Analysing Bilingual Collocational Expressions with *AntConc*

Adelina Racaj

June 2021

Abstract

The present paper focuses on a syntactic and statistical analysis to extract bilingual collocations from a parallel corpus. It describes how to obtain preferred syntactic patterns from collocational expressions in German and English using *AntConc* in order to extract word co-occurrences. Collocational knowledge is profoundly important for developing language skills, and successfully translating across languages. The aim of this paper is to analyse to what extend *AntConc* can be beneficial in extracting and compiling collocational expressions from parallel German-English corpora of Kafka's novels *The Trail* and *Metamorphosis*.

Research question:

To what extand can *AntConc* be used for extracting and compiling bilingual collocational expressions for the development of a bilingual collocation dictionary with the language combination German and English by analysing corpora containing Kafka's novels *The Trail* and *Metamorphosis*?.

Keywords: c*orpus linguistics, AntConc, corpus analysis, Kafka*

Contents

1 Foreword

This paper is the final project of the course *Technologies de l'information et de la communication* held by the professors Max De Wilde and Simon Hengchen in the second term of the academic year 2020/21 within the study programme *Multilingual Communication Technology* at the University of Geneva. The aim of the final project consists in analysing at least two methods discussed in class using bi- or multilingual corpora.
The focus of this analysis lies in exploring the advantages and disadvantages of each tool when carrying out a corpus-based analysis. The scope is working on a realistic issue that can be encountered in our daily working life. In order to meet the project's objectives, it is important to establish the scope of the project, refine the objectives, and define the course of action required to attain the objectives that the project was undertaken to achieve. The methods used for this corpus analysis are as follows:

- **AntConc**: With Laurance Anthony's toolkit *AntConc*, the present paper focuses on the acquisition of collocational knowledge from a corpus-based and corpus-aided analysis. Examples will be given in form of screenshots to depict how collocations matching the patterns are extracted with certain functions and features. The screenshots taken have been modified with the screenshot program *Snagit*. It is important to highlight that solely the functions and tools intrinsic to the research question will be analysed and evaluated. The aim of this analysis is to first, extract word co-occurrences for a bilingual collocation dictionary with the language combination German-English and secondly, to evaluate to what extend *AntConc* is efficient and beneficial for such an activity. The methodology and approach will be explained with a concrete step-by-step guide. Finally, I will offer a critical look at the development of *AntConc* by discussing its strengths and weaknesses.

- **HTML**: The HyperText Markup Language is used to display the main core points of this analysis in a web browser. There you will also find an additional chapter exploring the strengths and weaknesses of the markup language.

- **Github**: The HTML web page will be published together with a licence and readme file on Github in order to make it accessible to a wider audience. The link to the repository is as follows: *TIC Final Project*

- **Markdown**: The readme file contains a short summary of the analysis and the credentials of the author written in the lightweight markup language *Markdown*.

2 Introduction

In the past two decades corpora of language data have started to play an increasingly significant role in determining how languages are taught, perceived and used[5]. Corpora have been applied in a wide range of areas, such as translation studies, grammar and dictionary development [7] and language acquisition approaches. Translators, linguists and terminologists thrive in a cutting-edge working environment, where mastering good computer literacy skills and being well-versed in strategic information mining is essential for both being able to use the resources at their disposal and to find required and significant information in a short period of time. This involves, for instance, finding translation equivalents and domain-specific terms, extracting collocations, idioms, and phrasal verbs, and exploring the terms used in specific contexts in order to render future documents grammatically, semantically and stylistically appropriate for the target audience. In order to conduct such a cross-linguistic analysis of corpora, software tools are required to process them and display the results in a comprehensible way. For this analysis we will be using the corpus analysis program *AntConc*. This paper focuses on the analysis of small sized corpora with the language combination English-German containing Kafka's novels *The Trail* and *Metamorphosis*. This contribution aims at investigating, from a comparative corpus-assisted perspective, what are the salient features the textual analysis software provides that can be beneficial to translators, linguists, terminologists, and language learners when translating literary texts and learning a new language. Indeed, collocations are recurrent combinations of words that co-occur in a specific context, and are notoriously difficult to translate literally into another language because they are opaque. I will detail and analyse the software's limitations, depict the advantages and disadvantages, before concluding the paper with a discussion on its future development.

