universität wien

Diplomarbeit

Titel der Diplomarbeit

"Implementing computer corpora as a source in English Language Teaching (ELT)"

Verfasserin

Alexandra Hable

angestrebter akademischer Grad

Magistra der Philosophie (Mag.phil.)

Wien, im Dezember 2010

Studienkennzahl lt. Studienblatt:	A 190 344 299
Studienrichtung lt. Studienblatt:	Lehramtsstudium UF Englisch, UF Psychologie & Philosophie
Betreuerin:	Ao. Univ.-Prof. Mag. Dr. Christiane Dalton-Puffer

Abstract

The implementation of computer corpora into the field of language pedagogy has been an issue of increasing importance for applied linguists throughout the last two to three decades. Due to a lack of communication between theoreticians and practitioners, the potential of this application, however, is still hardly perceived by either experienced or future teachers.

As a consequence, this thesis aims at introducing language pedagogues to the different methods as well as the potential of using computer corpora in the language classroom. This will be done by providing them both with theoretical background information concerning corpora as well as with practically relevant remarks and examples illustrating the various ways in which corpora can be directly applied by both teachers and students inside as well as outside the language classroom. Finally, the sources of information used for this study, namely specialist literature and to a certain extent also personal practical experience with corpora, will be reflected critically, in order to evaluate the potential and limitations of the application of computer corpora in the area of language teaching.

This introduction to and evaluation of potential pedagogical uses of corpora in the field of language pedagogy aims at enabling teachers to make informed decisions regarding the questions whether, when and how to integrate corpora into their language classrooms. A more implicit objective underlying this survey is to draw attention to corpora as influential educational instruments which should not be neglected in the training of teachers any longer, as is frequently the case nowadays.

Acknowledgments

I want to take this opportunity to thank ...

... my family and friends their moral support, their angelic patience and for helping me to relax at times when I was too tense to sleep.

... Mag. Dr. Gunther Kaltenböck first of all for the introductory course on corpus linguistics which he offered and which inspired and motivated me to write this diploma thesis and following for his precious time, valuable opinion and practical guidance throughout the writing process.

... Ao. Univ.-Prof. Mag. Dr. Christiane Dalton-Puffer for supervising my thesis, helping me with organisational issues and giving me the opportunity to work on a topic which is of special interest to me.

All errors, of course, are mine and mine alone.

Table of contents

1. Introduction

At the second ‘Teaching and Language Corpora’ (TaLC) conference held in Lancaster almost fifteen years ago the announcement was made that while computer corpora are already well integrated as analysis tools in the field of research, “they are now being used increasingly for teaching purposes.” (Stewart, Bernardini & Aston 2004: 1) Since then a great number of articles and books regarding corpus linguistics in general and their implementation into the field of language teaching and learning have been published. Unfortunately, the question whether corpora have really found their way into common teaching practice still needs to be answered with ‘no’.

One of the reasons why computer corpora are not yet regarded as powerful pedagogical instruments in the language classroom may be that “many important developments in the field of corpus linguistics are not always communicated or usefully mediated in terms of their implications for language teaching.” (O’Keeffe, McCarthy & Carter 2007: xi) As a result, only a very limited number of language teachers is actually familiar with corpora as such as well as with their potential and even from these pedagogues only few apply corpora in their classrooms (cf. Mukherjee 2009: 161-162). Regrettably this is the case, even though numerous ways exist in which the use of corpora can impinge on language teaching in a positive way, as the following quotation taken from Kennedy (1998: 281-282) attempts to illustrate:

> First it can influence the content of language teaching by affecting selection of what to teach, the sequencing of pedagogy, and the weight given to items or parts of the language being taught, thus contributing directly to the content of instruction. Secondly, through the consciousness-raising of teachers about language and language use, it can show that likelihood of occurrence, or frequency of use, is an important measure of usefulness. Corpus studies can also contribute to language teaching methodology by influencing the approach to instruction and making available techniques and procedures which encourage self-access and individualized instruction through interaction with authentic, analysed text from a corpus database.

