was found that	121
previous work has	14	it is reported that	63	have shown the	51	has been linked to	120
has focused on the	14	have been reported in	63	have reported that the	49	were more likely to	117
have been proposed in	14	have been found to	62	have suggested that the	49	depending on the	117
have been explored	14	it has been	62	have reported the	48	for example a	115
have been used	13	indicated that the	61	suggested that the	47	it is possible that	104
it is known that	13	has been reported	61	it is widely accepted that	45	have shown that * cite-	103
has been proposed to	13	was shown to	61	there are several	44	has demonstrated that	103
previous work has focused on	13	have been reported for the	58	we have recently shown that	43	see also cite-	103
have been proposed to address the	13	demonstrated that the	57	it has been suggested that the	43	is characterized by	100

(Continued)

CL		Chem		Onc		Psy	
it has been shown that the	12	it has been suggested that	56	it was shown that	41	they found that the	100
it has been observed that	12	it is estimated that	55	recently it has been reported that	41	it has been	100
CF: Showing the aim of the paper							
in this paper we	667	in this paper we	210	the aim of this	197	the aim of the present	134
in this paper we propose a	311	the aim of this	205	the purpose of this	93	in this paper we	112
in this paper we present a	281	the aim of the present	90	the aim of the present	64	the aim of this	87
in this paper we focus on	136	in this paper we describe the	76	in this paper we	48	the aim of the current	51
in this paper we propose a novel	119	the aim of this work was to	72	here we describe the	44	the purpose of this	44
this paper presents a	110	the purpose of this	64	the aim of our	34	the purpose of the present	41
we present a	108	in this paper the	61	therefore the aim of this	31	this paper aims to	27
in this paper we present	94	herein we report the	55	of the present	30	in this paper	26
in this paper we present an	92	herein we describe the	55	the aims of this	24	in this paper we focus on	26
in this paper we describe a	85	we report the	44	we describe the	24	the aim of the	23
in this paper we address the	85	therefore the aim of this	44	the aim of this work was to	21	the present paper	21
in this paper we propose	83	the aim of the present work was to	43	in this work we	20	the aim of our	19
in this paper we focus on the	77	this paper describes the	42	we discuss the	20	in this paper we present a	18
in this paper we propose an	74	of the present	42	was to determine the	19	aims to explore the	16
in this paper we describe the	69	we describe the	36	was to determine whether	19	the purpose of the current	15
this paper describes a	68	in the present work we	33	here we present	18	our aim is to	15
in this paper we propose a new	67	here we describe the	31	here we describe a	17	of this paper is to	14
this paper proposes a	65	focuses on the	30	the aim of the	17	in this paper we focus on the	14
this paper describes the	64	of this work was to	30	the aim of the current	17	therefore the aim of the present	14
in this paper we describe our	58	the aim of the	29	here we present the	17	the second aim was to	13
we present a novel	53	herein we wish to	28	we report the	17	in this paper we aim to	13
in this work we propose a	53	of the new	28	therefore the purpose of this	16	in the current paper we	13
this paper focuses on	52	was designed to	27	the purpose of the present	15	finally we discuss the	13
in this paper we propose to	50	we report herein the	26	thus the aim of this	15	in this paper we present	13
in this paper we explore the	47	the aim of this work is to	26	the aim of this * was to examine the	14	the main aim of the present	12
in this paper we use	44	in this paper we present the	25	here we describe	13	in this paper we explore	12
in this paper we describe	44	we aimed to	24	therefore the aim of our	12	this paper presents a	12
this paper presents	44	was to determine the	23	was to identify	12	the aim of the present * was to examine the	12
in this paper we explore	41	will focus on the	22	the aim of this * was to identify	11	this paper focuses on	11
this paper describes our	41	in this paper	22	here we present a	11	the purpose of this * was to examine the	11
in this paper we present a novel	41	we present the	22	was to analyze the	10	was to examine the role of	11
this paper focuses on the	41	therefore the present	21	was to examine the	10	aims to address this	11
this paper presents an	39	we report on the	20	the purpose of our	10	the main aim of this	11
this paper describes	37	therefore in this	19	the primary aim of this	9	aims to fill this	11
in this work we present a	37	in the present * we report the	19	the aim of this * was to determine the	9	here we aim to	10
we present a new	36	thus the aim of this	19	the aim of the present work was to	9	the purpose of the present * was to examine the	10
in this paper we present a new	35	is focused on the	19	therefore the aim of the present	9	this paper focuses on the	10
we present an	35	the aim of this * was to determine the	19	the aim of the present * was to determine the	9	the aim of this paper is to	10
in this paper we aim to	33	we focus on the	18	herein we describe the	8	therefore the aim of this	10
we describe a	30	therefore the aim of the present	18	our aim was to	8	aims to examine the	10
in this paper we present the	29	in this paper a	17	the aims of the present	8	in this paper we are	9
in this paper we consider the	27	in this work we describe the	17	was to identify the	8	in the present paper we	9
in this paper we are	26	we discuss the	17	here we report our	7	the current paper	9

(Continued)

CL		Chem		Onc		Psy	
this paper presents the	26	was to develop a	16	the aim of this * was to assess the	7	in this paper we explore the	9
this paper addresses the	26	herein we present the	15	aimed to identify	7	in this paper we examine	8
in this paper we develop a	26	was to explore the	14	in the present work we	7	the aim of the current * was to examine the	8
this paper aims to	25	the present work	14	aims to explore the	7	in this paper we use	8
this paper describes an	24	the aim of our	14	aimed to determine the	7	in this paper we discuss the	8
this paper proposes a novel	24	here we present the	13	here we aim to	7	in this paper we propose to	8
this paper presents a new	24	in this paper we wish to	13	therefore the purpose of the present	7	the main aim of the	8
CF: Showing explanation or definition of terms or notations							
is defined as the	45	have been used as	186	are referred to as	34	refers to the	188
we refer to this	43	have been used to	171	is defined as	33	is defined as the	90
is defined as	40	has been used in	144	is defined as a	31	is referred to as	57
we call this	35	has been used as a	143	is defined as the	22	refer to the	51
we use the term	30	have been used in	139	refers to the	21	refers to a	51
is defined as a	29	has been used to	129	is referred to as	18	is defined as	49
refers to the	29	have been used for	116	are defined as	17	are referred to as	43
we refer to the	28	has been used for	86	is defined as an	10	is referred to as the	40
is called a	23	are used in	84	is also called	9	we use the term	31
is referred to as	22	have been used for the	72	is referred to as the	8	to the ability to	28
we refer to	21	are used as	63	refers to a	7	refers to the ability to	27
is defined as follows	19	are widely used in	62	has been termed	6	we refer to	27
are referred to as	18	are shown in	62	has been defined as a	6	we refer to the	25
to denote the	18	is widely used in	61	are defined as * longer than 200	6	has been referred to as	25
is called the	17	is used as a	60	to describe the	5	it refers to the	25
we will refer to the	16	has been widely used in	60	are generally defined as	5	we will refer to this	24
is said to be	16	is used to	56	they are referred to as	5	is called the	24
refer to the	15	has been used for the	54	is referred to as a	4	refers to an	22
we will refer to this	15	have been used as a	53	has been referred to as the	4	is defined as an	20
we will use the term	15	is used in	52	has been referred to as	4	we will use the term	18
can be defined as the	15	have been widely used in	51	we will refer to	4	we refer to this	17
we refer to such	13	have been used in the	51	is often referred to as	4	is often referred to as	15
we denote by	12	has been used in the	45	are often referred to as	4	can be referred to as	15
we mean that the	11	has been used as an	43	are defined as * more than 200	4	refers to the extent to which	13
we will use the	11	has been widely used to	40	broadly referred to as	3	are often referred to as	13
1 we use the	11	have been used	40	be referred to as	3	will be referred to as	12
are defined as	11	are used for	39	can be defined as a	3	is used to refer to	11
we will refer to	11	is used for	38	has been defined as	3	refers to a set of	11
1 we use	10	has been widely used for	34	will be referred to as	3	is referred to as a	10
denote the set of	10	was used as a	33	are defined as those	3	to the extent to which	10
is often referred to as	9	were used as	33	is referred as	3	has been referred to as the	10
refers to a	8	is defined as the	32	were defined as	3	has been termed the	9
we refer the	8	is used in the	31	are commonly referred to as	3	is used to describe	9
is defined to be	8	is defined as a	31	are defined as * of more than 200	3	we will refer to these	9
may refer to a	8	have been used in folk	30	are called as	3	is defined as the ability to	9
to refer to the	8	has been used as	30	is also referred to as	3	we will refer to the	9
is referred to as a	8	has been widely used for the	30	is commonly referred to as	3	this is referred to as	9
can be defined as	7	it has been used to	29	is defined as any bodily	3	it refers to a	9
we refer to these	7	is defined as	28			to the degree to which	9
1 we use the term	7	have been widely used for	27			this is referred to as the	8
in this paper we refer to	7	has been used	27			we will refer to	8
is defined in	7	has been widely used as a	26			to the tendency to	8
are defined as follows	7	is commonly used in	25			is commonly referred to as	8
we refer to this as the	6	is widely used in the	25			this is called the	8

(Continued)

CL		Chem		Onc		Psy	
to refer to	6	is widely used as a	24			is also referred to as	8
is defined by	6	it has been used as a	24			we refer to these	8
we refer to our	6	is shown in	23			we call this the	8
can be defined as a	6	has also been used to	23			to refer to the	8
throughout this paper we use the	6	are referred to as	23			are often used interchangeably	8
here we use the	6	it has been used in	23			has been termed	7
CF: Showing the importance of the research							
to the best of our	87	for the first time	90	this is the first	98	this is the first	55
we would like to	47	this is the first	60	for the first time	55	this allowed us to	52
this is the first	44	for the first time the	47	for the first time that	47	allows us to	48
for the first time	35	is the first	45	are needed to	47	should be able to	40
we aim to	33	was the first	44	for the first time the	46	is the first to	39
this is the first work to	32	for the first time in	25	it is important to	45	for the first time	38
we will show that	28	this is the first report on the	19	we show for the first time that	32	allowed us to	33
we are the first to	27	is the first to	14	we demonstrate for the first time that	32	were the first to	30
in this paper we will	23	was one of the first	14	we aimed to	32	was the first to	28
we present the first	22	may provide a	13	therefore it is important to	31	this would suggest that	25
in this work we aim to	21	it is expected that the	12	are required to	28	is expected to	25
this is the first work that	20	for the first time by	12	provide new insights into the	25	we will focus on	23
we would like to thank	19	this is the first report on	11	in the present * we aimed to	24	would suggest that	20
would like to thank	19	for the first time we	11	a better understanding of the	22	it should be possible to	20
is to build a	18	it is the first	11	is needed to	21	should be more	19
we will discuss the	17	will provide a	10	may help to	20	would be the	19
we will describe the	16	this is the first time that	10	therefore it is of great	20	can serve as a	19
would be to	16	in the first step	10	is the first to	20	could lead to	18
is to develop a	16	will contribute to the	10	we provide the first	20	makes it possible to	18
is the first	15	et al reported the first	9	are needed for	20	make it possible to	17
our aim is to	15	demonstrated for the first time that	9	this is the first study to	19	will allow us to	17
is to provide a	14	as far as we know this is the first	9	may provide a	19	would be a	16
this is the first attempt to	14	one of the first	9	for the first time we	19	would be able to	16
ideally we would like to	14	may be a promising	9	may be useful for	18	may help to	16
we will show that the	14	was the first to	9	thus it is important to	18	this is the first study to	16
will be described in	13	could provide a	8	demonstrate for the first time that	17	may serve as a	16
it will be	13	for the first time and	8	we demonstrated for the first time that	17	enables us to	16
is to create a	13	this is the first time that the	8	highlight the importance of	16	we believe that the	16
we will use	13	the first step in the	8	it is important to understand the	16	we would like to	16
we hope to	12	this is the first report of	8	shed light on the	16	there should be a	16
would like to	12	for the first time that	8	is important to	16	would lead to	15
to the best of our * we are the first to	12	we believe that the	8	are needed to improve	15	will be more	15
we will discuss	12	could be useful to	7	we aimed to identify	15	allows for the	15
will be used to	12	are the first	7	needs to be	15	we will focus on the	15
we expect that	11	is reported for the first time	7	our understanding of	15	may be useful for	15
we wish to	11	could improve the	7	provide novel insights into the	15	this allows us to	15
this paper is the first to	11	were the first to	7	here we show for the first time that	14	should be more likely to	15
we will focus on	11	for the first time from	7	it would be interesting to	14	a better understanding of the	15
our work is the first to	11	the first generation of	7	may provide a novel	14	for the first time the	15
it is the first	10	are expected to provide	7	therefore understanding the	14	should lead to	14
we hope that the	10	for the first time from the	7	a better understanding of	13	one of the first	14
we are currently	10	will be useful for	7	it is therefore important to	13	this enabled us to	14
we will show that this	10	reported the first	6	are needed to improve the	13	may be useful to	14
we will see that	10	basis for the	6	for the first time a	13	therefore we expect that	14

(Continued)

