Toward a computer-aided methodology for discourse analysis

Jörn Stegmeier
Ruprecht-Karls-Universität Heidelberg, Hauptstraße 207-209, D-69117 Heidelberg, Germany /
Department of General Linguistics, Stellenbosch University, Private Bag X1, 7602 Matieland, South Africa
E-mail: joern.stegmeier@gs.uni-heidelberg.de

Abstract
This paper describes and outlines a new project entitled "Applying computer-aided methods to discourse analysis". This project aims to develop an e-learning environment dedicated to documenting, evaluating and teaching the use of corpus linguistic tools suitable for interpretative text analysis. Even though its roots are in discourse analysis, the scope of the platform also covers topics pertaining to general text analysis, as is reflected in its name, Computer Aided Methodology for Text Analysis (CAMTA). The paper provides a discussion of some of the basics of corpus linguistics in relation to discourse analysis and a demonstration of some of the basic features of the software developed for the e-learning environment, Qua Mano.

Keywords: corpus linguistics; computational linguistics; discourse analysis; e-learning

1. Introduction

This paper describes and outlines a new project entitled "Applying computer-aided methods to discourse analysis". This project aims to develop an e-learning environment dedicated to documenting, evaluating and teaching the use of corpus linguistic tools suitable for interpretative text analysis. Even though its roots are in discourse analysis, the scope of the platform also covers topics pertaining to general text analysis, as is reflected in its name, Computer Aided Methodology for Text Analysis (CAMTA).

The idea behind the e-learning environment is simple: anyone trying to understand something as complex as corpus technology needs help that goes beyond a textbook or tutorial for a corpus tool like AntConc, WordSmith, MMax2 or others[1]. The aim, therefore, is to provide students and scholars with a platform where they can not only access customised tutorials and online help, but where they can also participate in video conferences headed by various

---

[1] There is a wide range of corpus tools available: AConcorde, ANNIS, AntConc, BitPar, ConcApp, Concorder, Corpus Search 2, Corpus Wizard, ETS (Event Tracking Summarizer), MMAX2, Linguistic Toolbox, Multilingual Concordancer, NLTK (Natural Language Tool Kit), Simple Concordance Program, Star-Tagger, TEXOR (Topic Extractor for chats), TextLab, TextStat, Topicalizer, TF/IDF Ranker, treetagger, UAM CorpusTool, UNIS Summarizer, wac-tk (Web As Corpus Tool Kit), WordSmith, Xaira, Y-Stemmer. This list is not exhaustive and it would be beyond the scope of this paper to describe all of them in detail. The sheer quantity of different software packages shows not only that the application of corpus linguistic methods is still a growing field but also that there is no standard software suite available.

teachers and tutors, depending on the subject. Furthermore, users can exchange ideas and share experiences with other users via chat rooms, forums and wiki-style glossaries.

The setting for the project is international, with the (initial) use of English and German as platform languages. The first trial launch is scheduled for November 2012 and the cooperating departments include the German Linguistics department of Heidelberg University[2], the German Linguistics department of Budapest University[3] and the General Linguistics department of Stellenbosch University[4]. Figure 1 below is a screen shot of the home page in its current beta state.
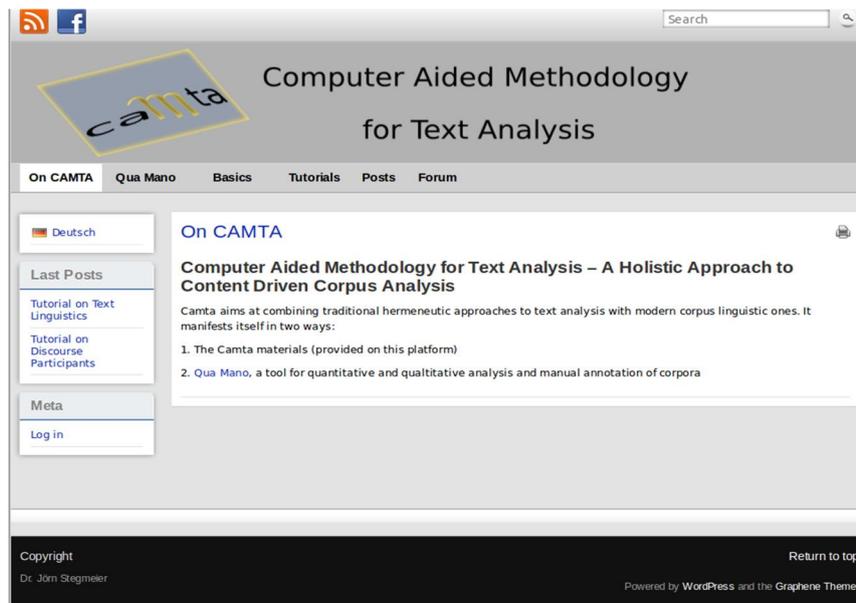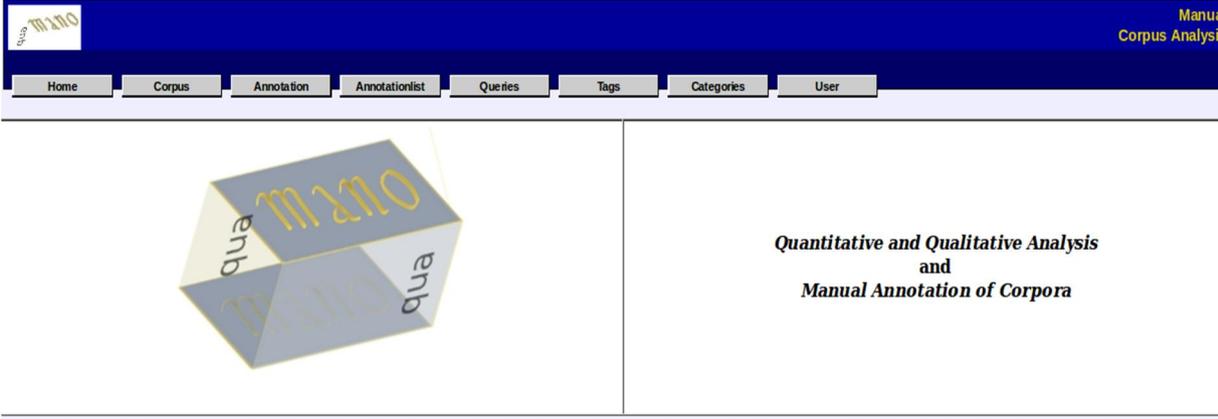

Figure 1. The CAMTA home page in its current beta state

In this paper, I will present excerpts from the materials provided on the platform and demonstrate the use of Qua Mano (or "Quantitative and Qualitative Analysis and Manual Annotation of Corpora"), a tool for corpus analysis which is currently being developed. Qua Mano aims at combining quantitative and qualitative approaches to corpus analysis. Figure 2 is a screen shot of Qua Mano's welcome page.

---

[2] See http://www.gs.uni-heidelberg.de/sprache02/
[3] See http://germanistik.elte.hu/lehrstuhlfursprachwissenschaft.htm
[4] See http://www.sun.ac.za/linguist/

Figure 2. Qua Mano's welcome page

A demonstration of Qua Mano's basic features will be provided in the penultimate section of this paper.

2.   Basics: Corpus linguistics in relation to discourse analysis

At present, corpus linguists are using "large bodies of naturally occurring language data stored on computers" as well as "computational procedures which manipulate this data in various ways" in order to find linguistic patterns (Baker 2007:1). The primary aim of discourse analysis is to establish an understanding of language in use, the "unit of analysis [being - JS] not an abstract 'language' but the actual and densely contextualized forms in which language occurs in society" (Blommaert 2009:15).