This thesis, however, will not deal with all three of the mentioned areas in which corpora can be applied in language pedagogy, as this would clearly go beyond

the scope of this paper. Leaving the question of what to teach aside, this paper will therefore mainly focus on the third way described by Kennedy and thus on the question of how corpora can be used as direct sources inside as well as outside the language classroom (Bernardini 2004: 15). In other terms, the major aim underlying this work is to introduce already experienced language teachers, teachers in training, student teachers as well as those interested in "the pedagogical aspects involved in using corpora", as Kaltenböck and Mehlmauer-Larcher (2005: 67) formulate it.

For this reason, the following sections are structured in a way so that also readers who are entirely unfamiliar with corpora and their application in the language classroom could be able to follow the provided line of reasoning. Therefore, the overall organisation of this thesis can be described as one which develops gradually from chapters providing theoretical background information to sections dealing with practically relevant subjects and thus the application of corpora in the field of language pedagogy in more detail.

Section 2 will be concerned with the introduction of the most important and thus basic facts which should be familiar to those interested in the ways in which corpora can be integrated into the field of language pedagogy. Therefore, it will provide a definition of the term *corpora* and information with regard to the development as well as the potential use of corpora in the field of linguistics in general. Afterwards, section 3 will concentrate on the implementation of corpora into the field of 'English Language Teaching' (ELT) from a historical and still quite theoretical perspective.

From section 4 onwards the focus will then shift towards more practical issues, as this chapter will deal with the question of how to apply corpora in the language classroom. This will be done by addressing two options of how corpora can be integrated into language teaching and learning, namely via teacher-corpus interaction (e.g. for reference, for materials design, for demonstration) (see section 4.1.) or via student-corpus interaction (e.g. for reference, for discovery learning) (see section 4.2.). Section 5 will finally be devoted to the potential and

limitations underlying the practical use of corpora in the field of language teaching and learning (see section 5.) and will consider both advantages and restrictions of corpora and retrieval software as such, as well as their application in the field of language pedagogy.

2. Corpora: definition and potential use

As a first step, this section attempts to define the term *corpora*, before giving a brief overview of the different types available and the overall historical development of corpora. Thereafter, the fourth subsection will focus on the potential use of corpora in the study of languages, while the last part of this section will be concerned with corpora as one means of identifying regularities in language use.

2.1. What is a corpus? – A definition

The question of what a corpus actually is, cannot be regarded as new, as it has already been discussed by numerous experts working in the field of linguistics generally or in the areas of corpus linguistics and applied linguistics more specifically. However, the answers provided by these individuals are not identical, but vary considerably. Thus, the aim of the subsequent paragraphs is to give an overview of the most prominent features characterising corpora.

"*Corpora* or *corpuses* (singular: *corpus*) are simply large collections or databases" containing texts or in other words, language, as Schmitt (2002: 68) notes. This first feature, which almost all definitions regarding the term *corpus* enclose, is even implicitly expressed by the term *corpus* itself, which is a word of Latin origin meaning *body*. "[H]ence a corpus may be defined as any body of text", as McEnery and Wilson (1996: 21) point out, even though the term *text* in this context does not necessarily refer to longer coherent passages of written language. Since texts compiled in order to make up a corpus can vary considerably with regard to language mode as well as length, they can consist of everything from only a few unconnected words to copies of whole books.

This already implies that not only written works of various genres can be included in corpora, as they may also contain spoken texts or a combination of both written and spoken language samples (e.g. Odlin 1994: 319). While written texts can be transformed into a machine-readable format quite easily “by scanning, typing [or] downloading [them] from the internet or by using files that already exist in electronic form” (O’Keeffe, McCarthy & Carter 2007: 2), the process of uploading spoken material to a database is a far more complex procedure. This is due to the fact that spoken input has to be transcribed by hand first (e.g. Aston 1997a: 205), even though these days programmes exist which make it possible “to add multimedia elements, such as video clips, to corpora of spoken language.” (O’Keeffe, McCarthy & Carter 2007: 2) Thus, naturally, the compilation of corpora containing spoken language samples is more time-consuming than that of written texts, even though the introduction of computers in corpus studies clearly simplified both processes.