2.1 General context and purpose of the analysis

The developer of *AntConc* Laurence Anthony states in his linguistic research *A critical look at software tools in corpus linguistics* that "the functionality offered by software tools largely dictates what corpus linguistic research methods are available to a researcher, and hence, the design of tools will become an increasingly important factor as corpora become larger and the statistical analysis of linguistic data becomes increasingly complex"[1].
The analysis depends on quantitative and qualitative analytical techniques to interpret the findings. The success of corpus linguistics is intrinsically related to the tools used to access, analyse, and display the results of corpus searches. Recent rapid advances in the development of computer technology, particularly the availability of storage of linguistic databases, is leading to a new approach in both the field of translation studies and language teaching and learning [8]. In fact, ample examples of specific word combinations or phrases are stored electronically in corpora which allow learners and researchers to discover and analyse cross-linguistic patterns by observing extensive naturally occurring phrases. At the First International Conference on Teaching and Language Corpora held at the University of Lancaster, Goeffrey Leech stated that an open-ended supply of language data encourages an exploration and discovery approach to learning languages [9].

2.2 Software tool: Introduction to AntConc

AntConc is a freeware concordance program for Windows, Macintosh OS X, and Lunix. The analysis type preformed is based on data mining methods and algorithms. The data import supports HTML, TXT and XML format. AntConc hosts a comprehensive set of functions including a powerful concordance, word and keyword frequency generators, tools for cluster and lexical bundle analysis, and a word distribution plot. Its vocabulary search function can be divided into basic search and advanced search. It runs on both Windows and Linux/Unix based system and includes an easy-to-use, intuitive graphical user interface and offers a powerful concordancer, word and keyword frequency generators, tools for cluster and lexical bundle analysis, and a word distribution plot [1].

3 Methodology

The main focus of this paper lies in extracting collocational expressions from German and English texts by analysing preferred syntactic patters obtained from the parallel corpora chosen. The approach lies in filtering the collocations based on both linguistic and statistical constraints. The bilingual collocations have lexical correlations between them and are rigid in both languages. Research papers and literature about similar studies have been consulted and referred in this analysis.

3.1 Materials

The materials used for this research paper are two online corpora containing two of Franz Kafka's novels selected from the online libraries of free eBooks Project Gutenberg and Deutsches Textarchiv. Both websites are dedicated to encourage the creation and distribution of eBooks by providing free access of resources to public use. The parallel corpora chosen are Kafka's novels *The Trail* and *Metamorphosis* and their language combination is German-English. They contain both arbitrary word pairs and fixed constituents, for instance, compound words, frozen expressions or idioms. Collocations are not predictably on the basis of syntactic or semantic rules, and can solely be learned through repeated usage [11]. A translator needs to be aware of the meaning of the expression in order to be able to translate it idiomatically into the target language. In this analysis I will identify collocational expressions and phrases which cannot be translated on a word-by-word basis.

4 Analysis

This chapter describes how bilingual collocations are obtained by using preferred syntactic patterns and associative information.

4.1 Concordancer Tool

The most commonly used tool in corpus analysis software is the concordancer. Several studies have shown that a combination of bilingual corpora and concordance tools leads to a better acquisition of collocational knowledge [8]. Moreover, according to a longitudinal case study strategies learners tend to experience different levels of success when using concordance strategies while studying a new language [12]. The concordancer indicates which words co-occur frequently with other words and how they combine within a sentence.

1. Navigate to the menu *File > Open File(s)* to import the respective corpus. The list of the selected files is shown in the *Corpus Files* column on the left of the main window.

2. Enter the search term on which to build concordance lines in the entry box *Search Term*. Here the search terms chosen are "möglichkeit" from the German corpus and "opportunity" from the English corpus.