CL		Chem		Onc		Psy	
will be a	10	provide insight into the	6	are needed to identify	12	will focus on the	14
will be able to	10	provide a basis for	6	is needed in order to	12	on the other hand if	13
we will also	9	allowed us to	6	are necessary to	12	can contribute to the	13
we would also like to	9	are reported for the first time	6	may be useful in	12	may provide a	13
we are the first to apply	9	was reported for the first time	6	may provide a new	12	this would mean that	13
we are also	9	will be helpful for	6	we are the first to	12	it allows us to	13
CF: Showing limitation or lack of past work							
there is no	48	it is known that	78	however the role of	189	to the best of our	97
it is difficult to	39	have not been	65	has not been	176	there is no	96
it is hard to	27	has not been	64	have not been	142	little is known about the	92
there are no	25	little is known about the	58	little is known about the	134	has not been	84
there has been little	22	there is no	55	is still unclear	118	can not be	78
it is not possible to	21	has not been reported	53	however little is known about the	116	is not a	75
has not been	18	there are no	49	remain largely unknown	93	have not been	56
however it is difficult to	15	have not been reported	48	remains largely unknown	93	has not yet been	48
however there is no	15	there are few	46	has not been reported	86	there was no	46
is not always	14	however there are few	44	are not fully understood	82	may not be	44
has not yet been	14	however there is no	42	has not yet been	81	there are no	43
it is impossible to	13	has been paid to the	40	is not fully understood	78	only a few	43
however to the best of our	13	has not yet been	33	remain poorly understood	74	it is unclear whether	41
it is not	13	it is known that the	30	remains poorly understood	71	there has been little	37
there are few	12	it is difficult to	30	however there is no	71	is not limited to	35
it is not clear how to	11	has been paid to	28	there is no	69	it is not clear whether	34
is it possible to	11	has been extensively studied	28	is not clear	69	has yet to be	34
there are a few	11	however little is known about the	27	has not been fully elucidated	66	however little is known about the	33
there has been little work on	11	is available on the	27	is not well understood	62	little is known about	33
it is challenging to	10	there are only a few	26	remains to be elucidated	60	there is little	33
are not always	10	is still unknown	24	remain to be elucidated	60	it is unclear whether the	32
there is little	10	there is little	24	is largely unknown	60	has examined the	32
have not been	10	are known for their	23	are still unclear	60	there has been no	31
there has been no	10	has been studied	23	remains to be determined	57	however there is no	29
it is unlikely that	10	there is a lack of	22	is still unknown	56	have not yet been	29
is not easy	10	have been extensively studied	21	are not well understood	55	are not always	28
none of the	10	have been studied	21	have focused on the	55	we are not aware of any	28
are not available	10	however to the best of our	21	however the underlying	55	has been paid to	27
none of these	10	are still unclear	20	there are no	55	there are only a few	27
it is not clear that	10	are still unknown	20	are poorly understood	54	there are few	27
it is not clear how	9	there are a few	19	have not been fully elucidated	52	little is known about how	26
however there is	9	is still unclear	18	are largely unknown	51	it remains unclear whether	26
it is not trivial to	9	have not been studied	18	is still poorly understood	48	is still unclear	25
are not suitable for	9	are not fully understood	18	is poorly understood	48	there were no	24
has been paid to the	9	have yet to be	18	have not been reported	46	none of the	23
there is no clear	8	there have been few	18	has not been studied	45	there is only one	23
however there has been little	8	have not yet been	18	there are few	45	could not be	23
there have been a few	8	there has been no	18	have not yet been	45	is not an	23
is not able to	8	is not well understood	17	is not known	44	however to our	23
there is a large	8	however there are no	16	has yet to be	41	this is not to say that	22
however none of these	8	however there is little	16	remain to be	39	however not all	22
it is not easy to	8	however there are only a few	15	has not been explored	38	has never been	22
we are not aware of any	8	has been extensively	15	remains to be	37	it is not	22
have not yet been	7	are not well understood	15	little is known about	36	are not necessarily	21

(Continued)

CL		Chem		Onc		Psy	
is not sufficient	7	has yet to be	15	has not been elucidated	36	has been paid to the	21
it is very difficult to	7	there are no reports on the	15	however the precise	35	it is not possible to	21
it is not clear whether	7	however there are	15	have focused on	35	none of these	20
little is known about how	7	has never been	14	have examined the	35	it is not clear whether the	20
there is a lack of	7	there have been no	14	are still poorly understood	35	is difficult to	20
most previous work on	7	however there is a	14	however the exact	34	however only a few	20
CF: Showing the main problem in the field							
is the lack of	45	is a serious	31	are urgently needed	125	there is a lack of	55
is a challenging	41	are urgently needed	27	is urgently needed	66	it is difficult to	50
one of the main	31	one of the main	26	therefore it is	45	need to be	41
there is a need to	22	therefore it is necessary to	21	remains a major	44	the need for	39
there is a need for	17	is urgently needed	19	has become a major	29	there is a need for	30
one of the major	17	therefore there is an urgent need to develop	18	therefore there is an urgent need to	28	there is a need to	30
is that they are	15	is still a	14	therefore there is an urgent need for	22	one of the main	27
is that it	13	therefore it is necessary to develop	14	it is necessary to	21	we need to	25
there are two major	12	there is a need for	13	remains a challenge	21	makes it difficult to	23
is how to	10	therefore there is a	12	are urgently required	20	make it difficult to	20
is that they	10	therefore there is an urgent need for	12	therefore it is necessary to	19	need to be able to	18
is still a challenging	9	has become a serious	12	there is a need to	19	needs to be	16
with this approach is that	9	there is an urgent need for the	12	therefore there is an urgent need to develop	19	making it difficult to	15
is one of the main	9	there is a need to	11	there is an urgent need to	18	was to examine whether	15
a key challenge in	8	is highly desirable	11	therefore it is necessary to identify	18	is a serious	14
is a very challenging	8	however there are still some	11	therefore there is a need for	17	the need for a	14
remains a challenge	8	is a major challenge	11	therefore there is an	17	was to explore the	14
is one of the major	8	there is an urgent need to	11	is the lack of	17	is the lack of	14
this is a difficult	8	is a challenging	11	there is a need for	16	was to determine whether	13
is one of the most challenging	7	therefore there is a need to	10	therefore it is urgent to	16	this makes it difficult to	13
the main difference is that	7	thus there is a	10	is still a	16	the need to	13
is that they do not	7	is one of the most serious	10	therefore there is a	15	and the need for	12
this is a challenging	7	there is a great need for	10	thus there is an	15	a need for	11
is a challenge for	7	therefore there is an urgent need to	10	thus it is	15	one of the most common	11
there is still a	7	there is still a need to	9	is urgently required	14	it is very difficult to	10
the challenge of	7	thus it is necessary to develop	9	therefore there is an urgent need to identify	14	is a difficult	10
a challenge for	7	therefore it is of great	9	is still a major	14	it is more difficult to	10
there is a need for a	7	represents a major	9	is a serious	13	is the need to	10
lies in the	6	has become a major	9	are urgently needed for	13	there is a need for more	10
is a challenge	6	is still needed	9	thus it is necessary to	13	one of the major	9
a major challenge in	6	however the main	8	remains a major challenge	13	it is hard to	9
of this approach is that	6	there is a growing need for	8	it is essential to	12	a need to	9
is the need to	6	therefore there is a need for	8	are desperately needed	11	the need for more	8
is still a	6	thus it is necessary to	8	it is therefore	11	which makes it difficult to	8
there is a pressing need for	6	is the lack of	8	represents a major	11	needs to be able to	8
of this approach is that the	6	need to be developed	7	remains an important	11	thus there is a need to	8
is that it only	6	and the need for	7	are urgently needed to	11	the main purpose of the present	8
is the lack of a	6	the main advantages of	7	thus there is a	11	may need to	8
is a hard	5	remains a major	7	is an urgent	11	it can be difficult to	8
is the high	5	remains a challenge	7	continues to be a major	11	we need to be able to	8
is a significant	5	there is an urgent need for	7	is of great importance	11	is one of the major	8

(Continued)

CL		Chem		Onc		Psy	
the main challenge in	5	the urgent need for	7	therefore it is urgent to identify	11	there is a need to develop	8
is that it does not	5	there is an urgent need for new	7	thus there is an urgent need to	11	they need to	8
there is a clear need for	5	therefore it is urgent to develop	7	are still urgently needed	11	was to determine whether the	7
is the large number of	5	therefore it is urgent to	7	therefore it is urgently needed to	11	one needs to	7
the main advantage of	5	therefore it is necessary to develop a	7	are urgently needed to improve	10	and the need to	7
one of the main advantages of	5	is an urgent need	7	thus there is an urgent need to identify	10	thus it is difficult to	7
the biggest challenge	4	thus there is a need to	7	therefore it is critical to	10	the need for further	7
pose a challenge to	4	it is necessary to develop	6	thus there is an urgent need for	10	is a lack of	7
we are faced with the	4	one of the biggest	6	thus there is a need for	10	this can lead to	7
CF: Showing controversy within the field							
it is important to note that	16	is a matter of	9	this has led to the	18	is still a matter of	14
it is not surprising that	15	it is not surprising that	7	is still a matter of	16	has been challenged by	10
this is in contrast to	10	by the fact that	6	has been paid to	13	have been raised	7
it should be noted that	8	it should be pointed out that the	6	has been focused on	13	have questioned the	7
are those of the * and do not necessarily reflect the	7	is still a matter of	6	has been limited by the	12	was introduced by	7
are not necessarily endorsed by the	7	therefore it is not surprising that	5	has been focused on the	11	was inspired by the	7
this is in contrast with	6	however it should be noted that	5	has been paid to the	10	was motivated by the	6
it should be noted that the	5	the need for new	5	has been challenging	10	has been questioned	6
are those of the	5	is still under	5	has been hampered by the	9	this raises the	5
this is especially true for	5	it is worth mentioning that the	5	this has led to	8	was inspired by	5
our work is inspired by	5	it is not surprising that the	5	is a matter of	8	it has been debated whether	5
this is also true for	5	has been a subject of	5	has been controversial	7	was also supported by	5
it is important to note that the	4	has been a hot topic	5	there has been growing interest in	7	there is an ongoing debate regarding the	5
it is often the case that	4	has prompted the	5	has been a subject of	6	arises as to what	5
is inspired by the recent work	4	it is important to highlight that	4	has made it difficult to	5	is still under	5
this is not to say that	4	has been a topic of	4	has been challenged by	5	as to what	5
it is important to note that this	3	it should be pointed out that	4	interest in the	5	as to whether	5
is a matter of	3	have been raised	4	has been increasingly recognized	5	there is debate about whether	5
is hard to justify	3	banned the use of	4	remains a matter of	5	arises as to whether the	4
this is in	3	has been hindered by its	4	has been hindered by the	5	was whether the	4
this is in contrast to previous	3	however it should be noted that the	4	is challenged by	5	or whether it is	4
was motivated by the	3	is complicated by the fact that	4	have received much attention	5	has been addressed by several	4
it should be noted however that	3	have been hampered by	4	has been limited by	5	was motivated by the fact that	4
are those of the * and are not necessarily endorsed by the	3	has prompted the search for	4	there has been much interest in	4	that arises is whether the	4
this is motivated by the fact that	3	have necessitated a search for	4	has been devoted to	4	has been debated	4
this is not always the	3	has been hampered by	4	has been hampered by the lack of	4	concerned the role of	4
this is in contrast with the	3	has been recognized as an important	3	has been debated	4	are still a matter of	4
for being a sense repository that often	3	aroused the interest of	3	has been attributed to the	4	there is an ongoing debate	4
is inspired by recent	3	has been driven by the	3	has attracted much attention in	4	have become more	4
is motivated in part by the	3	it should be emphasized that the	3	a matter of	4	however a number of	4
are those of the * and do not reflect the	3	the wide use of	3	has been limited due to	4	has been challenged	4
it should also be noted that	3	a matter of	3	is still debated	4	has been called into	4
this is particularly true for	3	it would not be	3	has given rise to	4	that arises is	4
we are inspired by	3	has been recognized by	3	has recently been challenged by	4	by a recent	3

(Continued)

CL		Chem		Onc		Psy	
in this paper are those of the	3	have limited the	3	has been hampered by	4	it is still a matter of	3
this is supported by	3	has driven the search for new	3	has been given to the	4	are replete with examples of	3
would be true if	3	a search for	3	has been questioned by	4	was put forward by	3
		it should be mentioned that	3	recently there has been	4	motivated by the fact that	3
		it should also be noted that	3	challenged by the	4	can be replaced by	3
		is a growing concern	3	is complicated by the	4	was motivated by two	3
		it must be noted that	3	have attracted much attention	4	has recently been challenged by	3
		has led to the use of	3	has been challenged	4	is a matter of ongoing	3
		it must be emphasized that	3	has been an area of	3	has been challenged by some	3
		it is no surprise that	3	have been hampered by	3	has been subject to	3
		however it was	3	a search for	3	has been adopted by	3
		have limited their	3	efforts towards the	3	remains a topic of	3
		however it is important to	3	there is an ongoing debate	3	arises as to whether	3
		it is important to note that in	3	efforts to develop	3	should be skeptical about the	3
		we reasoned that the	3	interest in the role of	3	was introduced by cite- to	3
		we were intrigued by the	3	there has been significant interest in	3	concerns the nature of the	3
CF: Showing the limitation of the research							
is not a trivial	12	is referred to	8			it is beyond the	9
is still an	9	is referred to the	6			is beyond the scope of this paper	8
is not an easy	7	can be found elsewhere	5			of this paper	8
is not trivial	5	is beyond the scope of this	5			which is the focus of the present	5
is still an open	5	is the focus of this	4			is the focus of the current	5
is an area of	4	are mainly focused on	4			is beyond the scope of this	5
has been the focus of	4	is provided in	4			is the focus of this	5
has been the topic of	4	are not included in this	4			of the current paper	4
is out of the	4	only focus on the	3			this is the focus of the present	4
has been the focus of much	3	we focus here on the	3			is the focus of the present	4
remains an open	3	which are the focus of this	3			not the main focus of the	3
is still in its	3	are discussed below	3			is the topic of the present	3
which is the focus of this paper	3	is mainly focused on	3			is the focus of this paper	3
is beyond the scope of this paper	3	will not be discussed	3			will be the focus of	3
it is not a trivial	3	here we focus on the	3				
has become an active	3	will focus only on the	3				
that is the focus of this paper	3	will be discussed in the following	3				
outside the scope of this paper	3	will be discussed in	3				
		will be discussed later in	3				
		of the present work	3				
		are excluded from this	3				
		briefly described in the	3				
		will be discussed below in	3				
		are referred to	3				
		are not included in the	3				
		can be found elsewhere cite-	3				
CF: Showing the outline of the paper							
is structured as follows	317					in the following	53
of this paper is	262					is as follows	37
the remainder of this	208					is structured as follows	29
the remainder of the	121					is presented in	27
are as follows	115					is shown in	27
is as follows	94					are as follows	24
are presented in	91					were as follows	23
consider the following	75					the remainder of the	17