So far, both definitions are inadequate as both fields naturally involve much more than this. Still, it is quite clear that the two disciplines are compatible and that corpus linguistics seems to be well suited for application to discourse analysis since "large bodies of naturally occurring language data" is exactly what Blommaert labels the "unit of analysis". In addition, corpus linguistics is often used as a methodology for discourse analysis; however, it could, and quite probably should, be used more often and with more ease. The reason that corpus linguistic methods do not constitute an inseparable methodology from those of discourse analysis is, in my understanding, due to the completely different scholarly traditions of the two disciplines which created a gap between them (cf. section 2.2 of this paper).

Corpus linguistics, which handles computer-aided analysis of extremely large compilations of texts, is closely related to computational linguistics which is generally closer to Information Technology (IT) than to linguistics proper. However, discourse analysis, which deals with the understanding of language in use, is closer to classical hermeneutics: often, intimate knowledge of the texts that are to be analysed is expected and even regarded to be a self-evident part of the methodology. Corpus linguists, on the other hand, regard it as no less self-evident that intimate knowledge of the texts to be analysed is impossible to gain since, for them, corpora do not consist merely of several texts but rather of several thousand texts

(Bubenhofer 2009:16). The aim of CAMTA (and this paper, albeit on a much smaller scale) is to contribute to bridging this gap.

The question then remains: why should one use a computer-aided methodology for discourse analysis at all? An obvious answer is that discourse analysis relies on the use of text corpora and should therefore be a part of, or at least benefit from, the rapid development of corpus linguistics. However, discourses have been successfully analysed without the need to develop extensive computational tools like automated annotation software or sophisticated retrieval programs, even though the use of corpora has been part of discourse analysis from the beginning. That discourse analysis can benefit from computational assistance is, in general, also true for the use of statistics: discourse analysts have been using corpora in a general, non-statistical manner and, by virtue of hermeneutic methods and profound linguistic training, have been able to gain insight into the workings of discourse. By doing so, a tradition of philologically motivated research into discourse has been established, which a computer-aided methodology for discourse analysis should not strive to substitute in any way with statistic measurements and/or automation. Rather, computer-aided methods of all kind should be seen as complementary, providing the analyst with an ever-growing range of tools which, in turn, will lead to a much better understanding of language use (Archer, Culpeper & Davies 2008:619).

The two disciplines, discourse analysis and corpus linguistics, will only converge if scholars make an effort for them to do so. In this paper, I will briefly outline why it is worthwhile to make this effort, as well as describe how this effort can be made without the need for scholars engaged in discourse analysis to become corpus linguists or computational linguists themselves.

The benefits of using a computer-aided methodology for discourse analysis are numerous. Using computational help in discourse analysis is aimed at combining qualitative approaches with quantitative ones, which aids the researcher in the following aspects:

- Corpus size, i.e. the number of texts analysed:

  Text corpora provide large databases of naturally occurring discourse, enabling empirical analyses of the actual patterns of use in a language, and, when coupled with (semi-)automatic computational tools, the corpus-based approach enables analyses of a scope not otherwise feasible (Biber, Conrad & Reppen 1994:169).

- Corpus quality, i.e. the sustainable enriching of corpus texts with additional information (annotation and mark-up):

  Whether it is actually possible to systematically find a language resource strongly depends on the existence of metadata referring to the nature of the resource and on its being stored in an accessible repository (Lehmberg & Wörner 2008:484).

- Application of statistical approaches, i.e. both the use of probabilistic models to assist in the detection of patterns of any kind and the use of statistical measures to describe the significance of these findings:

Statistical inference allows the linguist to generalize from properties observed in a specific sample (corpus) to the same properties in the language as a whole […]. Statistical inference requires that the problem at hand is operationalized in quantitative terms, typically in the form of units that can be counted in the available samples (Baroni & Evert 2009:777).

- Corpus handling, i.e. the use a scholar makes of a corpus and the ease with which he or she does so.

In this paper, I will focus on the non-statistical parts of the computer-aided methodology. The statistical part of the methodology will be developed and described at a later stage of the project. I will now briefly describe the basic idea of how the fields of discourse analysis and corpus linguistics can be, at least partly, reconciled.

2.1   Relevant aspects of discourse analysis

There are a great number of introductions and other works[5] on discourse analysis available which will not be reviewed in detail here. For the purposes of this paper, it will suffice to provide the basic notions and goals of discourse analysis, concentrating on those aspects which link easily with corpus linguistics.

The term "discourse" has various definitions. Firstly, in line with linguistic tradition, it is used to express "language above the sentence" or "language in use" (cf. Blommaert 2009, Brown & Yule 1983, De Beaugrande & Dressler 1981). Discourse analysts following this definition may ask questions about the discourse structure of texts, identifying certain elements as typical. Baker (2007:3), for example, illustrates this by analysing the discourse structure of recipes.

Other uses of the term refer to specific types of language use as in "political discourse", "media discourse" or "learner discourse". The conceptualizations of these discourse types focus on genre and style by studying, amongst other things, the vocabulary typical of a certain genre or the use of hedge words – like "possibly" or "perhaps" – in learner discourse.[6]

Foucault (1972:49) views discourse as "practices which systematically form the objects of which they speak". Focusing on this definition, discourse analysis – by way of language and text analysis – looks for any traces of the social interaction and interpretation which construct (part of) our world as a social practice. Discourse on a certain subject, like human rights, smoking or euthanasia, in this sense, is a phenomenon which is hypero-individual and can be described as a network of topics and positions which materialize in language. The formal units carrying traces of discourse are manifold, starting with lexical items ranging from morphemes to n-grams (phrases containing two or more words) like multi-word phrases. However, besides lexical items, certain constructions and communicative settings are also

---

[5] Cf. Gee (2010) and (2011), Teubert (2010), Wodak (2008) and (2009), van Dijk (2008) and (2009), Blommaert (2009), Blackledge (2009), Van Leeuwen (2008), Johnstone (2008), Fairclough (2007) and (2008), Bührmann (2008), Titscher (2007), Rapley (2007), Machin (2007), Bloor & Bloor (2007), Biber, Connor & Upton (2007), Paltridge (2006), Schiffrin (2005), and Brown & Yule (1983).

[6] A study by Flowerdew (2000), for example, showed that learners of English, in comparison to native speakers, under-use such hedge words, hence adopting too direct and informal a style.

prone to carrying these discoursal traces. These include, for example, verbal aspect and tense, page layout, and the semantic relation between writing and graphical illustration. They also include the purpose of a given text (whether it is a press release, an interview, a comment, etc.), the author and, most importantly, his or her role as a participant in a certain discourse. A list of the linguistic resources used for qualitative discourse analysis is given in Mautner (2008).

Any of the elements mentioned here can be made accessible by annotation techniques, making it possible to track traces of discourse in any kind of formal element. However, of all these elements, lexical items are probably among the most accessible. Frequent use of combinations of lexical items cannot only be interpreted as a grammaticalization process but also as a process that furthers the proximity of the concepts which those lexical items evoke.

2.2   Research perspectives in corpus linguistics

First and foremost, the use of a corpus is essential to engage in corpus linguistic research, which then makes it necessary to decide on a design and on the content (Hunston 2008). Furthermore, in order to employ the computational methods mentioned in my introduction, a way of computing is also vital. In this regard, the impact of personal computers on corpus linguistics and on the way corpora are being used has been tremendous.

With personal computers becoming more and more frequent, corpus linguistics developed rather rapidly into a diverse field where various interests led to various research perspectives. Considering the actual research questions and the academic background of the scholars involved, the most prominent of these perspectives are the technological/mathematical on the one end and the philological on the other. This also leads to the gap between state-of-the-art of corpus linguistics and the use non-corpus linguists – whom Partington termed "linguists who use corpora" (2003:258) – make of it.