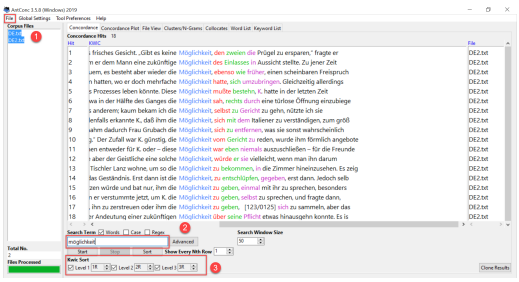
3. With the "KWIC" (KeyWord In Context) format, you can see how the search terms are used commonly in the corpora. Rearrange the concordance lines by using the buttons of the *Kwic Sort*. Select how many words should be highlighted to the right or left of the search terms. Click on the *Sort* button to start the sorting process (see Figure 1).

4. Select the check box *Word* if you want your search terms to be a word.
   Select the check box *Case* if you want your searches to be case sensitive or case insensitive.
   Select the check box *Regex* if you want your searches to be full regular expressions.

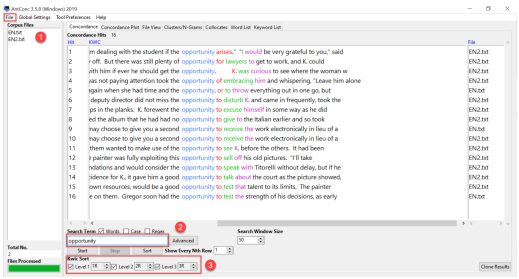
   →The check box *Case* is thoroughly important depending on the language of the corpus. Certain languages, such as German, are rich on upper and lowercase vocabulary whose meaning can drastically vary according to the capitalization of the term.
   →For further information about "regular expressions", consult Jeffrey E. F. Friedl's book *Mastering Regular Expressions* or websites with "Quick Start" guidelines like *RegexBuddy*.

5. Click on the *Advanced* button for more complex searches:

   - **Defining Search Terms:**
     You can add a set of search terms by selecting the *Use search term(s) from list below* check box, and typing them one per line or by uploading a file with search terms. The feature allows the user to search multiple terms simultaneously without having to type them individually.



(a) German figure

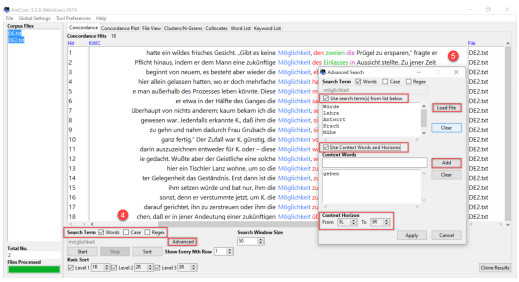


(b) English figure

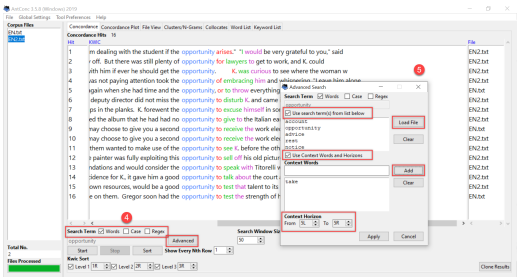
Figure 1: Generation of search results in *KeyWord In Context* format

→This feature can be useful when analysing the collocational patterns of synonyms.

- **Defining Context Words:**
  You can define the context within which the search term(s) must appear by selecting the *Use Context Words and Horizons* check box, and by adding the preferred words. Here we add the verbs "geben" and "take" which we want to appear within the context of the search term(s) chosen.
- **Defining Context Horizon:**
  You can define the range in which you want the search term(s) and context term(s) to appear. Here we want the word "geben" and "take" to appear at least five words the left or right of the search term(s) chosen (See Figure 2).



(a) German figure



(b) English figure

Figure 2: Generation of search results in *KeyWord In Context* format

4.2 Clusters/N-Grams Tool

The Clusters/N-Grams Tool is one of the most useful ones for this analysis as it allows you to find patterns in the corpora chosen and cluster the results in a list. It summarises the results generated in the *Concordance Tool*. You can find strings of text based on their length (number of tokens or words), frequency, and occurrence. By searching for a specific term, you can opt to see terms that precede or follow the search term.

1. The results can be sorted as follows:

   (a) *Frequency:* The *Frequency* feature shows recurring patterns from which can be deducted that the more frequent the pattern is, the more relevant it might be. Hence, the higher the figure in the frequency column, the more often the collocation occurs.