(Continued)

CL		Chem		Onc		Psy	
of this paper are as follows	64					we conclude with a	14
we conclude in	59					we describe the	14
is presented in	57					is organized as follows	13
of this paper is structured as follows	53					followed by the	13
of this paper are	53					the first is the	12
finally we conclude in	47					in the following we will	12
for future work	46					this is followed by a	11
we make the following	42					we discuss the	11
the related work	40					the remainder of this	11
in the following	39					is depicted in	11
related work in	39					is illustrated in	10
are summarized as follows	37					of this paper is	10
4 presents the	37					the first is	10
are given in	35					consider the following	9
can be summarized as follows	34					in what follows we	9
we conclude with a	33					in the following we	9
consists of two	33					we start with a	9
5 concludes the	33					this is followed by an	8
this paper makes the following	31					in the first step	8
is structured as follows in	30					are summarized in	7
are discussed in	29					addressed in the current	7
we discuss related work in	29					with a brief	7
we discuss the	29					can be summarized as follows	7
6 concludes the	28					in the following we first	7
of this paper is as follows	27					a summary of the	7
consider the following example	26					is divided into two	7
of this work are	26					this is followed by the	7
of this paper can be summarized as follows	26					addressed in the present	7
are the following	25					are described in	7
and future work	25					has two main	7
7 concludes the	24					has two aims	7
consists of three	22					the following two	7
5 presents the	22					this leads to the following	7
there are two main	21					will be addressed	6
finally we conclude the	21					we start with a brief	6
can be divided into two	20					begins with a	6
we first describe the	20					we provide a brief	6
has the following	20					next we discuss	6
we start with a	20					in a first step	6
of this paper	19					first we examined the	6
related work on	19					are discussed in	6
4 presents our	19					had two main	6
Section: method							
CF: Showing reasons why a method was adopted or rejected							
is used to	261	was used to	1331	was defined as the	720	was used to	1402
are used to	151	was used for	1006	was applied to	296	was designed to	111
can be used to	145	was used as a	876	was used to identify	212	was applied to	85
is used for	106	was used as the	773	were selected for	135	was employed to	68
are used as	68	were used for	501	was used to calculate the	129	is used to	64
was used to	63	were used to	485	was employed to	112	cite- was used to	62
is used as the	61	was used for the	478	were used to determine the	91	it is a	52
are used for	60	was used as	473	was used to identify the	89	was developed to	47
is used as a	52	were used as	354	were used to identify	89	was used to examine the	47
was used for	50	were used for the	241	was conducted to	80	can be used to	46
can be used for	41	was used to determine the	231	was used to test the	77	was used to determine the	45
were used to	37	was used as an	149	was defined as the time from	73	was used to analyze the	44
is used for the	37	in order to	138	was used to examine the	68	it is possible to	42
is that it	32	was applied to	129	was performed to identify	63	allowed us to	41
will be used to	31	was performed to	124	was also used to	59	was set to	41

(Continued)

CL		Chem		Onc		Psy	
is then used to	30	was employed to	121	was used to assess the	52	we decided to	41
is used in	30	were used as the	117	was designed to	52	was adopted to	40
is designed to	29	to determine the	107	was considered as	51	was chosen to	38
is that it can	29	is shown in	88	were calculated to	50	were selected to	37
are used in	28	was used in the	87	was considered as a	44	was used to test the	37
are used in the	27	was used to calculate the	84	was adopted to	43	is that the	36
can be used as	26	was used in this	76	was selected as the	43	is designed to	36
has been used to	26	is based on the	74	were used to calculate the	40	was used to identify	34
is used to represent the	25	due to the	73	was used to select	39	was used to examine	33
can also be used to	25	was used to determine	72	was also performed to	38	was found to be	32
can then be used to	24	were used in the	66	is defined as the	37	were chosen to	31
is that it is	22	were used to determine the	64	was calculated to	36	were designed to	31
is used to compute the	22	we used the	58	was used to explore the	36	was used to determine	31
is used to generate	20	was used in	55	was considered as the	35	we chose to	30
are then used to	20	was used to analyze the	53	were selected to	35	was used to assess	30
this allows the	20	were used as a	52	is used to	34	it is important to	28
has the advantage of	20	was employed for the	50	was used to estimate	34	it is necessary to	28
is also used to	19	was used for all	49	were conducted to	34	were employed to	27
is used in the	19	was utilized to	47	were also used to	32	allows us to	26
could be used to	19	was employed for	45	was used to define	31	was also used to	24
can be applied to	18	is used to	43	were used to assess the	31	was used to calculate the	24
are designed to	18	can be used to	43	was applied for the	31	were used to examine the	24
are used for the	18	was used to perform	41	was performed to identify the	31	on the other hand	23
is used as	18	was used to perform the	40	was used to define the	30	was used to explore the	23
is the ability to	17	was used with	40	was defined as the number of	29	was used with	21
can be used as a	17	was used to obtain the	40	was selected for	29	has shown good	21
is useful for	17	was used and the	40	we defined the	29	was utilized to	21
has been shown to be	16	were used to calculate the	40	were used to identify the	29	was set to 005	21
can be used in	16	was used to identify the	39	to explore the	29	was selected as the	20
is that we can	16	was used to obtain	37	the 2 test was used to	28	was considered to be	20
is used to find the	16	was also used to	37	was selected to	28	was chosen as the	20
this allows for	15	was used with a	36	were employed to	27	was used in order to	20
that can be used to	15	was conducted to	36	was applied to determine the	27	it was possible to	20
can be used	15	was performed to determine the	35	and their 95	27	were used to test the	20
is employed to	15	was used to identify	34	were used to examine the	25	we chose to use the	19
CF: Description of the process							
in order to	361	were obtained from	1540	were obtained from	5839	was approved by the	1762
we compute the	133	were recorded on a	958	was performed using	4527	were asked to	864
we need to	94	was obtained from	936	was used to	4153	in order to	478
we used a	94	was purified by	835	were performed using	3535	was carried out in accordance with the	403
this allows us to	84	was performed using	704	was used for	3504	at the end of the	387
we calculate the	83	were performed in	704	was obtained from	2708	were used to	378
we set the	83	was determined by	666	were used for	2278	in accordance with the	352
it is possible to	76	were obtained from the	543	were obtained from the	2203	as well as	324
for each of the	71	was added to the	522	was performed using the	2197	were presented in	314
we would like to	69	were performed using	459	was determined by	2051	was obtained from the	309
we train a	68	was washed with	454	at 4 c	1875	were performed using	300
in addition to	63	were determined by	436	supplemented with 10	1746	were approved by the	295
we create a	61	was performed on a	403	were stained with	1572	was obtained from all	285
according to their	56	was performed on	390	were used to	1559	prior to the	274
it is necessary to	54	was extracted with	387	was added to the	1552	in the present	267
we are able to	53	were dissolved in	383	were washed with	1517	were presented on a	254
to obtain the	52	were washed with	380	was performed by	1409	the number of	246
we want to	50	was performed by	379	was performed with	1369	gave written informed * in accordance with the	234

(Continued)

CL		Chem		Onc		Psy	
we count the number of	48	was performed using a	369	were performed using the	1331	consisted of a	227
need to be	48	was performed in	357	were subjected to	1237	were presented in a	226
to compute the	48	was dissolved in	355	at 37 c in a	1197	they were asked to	222
we train the	48	were carried out in	351	was extracted from	1185	were conducted using	221
we build a	48	were performed on a	317	were cultured in	1171	the order of	211
allows us to	47	was performed using the	315	were counter-stained with	1134	was conducted in accordance with the	207
to capture the	46	were performed using the	311	was used as a	1127	as well as the	207
we use two	46	was obtained from the	289	was performed using a	1055	were used as	194
were asked to	45	was subjected to	269	were performed with	1048	was used for	190
we decided to	44	was carried out using	267	were determined by	1019	in the current	184
we aim to	44	was performed with	262	were as follows	1002	the order of the	177
to predict the	43	were collected from	251	at 37 c	986	are presented in	172
to determine the	43	was determined by the	246	were washed twice with	946	were required to	167
in addition we	43	was obtained by	244	for 1 h at	927	were used for	165
so that the	42	were subjected to	243	are listed in	926	between the two	148
we do not	41	was determined using a	238	was used as	917	in front of the	146
we construct a	41	were added to the	235	was obtained from the	886	in front of a	146
we were able to	41	were obtained on a	230	was performed on	882	are shown in	141
to train the	41	were performed using a	223	at 4 c overnight	878	in the first	140
to get the	41	were obtained by	217	were fixed with 4	852	were presented with a	140
we found that	39	was determined using	212	was confirmed by	795	consisted of two	137
we obtain the	39	was carried out by	194	were seeded in	792	was performed using	137
we also use the	38	was added and the	189	at 4 c with	780	was counterbal-anced across	132
is applied to the	38	were stored at	186	were maintained in	758	was used as a	132
to generate the	37	was obtained as a	183	were used as	757	at the end of each	128
we take the	37	were recorded on	183	were fixed in 4	755	and approved by the	128
are added to the	36	at 37 c	176	were added to the	744	individually in a	126
we can see that the	35	and used without further	176	was performed in	743	of the two	126
based on their	35	was determined using the	175	was determined using the	727	in which the	121
table 3 shows the	35	were prepared in	172	was determined using	720	were performed with	119
in the second	35	for 1 h at	170	was performed to	700	all of the	115
we then use the	35	were obtained us-ing a	170	were plated in	662	were conducted to	114
CF: Using methods used in past work							
based on the	359	according to the	1233	according to the	8240	according to the	353
we use a	358	using the follow-ing	288	as previously de-scribed	2051	is shown in	118
is based on the	318	according to the following	227	as described pre-viously	1504	was based on the	111
is shown in	263	as previously de-scribed	224	was approved by the	1429	can be found in	96
is based on	229	was performed according to the	191	were approved by the	1261	as shown in	89
is given by	216	as described pre-viously	183	was performed as previously described	1253	is based on the	69
we propose a	177	was determined according to the	162	as previously de-scribed cite-	981	was adapted from	65
is as follows	143	as described by	110	was performed as described previously	923	is presented in	58
is defined as	127	as described above	107	was performed according to the	887	was developed by	44
is defined as fol-lows	125	was performed as previously described	84	were performed as previously described	698	was used in this	39
is based on a	104	as described in	82	as described above	670	as described in	39
is illustrated in	99	as previously de-scribed cite-	81	were performed according to the	636	is illustrated in	39
we use the same	87	was prepared ac-cording to	81	was extracted from * according to the	581	adapted from the	37
as described in	87	were determined according to the	77	in accordance with the	570	was adapted from the	37
is similar to the	79	was calculated using the follow-ing	76	were kindly pro-vided by	516	was conducted according to the	36
based on a	74	were prepared ac-cording to the	74	as described pre-viously cite-	465	is depicted in	32

(Continued)

CL		Chem		Onc		Psy	
we describe the	69	by the following	70	was performed as previously described cite-	437	according to the following	30
are based on the	69	was carried out according to the	70	were performed as described previously	406	as suggested by	29
are based on	68	was performed according to	69	as described cite-	398	as described above	29
we use an	67	is in accordance with that reported in	68	were performed in accordance with the	396	are as follows	28
we adopt the	67	was prepared according to the	66	was performed using * according to the	345	is as follows	27
we describe our	65	in accordance with the	63	were performed as previously described cite-	343	as described in the	27
is given in	64	was determined according to	61	was performed according to	328	we adapted the	27
is described in	64	was prepared from	61	was used according to the	324	is similar to the	23
by using the	60	were performed according to the	60	was extracted using * according to the	284	as implemented in the	22
we propose to	60	was performed as described previously	59	were performed according to	274	as recommended by	22
are as follows	57	was conducted according to the	53	were conducted in accordance with the	271	we followed the	21
consists of two	57	was prepared from * according to the	51	was performed as described previously cite-	251	was used in the present	19
is obtained by	57	according to a	50	were described previously	247	were adapted from the	19
we apply the	56	was performed as described by	49	was performed as described	242	as described below	18
can be obtained by	56	were prepared by the	48	was conducted in accordance with the	226	is given by	16
we follow the	55	was prepared using the same	47	and approved by the	219	was the same as in	15
we use a simple	54	11 40 ml was reacted according to	45	as described in	195	were the same as in	15
consists of a	54	as described previously cite-	44	were performed as described previously cite-	183	was calculated using the	14
is similar to	52	as described in the	41	have been described previously	177	we adopted the	14
of the proposed	52	were prepared according to	40	was used to * according to the	176	according to cite-	14
we present a	51	as previously reported	40	were conducted according to the	173	according to this	14
is given by the	46	was calculated by the following	39	was conducted according to the	172	we used an adapted	13
is defined by	44	was extracted from * according to the	35	was carried out according to the	168	was estimated using the	13
using the following	43	were prepared using the	33	as previously reported	167	in the present * we used the	13
is presented in	42	as previously described by	33	were in accordance with the	163	is described in	12
we extend the	42	were performed according to	33	approved by the	160	was adapted from a	12
by using a	42	were obtained from 462 mg 01 mmol of 10-o-propargylated	32	as per the	160	were estimated using the	12
similar to the	40	was determined following the	30	as described in the	159	as described by	12
is depicted in	39	was performed following the	30	has been conducted in * and has been approved by the	157	were adapted from	12
we introduce a	39	were determined according to	29	were performed using * according to the	155	used in previous	11
we propose a novel	39	was extracted using the	29	was performed in accordance with the	154	as in the previous	10
using the same	38	as reported previously	29	as described before	150	as shown in the	10
given a set of	37	was calculated using the	28	were carried out in accordance with the	149	was the same as	10
consists of the following	37	was prepared following the	28	was carried out as previously described	144	as implemented in	10
CF: Showing criteria for selection							
for example the	412	were approved by the	194	p 005 was considered	1034	was defined as the	188
is the number of	304	was approved by the	161	005 were considered	512	were selected from the	99
is the set of	234	were as follows	117	less than 005 were considered	338	was defined as	85
is a set of	203	was defined as the	108	005 was considered	323	is defined as the	78