Figure 3 shows some of the aspects of these different research perspectives. Each one will be briefly described in the following sections.
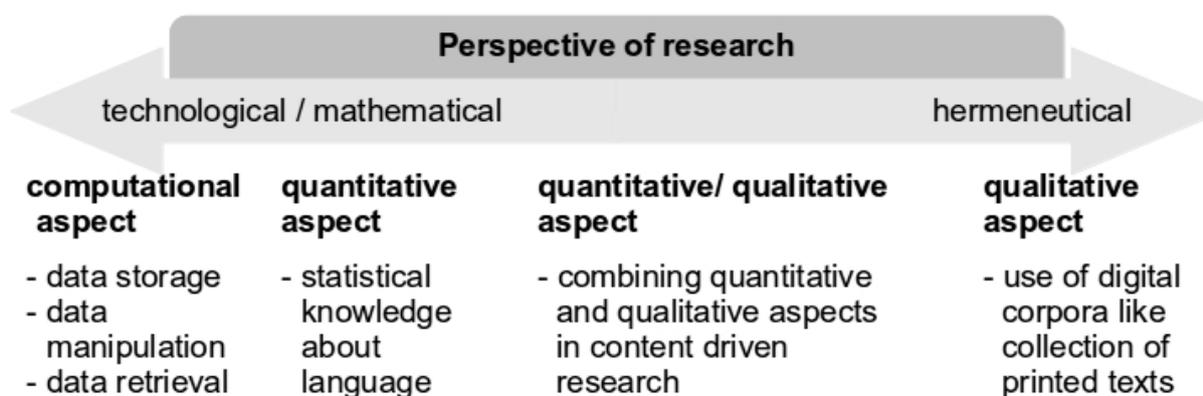

Figure 3. Perspectives of research in corpus linguistics

2.2.1   Computational aspect

There is without a doubt a computational aspect to corpus linguistics. Projects in this field are

mainly concerned with various types of data handling, especially data storage and data structure, data manipulation and data retrieval. Lüdeling and Kytö note that "in computer linguistics, corpus data are exploited to develop Natural Language Processing (NLP) applications" (2008:x).

Data storage and structure means that any information that is part of a corpus needs a format, like plain text or XML (Lehmberg & Wörner 2008). Most corpora are more than just huge folders containing thousands of plain text files. Language data in corpora come, more often than not, with markup, or additional data. These could take the form of, for example, the layout of the original (if it is written text), the date of recording or publication, the name of the author and other information of a descriptive nature. It is important to note that this information needs to be distinguishable from the language data proper, so the format chosen for the corpus must allow for this and also for another type of information to be added, namely annotation. Annotation is any linguistic information that is not obvious by the occurrence or co-occurrence of formal units.

Information of these kinds needs to be added to the texts and transcripts that are part of a given corpus. Corpus linguists then require a means for data manipulation. From a computational perspective, this quite often means that the researcher has to write his or her own program from scratch or customize freely available open-source programs to suit his or her needs. The result is a variety of corpus tools designed to either automatically tag each word of a corpus with its respective part-of-speech (POS) class, or use a so-called "parser" to automatically assign syntactic functions to the lexical units of a text. The software that does this needs to be rooted in linguistic knowledge to make sure the units which are found and labelled are those in which linguists are interested (cf. Schmid 2008, Fitschen & Gupta 2008, Rayson & Stevenson 2008, Kermes 2008, Archer et al. 2008).

The need for specific data manipulation software is also why some corpora cannot be accessed by any freely available software designed to compute frequency lists and such. Such corpora, which use their own format, also need their own retrieval software (Gries 2009:13).

Even though it seems quite obvious at first glance that software like parsers or taggers (which annotate different elements of language like POS, named entities or semantic fields) needs to be rooted in linguistic and/or grammatical theory and knowledge, corpus linguistics also strives to develop general tools that do not subscribe to any given theoretical framework. For example, POS differ in number and kind. Therefore, a POS tagger allowing for only one tag set is useless if you do not agree with the categories used by the tagger (Atwell 2008). Even though POS might not seem very interesting for interpretive text analysis, a corpus tagged with them allows for searches that retrieve co-occurrences of word classes, or of a given word, like 'euthanasia', with a certain word class like, for example, preceding adjectives. The result of a search like this shows semantically-rich combinations like 'active euthanasia', 'passive euthanasia', etc..

Corpus linguists also strive to look into patterns that derive from the data itself. This is why in corpus linguistic research projects it is often stated whether the approach is corpus-based or corpus-driven. The former refers to the deductive use of the corpus where a theory is employed or a hypothesis is tested. The latter, however, relies more on the data and the distribution itself. With this approach, corpus linguists try to find new and completely data-

compliant models to classify formal units of language. For example, the proximity with and frequency in which certain units co-occur with other units is a possible manner for unit description and classification.

2.2.2   Quantitative aspect

Quantitative research projects might be concerned with, for example, the number of foreign words in a given language, or the total number of prepositions in modern texts to be compared with findings from medieval texts. Frequency dictionaries are also a popular application of quantitatively-motivated research (Alekseev 2005).

There is always a quantitative aspect to data and there is always an intuitive, non-statistical way with which to handle the data. Very frequently occurring items are the most important ones. The most frequent words of a corpus are interesting to analyse from different perspectives and frequency lists usually generate further queries from the corpus. Nevertheless, large numbers alone do not do the quantitative aspect of language data any justice:

> […] corpus data must be evaluated with tools that have been designed to deal with distributional information and the discipline that provides such tools is statistics (Gries 2009:4).

Therefore, basic knowledge about standard statistical measures should be a part of any corpus linguistic activity. However, this does not mean that every corpus linguist requires a strong background in mathematics and/or statistics; it means that any statistical knowledge should be made accessible to linguists as easily as possible.[7] As has been mentioned previously, narrowing down the use of statistics in linguistics in general to the use of statistics in interpretive text analysis will be part of future work.

2.2.3   Quantitative/qualitative aspect

The quantitative/qualitative aspect is where linguistic theories and models are tested or set as framework for corpus analysis. There is a wide variety of linguistic research that makes use of corpora. This includes, but is not limited to, research on first and second language acquisition (Diessel 2009), grammar and corpora (Stefanowitsch & Gries 2009), productivity of morphological processes (Baayen 2005), and the use and limits of corpora for syntactic research (Meurers & Müller 2009).

The researchers engaging in studies which are primarily interested in learning something about their linguistic field of research are Partington's "linguists who use corpora". They do not engage with and are not concerned with writing their own scripts, let alone software. Nevertheless, as soon as they utilise such tools, they realize that they are now dealing with distributional and frequency information and treat it accordingly, meaning that they put the statistics toolbox to use.

2.2.4   Qualitative aspect

The qualitative aspect of corpus linguistics uses digital corpora not for their merits with

---

[7] This is increasingly often the case, as has been indicated by numerous introductory and intermediate textbooks such as Baayen (2008), Gries (2008), Best (2006), Bubenhofer (2006), Johnson (2008), Köhler, Altmann and Piotrovskiĭ (2005), Rietveld & Hout (2005), Manning & Schütze (2003) and Oakes (2003).

regard to information on frequency and such, but as a substitute for a collection of, for example, printed newspaper articles. The approach of Partington's "linguists who use corpora" to their subject is that of a traditional philology: they read as much as they think they need to in order to interpret the data against the background of the theory they wish to test or prove.