The fact that language databases can nowadays be processed electronically is worth mentioning, as corpora were formerly only available in printed form (Mindt 1988: 11) and were collected as well as analysed manually (see section 2.5.). These days, however, corpora are normally stored on electronic devices, like, for instance, on hard disk or on CD-ROM and can also be accessed in this way, for example, via the internet (Hunston 2002: 2). They are therefore referred to as being modern and machine-readable databases which “can be accessed [...], automatically searched, copied [and] transferred to another computer” easily (cf. Leech & Fligelstone 1992: 115). The advantages of computer usage in corpus studies are enormous, as McEnery, Xiao and Tono (2006: 6) note, when mentioning “the speed of processing” which computers allow, as well as the simplicity by which language data can be dealt with in an accurate, consistent, reliable and low-cost way. Furthermore, corpora can be accessed and manipulated by various specifically designed computer programmes in order to reveal “patterns and regularities of language use” (Cook 2003: 73). However, the actual process of analysing a corpus is a feature which is not included in all definitions of the term *corpus* provided by linguists, as the description of the

concept is often restricted to the structure of corpora without mentioning any potential uses.

All remarks noted up until this point imply that any random set or collection of texts can be referred to as a corpus. However, this is not the case, as one specific feature of corpora is that the compilation of texts has to be a *sampled* or as the majority of authors calls it, a *principled* one. In other terms, in order to avoid confusingly large and unstructured amounts of language material (Mindt 1975: 9) as can be found in text archives, each corpus is "designed to represent a language, language variety, register or genre" (cf. Anderson & Corbett 2009: 4) and is thus compiled to suit a specific purpose. Depending on the aim underlying the composition of a corpus, the quantity, the quality as well as the diversity of text types included can vary enormously. An analysis of any corpus can only provide linguistically relevant results if the corpus under investigation allows for conclusions to be drawn not only with regard to the specific language data it contains, but also regarding the language, variety or genre it represents (cf. Mindt 1975: 9). McEnery and Wilson (1996: 24) even go as far as claiming "that a corpus constitutes a standard reference for the language variety which it represents", especially if the corpus under investigation includes all language material available regarding a specific field of application (Barnbrook 1996: 24).

Concerning the overall size of a corpus, it has to be noted that no clear boundaries regarding the minimum or maximum word count exist. As Leech and Fligelstone (1992: 117) mention quite appropriately, "[a] computer corpus can consist of anything from a few thousand words of text typed into a PC by its owner, to hundreds of millions of words of text acquired through 'data-capture' by a large-scale research institution." Moreover, corpora do not necessarily have to be of finite size (cf. Sinclair 1991: 9), as can be seen when having a closer look at the different kinds of corpora available for analytical investigation provided in section 2.2.

Another typical quality of corpora has already been mentioned in the preceding paragraphs. They can also be referred to as "extensive databanks of language

which has actually occurred in use" (Cook 2003: 126). This allows people interested in language studies access to authentic language samples and information concerning which patterns occur frequently or less often in naturally-occurring spoken or written language (cf. Bernardini 2000: 105). Consequently, invented or modified language samples mirroring how language should be used correctly or "how it is *commonly* and *typically* used" (Anderson & Corbett 2009: 2) are not comprised in corpora.

Concluding, it can thus be noted that a corpus is a principled, nowadays machine-readable collection of naturally-occurring written or spoken language. Variable with regard to size and finiteness, a corpus contains authentic texts representative of "a state or variety of a language" (Sinclair 1991: 171), which can easily and quickly be manipulated by search programs for qualitative or quantitative analysis. "[I]ntended to represent something larger, a corpus is motivated, created with a linguistic purpose in mind" (Anderson & Corbett 2009: 4) and can thus function as a standard reference for the language, language variety, register or genre it mirrors.

As can be seen when having a closer look at the definition just provided, a number of variables are included which may vary considerably with regard to different types of corpora. Therefore, the following section will present the most prominent types of corpora existing, in order to display the potential underlying their analysis.

2.2. Different types of corpora

Researchers may, depending on the purpose they pursue in analysing a corpus, prefer corpora of varying "design, size and nature" (Kennedy 1998: 3) for their projects. As corpora themselves cannot be described as being good or bad, the only statement to be made about them is whether they actually suit or do not suit the purpose one follows in analysing them (Hunston 2002: 26). Therefore, when aiming at a detailed analysis regarding a specific language issue, one should always go for the corpus which suits the needs of the analyst best, as O'Keeffe,

McCarthy and Carter (2007: 3) stress. The following section will give a brief overview of various types of corpora.