   (b) *Word:* The *Word* feature allows you to sort the recurring pattern alphabetically.

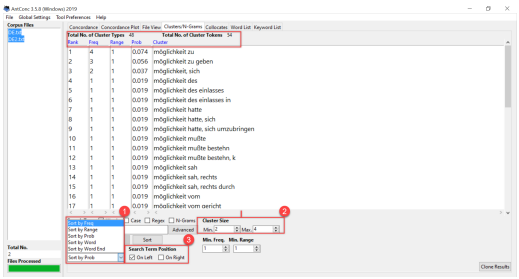
   (c) *Word end:* The *Word end* feature allows you to sort the recurring pattern alphabetically based of the last word in the string.

   (d) *Range:* The *Range* feature allows you to analyse in which of the chosen files the search terms appear.

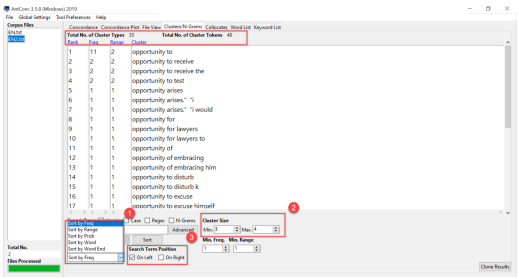
   (e) *Transitional probability:* The *Transitional probability* feature allows you to analyse how likely it is that word2 occurs after word1.

2. Adjust the range of the feature *Cluster Size* to define the clusters range. Here we chose clusters that are 3 to 4 words long.

3. Adjust the feature *Search Term Position* to determinate if your search terms should be the first word in the cluster (left) or the last one (right).



(a) German figure



(b) English figure

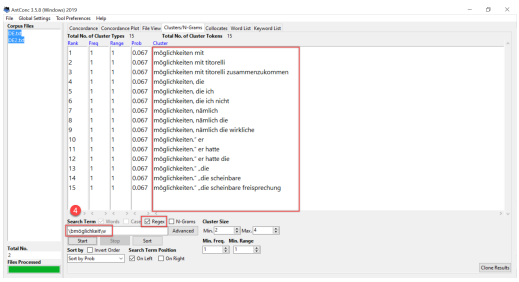
Figure 3: Generation of search results with the *Clusters Tool*

4. Using the *Regex* option: Add a boundary *slash+b* before the term in the *Search Term* feature to generate 2-word clusters that start with a certain term. In order to find out the terms' collocational pattern, you need to set the second word as "any word" by adding *slash+w* at the end of the term in the *Search Term* feature. The results will follow the form *möglichkeit+any word* or *opportunity+any word* (See Figure 4).

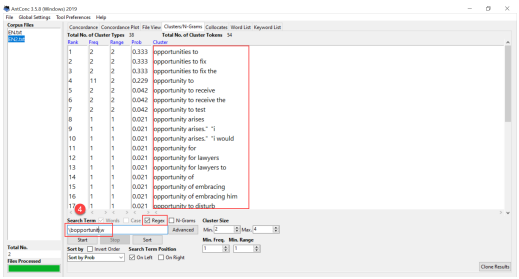
5. Searching with **N-Grams**: The N-Gram defines the length of a string rather than its content. This is a potent feature that allows you to find recurring collocations without specifying any search terms. It extracts all common expressions and repeated phrases from your corpus and reports them in a list.
In the *Search Term* feature select the check box *N-Grams*.

6. Select the minimum and maximum size of each N-Gram in the *N-Gram Size* feature.

7. Select how many times your N-Gram should occur and in how many files of your corpus using the *Min. Freq.* and *Min. Range* feature.