(Continued)

CL		Chem		Onc		Psy	
can be found in	188	was defined as the lowest	74	p005 was considered	245	was defined as a	73
note that the	178	were selected for	67	were as follows 1	241	were defined as	55
1 is the	168	was selected as the	63	a p value 005 was considered	202	is defined as	40
be the set of	139	was defined as the amount of	52	were considered to be	193	were selected for the	34
the set of	133	was chosen as the	42	of p 005 were considered	174	were selected based on the	34
for example in	116	were selected for the	39	p 005 was considered to be	167	were selected for	33
are shown in	112	and approved by the	39	when p 005	157	were selected from a	29
corresponds to the	104	were selected for further	35	was defined as p 005	154	were defined as the	25
for example in the	101	were selected as the	35	and p 005 was considered	151	were selected based on	25
we denote the	96	were selected as	34	a value of p 005 was considered	150	was defined by the	25
is the total number of	96	were performed in accordance with the	33	of 005 was considered	142	was defined as an	24
is represented as a	94	was defined as	32	was set at p 005	132	were as follows 1	23
we call this	92	were conducted in accordance with the	31	p 005 was considered as	130	is defined as a	21
1 is a	92	was selected for	30	005 were considered significant	128	we selected the	19
this is the	90	is defined as the	30	a p value of 005 was considered	125	were defined as follows	15
we refer to this	90	were selected from the	26	were selected	106	was chosen for the	15
we refer to the	90	were selected and	25	p value 005 was considered	102	was to examine the	13
extracted from the	86	were in accordance with the	24	were selected for further	95	based on the following	13
is represented by a	86	were chosen for	23	a p value of less than 005 was considered	93	was defined as the number of	13
is defined as the	86	was selected for the	22	a p value less than 005 was considered	93	were selected on the basis of	12
corresponds to a	83	were chosen as the	21	of 005 were considered	88	were chosen from the	12
used in the	82	was conducted in accordance with the	20	at p 005	84	was defined as any	12
are extracted from the	81	were defined as the lowest	20	005 were considered to be	78	were selected for this	12
for example consider the	77	was selected as a	20	a p 005 was considered	75	were selected based on a	12
is available at	77	were defined as the	19	005 were considered as	74	defined as the	11
corresponding to the	74	were selected based on the	19	005 was considered significant	74	criteria for the	11
is the set of all	72	used were as follows	18	of less than 005 was considered	72	selected on the basis of	10
this is a	71	were selected based on	18	less than 005 was considered	69	was defined by	10
an example of	70	were seeded at a	17	were also excluded	69	between 18 and 65	10
there is a	68	was selected as	15	were selected based on the	68	were selected from a larger	10
table 1 the	68	approved by the	15	005 was considered to be	64	were defined as those	10
used in our	65	were performed in accordance with	15	with p 005 were considered	61	were chosen based on	10
of the form	64	is defined as the amount of	15	of less than 005 were considered	61	was defined as p 005	9
is drawn from a	63	were plated at a	14	were selected based on	59	are defined as	9
are given in	63	was chosen as a	14	if p 005	57	was selected based on	9
for example a	58	was defined as a	13	p005 was considered to be	57	was based on previous	9
refer to the	58	were chosen for further	13	were selected from	55	was defined using the	9
is the sum of the	56	were randomly selected for	13	005 was considered as	54	were selected based on their	8
be the number of	54	were chosen as	13	of p 005 was considered	53	was chosen for this	8
1 is the number of	54	was chosen for the	13	less than 005 were considered to be	52	was that the	8
is defined as a	53	was chosen based on the	12	were chosen for	47	had to meet the following	8
to denote the	53	were chosen for the	12	we selected the	46	were chosen based on previous	8
is a list of	52	and use of	12	a p value 005 was considered to be	46	were selected for the final	8
0 is the	51	were approved by	12	p005 was considered as	45	we considered the following	8
we refer to	50	were randomly selected and	12	a value of p 005 was considered to be	45	was selected for this	8
word in the	50	used for this	11	of 005 was considered significant	44	was selected for	7

CF: Showing methodology used in past work

(Continued)

CL		Chem		Onc		Psy	
there are several	33	it is possible to	19	is based on the	77	eg cite-	46
there are many	25	is one of the	18	is based on	32	has shown that	32
we consider two	24	have been reported	16	is defined as	26	has been used in	31
have been proposed in the	22	a number of	15	have been described	24	has been shown to be	29
there are a number of	21	have been developed	14	has been shown to	19	has been shown to	28
have been proposed to	20	is the most	12	has been described	18	have been shown to	28
is to use	19	can also be	12	have been described cite-	18	has been used to	27
have been used in	18	have been shown to	12	is based on a	15	have been used to	25
have been proposed	18	the most common	11	is directly proportional to the number of	13	is a widely used	24
have been used to	18	is a widely used	11	is one of the	13	has been used in previous	23
have been used	17	there are several	11	have been reported	12	has been shown to have good	21
has been used in	17	a wide range of	10	it is a	12	it has been shown that	21
is closely related to	16	and can be	10	are referred to as	11	cite- is a	21
is closely related to the	16	has been applied to	10	is one of the most	11	have been shown to be	20
previous work on	16	it is known that	10	have been previously reported	10	have shown that	20
rely on the	16	have been developed to	9	has been used to	10	has been found to be	20
in two ways	15	there are two	9	has been reported	10	is a commonly used	20
is a widely used	14	has been used	9	has been described cite-	9	have shown that the	18
it is well known that	14	more and more	9	which can be	9	has been found to	17
there are many ways to	14	some of these	9	has been previously reported cite-	9	has been shown to have	17
there are two main	14	one of the most	9	we have previously shown that	9	has been validated in	17
have been used for	13	the most popular	9	has been previously reported	9	have been found to be	16
there has been a	12	have also been	9	have shown that	9	in contrast to	16
have been developed	12	has been shown to	9	is deemed the least * that can always yield a unique	9	tend to be	15
have been proposed for	12	is widely used in	9	has been widely recognized and increasingly used by	9	has not been	15
is widely used in	12	a series of	9	is defined as follows	9	it has been	15
have been shown to be	11	a large number of	8	and has been	8	has been widely used in	15
have been widely used in	11	the most widely used	8	are derived from	8	has been shown to be a	14
is commonly used in	10	have been used as	8	is referred to as	8	is one of the	13
in different ways	10	have been used	8	is proportional to the	8	has been reported to be	13
there are some	10	have been studied	8	is a widely used	8	have been reported	13
is known to be	10	has been developed to	8	has been shown to be	8	there are several	12
is a common	10	has been widely used in	7	it has been reported that	8	is referred to as	12
are widely used in	10	can be used	7	we have previously	8	it has been suggested that	12
is known as	9	over the past	7	is given by	8	has been reported	12
a wide variety of	9	have been proposed	7	has been used for	7	many of the	11
there are various	9	have shown that	7	a variety of	7	have been reported to be	11
is a popular	9	need to be	7	was previously reported cite-	7	have been used	11
is different from	9	have been used to	7	and can be	7	the most commonly used	11
is a commonly used	9	have been described in	7	can be divided into	7	there are a number of	10
have been applied to	9	is a commonly used	6	are defined as follows	7	has shown that the	10
are often used in	9	is a common	6	has been previously shown to	7	have been used in	10
have been shown to	9	is widely used for	6	can be formulated as	7	is commonly used in	10
the most commonly used	9	the most important	6	have been shown to be	7	has been used in several	10
there has been	9	have been reported to	6	is a well-known	6	have been widely used in	10
has shown that	8	there is no	6	is a unique	6	is widely used in	9
has been used	8	have been applied to	6	has been widely used in	6	have been used in previous	9
approach is to	8	has been shown to be	6	is the difference between the	6	is a common	9

(Continued)

CL		Chem		Onc		Psy	
widely used in	8	has been successfully applied in	6	is widely used for	6	for example in	9
there are many different	8	have been used for	6	it has been shown that	6	have been found to	9
CF: Showing the characteristics of samples or data							
are included in the	33	are listed in	153	were included in the	388	participated in the	576
included in the	29	were used in this	101	as the mean	374	were included in the	502
in total there are	27	were included in the	60	none of the	328	were recruited from the	293
is divided into two	25	were randomly divided into four	54	were divided into two	314	were excluded from the	284
we split the	24	were considered as	44	were included in this	282	took part in the	253
is divided into	24	served as the	41	are presented as the mean	267	participated in this	243
is split into	20	were randomly divided into two	39	were randomly divided into two	205	were recruited from	234
is divided into three	20	were divided into three	37	were excluded from the	199	a total of	186
are classified as	20	were divided into	35	were classified as	187	with a mean	156
is included in the	19	005 were considered	35	were presented as mean	177	none of the	145
there are a total of	19	were randomly divided into three	35	were repeated at least three	165	were recruited through	134
we divided the	17	were divided into two	35	were divided into four	164	were not included in the	130
can be divided into	17	were randomly divided into five	34	were divided into	160	were recruited from a	128
participated in the	17	were listed in	33	were enrolled in this	158	included in the	107
with a total of	15	of p 005 were considered	31	were divided into three	154	the majority of	102
are more likely to	15	used in this * are listed in	31	was repeated three	151	were excluded due to	100
can be divided into two	15	were used for each	31	were randomly divided into four	143	were recruited via	97
with an average of	14	served as a	30	were randomly divided into three	141	the majority of the	91
are split into	13	were randomly divided into	28	were repeated three	132	most of the	88
the majority of the	13	were randomly divided into six	27	was repeated at least three	126	were divided into two	81
in total the	13	was divided into	26	were randomly divided into	117	were excluded from	79
can be divided into three	13	were excluded from the	26	were performed at least three	108	half of the	79
are divided into	12	were divided into four	26	were used as a	105	at the time of	79
are divided into two	12	was divided into two	25	were presented as the mean	103	were included as	79
there are four	11	were classified as	25	were enrolled in the	101	were excluded from further	78
were excluded from the	10	are described in	22	of at least three independent	100	was composed of	78
was split into	10	were defined as	22	were excluded from this	99	had a mean	74
are not included in the	10	was included in the	21	were used for each	98	at the time of the	74
was divided into	10	were divided into five	21	were performed for each	97	were included in this	70
are included in	9	consisted of two	18	were included in each	97	was included in the	69
contains a total of	9	p 005 was considered as	18	are presented as the means	95	were recruited for the	68
has a total of	9	were used in the present	16	as the means	91	took part in this	62
it consists of two	8	less than 005 were considered	16	at least three times	81	was excluded from the	62
are not included	8	were regarded as	15	were performed three	81	was divided into two	60
were labeled as	8	were randomly divided into 5	14	were randomly assigned to	80	with an average	60
were not included in the	8	which were used in all	14	was repeated in	74	were invited to	59
we randomly split the	8	included in the	13	were recruited from the	71	was included as a	57
we also included	7	were divided into six	13	were repeated three times	71	or corrected to	57
to be included in the	7	were included in	13	were included as	71	were recruited through the	52
is not included in the	7	were randomly assigned to	13	served as the	69	were recruited at the	52
tend to be more	7	is summarized in	12	are presented as the	67	were enrolled in the	51
is divided into four	7	were randomly divided into a	12	were randomly divided into 4	67	there were three	48
we found that in	7	were taken into account	12	were repeated in	61	there were four	48
there are about	7	was divided into four	12	none of these	60	were recruited by	47
were removed from the	7	are listed in the	12	from at least three	58	participated in the present	46
is split into two	7	with p 005 were considered	11	from at least three independent	57	were divided into three	46
out of these	7	as follows s	11	at least three	57	were recruited to	46

(Continued)

CL		Chem		Onc		Psy	
were randomly selected from	7	of less than 005 were considered	11	was divided into two	56	were included in the final	45
two of the	7	were included in each	11	were divided into 3	56	recruited from the	45
should be included in the	7	alone were used as	11	were repeated at least three times	55	were recruited via the	44
Section: result							
CF: Reference to tables or figures							
are shown in	1576	are shown in	4476	as shown in	9803	are presented in	1425
table 3 shows the	698	as shown in	2857	are shown in	3816	are shown in	1377
table 1 shows the	688	is shown in	1661	are summarized in	1503	cite- shows the	747
table 4 shows the	532	are presented in	1634	is shown in	1290	as shown in	526
as shown in	481	are summarized in	1211	are presented in	982	cite- presents the	505
are presented in	471	are listed in	794	were shown in	944	are reported in	459
is shown in	441	are given in	511	cite- shows the	851	are displayed in	373
are shown in table 1	372	is presented in	421	are listed in	743	are summarized in	356
table 5 shows the	359	were shown in	352	as shown in fig	663	as can be seen in	278
we can see that the	335	are reported in	330	cite- shows that	472	is shown in	270
are given in	319	is an important	277	was shown in	394	cite- displays the	252
we can see that	293	it has been reported that	256	were summarized in	373	see table cite-	241
are reported in	290	cite- presents the	247	is presented in	273	can be found in	223
6 shows the	229	it can be seen that the	238	were listed in	255	is presented in	208
results on the	208	it is known that	233	as shown in the	230	are depicted in	184
are shown in table 4	200	it is well known that	227	cite- shows that the	211	can be seen in	155
4 shows the	168	it should be noted that the	226	are reported in	187	cite- shows that the	128
can be found in	150	shows that the	222	is summarized in	183	are given in	127
we show the	142	cite- summarizes the	215	as seen in	180	are provided in	123
are shown in table 3	141	cite- a shows the	209	are described in	173	are listed in	99
are listed in	137	as can be seen in	197	were presented in	164	are illustrated in	95
table 1 the	137	is based on the	183	are provided in	141	cite- shows that	79
table 3 the	137	are depicted in	177	are given in	140	as can be seen from	63
we present the	134	are displayed in	175	cite- presents the	138	cite- show the	59
table 3 shows	132	it should be noted that	166	are displayed in	134	can be found in the	51
are summarized in	123	can be attributed to the	164	cite- shows a	129	as can be seen	50
table 7 shows the	121	are illustrated in	164	were obtained in	124	cite- summarizes the	49
shows the number of	118	is illustrated in	149	are illustrated in	111	is depicted in	48
table 2 presents the	116	was shown in	137	were showed in	110	are plotted in	48
table 1 shows	115	is depicted in	135	are depicted in	105	cite- shows a	47
are summarized in table 1	110	is one of the	132	cite- showed the	102	is illustrated in	47
table 1 presents the	107	it has been reported that the	127	can be found in	97	is provided in	46
is shown in table 1	106	in cite- the	124	as shown in * was observed in	94	as seen in	45
5 shows the	106	is related to the	121	as indicated in	92	is displayed in	44
we can see that our	105	it is clear that the	120	as demonstrated in	89	are presented in the	43
are shown in table 5	104	as seen in	119	cite- summarizes the	88	cite- provides the	42
table 4 the	104	is summarized in	119	as depicted in	85	cite- contains the	39
is given in	99	this is in	114	are shown in the	84	are shown in the	37
table 4 shows	98	cite- illustrates the	114	as presented in	82	as can be seen in the	34
are shown in the	91	as can be seen from	112	is depicted in	81	as shown in the	34
with and without	90	it can be observed that the	112	is illustrated in	79	were presented in	32
results for the	90	can be explained by the	112	as showed in	79	we present the	32
table 3 presents the	88	it is well known that the	111	as shown in * we found that	69	cite- for the	32
we describe the	86	were summarized in	109	this is in	69	is summarized in	32
table 2 summarizes the	82	it can be seen that	108	were obtained with	67	as depicted in	31
are shown in table 6	79	it is known that the	105	cite- displays the	66	are reported in the	30
it can be seen that	75	can be used to	104	is provided in	65	were shown in	30
it can be seen that the	75	is given in	104	was presented in	64	are described in	28
in table 1	74	can be found in	103	was summarized in	60	is reported in	28
as can be seen in	73	is due to the	102	were displayed in	60	as can be seen the	25
CF: Restatement of the aim or method							
we use the	2470	in order to	730	was used to	1452	in order to	309
we used the	1524	was used to	707	to determine the	1447	was used to	211
- 2 -	625	was used as a	448	to determine whether	1409	we conducted a	166
we use a	620	was determined by	370	in order to	1369	we used the	165
based on the	471	were used to	340	the role of	1337	based on the	153