The first and relatively general distinction to be made with regard to corpora is that between written, spoken and mixed collections of data. The first modern machine-readable corpus was a written one, namely the “Brown University Standard Corpus of Present-Day American English”, as McEnery, Xiao and Tono (2006: 61) note. Then corpora focusing on spoken authentic language data came into being, like, for instance, “the London-Lund Corpus (LLC) [... or] the Cambridge and Nottingham Corpus of Discourse in English (CANCODE)” (McEnery, Xiao & Tono 2006: 62), before finally mixed corpora consisting of both written and spoken texts were created, for example, ‘The British National Corpus’ (BNC).

Another differentiation which can be made with regard to corpora types is that between general and specialised databases of language. While general corpora “have been assembled simply to make available a text base for unspecified linguistic research” (Kennedy 1998: 19), specialised corpora are mainly created to represent relatively specific areas of language use and thus allow for data to be analysed on a different and more detailed level (cf. Bernardini 2000: 119). Therefore, specialised corpora are often created by linguists themselves “with particular research projects in mind”, as Kennedy (1998: 20) claims, and differ from general corpora frequently with regard to their size and also the text types, registers and genres they contain. In contrast, general corpora are characteristically not designed for a specific purpose, but are “[p]re-packaged corpora [which] typically offer advantages” (Aston 2002: 12) compared to self-made specialised ones. Especially with regard to reliability, documentation, corresponding analysis software, convenience and representativeness (cf. Kennedy 1998: 20) general corpora, like the ‘Brown Corpus’ or the ‘BNC’ already referred to above, are likely to be more trustworthy and easier to handle than homemade corpora (Bernardini 2000: 113). The aforementioned qualities of general corpora, however, do not necessarily have to imply that small corpora are less reliable with regard to their representativeness. As Barnbrook (1996: 25) points out, “[t]he most common features of the language will be well represented

even in relatively small quantities of text". Hence, also small, but specialised corpora are able to provide enough language data in order to allow for meaningful analysis results to be gained, although it is "actually impossible, to know in advance what size of corpus will meet the requirements of any given research project." (Barnbrook 1996: 25)

Apart from the distinction between general and specialised corpora, one can also distinguish between corpora on a temporal basis, as is the case with diachronic and synchronic corpora. These two types of corpora allow the researcher to focus on changes in language use over time. Diachronic or historical corpora (McEnery, Xiao & Tono 2006: 65) contain texts of a specific language or text type which have been produced over a longer period of time, "as would be the case for a corpus of personal correspondence between 1700 and 1900 [...]." (Anderson & Corbett 2009: 7) Therefore, diachronic corpora enable the researcher "to track changes in language evolution" by covering at least three to four decades (McEnery, Xiao & Tono 2006: 65). Synchronic corpora, on the contrary, only "contain texts from a particular time period (such as English from the 1990s)" (Anderson & Corbett 2009: 7) and can thus help linguists to gain insight into language usage at a certain point in time (Kennedy 1998: 22). Examples for both diachronic as well as synchronic corpora are, for instance, the Brown and Frown Corpus of American English as well as the LOB and FLOB Corpus for British English. When treating each related pair as one corpus only, each pair can be regarded as a diachronic corpus, as both comprise approximately thirty to forty years of language evolution. Taken individually, however, each of the four aforementioned electronic language databases covers a particular point in time and can thus be referred to as a synchronic corpus (cf. McEnery, Xiao & Tono 2006: 64-65).

Another type of corpus is the so-called monitor or dynamic corpus, like the so-called Bank of English (BoE), which is a specific type of the diachronic corpus (Anderson & Corbett 2009: 7-8). What makes this type of corpus stand out, as opposed to sample corpora of finite size, is that it grows as language data is added to the corpus regularly (e.g. daily, weekly, monthly or annually), even "though the proportion of text types included in the corpus remains constant"

(McEnery, Xiao & Tono 2006: 67). This means, as Sinclair already noted in his work from 1991(: 9) that “the whole idea of a corpus of finite size” can no longer be sustained, as no principled collection of naturally-occurring language data can ever be regarded as completely balanced, structured and thus finite. Kennedy (1998: 22) even goes as far as claiming that monitor corpora “are open-ended language ‘banks’ which are limited only by the financial resources and technology needed to maintain them.”