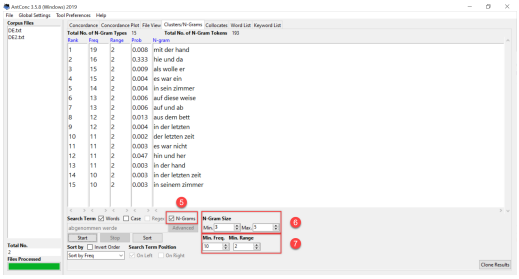


(a) German figure

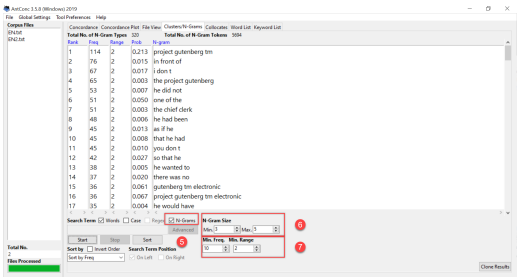


(b) English figure

Figure 4: Generation of search results with 2-word clusters using the *Regex* option



(a) German figure



(b) English figure

Figure 5: Generation of search results using the *N-Grams* option

4.3 Collocates Tool

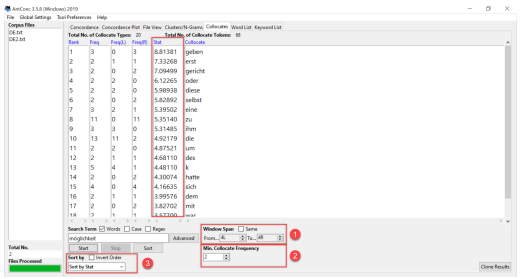
The Collocates Tool is by far the most potent tool for this analysis as it allows you to investigate non-sequential patterns in the corpus.

1. Select the span of words to the left and right from the search term in which to find

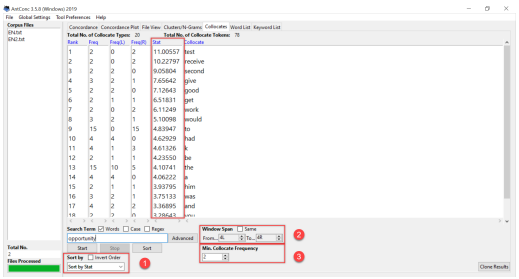
collocates using in the *Window Span* feature. Here we chose a span of 5.

2. Select the minimum frequency of collocates displayed in the *Min. Collocate Frequency* feature. Here we chose a minimum frequency of 2.

3. Sort the results *by Statistics* in order to analyse what collocation occurs. The selected statistical measure is the *MI score*.
   →The *MI score* is a measure of collocational strength. It uses a logarithmic scale to express the ratio between the frequency of the collocation and the frequency of random co-occurrence of the two words in the combination. The higher the MI score, the stronger the link between two terms. The closer to 0 the MI score gets, the more likely it is that the two terms co-occur by chance. For further information about the the *MI score*, please consult Kenneth Ward Church and Patrick Hanks' paper *Word Association Norms, Mutual Information, and Lexicography*.



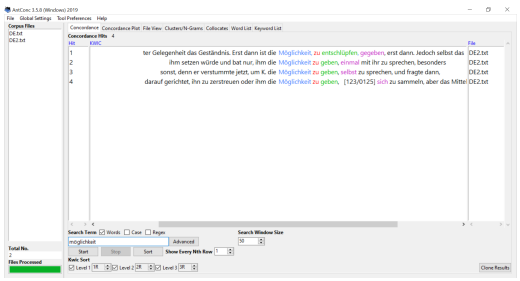
(a) German figure



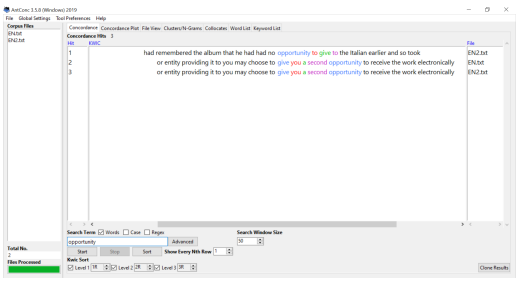
(b) English figure

Figure 6: Generation of search results using the *Collocates Tool*

4. Click on the collocate in order to generate a set of KWIC lines and analyse the context in which the collocational expression occurs (See Figure 7).



(a) German figure



(b) English figure

Figure 7: Generation of KWIC lines using the *Collocates Tool*

4.4 Additional features

Wildcards: *AntConc* offers a wide range of wildcards in which the search term can be embedded and which you can assign to any particular character or string of characters via menu option [2]. You can vary the kind of concordances and searches according to the results you are opting to obtain.