(Continued)

CL		Chem		Onc		Psy	
in order to	391	to determine the	301	we examined the	1207	were used to	146
we use the same	389	were used as	286	was confirmed by	1114	were included in the	135
we used a	383	were determined by	270	we next examined the	714	to examine the	122
we set the	344	was used as the	238	were used to	700	to test the	105
as described in	296	was used as	183	was performed to	685	was conducted to	97
according to the	280	was used for the	178	we performed a	678	we performed a	96
is used to	247	was selected as the	173	to examine the	651	we used a	83
as well as	223	were subjected to	172	to explore the	630	was conducted on the	83
is used for	213	was chosen as the	167	we used the	611	we examined the	81
is the number of	209	was used for	158	based on the	611	were conducted to	81
is based on the	206	we examined the	151	to test this	548	were excluded from the	78
we used the same	202	was performed to	151	to determine whether the	547	was performed on the	74
we compute the	172	was determined by the	135	was determined by	481	for each of the	70
we follow the	170	were used for	134	we used a	479	was conducted on	60
as well as the	168	was subjected to	132	was used as a	453	were used as	56
we use two	166	was carried out	127	we determined the	444	was conducted with	55
we train the	164	we used the	125	and found that	437	were included as	54
are used for	162	was performed by	125	with or without	434	was used as a	53
is based on	156	was used to determine the	122	were subjected to	433	in addition to	50
was used to	138	were used for the	122	in addition we	430	were entered as	48
in addition to the	137	was carried out by	119	to test the	407	was performed to	48
a set of	136	was performed using	118	were included in the	405	were asked to	47
is used as the	135	was added to the	113	we first examined the	403	we ran a	47
was used for	130	in order to determine the	112	were confirmed by	399	were removed from the	46
are used as	130	were prepared by	112	to determine if	395	was used as the	45
we split the	124	was applied to	104	were used as	388	was included as a	45
we train a	122	were used as the	102	next we examined the	387	in addition we	44
used in the	121	to examine the	100	to identify the	363	were performed to	42
are used to	121	to explore the	99	to this end we	341	were performed on the	40
for each of the	120	we determined the	96	we next examined whether	320	were conducted on the	40
trained on the	118	was carried out using	95	we further examined the	313	was performed on	39
were used for	115	was determined as	94	to validate the	296	to this end we	39
is trained on the	110	by using the	94	was performed in	288	was based on	39
with the same	107	was applied to the	92	were performed to	281	we examined whether the	39
using the same	107	in order to further	88	we examined whether	276	we predicted that	38
we apply the	106	was based on the	88	were divided into two	270	we decided to	38
in the first	106	were characterized by	84	was performed on	269	to explore the	37
we consider the	105	was employed to	82	was examined by	263	we compared the	37
we train our	104	to determine whether	82	to understand the	262	focused on the	36
in addition we	103	to identify the	82	to examine whether	262	was applied to the	36
we use the following	103	were performed to	81	a total of	257	we conducted a 2	36
is defined as	97	were selected for	80	the ability of	256	was conducted to examine the	36
by using the	95	was obtained by	80	we then examined the	254	we conducted an	35
we adopt the	94	was carried out to	80	was performed using	252	to determine whether the	35
in addition to	93	was evaluated by	79	we compared the	249	on the basis of	35
CF: Description of the results							
we found that the	208	showed that the	985	we found that	4273	there was a	480
we found that	195	compared to the	753	showed that the	2343	was not significant	380
we find that	192	in addition the	660	was observed in	1722	showed that the	362
we find that the	178	due to the	641	we found that the	1642	there were no significant	357
show that the	151	as well as	586	compared to the	1564	there was no significant	352
we observe that the	137	indicated that the	585	as well as	1205	there was a significant	351
compared to the	135	was obtained as a	570	we observed that	1075	p 0001 and	343
achieves the best	117	on the other hand the	562	compared with the	1057	revealed a significant	290
the average number of	117	was observed in the	560	was observed in the	1013	there was no	272
we see that the	102	was found to be	519	the number of	1011	compared to the	254
we observe that	96	were observed in the	506	as compared to	948	showed a significant	245
indicates that the	87	on the other hand	422	revealed that the	907	there was a main	229
there is no	86	revealed that the	419	were observed in	898	there was also a	228
the total number of	84	were found to be	418	but not in	816	indicated that the	225

(Continued)

CL		Chem		Onc		Psy	
most of the	79	was observed in	397	there was no	809	revealed a significant main	215
there is a	69	it was found that the	341	in addition the	720	were not significant	212
is able to	61	compared with the	335	indicated that the	715	none of the	210
indicate that the	58	led to the	331	was significantly higher in	704	revealed that the	206
we note that the	58	we found that the	323	did not affect	639	were found for	175
showed that the	57	was confirmed by	322	were observed in the	610	was found between	169
is significantly better than	57	was the most	317	showed a significant	610	in terms of	164
we observed that	55	as well as the	308	in contrast the	561	revealed a main	157
are due to	55	in contrast the	307	did not affect the	556	were found between	149
none of the	53	in the present	306	was found in	548	revealed a significant main effect of	148
better than the	53	was observed for	293	resulted in a	535	we found that	146
it shows that the	52	the number of	292	we also found that	515	we found a	141
show that our	52	at the same	288	was found to be	508	was a significant	140
we observed that the	51	we found that	288	have shown that	505	was found to be	133
achieved the best	51	in the range of	286	were found to be	496	there was no main	131
in the number of	50	it was found that	281	there was no significant	485	were found in the	127
we observe a	48	indicating that the	280	p 0001 and	461	was found in the	126
is the best	46	in the presence of	269	has been shown to	456	there were no	125
the best performing	46	resulted in the	265	resulted in a significant	449	was found for	125
we are able to	44	showed the highest	263	we observed a	449	showed a significant main	124
is better than	43	were found in the	261	we observed that the	445	we found that the	122
we also find that	42	corresponding to the	259	were found in	442	between the two	120
is not significant	42	were found in	254	there was a	439	in addition the	117
performs better than	41	were observed in	251	none of the	438	showed no significant	113
performs better than the	41	on the other	249	on the other hand	436	p 005 and	112
significantly better than	41	did not show any	248	demonstrated that the	434	did not differ between	112
suggests that the	41	resulted in a	224	it has been reported that	431	was also significant	112
by a large	40	of the two	223	was observed between	408	had a significant	109
we see a	40	most of the	222	at the time of	408	p 0001 and the	108
we note that	40	the addition of	220	there were no significant	391	the majority of	107
we find that our	39	with respect to the	218	was found between	387	than in the	106
it shows that	37	than that of	212	as compared to the	372	p 001 and	106
means that the	37	as compared to	210	there was no significant difference in	370	with respect to	103
we also found that	36	was found in the	209	as well as the	363	was found for the	100
we obtain a	36	was observed for the	207	fig cite- and	358	there was no significant difference between	99
achieves the highest	35	as compared to the	204	there was a significant	347	revealed a main effect of	96
CF: Describing interesting or surprising results							
is that the	305	it is interesting to note that	70	the most common	249	on the other hand	73
for example the	219	it is interesting to note that the	60	interestingly we found that	198	note that the	59
on the other hand	177	it is worth mentioning that	36	of note the	90	as expected the	58
on the other hand the	142	it is worth mentioning that the	29	in particular the	73	is that the	56
this is because the	103	it is interesting that	29	interestingly we observed that	70	it should be noted that	54
this is because	93	it is interesting that the	25	interestingly we found that the	66	on the other hand the	51
in contrast the	84	interestingly we found that	17	interestingly we observed a	55	for example the	50
it is worth	77	it is remarkable that	16	for example the	45	it should be noted that the	50
it is difficult to	65	interestingly in the	12	interestingly there was a	36	it is important to note that	43
on the other	64	it is not surprising that	12	in fact the	36	on the other	37
can not be	59	it is not surprising that the	11	in line with this	35	for example one	37
it is interesting to note that	52	it was interesting that	10	interestingly we observed	32	a number of	35
there are several	51	is the fact that	9	more importantly the	30	it is possible that	31
as expected the	50	it is also worth	9	the most frequently	28	it is possible that the	31

(Continued)

CL		Chem		Onc		Psy	
this is due to the	49	it is also interesting to note that the	9	interestingly we found	28	it is important to	28
it is important to	47	this is not	7	similarly in the	28	the most common	27
in particular the	47	it is remarkable that the	7	interestingly we observed that the	27	the importance of	27
for example in the	45	it was notable that the	7	surprisingly we found that	25	it is interesting to note that	26
in fact the	44	it is worth noticing that	7	was even more	23	it is important to note that the	26
this is an	42	it is also interesting to note that	7	interestingly in the	22	it is worth	25
this is due to the fact that	42	interestingly we found that the	7	one of the most	21	in particular the	24
it is interesting to note that the	42	interesting to note that	7	notably we found that	21	for example a	22
it seems that the	42	it was interesting to note that	7	interestingly we found a	19	on the one hand	21
it should be noted that	40	is the presence of the	6	interestingly we also observed	19	seems to be	21
it should be noted that the	39	it is interesting to observe that	6	importantly we found that	19	in fact the	20
it is clear that	39	to find that the	6	first we found that	19	for example in the	20
it is possible that	38	this is not surprising as	6	is that the	18	many of the	20
this is not	37	is the fact that the	6	was also observed when	18	it is also	20
seems to be	36	it is interesting to note	6	notably there was an	18	it seems that	20
in general the	35	it was not surprising that	6	the most important	17	it is clear that	19
what is the	35	as a matter of fact the	6	moreover the number of	17	it is likely that	19
for this is that	34	is presumably due to the	6	the most commonly	17	for example in	18
tend to be	33	of note the	5	moreover in the	16	seemed to be	17
there are many	33	it is interesting to note the	5	was even more pronounced	16	on the contrary	16
for this is that the	33	therefore it is not surprising that the	5	interestingly we also found that	15	this is not	16
a large number of	33	it is also worth mentioning that	5	for example in the	15	as would be expected	15
is that it	33	unfortunately none of the	5	this was also the	15	however it should be noted that	15
there are some	32	one of the most interesting	5	for example in	15	it is interesting that	15
for example a	32	it is very interesting that the	5	similar to our	14	this is the	14
it is important to note that	31	more importantly the	5	was also largely	14	the fact that	14
it is possible that the	31	interestingly none of the	4	intriguingly we found that	14	it is interesting to note that the	14
in contrast to	30	interestingly there is a	4	a similar trend was observed in	14	it seems that the	14
this is due to	30	it was notable that	4	was enriched in	13	the most important	13
this is due to the fact that the	30	contrast to the	4	is their ability to	13	it is also worth	13
it is important to note that the	30	interestingly we found a	4	of note in	13	for the first time	13
many of the	29	it is of interest that	4	interestingly in contrast to	12	to cope with	12
this means that the	29	is not surprising	4	in particular in	12	the fact that the	12
it is interesting to	29	it is interesting to note that despite	4	importantly we observed that	12	this is a	12
it is hard to	28	is not surprising because	4	notably we found that the	12	it should be noted that this	12
is that our	28	it is interesting to note that these	4	most importantly the	12	it is also possible that	12
CF: Comparison of the results							
we compare our	38	it could be seen that the	28	it can be seen that	17	it can be seen that	37
table 3 compares the	37	cite- compares the	18	it can be seen that the	9	it can be seen that the	28
we compare the	22	one can see that the	16	we can see that the	8	we can see that	17
table 1 compares the	21	it could be seen that	14	an example of the * is shown in	4	it can be observed that	13
it can be observed that the	20	one can see that	8	a search of the	4	we can see that the	12
in table 5 we	15	and this included	5	a search for	4	we now turn to	11
table 3 compares our	14	for the 50 different	5	is shown for	4	we see that the	11
it can be observed that	14	it was possible to observe that the	3	comparison of the mean of each * multiple comparisons test indicated that	4	it can be observed that the	10
with previous work	13	it can be seen that there is no	3	revealed that the * was significantly lower in the	4	point showed that at time 3 f1 3	10
it has been shown that	13	it is possible to notice that	3	indicating there were no substantial	4	we report the	8