While most corpora actually include language data provided by native-speakers in naturally-occurring communicative situations, learner corpora contain a collection of language data produced by second language learners in their process of acquiring or learning a foreign language. This type of corpus can help language researchers to gain closer insights into problems in second language acquisition as well as learning and “can be used for either cross-sectional or longitudinal analysis.“ (McEnery, Xiao & Tono 2006: 65)

2.3. A brief history of corpora

John Sinclair (1991: 1), an influential personality regarding the implementation of corpora in the study of languages, describes the introduction of corpora and their development in the field of linguistics over the last thirty years as follows:

> Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular.

The earliest corpora available emerged “in the first third of the 1900s” (Schmitt 2002: 68) and contained specific text passages or whole books being copied in laborious handwork onto index cards or dictionary slips and from the early 1960s onwards on punch cards which were then frequently stored in shoeboxes (Leech & Fligelstone 1992: 116). Since “[e]arly corpus linguistics required data processing abilities that were simply not readily available at the time” (McEnery & Wilson 1996: 10-11), the compilation and analysis of corpora was an expensive, time consuming and quite inaccurate way of approaching real language data, as

researchers did not only have to collect printed texts and read them through closely, but also had to analyse them manually (cf. Cook 2003: 73). Despite the fact that corpora were generally of quite small size, as it was “virtually impossible to collate and analyse large bodies of language data” (McEnery, Xiao & Tono 2006: 4), they already functioned as basis “for research into the areas of lexicography, dialectology, anthropology and grammar” (Anderson & Corbett 2009: 9) as well as language pedagogy (Kennedy 1992: 336).

The preceding paragraph clearly illustrates that “important corpora of English were assembled long before the computer was invented”, as Nelson Francis (1992: 17) notes, who finds himself referred to as a pioneer regarding the compilation of electronic corpora quite frequently. Together with Henry Kučera, Nelson Francis created the first computer-based corpus in the 1960s, the so-called ‘Brown University Standard Corpus of Present-Day American English’. However, their work was only welcomed by few linguists, while the majority of members of “generative grammar dominated linguistics” (Meyer 2002: 1) observed this new development critically. Noam Chomsky’s thinking as well as his theories clearly influenced the way in which corpora were perceived by his colleagues, as he advocated the rationalist theory which is “based on artificial behavioural data, and conscious introspective judgements” and thus invalidated empirical corpus data (McEnery & Wilson 1996: 4). As a result, Chomsky regarded corpora as being skewed (Gavioli 2005: 17) and considered them as inappropriate instruments for linguistic studies, as observed data focus on language performance, instead of competence and can thus never represent language in its infinity (McEnery & Wilson 1996: 5-10).

Nevertheless, even though Chomsky’s criticism clearly affected the linguistic community strongly and many researchers excluded corpora entirely from their linguistic methods of investigation, minority groups still used them in the 1960s and 1970s (McEnery & Wilson 1996: 11). Francis and Kučera, for example, started their work on the ‘Brown Corpus’ during these years, a project finished roughly twenty years later. In the early 1980s corpora then experienced a sudden

revival “after some intrepid explorer-style linguists rediscovered” (McEnery & Wilson 1996: 17) them.

At approximately the same time shoebox corpora were replaced by electronic databases of varying size as well as type. Years later, linguists were even able to approach ”virtually unlimited collection[s] of data on the Internet” (cf. Johansson 2007: 19), with some corpora being publicly available, while others had to be paid for (O’Keeffe, McCarthy & Carter 2007: 5). This change in format, however, did not change the method, but only the means of corpus analysis (Anderson & Corbett 2009: 9) and thereby made the analysis of real language data gradually cheaper, more accurate and reliable as well as less time-consuming (Kennedy 1992: 336), as texts could from then onwards be captured by the computer itself or uploaded quite simply (cf. Leech & Fligelstone 1992: 116). Further enhancements were that computers became more accessible for home as well as public use (McEnery & Wilson 1996: 17) and that their memory capacity as well as existing retrieval programmes improved gradually. As a result, “an increase in the construction of corpora, the publication of corpus-based studies, and a wide-spread recognition of the validity of the corpus as a tool in the analysis of language” (McEnery & Wilson 1996: 169) could be recorded.