3.   Using corpus linguistics as a methodology for discourse analysis

Discourse analysis does not depend upon corpora in the sense of large bodies of digitalized text. It is quite common, as has been previously mentioned, to analyse discourse by traditional means, namely reading the most appropriate or promising texts for your study, which then renders the use of large digital corpora unnecessary. Baker deflects critique based on this observation:

> One criticism of corpus-based approaches is that they are too broad – they do not facilitate close readings of text. However, this is akin to complaining that a telescope only lets us look at faraway phenomena, rather than allowing us to look at things close-up. (Baker 2007:7)

The differences in scope do not necessarily mean that corpus linguistic approaches are completely useless for a given project – on the contrary, since corpus linguistics provides insight into discourse in a completely different way, it might just be a change in perspective that is needed for an all-encompassing interpretation.

Baker also notes that the traditional way of reading might be tricky. He believes that analysing discourse from an individual perspective harbours the risk of completely excluding other possible perspectives. It is my belief that using a large body of texts should be a way around that. (The risk of being selective also applies to the corpus-building process which I will address very briefly in sections 2.3.2 and 2.3.3.)

> By using a corpus, we at least are able to place a number of restrictions on our cognitive biases. It becomes less easy to be selective about a single newspaper article when we are looking at hundreds of articles – hopefully, overall patterns and trends should show through. (Baker 2007:12)

Therefore, as is the case with all methods, it is sensible to find out what a corpus-based or corpus-driven approach to any subject can yield and choose another method if the use of corpus linguistics does not seem appropriate. Assisting students and scholars in determining which approach seems appropriate with regard to a given research question is therefore another goal of the project described here.

Besides a corpus and the means to manipulate the data therein, corpus linguistics relies on some basic assumptions, notions, and methods. The basic assumptions are that "formal distributional differences reflect functional differences" and "corpus-linguistic analyses are always based on the evaluation of some kind of frequencies". The basic notions are "corpora, representativity and balancedness, markup and annotation" and the basic methods are "frequency lists, concordances, collocations" (Gries 2009:1).

The basic assumptions concern the whole design of a research project (cf. section 2.3.1). The

basic notions all relate to corpora in one way or another. Firstly, there is the question of what exactly a corpus is. Here, a decision must be made regarding what to include in and exclude from the corpus, and how to handle the information about those texts like, for example, authorship or the date of publication. This is briefly elaborated upon in section 2.3.2. The basic methods refer to the ways of going about analysing the data which will be explained in a little more detail in section 2.3.3.

3.1.   Basic assumptions

The first basic assumption is that "formal distributional differences reflect, or correspond to, functional differences." Gries (2009:4) notes that the use of "functional" here is used in a general sense to refer to "anything that is intended to perform a particular communicative function." This basic assumption subscribes to the view that there is no true synonymy in language; in other words, lexical units or formal elements, whether morphemes, words or phrases, always differ slightly in their meaning. Therefore, a change in the usage of these formal elements translates to a change in meaning.

The second basic assumption is that "corpus-linguistic analyses are always based on the evaluation of some kind of frequencies." This assumption takes into account the basic nature of corpora, namely that they consist of data that are meaningless without someone to interpret them. Therefore, analyzing a corpus computationally results in numbers showing the frequency of occurrence or co-occurrence of formal elements.

To determine what those frequencies mean in terms of a linguistic framework or theory remains the concern of the researcher who is conducting the corpus linguistic research. This is especially true if the main interest in using a corpus for one's study is a linguistic one, as one can then view corpus linguistics as a methodology for linguistic research.

Using frequency information is feasible in order to find and investigate lexical units dealing with topics within a discourse, as a finding of Biber et al. (1999:53) suggests:

> In longer texts there is a greater chance that words which have already been used will be repeated. This is true both of the most frequent words which recur in all kinds of texts (*the, of, and,* etc.) and of the words which are connected with the topic of a particular text.

3.2.   Basic notions

The first question which needs to be addressed in corpus linguistics is "What is a corpus and how is it constructed?" After a short answer to this question, I will briefly introduce the notions of corpus 'representativity', 'balancedness' and text 'markup' and 'annotation' as a means to enrich the data compiled in a corpus.

The term "corpus" today almost always refers to a digital collection of language recordings in various formats. This means that a corpus can consist of digitalized text documents as well as audio and video files. The ways in which this is technologically realized vary, but it is safe to say that the various files are usually part of a framework structure like a database or an XML document or simply within the same folder.

This indicates that building a corpus can be simple provided that all of the data to be

investigated and analysed are stored as plain text files on one's computer. However, this is not usually the case, so it is necessary to gather the texts and/or recordings before they can be formatted for use in a corpus. Nevertheless, the purpose of a corpus is, amongst others, to provide the simplest form of access to written texts in digitized form. For example, if you are working on a project centred on newspaper articles of the 1980s and, after a quick online search, it becomes apparent that all of the necessary texts are archived online, all you have to do then is decide whether or not to keep a copy of the original web page in order to access the co-text and the pictures, and then convert the files into plain text.

Unfortunately, research projects are not always this simple. The researcher may find that he or she will need to scan the texts involved using an Optical Character Recognition (OCR) program which outputs plain text files, scrutinize those files for recognition mistakes and then import them into the corpus structure. Audio and video recordings are more labour intensive in that both will need to be transcribed before they can be put into plain text format. Here, certain decisions need to be made in terms of how detailed a transcription is necessary for the purposes of the research. If a corpus is being built to allow for various types of analyses, it is essential that the transcriptions are as thorough as possible. However, this is not always possible as, as is often the case, issues of time and funding sometimes pose restrictions on the quality of the transcriptions and researchers often end up transcribing only what is absolutely necessary. Ultimately, availability and copyright issues are the factors that decide whether the corpus that a researcher needs can be put together or not (Partington 2003:258). Helpful advice on corpus building can be found in Wodak and Krzyzanowski (2008), Mautner (2008) and Gruber (2008), amongst others.

To make a corpus representative of what is under investigation, the researcher has to think as exhaustively as possible about what he or she already knows. For example, if the vocabulary of business reports is under investigation, the researcher has to decide which businesses are representative for this kind of language use.

In order for the corpus to be balanced, the researcher's corpus should mirror the proportion of different textual categories in real life. Continuing with the previous example of business reports, this means that the researcher should know beforehand by which textual categories business reports can be classified as well as knowing the proportion of these categories. Both notions are "a theoretical ideal corpus compilers constantly bear in mind, but the ultimate and exact way of compiling a truly representative and balanced corpus has eluded us so far" (Gries 2009:8).

As was previously mentioned, "markup" refers to the information which describes the text in question, whereas "annotation" is the linguistic information on the text in question. Both of these notions are very important for corpus linguistic research. Without annotation, there would be, for example, no way in which to acquire information on the distributional frequencies of certain n-grams, no POS-tagging, etc. Without markup, all information about the context would be lost – it would be impossible to know who uttered or wrote a certain text and on which day, etc.

Markup and annotation may well be the key to using corpus linguistics as a methodology for discourse analysis, since annotating a corpus denotes assigning an interpretation to a certain portion of text. This enables us to systematically annotate sociolinguistic, cognitive and

pragmatic categories in the broadest sense, like language variety, semantic frames and scripts, speech acts, and presuppositions, to name a few.

3.3.   Basic methods

What follows is a brief introduction to and demonstration of the basic methods of corpus linguistics, namely frequency (or word) lists, concordances, clusters and lists of keywords. The textual basis for this demonstration will be a statement from the website of the organization Doctors For Life International[8] (cf. Appendix).

To make full use of the interpretative value of any result made available by corpus linguistic tools one needs to be able to relate those results to one's own understanding of the weight carried by the various units of analysis. For this reason, it is essential to start out and remain in control as far as possible during the analytical process. It is therefore suggested that one begins an analysis by using only one, carefully-read text.