(Continued)

CL		Chem		Onc		Psy	
comparison on the	13	it was possible to observe that	3	an example of a * is shown in	3	we now turn to the	7
our approach with two	12	it can be seen that the addition of	3	highlights the role of	3	we will focus on the	6
with that of the	12	it can be seen that there is a	3	did not suggest inconsistency between	3	and the 4757	6
our approach with the	11	shows that this is due to the fact that this	3	can be seen from	3	we will return to this	5
with previous work on	10	and rm2test for the 50 different	3	one can see that	3	it can also be seen that	5
with those obtained by	10	one can see that there is	3	levels that were the most strongly up- or * is shown in	3	it can be seen that all	5
our approach with	9	mgkg it was observed that this	3	shows the presence of	3	we return to this	5
our approach to the	9	and it can be seen that the	3	showed a main effect of	3	did not reveal a significant difference	5
table 5 compares the	8	it can be seen how the	3			we start by	4
with three other	8					let us consider the	4
with those of	7					for ease of	4
we also compare our	7					there was an * cite- for result of other	4
our approach with the following	7					it can clearly be seen that the	4
table 1 compares our	7					in the following we will	4
we compare our proposed	7					we will discuss the	4
with two other	7					we discuss the	4
comparison we use the same	7					first we present the	4
of our approach with	6					it could be seen that	4
it is shown that	6					it can observe the	4
are comparable to	6					we turn now to	3
related work on the	6					it is important to see	3
comparison we adopt the same	5					to allow for a	3
comparison between the	5					see cite- can be found in the	3
is shown for	5					it can be seen that for	3
results comparing the	5					reveals that the	3
with the existing	5					it becomes clear that	3
with the recent	5					we can observe that	3
table 4 compares our	5					we will first	3
with those reported in	5					finally we also	3
this can be seen as a	4					it can be seen that most	3
with several existing	4					one can see the	3
table 5 compares our	4					cite- plots the	3
to two other	4					we can see how the	3
table 7 shows a	4					can be gained by	3
generated using the same * types are shown	4					however looking at the	3
as a point of	4					we turn to the	3
in table 8 we	4					in the following we present	3
each of which uses a single	4					to summarize the	3
we compare the proposed	4					is provided by	3
it can be noticed that	4					weights indicated a significant	3
CF: Summary of the results							
this suggests that the	186	indicate that the	342	taken together these	2246	this suggests that	127
this indicates that the	145	suggest that the	316	suggest that the	903	suggest that the	113
this suggests that	133	this indicates that the	203	indicate that the	631	this suggests that the	113
this indicates that	113	this suggests that the	184	taken together our	475	this indicates that	88
this shows that	95	show that the	182	this suggests that	410	this indicates that the	79
this shows that the	94	this suggests that	121	this indicates that	239	indicate that the	67
this suggests that our	48	this indicates that	113	strongly suggest that	200	this means that	65

(Continued)

CL		Chem		Onc		Psy	
this indicates that our	47	suggests that the	100	this suggests that the	196	this means that the	65
this shows that our	40	taken together these	93	demonstrate that the	194	suggests that the	47
suggest that the	30	this result indicated that	78	taken together the	174	taken together these	32
this demonstrates that the	28	this result indicated that the	68	show that the	173	taken together the	29
this demonstrates that	25	demonstrate that the	60	we conclude that	151	in summary the	29
this demonstrates the	20	based on these	50	this indicates that the	123	in sum the	29
this suggests that a	19	it seems that the	47	in summary these	111	this indicated that the	20
we conclude that	18	are in agreement with previous	44	suggests that the	90	this shows that	19
this demonstrates that our	18	this shows that the	43	taken together these * suggest that the	79	this shows that the	19
this confirms that	16	this means that	33	may contribute to the	66	demonstrate that the	16
this confirms that the	13	taken together the	31	thus we conclude that	64	we can conclude that the	15
this suggests that for	12	confirm that the	30	taken together these * indicate that the	64	this indicated that	14
this shows that a	12	we suggest that the	30	we conclude that the	63	are in line with the	13
this indicates the	11	are in accordance with the	30	strongly suggest that the	63	this would suggest that	12
we conclude that our	9	this suggested that	26	clearly indicate that	61	it appears that	12
this suggests that there is	8	we can conclude that	26	may contribute to	59	we conclude that the	12
this confirms our	7	we conclude that the	26	all together these	58	in sum these	12
thus we conclude that	7	this suggested that the	25	may be a	52	we can conclude that	11
in summary we can conclude that	7	we speculate that the	23	further support the	51	provide support for the	11
this indicates that when	7	are in accordance with	23	clearly demonstrate that	50	confirm that the	10
we thus conclude that	6	taken together our	22	and that the	48	are in line with	10
are in line with	6	in conclusion the	20	confirm that the	47	it shows that the	10
this suggests that in	6	we propose that the	20	strongly indicate that	47	are in line with previous	10
in summary our	5	this indicates the	20	therefore we conclude that	46	imply that the	10
this shows that by	5	this confirms the	19	suggest that a	45	provide partial support for	10
this indicates that using	5	this may suggest that	19	suggest that both	44	support the idea that	9
seem to suggest that the	5	this shows that	18	support the idea that	44	this confirms that the	9
this suggests that using	5	clearly show that the	18	is sufficient to	43	this suggests a	9
this suggests that we	5	we suggest that	17	in summary our	43	support for the	9
this supports our	5	are in accordance with previous	17	may be involved in the	43	therefore we can conclude that	9
this example shows that	5	are in line with the	17	is capable of	42	this supports the	9
we therefore conclude that	5	we conclude that	16	taken together these * strongly suggest that	41	suggests that a	8
this suggests that most of the	4	also suggest that	16	together suggest that	41	we concluded that the	8
this supports the	4	we believe that the	16	is critical for the	36	provide support for	8
seem to suggest that	4	may indicate that the	15	taken together we conclude that	36	in summary these	8
this further suggests that	4	clearly show that	15	and suggest that	36	is in line with the	7
this suggests that most	4	clearly indicate that	14	suggest that in	36	show that both	7
suggest that our	4	imply that the	14	taken together the above	36	this indicates a	7
demonstrate that the proposed	4	we speculate that	14	clearly show that	35	this confirms that	7
this confirms that our	4	on the basis of these	13	indicate that in	35	this indicates that a	7
thus we conclude that the	4	these indicated that	13	strongly suggested that	34	this means that in	7
clearly show that the	4	it suggests that	13	this demonstrates that	33	we conclude that	7
thus we can conclude that	4	this suggests a	13	clearly demonstrated that	33	suggest that both	7
Section: discussion							
CF: Suggestion of hypothesis							
can be used to	67	suggest that the	163	in conclusion our	802	this suggests that	320
this suggests that	50	indicate that the	108	suggest that the	562	suggest that the	303
this suggests that the	44	suggested that the	66	in summary our	540	this suggests that the	286
suggest that the	40	this suggests that	64	this suggests that	430	this is the first	188
we can see that	34	it can be concluded that	62	taken together our	365	suggests that the	140
this indicates that the	34	it can be concluded that the	61	indicate that the	327	indicate that the	114

(Continued)

CL		Chem		Onc		Psy	
we can see that the	33	suggests that the	61	taken together these	307	this indicates that	104
this means that the	32	this suggests that the	48	we show that	270	this indicates that the	89
suggests that the	27	suggesting that the	47	here we show that	262	taken together these	75
indicate that the	22	taken together these	37	in conclusion this	215	support the idea that	67
we can conclude that	22	we conclude that	32	we demonstrate that	214	we suggest that the	67
we conclude that	22	this indicates that	27	suggested that the	183	we suggest that	63
this shows that the	21	we can conclude that the	27	this suggests that the	182	in conclusion the present	62
can be used for	21	we believe that the	27	suggests that the	176	we conclude that the	51
this allows us to	21	demonstrate that the	26	we speculate that	169	in summary the present	47
this indicates that	21	we can conclude that	26	we propose that	160	taken together the	46
this shows that	20	taken together our	25	in conclusion we	157	in sum the present	46
this means that	18	this indicates that the	25	this indicates that	145	supports the idea that	45
we can conclude that the	17	it could be concluded that	24	show that the	140	it can be concluded that	44
we conclude that the	17	we suggest that the	23	here we demonstrate that	133	we conclude that	43
it is clear that	15	may be a potential	22	in conclusion the present	131	we propose that the	42
can be used as a	12	we speculate that the	21	we suggest that	123	taken together our	42
indicates that the	12	we speculate that	20	in conclusion the	122	is the first to	40
we argue that	10	we conclude that the	19	in summary this	121	we propose that	39
we believe that these	10	it is concluded that	18	in summary we	115	do not support the	39
it is clear that the	10	we concluded that	17	in conclusion we have	104	in sum our	38
demonstrate that the	10	we suggest that	17	we propose that the	102	in conclusion our	37
can be used as	10	which suggests that the	16	we conclude that	102	we can conclude that the	37
can be viewed as a	10	this indicated that	15	we speculate that the	101	provide support for the	36
this demonstrates that	10	we concluded that the	15	in conclusion we demonstrated that	100	this is the first study to	34
it can be	9	suggest that these	15	demonstrate that the	100	in summary this	33
also suggest that	8	taken together the	14	strongly suggest that	96	in sum the	33
this enables us to	8	we propose that	14	in addition our	95	we can conclude that	32
can be viewed as	8	we believe that	14	based on these	90	in summary our	31
it seems that the	8	it appears that the	14	in conclusion we found that	87	also suggest that	31
based on these	8	may be a promising	13	we believe that the	87	support the view that	30
this suggests that our	8	which indicates that the	13	in summary we have	77	also suggest that the	29
we argue that the	7	is a potential	12	we believe that	76	our results suggest that	28
this means that our	7	we propose that the	12	here we demonstrated that	76	support the idea that the	27
suggest that a	7	based on these	12	may serve as a	73	this supports the	27
this indicates that our	7	it was concluded that the	11	we hypothesized that	72	provides the first	27
can be regarded as	7	may serve as a	11	this suggested that	72	this shows that the	26
it provides a	7	it is speculated that	11	this indicates that the	71	it can be	26
this shows that our	7	which suggests that	11	we suggest that the	66	this suggests that a	26
this suggests that a	7	this suggested that	11	in summary the present	63	in summary the	26
this suggests the	7	reveal that the	10	we show here that	62	highlight the importance of	26
seem to suggest that	7	we believe that this	10	therefore we suggest that	62	suggest that a	26
it is likely that	7	might be a potential	10	in summary we demonstrated that	62	thus it appears that	24
it appears that	7	therefore we propose that	10	in summary we have shown that	60	thus it seems that	24
can be considered as	7	also suggest that	10	it is believed that	59	it can be concluded that the	23
CF: Restatement of the results							
show that our	128	in the present	504	we found that	3635	we found that	489
are shown in	81	showed that the	352	in the present	1692	we found that the	255
show that the	70	as well as	325	as well as	1293	showed that the	249
showed that the	61	we found that	283	we found that the	1083	cite- found that	163
note that the	47	was found to be	159	this is the first	675	revealed that the	121
as shown in	45	based on the	148	we also found that	664	was found to be	121
show that the proposed	36	we found that the	147	we observed that	661	we also found that	114
is shown in	34	revealed that the	142	we showed that	657	were more likely to	104

(Continued)

CL		Chem		Onc		Psy	
table 3 shows the	34	according to the	138	in addition the	646	indicated that the	94
table 4 shows the	33	were found to be	136	showed that the	617	we found a	88
showed that our	31	it was found that	127	was observed in	546	it was found that	79
are presented in	29	was used to	118	et al found that	537	were found to be	73
we observe that the	28	for the first time	118	was found to be	532	was related to	69
table 5 shows the	26	compared to the	117	we demonstrated that	522	they found that	67
we note that the	22	was found to	114	was shown to	478	was found in the	66
it is interesting to note that	21	it was found that the	106	in addition to	477	it is interesting to note that	64
was found to be	21	on the other hand the	97	in the current	472	was observed in the	61
it is worth	20	most of the	97	et al showed that	451	we did not find any	60
table 1 shows the	20	led to the	87	on the other hand	447	there was a	57
it should be noted that the	19	was the most	79	was found to	391	was found between	57
show that our proposed	19	was able to	77	to the best of our	375	we also found	56
as can be seen in	19	on the other hand	76	the number of	330	there was no	55
it is interesting to	18	in summary we have	76	as well as the	326	was found for	55
showed that the proposed	17	we demonstrated that	74	were found to be	313	showed a significant	54
we note that	16	we observed that	74	based on the	308	it is important to note that	53
shows that our	15	resulted in the	73	revealed that the	302	were found in the	51
it is also worth	15	were used to	73	according to the	300	we also found that the	51
are shown in table 4	15	involved in the	72	and found that	298	were related to	50
it is important to note that the	15	in terms of	71	but not in	291	we did not find a	49
6 shows the	15	was observed in	71	here we found that	284	cite- found that the	49
are shown in table 1	14	in order to	70	in terms of	267	did not differ between	48
it is also interesting to	14	was shown to	70	on the other	267	than in the	48
it should be noted that	14	in addition to	68	as shown in	261	was found to	46
it is important to note that	14	show that the	67	was reported to	257	we also found a	46
is better than	12	in conclusion the	67	et al demonstrated that	253	also showed that	45
of up to	12	was observed in the	64	was found in	253	was found in	45
we can observe that the	12	were found to	63	due to the	250	were found to	45
as can be seen	12	it was observed that	63	was shown to be	237	was found for the	45
achieved the best	12	of the present	61	compared to the	236	and found that	44
have shown that the	12	showed the highest	61	in order to	235	were observed in the	44
are given in	12	the number of	60	were observed in	220	we observed that	42
are shown in table 3	12	we have shown that	60	we observed that the	213	it is important to note that the	42
as shown in the	11	depending on the	58	was able to	209	for example cite- found that	41
are shown in table 5	11	were the most	57	we showed that the	207	were found for	41
is found to be	11	of the two	56	was observed in the	206	it was found that the	41
of our proposed	11	in the current	56	we confirmed that	206	were found in	40
are found to be	10	in summary the	55	is able to	202	did not show any	40
achieves the best	10	shows that the	54	we also observed that	201	specifically we found that	40
we can observe that	10	in this work	54	we observed a	196	we found no	39
it has been shown that	10	in this work we	52	we also showed that	196	was observed in	38
CF: Comparison of the results and past work							
this is in	9	this is the first	105	this is in	282	this is in	266
is based upon work supported by the	9	et al reported that	83	in contrast to	132	is in line with the	191
is based upon work supported in part by the	8	was confirmed by	71	in contrast to the	124	are in line with the	169
is in line with	7	also showed that	47	also demonstrated that	115	are in line with previous	149
is supported by the	6	confirmed that the	39	similar to the	109	are in line with	137
this is in contrast to the	6	were confirmed by	37	in line with this	107	is in line with	135
this is similar to the	6	than that of	37	are in line with the	103	in contrast the	125
is similar to the	6	similar to the	36	in line with these	85	is in line with previous	109
is based in part on	6	reported that the	33	are in agreement with the	72	in contrast to the	91
is in line with the	5	et al showed that	27	are in agreement with previous	72	in line with the	84
in line with the	4	are in agreement with the	26	et al also reported that	69	is supported by the	82
are in line with	4	are in agreement with	25	it was also reported that	68	is consistent with the	74