While the ‘Brown Corpus of American English’ and the ‘LOB Corpus for British English’ were the first electronic corpora created (cf. Gavioli 2005: 17) in the 1960s, “it took almost three decades for the use of corpora to spread beyond the inner circle of corpus linguists.” (Granath 2009: 47) The real breakthrough of corpora in the study of language took place in the 1980s when the ‘Collins Birmingham University International Language Database’ (COBUILD) organised by John Sinclair (Mukherjee 2006: 7) was released. In addition to a dictionary called the ‘Collins COBUILD English Language Dictionary’ first published in 1987, numerous teaching materials were designed based on Sinclair’s corpus, which allowed for “’more realistic’ descriptions of English for teaching purposes” (Gavioli 2005: 17-18). As can be seen from this example, especially dictionary makers and language learners benefited from the introduction of corpora in the field of

linguistics, as the usage of third-person data “seemed to guarantee relevance and authenticity” (Anderson & Corbett 2009: 2) of any language material based on it.

Overall, it can be said that since the time when computers were first introduced as instruments into the study of languages, they clearly underwent a rapid development regarding the amount of data they could process. While most *first-generation corpora* only consisted of approximately one million words and were “set out to represent a particular variety of a language” (e.g. Brown Corpus, LOB Corpus), *second-generation mega-corpora* available from the 1990s onwards already contained about hundred million words, like, for example the British National Corpus (Kennedy 1998: 45). Nowadays, *third-generation mega-corpora* are created comprising up to several hundred million of words (Schmitt 2002: 69)

Coming back to the opening statement of this section provided by John Sinclair, one can clearly agree with him that today “the corpus methodology enjoys widespread popularity” (McEnery, Xiao & Tono 2006: 4) in various areas of linguistic research. At present corpora are generally used in the fields of lexicography, grammar, stylistics, translation, forensic linguistics and sociolinguistics (cf. O’Keeffe, McCarthy & Carter 2007: 17-21), just as they have gained enormous influence in the field of second or foreign language teaching (cf. Braun, Kohn & Mukherjee 2006: 1). But before going into more depth regarding the role of corpora in the language teaching (see section 3.), the subsequent section will address the potential of corpora in linguistics in general.

2.4. The potential of corpora for the study of languages

As all of the aforementioned types of corpora are “nothing but [...] store[s] of used language,” as Hölzl (2003: 14) remarks, they need to be analysed either manually or with the aid of special computer software in order to provide useful results for the study of languages. Therefore, the subsequent sections will investigate in which ways corpora can actually be used and take a closer look at corpora programmes as instruments for different types of analysis.

2.4.1. Retrieval software

As could already be seen so far, the process of corpus analysis is strongly linked to the usage of electronic devises, as retrieval software allows for user-friendly involvement with "'raw' corpora", as Aijmer (2009: 1) points out. The computerised analysis of corpora has become a tool more and more appreciated by linguists, as the rapid "development of computer technology over recent years has made observation possible" (Widdowson 1996a: 73) which is not available to researchers when examining data by means of introspection or elicitation (see section 2.4.). Nevertheless, it also needs to be pointed out that the analysis of corpora "cannot yet be fully computerized" (Kennedy 1992: 367), as certain linguistic analyses of language data still need to be prepared or at least complemented tediously by hand, as Kennedy (1992: 367) remarks and even after almost twenty years this is at least to some extent still true nowadays.

According to Mukherjee (2009: 65), three different kinds of software allow for corpora to be accessed and manipulated electronically, namely corpus-specific programmes, general corpus-linguistic programmes or programming languages and universally usable programming packages. While corpus-specific programmes are particularly designed for designated corpora and firmly connected to these (e.g. the 'International Corpus of English' with the 'ICE Corpus Utility Program'), general corpus-linguistic programmes can be used in order to analyse different kinds of corpora (e.g. 'antconc'*)*. Apart from these two types of software specifically designed to analyse corpora, programming languages and freely accessible programming packages can also be made use of (cf. Mukherjee 2009: 65).

All of these three types of computer programmes allow for corpora to be analysed quantitatively as well as qualitatively via the usage of different analytical tools (see section 2.4.3.). How detailed the results of any analysis are, however, does not depend on the retrieval software as such, but on the corpus and its characteristics and exactly these will be discussed in more detail in the following subsection.