The basic methods of corpus linguistics provide a different view of the texts linguists investigate. By that, they enrich the analytical and interpretative toolbox and make it possible to base interpretations on much more data than could possibly be yielded by the traditional approach of simply reading the texts.

The output of corpus linguistic tools consists of numbers and text, the numbers providing information on e.g. frequency (of characters, words, sentences etc., per text, per paragraph etc.), type-token ratio (TTR), and/or lexical density.[9] Textual output from corpus linguistic tools usually consists of parts of the analyzed text, often combined with frequency information on the listed lexical units.

Those data help to break up the purely linear processing of text, thus freeing the analyzing researcher from the need to read every single word and enabling him or her to look at the text from another perspective than would a normal reader. Any of those "views" – be it a (frequency) list of all the words or of keywords chosen by the user or a list of automatically computed keywords – is the result of computing carried out by a certain software program or scripts. Therefore, it seems most appropriate to amass as much data to any given text as reasonable and use these data as a means to focus on different aspects of the texts without getting lost in the detail.

The following demonstration will make use of AntConc as well as some of the other tools listed in footnote 1. AntConc was developed by Laurence Anthony from Waseda University, Japan[10]. It is free of charge and is one of the most frequently used concordance programs today.

For this demonstration, the mock research question is "What are the positions on euthanasia in the discourse on euthanasia?". The researcher's first step is to familiarize him- or herself as far as possible with the content of his or her corpus, i.e. to get an overview of the participants and

---

[8] www.doctorsforlifeinternational.com/issues/euthenasia.cfm

[9] While the type-token ratio does not take into account whether words are content words or function words, the measure of lexical density is usually understood as the "proportion of lexical items (content words) to the total discourse" (Halliday & Webster 2005:329); Castello (2008) provides a useful overview.

[10] www.antlab.sci.waseda.ac.jp/antconc_index.html

the positions. By taking notes on anything the researcher deems relevant, he or she builds up knowledge both about the discourse and the interpretations thereof.

Combining this first step with the help from a corpus tool translates into an attempt to automatically compute the topics discussed in your corpus. In the euthanasia example, this means that a software program should be able to extract lexical units that refer to topics like "active euthanasia", "patient's rights" and "passive euthanasia" which are all mentioned in the first paragraph. The software tool TEXOR, which "is supposed to process any texts to detect their main topics"[11] , yielded the results in Figure 4:

Figure 4. Topic extraction using TEXOR

While the results reflect parts of the text in question, they are not sufficiently meaningful to provide insight into the text without further reading or making use of another means of text analysis. In contrast to this, the output of the software tool "Topicalizer" (cf. www.topicalizer.com) is much more usable: Not only does it compute some statistical meta-data, like the number of tokens and types, the average sentence length in words, the average number of sentences per paragraph etc., but it also outputs some very useful keywords and 2-grams which provide suitable starting points. Topicalizer's keywords and 2-grams are very alike in the euthanasia example which is why they are presented together here in one slightly edited list.

---

[11] http://cll.khsu.ru/chatsumm/main.aspx

active euthanasia	intentional killing
contemporary healthcare/medicine	patient's right
extraordinary intervention	set secure limits

The drawbacks on Topicalizer are that it can only be used via its web interface and that its use is restricted to 10 requests per hour. This means that using it to analyse a large corpus would be very tedious and time consuming. There are many other tools and programs available to compile lists of phrases. The annotation software adds a POS-tag to every token of a text, enabling the script to find, for example, any combination of adjective and noun. Attributive adjectives followed by nouns is a good starting point to look for interesting topics (cf. Stefanowitsch & Gries 2009). The following list is the result of utilizing this approach (using the software Treetagger – cf. www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ and Schmid (1994) and (1995) - and having deleted less informative phrases and concatenated some entries):

active euthanasia	involuntary / voluntary euthanasia
incurable disease	medical community / treatment
intentional killing	ordinary treatment
contemporary healthcare/ medicine	secure limits
contemporary social evaluations	society's prohibition
general prohibition	truly voluntary

A list of keywords can be easily computed by feeding the corpus (i.e. the statement of Doctors for Life) and a reference corpus (i.e. another text of comparable length dealing with a completely different subject) into a program that can compare weighted word lists of different corpora. Using AntConc and an excerpt of the treaty of the European Union, the ten most important lexical keywords, in this case, are:

euthanasia	cure
active	dfl
killing	die
intentional	law
pain	life

Upon first glance, a list of single words might not seem very helpful. Considering, however, that not all expressions typically used in the discourse on a certain subject are n-grams, the value of such a list is evident: just like phrases, they serve as non-linear entry points into the discourse by their use as search terms. With AntConc, this either results in so-called "clusters", consisting of the search term and co-occurring lexical items, or in a list of the collocates of the search term. Taking "euthanasia" as the search term, the results are very similar to the keywords and phrases already presented (which is not surprising considering that they were taken from the exact same text as the search term). In addition to these, the following two phrases are of interest:

euthanasia starts
is euthanasia

A frequency list of the complete text from the respective corpus concludes the collection of possible search terms. From the example text, the most frequent words are the same as the lexical keywords (cf. previous list of lexical keywords). The words and phrases we have found so far are prominent in the discourse to be analysed which is why it is economic to use them as search terms in order to find conclusive parts of the text or corpus, i.e. concordances of the search terms we are most interested in. The concordances for "euthanasia" are presented in Figure 5 below:

```
IA .·Doctors·for·Life·(DFL)·is·opposed·to·'active'·euthanasia,·that·is·euthanasia·that·involves·the·intentional·
·(DFL)·is·opposed·to·'active'·euthanasia,·that·is·euthanasia·that·involves·the·intentional·killing·of·a·person·
's·right·to·refuse·medical·treatment·or·'passive'·euthanasia.··DFL·supports·the·present·ban·on·active·euthanas·
thanasia.··DFL·supports·the·present·ban·on·active·euthanasia·in·South·Africa.·We·believe·that·it·should·be·kep·
use·any·relaxation·would·be·open·to·abuse.·Active·euthanasia·could·put·pressure·on·the·old·and·the·sick·to·see·
table.··In·general,·arguments·in·favour·of·active·euthanasia·are·insufficient·to·weaken·society's·prohibition.·
lue·judged·by·a·second·party.·We·look·upon·active·euthanasia·prohibition·as·the·cornerstone·of·law·and·of·soci·
e·should·be·no·change·in·the·law·to·permit·active·euthanasia.··Finally,·we·at·DFL·do·not·believe·that·it·is·po·
·it·is·possible·to·set·secure·limits·on·voluntary·euthanasia.·Involuntary·euthanasia·based·on·contemporary·soc·
ecure·limits·on·voluntary·euthanasia.·Involuntary·euthanasia·based·on·contemporary·social·evaluations·of·age·(·
·be·next·to·impossible·to·ensure·that·all·acts·of·euthanasia·were·truly·voluntary,·and·that·any·liberalization·
he·limits·of·any·regulation.····The·road·to·active·euthanasia·starts·with·an·attitude·shift·in·the·medical·comm·
```

Figure 5. Concordances of the search term "euthanasia"