- **Star Symbol**:

  1. Go to the menu bar and select *Global Settings*. A window will pop up.
  2. Select *Wildcards* from the column on the left of the window. Here you can see for what each of the wildcards stands for (See Figure 8).

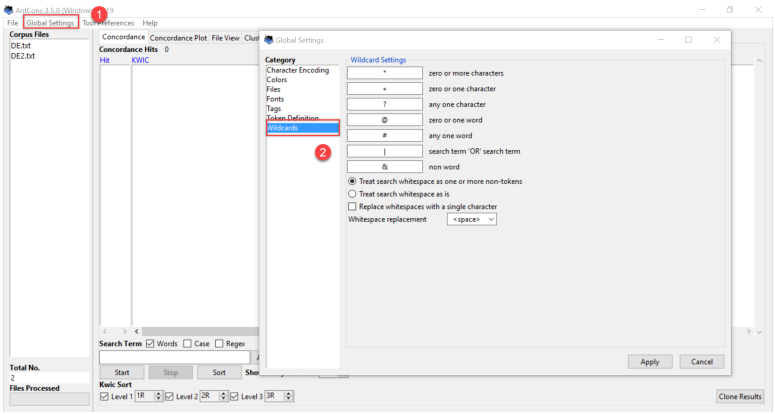
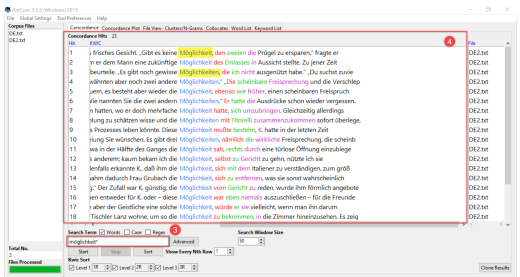
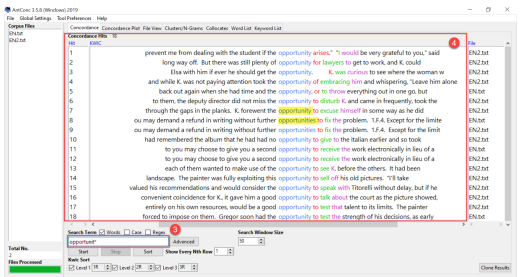


Figure 8: Finding *Wildcard* description in *Global Settings*

  3. Return to the main page and type "möglichkeit*" for the German corpus and "opportunit*" for the English corpus in the *Search Term* box. Click on *Start*.
  4. The results displayed show the terms ending with *zero or more characters*. For instance, in both corpora you can find the singular and plural form of the search terms which helps you to broaden your search range when analysing collocations. The star symbol can be particularly useful when exploring semantically related words or word forms of a lemma.



(a) German figure

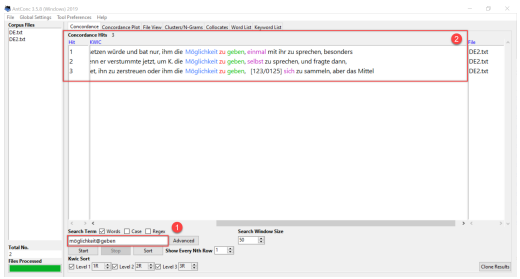


(b) English figure

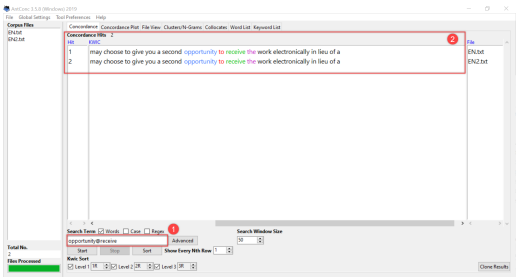
Figure 9: Generation of KWIC lines using the *Star Symbol* Wildcard

- **At Symbol**:

  1. Type "möglichkeit[at]geben" for the German corpus and "opportunity[at]receive" for the English corpus in the *Search Term* box. Click on *Start* (See Figure 10).
  2. The results displayed show the terms with *zero or one word* in between.