(Continued)

CL		Chem		Onc		Psy	
this is in contrast with	4	than that of the	24	in agreement with the	66	in contrast to	69
is supported by the fact that	4	also demonstrated that	24	are in line with previous	66	in line with this	68
with previous work	3	et al reported that the	23	in agreement with this	63	in accordance with the	51
this is similar to	3	also showed that the	23	similar to our	62	is supported by	47
is different from the	3	was confirmed by the	22	in accordance with the	62	in line with our	45
is compatible with	3	in accordance with the	21	in agreement with our	60	are consistent with the	44
this is comparable to the	3	et al demonstrated that	21	is in line with the	57	the idea that	41
is based upon work supported by	3	also indicated that the	20	also reported that	57	is similar to the	38
are comparable with the	3	this is in	20	is in agreement with previous	56	in line with previous	37
which is in accordance with	3	was similar to the	19	also found that	55	is also consistent with the	37
is confirmed by	3	which indicated that the	19	is in agreement with the	55	according to this	37
this is in contrast to	3	is in agreement with the	19	are in agreement with	54	in line with	35
this corresponds to a	3	we confirmed that	18	in line with the	49	with the idea that	32
are in line with the	3	also reported that	18	are in line with	48	this is in contrast to	32
in line with previous work	3	similar to that of	18	in line with our	48	this is also in	30
this corresponds to the fact that	3	is in agreement with	17	in accordance with previous	46	is compatible with the	29
are consistent with the	3	higher than that of	17	is similar to the	45	by contrast the	29
		are in accordance with	17	in contrast to our	39	in line with these	28
		et al indicated that	17	was similar to that of	36	this is supported by the	27
		is supported by the	16	is in line with	36	is in line with other	26
		are in accordance with the	16	is in agreement with	36	is also supported by the	26
		was higher than that of	16	in line with previous	35	is also consistent with	25
		than those of	16	cite- we found that	34	who found that	25
		well with the	16	are in accordance with previous	33	are in accordance with the	25
		is similar to the	15	is consistent with the	32	which is in line with the	25
		we confirmed the	15	in agreement with previous	31	it is reasonable to	24
		in agreement with the	15	was similar to the	31	are also in	24
		are in line with	15	this is in contrast to	31	in contrast to previous	24
		which indicated that	14	is in line with previous	31	this is supported by	24
		have confirmed that	14	it is also reported that	30	this is in accordance with	23
		also indicated that	14	which showed that	30	is also supported by	23
		was supported by the	14	with the previous	30	are in accordance with	23
		also confirmed the	14	which is in line with the	30	the idea that the	22
		also demonstrated that the	13	are consistent with the	29	in contrast to our	22
		is similar to that of	13	in contrast to previous	29	is in agreement with the	22
		was similar to that of	13	with a previous	29	are in agreement with the	22
		were similar to those of	12	also suggested that	29	showing that the	21
		is in accordance with the	12	similar to other	28	this is in contrast to the	21
CF: Showing background provided by past work							
in this paper we	237	it has been reported that	216	have shown that	1546	have shown that	243
we have presented a	138	have shown that	178	has been shown to	1361	it has been suggested that	191
we proposed a	138	has been shown to	118	it has been reported that	1054	has shown that	184
in this paper we presented a	117	it is well known that	93	et al reported that	714	it has been shown that	154
we presented a	115	as shown in	89	have demonstrated that	686	has been shown to	130
in this paper we proposed a	108	it has been shown that	86	has been reported to	600	have been shown to	101
in this paper we have	99	have demonstrated that	75	it has been shown that	555	have shown that the	96
in this paper we propose a	97	it has been demonstrated that	71	have been shown to	464	have found that	89
in this paper we have presented a	89	have been reported	70	have reported that	429	have suggested that	82

(Continued)

CL		Chem		Onc		Psy	
we propose a	72	it is known that	65	has been shown to be	404	has been shown to be	78
in this paper we present a	67	have been shown to	65	it is well known that	374	it has been	75
in this work we	62	has been reported to	65	we have shown that	344	it is known that	73
is based on	57	it was reported that	65	have shown that the	335	it has been proposed that	73
in order to	56	it has been reported that the	63	has shown that	327	it has been suggested that the	68
we have proposed a	56	have reported that	53	it is known that	326	have demonstrated that	66
we use the	52	was reported to	47	has been reported to be	304	has been found to	63
we introduced a	51	have shown that the	47	is known to	297	have been found to be	56
in this paper we propose a novel	47	have been reported to	45	has been reported in	290	has demonstrated that	54
in this paper we presented	47	is known to	45	have suggested that	270	have suggested that the	53
we have described a	46	has been reported	44	it was reported that	241	has been linked to	50
to the best of our	43	have been used to	43	it has been demonstrated that	238	has been found to be	48
we used the	42	has been shown to be	43	have been reported to	229	it has been reported that	48
in this paper we proposed a novel	38	to the best of our	42	we have previously shown that	220	has shown that the	48
we proposed an	38	plays an important role in	37	have indicated that	215	have been found to	48
is based on a	38	in our previous	37	has been implicated in	212	it is well known that	45
this paper proposes a	38	has been reported to be	35	have been shown to be	203	in a recent	44
in this paper we have proposed a	38	it has been suggested that	35	it has been suggested that	194	it has been demonstrated that	41
we present a	38	are known to	33	we previously reported that	181	have been shown to be	41
we proposed a novel	36	have been reported to be	32	have been reported	177	has found that	41
is based on the	36	has been developed	32	we have demonstrated that	175	most of the	38
in this paper we presented a novel	35	has been used to	31	has also been shown to	174	it has been shown that the	37
we propose a novel	35	have been reported in	31	have been reported to be	174	have found that the	37
we developed a	34	it has been shown that the	30	it has been reported that the	172	it has been found that	37
as described in	34	in a previous	30	have been reported in	171	have reported that	36
we have introduced a	33	is a key	29	are known to	168	have been reported in	32
in this paper we presented an	32	plays an important role in the	29	is a key	166	has been reported in	32
we proposed a new	31	has been found to	28	has been found to	163	have been observed in	32
in this paper we have presented	31	it is reported that	28	can lead to	158	as described in the	31
in this paper we proposed	30	it has been	27	have revealed that	155	have examined the	31
we presented a novel	30	has been reported in	27	are involved in	153	has suggested that	31
in this paper we described our	29	it is well known that the	27	it is well established that	151	is known to be	30
we use a	28	it is known that the	26	has demonstrated that	149	little is known about the	30
in this paper we present a novel	28	have reported the	26	has been linked to	149	in a previous	30
we used a	27	has been developed for the	26	has been demonstrated to	147	have shown that * cite-	30
in this paper we propose	27	it has been proposed that	26	has not been	142	reported that the	29
in this paper we described the	25	has been shown to have	25	has been observed in	142	however in the	29
we have presented an approach to	25	has shown that	25	it is reported that	140	in the previous	29
this paper presents a	25	were reported to	23	have demonstrated that the	139	it has been observed that	29
we described a	25	have indicated that	23	in our previous	134	it has been proposed that the	28
this paper presented a	25	have been shown to be	23	has been found to be	131	there has been a	28
CF: Explanation for findings							
this is due to the	18	may be due to the	52	it is possible that	578	it is possible that	616
it may be possible to	16	may be attributed to the	41	it is possible that the	286	it is possible that the	503
this is because	16	is due to the	37	it is likely that	175	is that the	350
is due to	15	can be attributed to the	35	may be due to	160	it is also possible that	144
are due to	15	can be explained by the	33	may be due to the	156	it may be that the	120
this is because the	15	could be attributed to the	32	it is likely that the	132	it may be that	115
this may be due to the	14	may be explained by the	31	may explain the	119	it is likely that	113

(Continued)

CL		Chem		Onc		Psy	
is due to the	14	may be related to the	29	therefore it is possible that	97	it is also possible that the	112
for this is that the	10	may be due to	26	we can not exclude that	95	it is likely that the	106
this is due to the fact that	10	could be attributed to	26	may be the	95	could be that the	94
it may be that	9	could be due to the	25	we can not exclude the	93	it should be noted that the	89
it is possible that the	9	might be due to the	23	can not be	88	may be due to	85
this can be	9	could be due to	23	may be related to	81	can be explained by the	85
this may be	8	could be explained by the	23	this may be due to the	79	it seems that the	77
this may explain why	8	might be attributed to the	21	may be related to the	78	can not be	76
there may be	8	can be attributed to	21	thus it is possible that	76	may be related to the	75
are due to the	8	is attributed to the	20	could be explained by the	76	can be explained by	74
can be attributed to	8	might be due to	19	may be explained by the	75	it could be that	73
is due to the fact that	7	attributed to the	19	could be due to the	74	may be due to the	72
can be attributed to the	7	this may be due to	19	may be attributed to the	72	it seems that	72
may be due to the	7	may be attributed to	19	might be due to	72	may be related to	71
this is likely because	7	was attributed to the	18	can be explained by the	71	it is possible that this	70
this could be	6	can be explained by	18	could be due to	69	it should be noted that	69
which might be	6	this may be due to the	17	can not be excluded	67	may be the	67
it might be possible to	6	this is due to the	16	we can not rule out	67	might be that the	64
may lead to	6	could be related to the	15	may be more	66	it is possible that these	64
this is mainly due to the fact that	6	can be ascribed to the	15	might be due to the	66	could be that	63
we attribute this to the	6	may be explained by	15	it is conceivable that	63	may have been	63
from the fact that	6	which may be due to the	14	could explain the	63	may be explained by the	60
may have been	6	due to the presence of	14	could be explained by	62	could be explained by the	60
due to the fact that	6	may be caused by the	14	may reflect the	61	it may be the	58
could be attributed to the	6	could be explained by	13	may account for the	59	could be due to	56
this can be explained by the fact that	6	might be explained by the	12	should be considered	58	may be that the	55
for this is that	6	may result from the	12	could be attributed to the	58	may be explained by	54
could be due to	6	has been attributed to	12	may be explained by	58	it is possible that a	52
we attribute this to the fact that	6	is due to	12	can not be ruled out	56	it might be that	51
this can be explained by the	6	were attributed to the	12	may not be	55	not be ruled out	48
we believe this is because the	6	is probably due to the	12	could be the	54	this may be due to the	47
it may be that the	6	may be caused by	12	may be attributed to	54	it might be that the	47
this may be due to	6	this is because the	11	might be the	54	might be that	45
can be explained by the	5	might be related to the	11	therefore it is likely that	54	it is possible that our	45
this may be because	5	was due to the	10	might contribute to the	53	might be related to the	45
it may be better to	5	was due to	10	we can not rule out the	52	might be due to the	45
this could be due to the	5	could be attributed to its	10	there are several possible	52	could be related to the	44
it is also possible that	5	is likely due to	10	might be explained by the	51	might be explained by the	44
can be handled by	5	might be attributed to	10	this may explain the	49	might be the	43
it could be that the	5	be explained by the	10	it is plausible that	48	there are several possible	43
this is partly due to the	5	mainly due to the	9	this could be due to	47	may lead to	42
this is why the	5	was attributed to	9	may have been	46	might be related to	42
this can be done by	5	this can be explained by the	9	may be responsible for the	45	could be explained by	42
CF: Suggestion of future work							
in future work we	247	are needed to	63	are needed to	279	it would be interesting to	142
we plan to	125	are required to	51	there are several	203	is needed to	115
we would like to	119	need to be	41	are required to	167	are needed to	107
in future work we will	83	it is likely that	36	there are some	146	there are several	90
as future work we	81	is required to	33	remains to be determined	144	need to be	60
we would also like to	69	it is necessary to	33	remains to be elucidated	114	it would be	55

(Continued)