The mock research question, "What are the positions on euthanasia in the discourse on euthanasia?", can now be answered. Furthermore, positions and arguments regarding the types of euthanasia mentioned in the corpus can be described or visualized using, for example, the software tool GraphViz, as demonstrated in Figure 6.
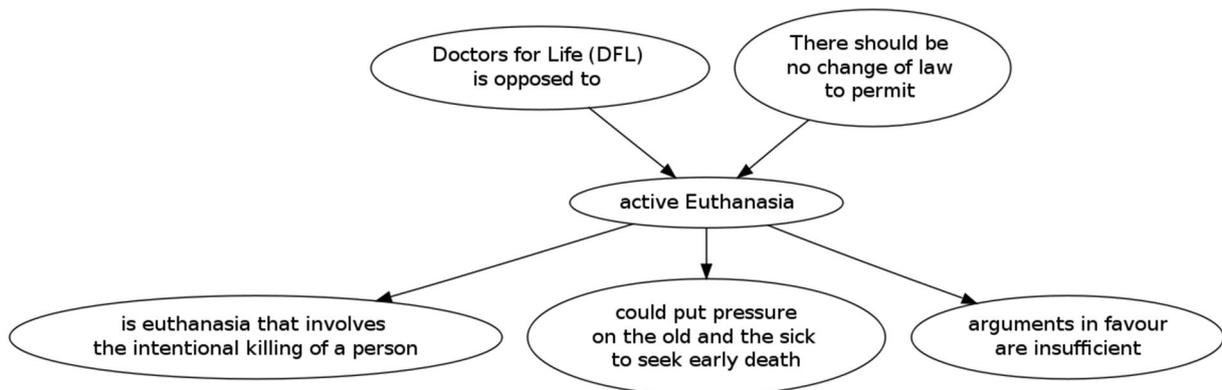

Figure 6. Visualized concordances of 'active Euthanasia' using GraphViz

The participant "Doctors for Life" and the organisation's positions towards active euthanasia can now be saved as a graphic representation. Subsequently, the procedure can be carried out for every multi-word phrase, keyword and highly frequent token that is of interest. In this way, much valuable information on the discourse can be gathered, analysed and interpreted. The example shows that even if one cannot read a certain text because the size of the corpus does not allow for every text to be read, this combination of tools can provide the researcher with a range of information from which to draw conclusions.

4.   Putting Qua Mano to use

In order to be able to do the kind of analysis I briefly demonstrated above, one needs a basic understanding of both corpus linguistics and discourse analysis and also some practical

training in the use of the software involved in corpus linguistics. The high number of available software tools and the need to combine the various methods according to a specific research question is much easier and faster to master if there is help available that is tailored to your needs.

One master principle of this project is that any methods and approach will be first introduced in seminars, including video seminars, since familiarising oneself with software solely through a textbook is rather tedious work unless you are an experienced computer user. Therefore, the features of the e-learning platform include customized documentation on corpus linguistic tools with very specific instructions (made available in writing and as video files) that show exactly how things introduced in a seminar were done.

The e-learning platform not only gives access to tutorials but also to basic software tools which can be used directly on the server. It also offers the use of customized scripts using the software available on the platform, namely Qua Mano.

In addition to the analytic possibilities discussed in section 3, Qua Mano allows for the manual annotation of any word or combination of words in every text of the corpus. This is accomplished by storing the texts in a tokenised form with every token being uniquely identified by an identification number. Any enriching information, be it POS-tags or manually-entered clues for interpretation is stored with reference to the word's identification number, enabling the scholar to search on three levels: words, tags from tagging tools like Treetagger, and manual annotation.

After pre-processing and storing the example text in Qua Mano's database, it can be searched by using either predefined or manually-entered SQL[12] statements on the query page, as in Figure 7 below.

Figure 7. Selecting a predefined SQL statement in Qua Mano

Selecting the first predefined statement results in showing all available predefined statements and an explanation of what the statements are for, as in Figure 8.

---

[12]SQL is a special-purpose programming language designed for managing data in relational database management systems.

Figure 8. List of all predefined SQL statements

In Figure 9, the third statement produces a list of co-occurrences of adjectives with nouns, similar to the one generated in Figure 8.

Figure 9. Selecting an SQL statement that will generate a list of co-occurrences of adjectives and nouns and also provide information on the frequency of the pairs

The list generated by this query not only shows the co-occurrences themselves but also how frequently they appear. The pair 'active euthanasia' occurs seven times in the text used for this demonstration.
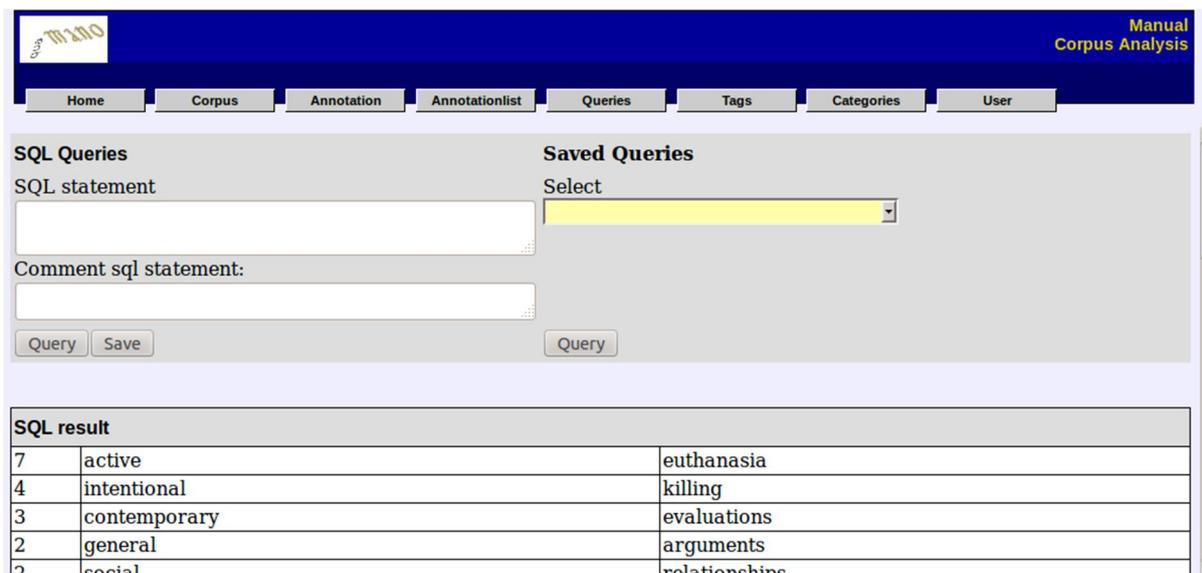
Figure 10. List of adjective/noun pairs with frequency information

Manual annotation is carried out in Qua Mano's annotation window using combinations of tags and categories which are able to be customised. Figure 11 shows the manual annotation of discourse participants. After having created the tag 'Discourse Participant' and several applicable categories like 'active (speaks about)', 'passive (is spoken about)' (cf. Figure 12), combinations of tags and categories can be saved to speed up the annotation process (cf. Figure 13). Tags and categories are created in respective panes and can be modified at any time.
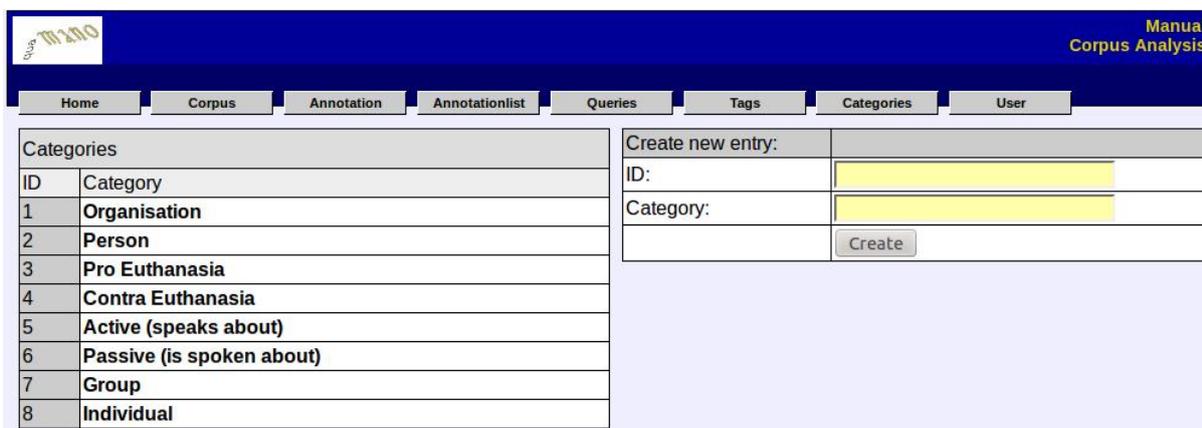