(a) German figure

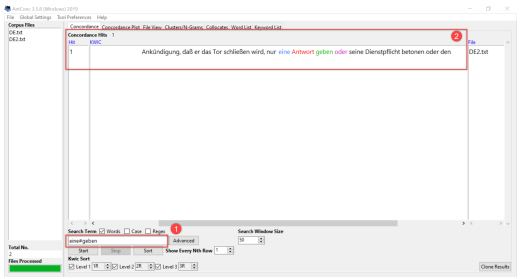


(b) English figure

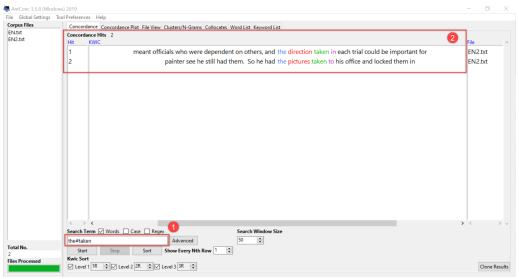
Figure 10: Generation of KWIC lines using the *At Symbol* Wildcard

- **Hashtag Symbol**:

  1. Type "eine[hashtag]geben" for the German corpus and "the[hashtag]taken" for the English corpus in the *Search Term* box. Click on *Start* (See Figure 11).
  2. The results displayed show the terms with *any one word* in between. The hashtag symbol can be particularly useful when exploring N-Grams. Rearrange the concordance lines by using the buttons of the *Kwic Sort*. Select how many words should be highlighted to the right of the search terms.



(a) German figure



(b) English figure

Figure 11: Generation of KWIC lines using the *Hashtag Symbol* Wildcard

5 Results

The results suggest that the functions and features examined are very suitable for collocational term extraction. *AntConc* allows the user to automatically extract collocations from already translated texts wrapped in corpora, and plays a crucial role not only in the development of a bilingual collocation dictionary, but also for many applications such as natural language generation, cross language information retrieval[4], and foreign language acquisition. The greatest advantage of corpora-based software is the ability to derive lexical knowledge from large-scale corpora via automated procedures. *AntConc* is a easy-to-use and user-friendly software with a simple interface design and convenient operation, which makes it suitable for beginners. This is essential for successful software design to avoid additional unnecessary pull-down menus or windows [10]. The software does not need to do any pre-processing of data. This allows users to work more rapidly as the waiting times are distinctively shorter than on other software. *AntConc* is designed so that all functions and features are accessible directly on the main screen in order to provide access to all environments necessary to perform each task. The tools are ranked according to significance with the *Concordance Tool* being qualified as the main tool. Search terms can be defined as full regular expressions offering the user access to highly powerful and complex searches. *AntConc* offers the ability to see the collocational patterns of a search term listed in a table, where the frequency of the most common words to the left or right of the search term are indicated. When using any tool, the user can click on the search term in the KWIC results display in order to automatically open the *View File* tool which depicts the search term hit in the original data file. Multi-word units can be investigated by using the *Cluster/N-Gram* tool. When exploring the clusters of words that surround the search term, the user can analyse frequency patterns of word sequences. This function is particularly useful after having established high-frequency words for the collocation search strategy and for increasing the precision of the search by establishing good collocational expressions. An alternative is the *N-Gram* function where n can vary according to the selected range. *AntConc* can be used as collocation compiler to automatically produce several lists of collocational expressions. It enables the learning process to be tailored to the user's needs as it provides a body of evidence for the contextual function and usage of words and expressions [8].

5.1 Suggestions for development for corpora analyses

AntConc preforms all operations directly on the raw corpus with data solely encoded in HTML, TXT, or XML format. This is a major weakness of the software as more formats need to be supported, such as TMX in order to be able to import the results into translation memory tools . For instance, it could be developed a function where the user could fully align the results obtained from one corpus to another corpus in order to create a bilingual term extraction. Like this the user could apply the cloned files directly into a machine translation system that uses a transfer approach as it relies on correspondences between segments in the source and target language. Although *AntConc* offers an easy way to save and copy and paste results into a spreadsheet program for further analysis using simply keyword shortcuts, yet there is no function to import and