CL		Chem		Onc		Psy	
we are currently	65	needs to be	30	is still unclear	106	it would also be interesting to	52
in future work we would like to	64	should be carried out to	28	remains to be	105	there are some	52
for future work	63	are necessary to	28	are warranted to	100	needs to be	46
for future work we	62	is needed to	24	need to be	92	we suggest that future	46
we intend to	57	will be needed to	22	are needed to confirm the	88	should examine the	44
in future work	51	we are currently	21	however the role of	85	will be needed to	43
we are also	51	need to be further	21	are needed to confirm our	81	is needed in order to	36
in the future we	48	should be further	20	should be further	79	is needed to examine the	35
we will also	48	remains to be	20	is needed to	77	should focus on	33
it would be interesting to	47	and will be reported in due	20	is still unknown	74	would be to	33
would be to	46	it is likely that the	19	needs to be further	74	is necessary to	33
it would also be interesting to	41	it is possible that	19	are necessary to	71	should be addressed in future	33
work will focus on	41	are currently underway in our	18	needs to be	67	it would be important to	32
there are a number of	41	therefore it is necessary to	18	are needed to determine the	67	are needed in order to	31
we hope to	40	is still in its	18	there were several	67	is needed to explore the	31
in future work we intend to	38	should focus on	18	will be required to	65	we recommend that future	31
future work includes	34	are currently in	17	need to be further	65	it would be useful to	29
we need to	29	should be considered	17	are needed to clarify the	61	is needed to determine the	29
we also intend to	29	still need to be	17	are urgently needed	58	it is necessary to	29
we plan to explore	27	should focus on the	17	are needed to elucidate the	57	are required to	29
it would be	27	are needed to elucidate the	16	remain to be elucidated	54	future work should	28
for future work we would like to	26	it is expected that	16	will be needed to	53	it remains to be	27
in future we	26	in the future	15	remains largely unknown	51	should be conducted to	27
there are several	25	are needed in order to	15	we are currently	50	it will be interesting to	26
in future work we hope to	25	needs to be further	14	there were some	49	will need to	26
in our future work we	25	should be done to	14	however the underlying	49	it would be necessary to	26
in future work we will explore	21	is necessary to	14	are needed to confirm these	49	would be needed to	26
we will try to	21	remain to be	14	remain to be	48	should be considered	25
needs to be	20	are needed for	14	are needed to explore the	46	it would be worthwhile to	25
we plan to extend the	20	it is therefore necessary to	14	should be conducted to	45	remains to be	25
finally we would like to	20	will be useful for	13	are needed to further	45	it will be important for future	25
we plan to extend our	19	should be carried out	13	are required to elucidate the	44	should address this	25
we plan to further	19	will be required to	13	are needed to validate our	44	there is a need to	25
we will explore the	19	are warranted to	13	however the exact	44	is required to	24
we would like to extend our	18	should be conducted to	13	are still unknown	43	will be required to	23
is needed to	18	are required to elucidate the	13	remain to be determined	42	is needed to understand the	23
to explore the	18	are still needed to	12	are needed to validate the	39	are needed to further	23
future work will include	18	should be further studied	12	are required to determine the	38	it would be of interest to	23
in future work we aim to	18	will be reported in due	12	in the future	38	need to be considered	23
for future work is to	18	there are still	12	remains to be explored	37	is needed to clarify the	22
we would like to explore	17	are underway in our	12	are required to confirm the	37	are needed to explore the	22
future work should	17	are expected to be	12	remain largely unknown	37	should be examined in future	22
in our future work	17	will be the	12	are needed to confirm this	37	should examine whether the	22
as future work we would like to	16	will be necessary to	12	remains to be clarified	37	should be considered in future	22
CF: Comments on the findings							
is a promising	9	it is clear that	28	are currently in	20	we were able to	112
this is an encouraging	5	was successfully applied to	26	was well tolerated	19	were able to	61
we are encouraged by the	4	it is clear that the	22	is currently in	18	can be used to	59
the most successful	4	it is suggested that	21	has shown promising	17	we have shown that	53
is easy to	4	it is believed that	20	have shown promising results in	14	it is possible to	52
is effective for	4	was successfully applied to the	18	have shown promising	13	we were not able to	41

(Continued)

CL		Chem		Onc		Psy	
are very promising	4	it was suggested that	17	is a promising strategy for	12	could be used to	40
it is our hope that	4	it was suggested that the	17	we successfully established a	12	allowed us to	36
is promising as it	4	it is believed that the	16	has emerged as a promising	11	in this way	34
are very encouraging	4	was achieved by	15	represents a promising	10	we believe that	31
seems to be promising to	3	it was proposed that	13	we have successfully	10	it would be possible to	29
the most promising	3	is expected to	13	was well tolerated and	10	in this way the	26
has been successfully applied to	3	have been successfully	13	was well tolerated in	9	we are able to	25
is a promising approach	3	we have successfully	13	are a promising	8	was able to	25
it is encouraging to note that	3	was successfully used to	12	showed promising results in	7	we have demonstrated that	25
are significant at 1	3	was developed for	12	may be a promising strategy for	7	should be able to	25
it is encouraging that	3	it was proved that	12	may be a promising strategy to	7	we have shown that the	24
is more successful	3	were obtained in	12	has shown promising results in	7	we would like to	23
this is good	3	was successfully applied for the	11	are not satisfactory	7	was not significant	23
this is encouraging as it	3	in good to	11	are promising candidates for	7	it was possible to	22
and easy to	3	has been achieved	10	is a promising approach for	7	we were able to show that	22
it is encouraging to see that	3	has the advantages of	10	and is a promising	6	we argue that	21
seems to be a promising approach for	3	in good yield	10	have been successful in	6	however when the	21
are promising as	3	it was demonstrated that the	9	appears to be a promising	6	would be more	20
		were obtained with the	9	seems to be a promising	6	they were able to	20
		is a very important	9	is a reasonable	6	in this paper we have	20
		it is evident that	9	was successful in	6	we tried to	20
		it is suggested that the	9	is a promising approach to	5	made it possible to	19
		can be achieved	9	is currently in a	5	would be able to	18
		was suggested to be	9	is a promising strategy to	5	it should be possible to	18
		it is obvious that	8	may be a promising approach to	5	are expected to	17
		were obtained for	8	it is hoped that	5	we would expect that	17
		it is proposed that	8	is currently undergoing	5	to the extent that	16
		in summary we have successfully	8	are promising for	5	we would expect to find	16
		is believed to be	8	holds great promise as a	5	it may be possible to	16
		has proven to be	8	could be promising	5	we have seen that	16
		were proven to be	8	we are convinced that	5	we demonstrated that	16
		is a good	8	are considered promising	4	to achieve this	16
		have proven to be	8	were not successful	4	one would expect that	16
		it should be emphasized that	8	is emerging as a promising	4	enabled us to	15
		in summary we have successfully developed a	8	could be a promising strategy for the	4	we have also shown that	15
		was developed and validated for the	7	could be a promising strategy to	4	can be expected to	15
		was developed to	7	have shown encouraging	4	we hope that our	15
		is considered a	7	was safe and	4	we have presented a	15
		were successfully prepared and	7	are now in	4	allows us to	15
		was successfully performed	7	we are convinced that the	4	we would expect	15
		was established for the	7	a very promising	4	may help to	15
		was proven to be a	7	have yielded promising results in	4	we have found that	15
		is believed to	7	might be promising	4	we hope that the	15
		was proven to be	7	is a promising approach	4	could be used in	15
CF: Unexpected outcome							
for example the	111	on the contrary	40	it is not surprising that	62	it is not surprising that	51
we have shown that	82	it is not surprising that	15	as expected the	33	as expected the	43
the number of	80	as expected the	12	it is expected that	27	this was not the	33

(Continued)

CL		Chem		Onc		Psy	
we found that	68	on the contrary the	12	surprisingly we found that	24	on the contrary	29
on the other hand	62	more importantly the	7	therefore it is not surprising that	24	it was expected that	27
we show that	59	interestingly we found that	6	would be expected to	24	it is not surprising that the	26
we showed that	57	interestingly we found that the	5	it is not surprising that the	19	this is not	18
we find that	46	it is not surprising that the	5	as expected we found that	17	interestingly we found that	16
we find that the	45	was prevented by	3	thus it is not surprising that	12	on the contrary the	15
on the other hand the	40	interestingly we found	3	would be predicted to	12	as expected we found that	15
we have shown that the	39	was observed only in the	3	it is therefore not surprising that	12	we expected to find	13
with respect to the	38	most importantly the	3	this is not	11	we expected that	11
we showed that the	38	by contrast the	3	this is not surprising as	9	thus it is not surprising that	10
we also showed that	37	this is not surprising since the	3	therefore it is not surprising that the	9	it is perhaps not surprising that	10
we show that the	37	it is thus not	3	is not surprising	8	therefore it is not surprising that	10
is available at	35	as it was expected	3	as we expected	8	is not surprising	10
we also show that	29	was not accompanied by a	3	we wondered whether	8	is not surprising given the	8
on the other	26			it was not surprising that	8	interestingly we also found that	7
most of the	26			it was expected that	8	this was not observed	7
for example in	26			we expected that the	7	it would not be	7
we show that our	25			we were surprised to find that	7	as expected we found	7
we have also shown that	24			we surprisingly found that	7	it is therefore not	7
in contrast the	24			it is perhaps not surprising that	7	it is not surprising that we found	6
of the two	24			this may not be	7	is not surprising given that the	6
is able to	23			one would expect	7	as expected we found a	6
for example in the	23			was expected to	7	as expected our	6
in the same	23			we expected that	7	interestingly we found that the	6
in particular the	23			one would expect that	7	might have been expected	6
we also found that	23			unexpectedly we found that	7	as might be expected	6
this is not	22			it is not surprising that a	7	was expected to	6
we see that the	22			this was not	6	it is therefore not surprising that	6
we observe that	21			it would not be	6	is not surprising given that	6
as the number of	20			would be expected to be	6	it was not surprising that the	6
of the same	20			it is reasonable to expect that	6	interestingly we observed a	6
we also find that	20			is not surprising as	6	unfortunately we did not	6
depending on the	18			would be expected to have	6	it was expected that the	6
for a given	18			it would be expected that	6	it is surprising that	6
many of the	17			would not be expected to	6	this is not surprising given that	6
we have demonstrated the	17			is not surprising since	6	we did not expect	6
we also observe that	17			this is not surprising given the	6	is not surprising as	6
we have shown that a	17			it is expected that the	6	interestingly in the	6
is not a	17			as expected we found	5	as we expected	6
and the number of	17			cite- it is not surprising that	5	as expected a	5
between the two	17			we would expect that	5	would be expected in	5
it is not	17			we would have expected	5	on the contrary in	5
for example a	17			cite- prompted us to	4	we expected a	5
on the same	16			surprisingly we observed that	4	would be expected if	5
we have also	16			would be expected	4	as would be expected	5
we observed that	16			we asked whether	4	however contrary to our	5
seem to be	16			it is not surprising that we	4	however there were no	5
CF: Implications of the findings							
it is important to	21	the possibility of	11	raise the possibility that	43	it is important to	74
this is an important	16	there is a possibility that	10	raises the possibility that	27	have important implications for	47

(Continued)

CL		Chem		Onc		Psy	
is an important	14	have the potential to be used as	8	this raises the	26	contributes to the	47
is important for	12	this is of	4	the possibility of	23	have implications for	40
is useful for	8	shed new light on the	4	implications for the	20	this is an important	38
can be applied to other	8	this could lead to	4	may have important	19	highlights the importance of	37
has the potential to	7	the need for further	4	have important implications for the	19	it is important that	34
may be useful for	5	have the potential to	4	may have important implications for	19	it is therefore possible that	31
is an important step towards	5	may find applications in	4	the possibility that	18	has important implications for	31
may be useful in	5	raise the possibility that	3	raising the possibility that	16	adds to the	29
will be useful for	5	is of crucial importance	3	has important implications for	14	it can be assumed that	29
is also useful for	4	this does not exclude the	3	underscore the importance of	13	it is also important to	29
we believe that this work	4	this is especially important for	3	raise the possibility that the	12	may have important implications for	26
it may be useful to	4	moreover there is a possibility that	3	have important implications for	12	it is important to consider the	24
can play an important role in	4	implications for the	3	have implications for	12	this is important because	24
up the possibility of	4	has a great	3	may have implications for	11	implications for the	22
are crucial for	4	offers the possibility to	3	raises the possibility that the	11	it is therefore important to	21
have a significant	4	open the possibility to	3	may have implications for the	11	it is therefore possible that the	21
it is important to be	4	has the advantage of	3	there is a possibility that	10	also have implications for	21
may prove useful for	4	is the possibility of	3	the possibility of a	10	is important for	19
also be useful for	4	this makes it a	3	this highlights the importance of	9	has implications for	19
some light on the	4	this may lead to	3	may have significant	8	therefore it is important to	19
this is important for	4	this is particularly important in	3	suggest the possibility that	8	it can be assumed that the	19
is applicable to	4	highlights the importance of	3	highlight the need to	8	it is therefore	18
is important to	3	important insights into the	3	support the possibility that	8	it is thus possible that	17
our work has implications for	3	highlights the need to	3	suggest a possibility that	7	it is crucial to	17
may be useful in other	3	have important implications for	3	have potential implications for	7	highlights the need to	17
this paper addresses the	3			has important implications for the	7	it is important for	17
this is especially important for	3			the implications of	7	have implications for the	16
could be useful in	3			limits our ability to	7	has important implications for the	16
also shed light on the	3			have several important implications	6	this raises the	16
our understanding of the	3			highlights the need for	6	raises the possibility that	14
this is an important point	3			implications in the	6	this highlights the	14
is particularly important for	3			may have important implications for the	6	highlights the need for	14
should be useful for	3			suggest the possibility of	6	have important implications for the	14
this will help in	3			may have important implications in	6	it is essential to	13
it is crucial to	3			raising the possibility that the	5	is important because it	13
shed some light on the	3			underscore the need for	5	are important for	13
could be applied to other	3			raise the possibility of using	5	it is important to understand the	13
highlights the importance of	3			limits the use of	5	may shed light on the	12
shed light on the	3			highlight the need for	5	important implications for	12
it is therefore important to	3			therefore there is a possibility that	5	this highlights the importance of	12
may prove to be	3			has great potential for	5	it is important to understand	12
it could also be used to	3			may have potential	5	it is therefore likely that	12
have important implications for the	3			open a new	5	have several implications for	12
may also be useful in other	3			offers the possibility of	4	this leads to the	11

(Continued)

CL		Chem	Onc		Psy	
this has two	3		which raises the possibility that	4	it may be assumed that	11
it is also important to consider	3		has the potential for	4	may have implications for the	11
is very important to	3		could have significant	4	emphasizes the importance of	11
may also be useful	3		may have significant implications for	4	is crucial to	11