Figure 11. Creating the tag 'Discourse Participant'

Figure 12. Creating categories by which discourse participants can be described in detail

Figure 13. Creating combinations of tags and categories

After creating tags, categories, and combinations of the two, the annotation itself is a simple matter of selecting the portion of the text that is to be annotated by clicking on it and choosing the tag, category or combination that is to be attributed to the portion of text. By clicking on a single token, it is logged as a point of reference for the current annotation; by expanding the range, parts of the context will also be saved with the annotation, resulting in manual annotations with concordances, as in Figure 14.
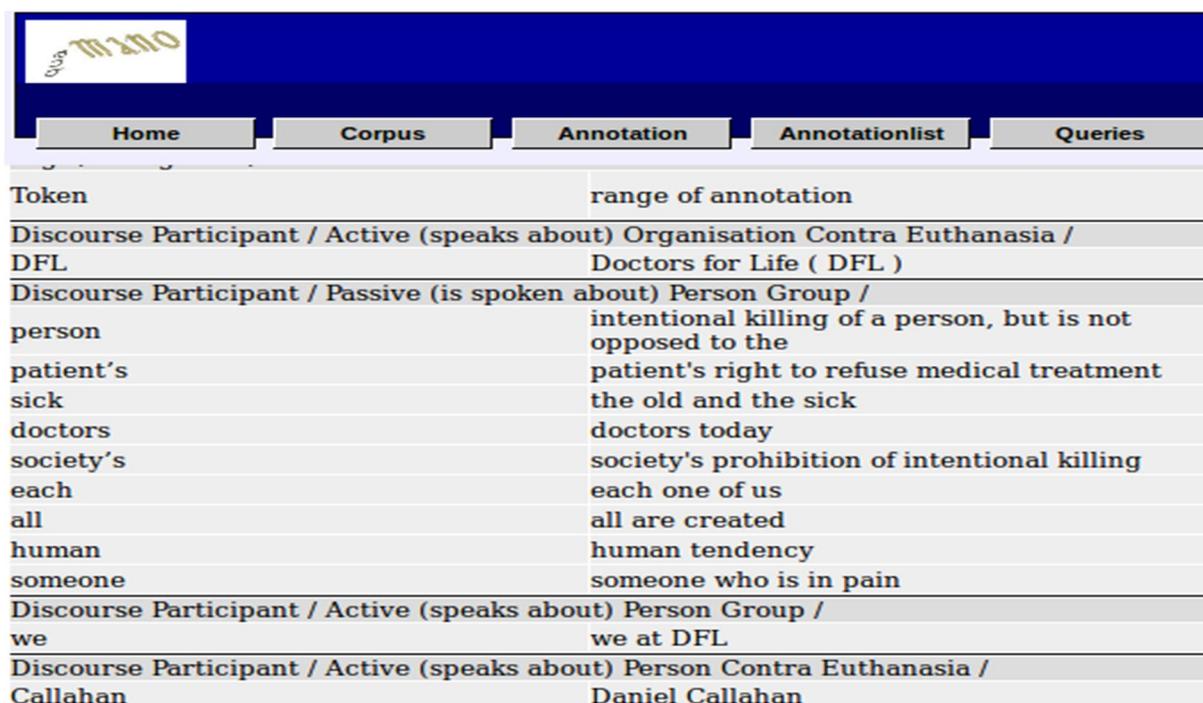

Figure 14. Annotating 'DFL'

After completing the annotation process, the results can be viewed as a list, as in Figure 15:

Token	range of annotation
Discourse Participant / Active (speaks about) Organisation Contra Euthanasia /	
DFL	Doctors for Life (DFL)
Discourse Participant / Passive (is spoken about) Person Group /	
person	intentional killing of a person, but is not opposed to the
patient's	patient's right to refuse medical treatment
sick	the old and the sick
doctors	doctors today
society's	society's prohibition of intentional killing
each	each one of us
all	all are created
human	human tendency
someone	someone who is in pain
Discourse Participant / Active (speaks about) Person Group /	
we	we at DFL
Discourse Participant / Active (speaks about) Person Contra Euthanasia /	
Callahan	Daniel Callahan

Figure 15. Result of manual annotation displayed as list ordered by combination

The list in Figure 15 shows the discourse participants and to which combination of tags and categories they belong. Complementing any purely frequency-based automated annotation, it provides the researcher with additional information helping him or her to reach a substantiated interpretation. In this example, the manual annotation reveals that the authors of the text in question try to establish their position as universally valid by referring to groups rather than individuals.

5. Conclusion

The main goal of CAMTA is to enable students and scholars alike to gain exactly that basic understanding and practical training that is needed for good practice in corpus linguistics and discourse analysis. Of course, it will also enable participants to employ corpus linguistics for other projects than discourse analysis since the tools employed in corpus linguistics are widely applicable.

Users of the e-learning platform will be individually supported in their research question with the aim of attracting users of the whole range of perspectives of corpus linguistic interest. Any user's project will be carefully documented with a special focus on the research questions and the methods that are employed to pursue them, which will result in a growing collection of well-proven ways of doing computer-aided discourse analysis.

References

Alekseev, P.M. 2005. Frequency dictionaries. In R. Köhler, G. Altmann & R.G. Piotrovskiĭ (eds.) *Quantitative linguistics*. Berlin: de Gruyter. pp. 312–324.

Archer, D., J. Culpeper & M. Davies. 2008. Pragmatic annotation. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics.*. Berlin: de Gruyter. pp. 613–642.

Atwell, E. 2008. Development of tag sets for part-of-speech tagging. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 501–527.

Baayen, H.R. 2005. Morphological productivity. In R. Köhler, G. Altmann, & R.G. Piotrovskiĭ (eds.) *Quantitative linguistics*. Berlin: de Gruyter. pp. 243–255.

Baayen, H.R. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baker, P. 2007. *Using corpora in discourse analysis*. London: Continuum.

Baroni, M. & S. Evert. 2009. Statistical methods for corpus exploitation. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 777–803.

De Beaugrande, R. & W.U. Dressler. 1981. *Einführung in die Textlinguistik*. Tübingen: Niemeyer.

Best, K.-H. 2006. *Quantitative Linguistik*. 3. Auflage. Göttingen: Peust & Gutschmidt.

Biber, D., U. Connor & T.A. Upton. (eds.) 2007. *Discourse on the move*. Amsterdam: John Benjamins.

Biber, D., S. Conrad & R. Reppen. 1994. Corpus-based approaches to issues in applied linguistics. *Applied Linguistics* 15(2): 169 –189.

Biber, D. (ed.) 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Blackledge, A. 2009. *Discourse and power in a multilingual world*. Amsterdam: John Benjamins.

Blommaert, J. 2009. *Discourse*: *A critical introduction*. Cambridge: Cambridge University Press.

Bloor, M. & T. Bloor. 2007. *The practice of critical discourse analysis*. London: Hodder Arnold.

Brown, G. & G. Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Bubenhofer, N. 2006. Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge. [Online] Available: http://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=cosmas_client_kookk.html [Accessed 4 June 2012].