work on two corpora simultaneously in different languages. The software does not have efficient features when working on bilingual corpora because correspondences between collocations in two languages are largely unexplored. The software is designed to analyse only one corpus at the time. The ability to compile a set of translated collocational expressions automatically would increase the portability of machine-translation systems and both facilitate and accelerate the development of a dictionary. I propose a bilingual collocation concordancer - a tool that provides the user with collocation correspondences between two languages. Therefore, more sophisticated methods need to be implemented in order to be able to develop an collocation dictionary by using *AntConc*. On a personal note, it is apparent that nowadays cross-linguistic corpus-assisted discourse studies are still a niche, especially those focusing on translation studies. Thus, I want to expand this promising field of research by contributing with the present analysis. Indeed, researchers such as Biber, Gries and Weisser suggest that corpus linguists should embrace acquiring skills in computer science in order to be able to develop analytical tools. In doing so, translators and linguists will develop their own tools in order to tailor the output to their own research needs [3] and take control of their own research agenda [6]. It would provide them with more flexibility to develop tools for a particular task and gives them an insight into the issues the developers need to tackle when developing the tool. The suggestions advocated in this paper have been derived from an empirical validation.

6 Summary and conclusion

The aim of this paper was to provide an insight as to what extend *AntConc* can be a useful corpora analysis toolkit when it comes to creating a bilingual collocation dictionary. With the growing availability of large textual resources, the corpus-based or corpus-aided approach is gaining significant importance and attention in translation studies [8]. The ability to use a corpus-based software to investigate collocational patterns is a promising method to accelerate translation processes and optimize standards. Collocations are opaque constructions that cannot be translated on a word-by word basis. Their meaning strongly differs and depends on the context and the phrase they are embedded in. As collocations are domain dependent, a variety of phrases can embrace a specific meanings and translations that apply only in the given domain. Through the lens of a real-life translation scenario, this paper discusses to what extend *AntConc* can be beneficial when creating a bilingual collocation dictionary with the language combination German-English. It investigates the performance of certain functions and features in *AntConc* on corpora of Kafka's novels *The Trail* and *Metamorphosis*. The methodology lies in extracting bilingual noun phrases using syntactic and statistical analysis of the co-occurrence of phrases with highly reliability. Corpora analysis software present an efficient method for extracting collocations in the respective language [8]. As highlighted in the examples provided in the *Analysis* chapter, *AntConc* incorporates a wide array of features and functions that are very useful. Overall, *AntConc* preformed well and can definitely be classified as a powerful tool to extract collocational patterns within a monolingual corpus. I hope this paper has provided a new perspective on corpus tools that leads to an increased awareness of the enormous potential of corpus analysis, and to continued growth of corpus linguistics tools and the field as a whole.

References

[1] Laurence Anthony. A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2):141–161, 2013.

[2] Laurence Anthony, Shinichi Fujita, Yasunari Harada, and Waseda Daigaku. Proceedings of iwlel 2004: An interactive workshop on language e-learning. 2005.

[3] Douglas Biber, Susan Conrad, and Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge Approaches to Linguistics. Cambridge University Press, 1998.

[4] Wu Chien-Cheng and Jason S. Chang. Bilingual collocation extraction based on syntactic and statistical analyses. 9(1):1–20, 2004.

[5] David Coniam. Concordancing oneself: Constructing individual textual profiles. *International Journal Corpus Linguistics*, 9(2):271–298, 2004.

[6] Stefan Th. Gries. *Quantitative Corpus Linguistics with R: A Practical Introduction*. Routledge, 2016.

[7] Susan Hunston. Corpora in applied linguistics. 2002.

[8] Kita Kenji and Ogata Hiroaki. Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. *Computer Assisted Language Learning*, 10(3):229–238, 1997.

[9] G. Leech. Text corpora in education: the grand design. *Conference Handbook of the 1st International Conference on Teaching and Language Corpora*, pages 24–5, 1994.

[10] Colin Lonfils and Johan Vanparys. How to design user-friendly call interfaces. *Computer Assisted Language Learning*, 14(5):405–417, 2001.

[11] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computer Linguistics*, 22(1):1–38, 1996.

[12] J. Turnbull and J. Burston. Towards independent concordance work for students: Lessons from a case study. 12(2):10–21, 1998.