Bubenhofer, N. 2009. *Sprachgebrauchsmuster*: *Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin: de Gruyter.

Bührmann, A. (ed.) 2008. *Discourse analysis in the social sciences*. Köln: Zentrum für Historische Sozialforschung.

Castello, E. 2008. *Text complexity and reading comprehension tests*. Frankfurt am Main: Peter Lang.

Diessel, H. 2009. Corpus linguistics and first language acquisition. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 1197–1212.

Fairclough, N. 2007. *Analysing discourse*. London: Routledge.

Fairclough, N. 2008. *Critical discourse analysis*. Harlow: Longman.

Fitschen, A. & P. Gupta. 2008. Lemmatizing and morphological tagging. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 552–564.

Flowerdew, L. 2000. Investigating referential and pragmatic errors in a learner corpus. In L. Burnard & T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Frankfurt am Main: Peter Lang. pp. 145-154.

Foucault, M. 1972. *The archaeology of knowledge*. New York: Pantheon.

Gee, J.P. 2010. *How to do discourse analysis*. London: Routledge.

Gee, J.P. 2011. *An introduction to discourse analysis*. London: Routledge.

Gries, S.T. 2008. *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Gries, S.T. 2009. What is corpus linguistics? *Language and Linguistics Compass* 3(5): 1–17.

Gruber, H. 2008. Analyzing communication in the new media. In R. Wodak (ed.) *Qualitative discourse analysis in the social sciences*. Basingstoke: Palgrave Macmillan. pp. 54–76.

Halliday, M.A.K. & J. Webster. 2005. *On grammar*. London: Continuum.

Hunston, S. 2008. Collection strategies and design decisions. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Handbücher zur Sprach- und Kommunikationswissenschaft Vol. 29/1. Berlin: de Gruyter. pp. 154–168.

Johnson, K. 2008. *Quantitative methods in linguistics*. Malden: Blackwell.

Johnstone, B. 2008. *Discourse analysis*. Malden: Blackwell.

Kermes, H. 2008. Syntactic preprocessing. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Handbücher zur Sprach- und Kommunikationswissenschaft Vol. 29/1. Berlin: de Gruyter. pp. 579–598.

Köhler, R., G. Altmann & R.G. Piotrovskiĭ (eds.) 2005. *Quantitative Linguistik: ein internationales Handbuch*. Berlin: Mouton de Gruyter.

Lehmberg, T. & Wörner, K., 2008. Annotation standards. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Handbücher zur Sprach- und Kommunikationswissenschaft Vol. 29/1. Berlin: de Gruyter. pp. 484–501.

Lüdeling, A. & Kytö, M. 2008. Introduction. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Handbücher zur Sprach- und Kommunikationswissenschaft Vol. 29/1. Berlin: de Gruyter. pp. v–xii.

Machin, D., 2007. *Global media discourse*. T. van Leeuwen, ed., London [u.a.]: Routledge.

Manning, C.D. & H. Schütze. 2003. *Foundations of statistical natural language processing*. Massachusetts: MIT Press.

Mautner, G. 2008. Analyzing newspapers, magazines and other print media. In R. Wodak and M. Krzyzanowski (eds.) *Qualitative discourse analysis in the social sciences*. Basingstoke: Palgrave Macmillan. pp. 30–53.

Meurers, W.D. & S. Müller. 2009. Corpora and syntax. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 920–933.

Oakes, M.P. 2003. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Paltridge, B. 2006. *Discourse analysis*. London: Continuum.

Partington, A. 2003. *The linguistics of political argument: The spin-doctor and the wolf-pack at the White House*. London: Routledge.

Rapley, T. 2007. *Doing conversation, discourse and document analysis*. Los Angeles: Sage.

Rayson, P. & M. Stevenson. 2008. Sense and semantic tagging. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 564–579.

Rietveld, T. & R. van Hout. 2005. *Statistics in language research*. Berlin: de Gruyter.

Schiffrin, D. 2005. *The handbook of discourse analysis*. Malden: Blackwell.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. [Online] Available: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf [Accessed 17 January 2013].

Schmid, H. 1995. Improvements in part-of-speech tagging with an application to German. [Online] Available: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf [Accessed 17 January 2013].

Schmid, H. 2008. Tokenizing and part-of-speech tagging. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 527–552.

Stefanowitsch, A. & S.T. Gries. 2009. Corpora and grammar. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 933–952.

Teubert, W. 2010. *Meaning, discourse and society*. Cambridge: Cambridge University Press.

Titscher, S. (ed.) 2007. *Methods of text and discourse analysis*. London: Sage.

Van Dijk, T.A. 2008. *Discourse and power*. Basingstoke: Palgrave Macmillan.

Van Dijk, T.A. 2009. *Society and discourse: How social contexts influence text and talk*. Cambridge: Cambridge University Press.

Van Leeuwen, T. 2008. *Discourse and practice*. Oxford: Oxford University Press.

Wodak, R. and M. Krzyzanowski (eds.) 2008. *Qualitative discourse analysis in the social sciences*. Basingstoke: Palgrave Macmillan.

Wodak, R. (ed.) 2009. *Methods of critical discourse analysis*. Los Angeles: Sage.

Wynne, M. 2008. Searching and concordancing. In A. Lüdeling & M. Kytö (eds.) *Corpus linguistics*. Berlin: de Gruyter. pp. 706–737.

Appendix

Euthanasia

Doctors for Life (DFL) is opposed to 'active' euthanasia, that is euthanasia that involves the intentional killing of a person, but is not opposed to the patient's right to refuse medical treatment or 'passive' euthanasia.

DFL supports the present ban on active euthanasia in South Africa. We believe that it should be kept because any relaxation would be open to abuse. Active euthanasia could put pressure on the old and the sick to seek early death such that the 'right to die' would become the 'duty to die'. To distinguish between murder and 'mercy killing' would be to cross the line that prohibits any intentional killing, a line that is essential to preserve. We can respect a patient's right to request to die especially when the patient has an incurable disease that would require extraordinary intervention. We do not, however, support the termination of ordinary treatment such as pain modifiers, antibiotics, food or water. With pain medications available to doctors today, 99% of pain is treatable.

In general, arguments in favour of active euthanasia are insufficient to weaken society's prohibition of intentional killing. We look at quality of life as an intrinsic value and not a qualitative value judged by a second party. We look upon active euthanasia prohibition as the cornerstone of law and of social relationships. It protects each one of us impartially, embodying the belief that all are created equal. We do not wish for protection to be diminished, and we therefore recommend that there should be no change in the law to permit active euthanasia.

Finally, we at DFL do not believe that it is possible to set secure limits on voluntary euthanasia. Involuntary euthanasia based on contemporary social evaluations of age (young or old), competence, illness, race, genetic differences would be subject to rampant abuses. It would be next to impossible to ensure that all acts of euthanasia were truly voluntary, and that any liberalization of the law was not abused. Moreover, to create an exception to the general prohibition of intentional killing would inevitably open the way to its further erosion, whether by design, by inadvertence or by the human tendency to test the limits of any regulation.

The road to active euthanasia starts with an attitude shift in the medical community. Daniel Callahan in his book "What Kind of Life: The Limits of Medical Progress" criticized contemporary medicine for shifting its focus from care to cure: 'The primary assurance we all require is that we will be cared for in our sickness regardless of the likelihood of cure. The greatest failure of contemporary healthcare is that it has tended to overlook this point, has become

distracted from it by the glamour of cure and the war against illness and death. At the center of caring should be commitment never to avert its eyes from, or wash its hands of, someone who is in pain or is suffering, who is disabled or incompetent, who is retarded or demented.