2.4.2. Annotation

Corpora have been described as consisting of plain text in this paper up until now and can thus be described as being *unannotated*. A common characteristic of many machine-readable corpora nowadays, however, is that they are *annotated*, which means that linguistic information of various kinds is added to them, as McEnery and Wilson (1996: 24) remark. These two authors also emphasise that

> [t]he important point to grasp about an annotated corpus is that it is no longer simply a body of text in which the linguistic information is implicitly presented. [...] By contrast, a corpus, when annotated, may be considered to be a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete annotation.

Corpora, if annotated, even offer researchers the opportunity to classify gained results of a corpus analysis according to specific variables, like age, gender, language level and nationality of the speaker as well as regarding the register, text type, genre and language variety used. However, not all types of corpora include this additional information and therefore, "[i]t is important to scrutinise how a corpus is designed when considering buying or accessing one, or when evaluating any finding based on it." (O'Keeffe, McCarthy & Carter 2007: 1)

The term *annotation*, as Hunston (2002: 19) notes, is an umbrella term which comprises both the processes of *tagging* as well as *parsing*. While tagged corpora provide the analyst with information regarding "the linguistic properties of the individual words", parsed corpora make "the functions of the words in relation to each other" (Barnbrook 1996: 109) available and thus present the researcher with information concerning the syntactic functions of single lexical items. Even though these two forms will be elaborated on in greater detail subsequently, it needs to be mentioned that in addition to these more traditional forms of annotation, corpora can also be *error-tagged*, which means that mistakes detected in the presented data are indicated (Meunier & Gouverneur 2009: 189).

While tagged versions of corpora allow for deeper linguistic analysis of language data than unannotated ones, it is problematic that each single word can only be

identified as belonging to one specific word class, as Dieter Mindt (1988: 12-13) remarks critically. As each part-of-speech in a corpus is labelled with the abbreviation of the word-class it belongs to in capital letters, word and tag can be easily identified as one unit by the analytical software (Mindt 1988: 12). However, a difficulty frequently encountered when corpora are tagged automatically via special software is that, although the process does not take long, it cannot be described as being totally accurate (Hunston 2002: 18). This is due to the fact that words appearing to be identical on a surface level can sometimes not be identified as belonging to different word classes by the computer. This may be the case, for instance, with the word *help,* which depending on the textual environment in which it occurs can be regarded both as a verb as well as a noun (Mindt 1988: 11).

Similar to the process of tagging described above, corpora can also be parsed automatically, even though this type of analysis may also be slightly inaccurate if solely conducted by a computer. However, "[a]ccuracy can be improved by 'training' the automatic parser, that is, by setting up the parser to learn from past examples," as Hunston (2002: 19) notes. Moreover, two different types of parsing can be distinguished. While full parsing provides the researcher with an analysis of the text constituents which goes into as much depth as possible, skeleton parsing is "a less detailed approach which tends to use a less finely distinguished set of syntactic constituent types" (McEnery & Wilson 1996: 44).

2.4.3. Quantitative and qualitative analysis

Even though quantitative as well as qualitative analyses of language data can be conducted by hand, the application of computer software allows for corpora to be analysed in a less time-consuming and more user-friendly way. While the quantitative approach, as Bernardini (2000: 121) points out, predominantly builds on computer software in order to analyse corpora, the qualitative approach relies on the ability of the researcher "to make inferences on the basis of the evidence provided." These two types of analysis also follow different aims. The quantitative analysis of naturally-occurring language data deals with the counting and identification of frequencies and distributions of features to be found within a corpus and allows for "reliable and generalizable statements about how language

works” (Anderson & Corbett 2009: 22) to be made. Qualitative investigations exploit language data in order to explore in which varying ways a specific lexical item is used within a corpus (O’Keeffe, McCarthy & Carter 2007: 2) and therefore in real-life situations (McEnery & Wilson 1996: 62). Consequently, based on the results obtained from both quantitative as well as qualitative analysis, linguists are able to gain information regarding the “elements and structural patterns which make up the systems we use in a language” (Kennedy 1998: 4). Subsequently, a number of tools will be introduced which can be very helpful if aiming at analysing language data both quantitatively and qualitatively.

2.4.3.1. Concordancing

Concordancing stands for “using corpus software to find every occurrence of a particular word or phrase” (O’Keeffe, McCarthy & Carter 2007: 8) contained in a corpus and can be used both as a quantitative as well as a qualitative analysis tool. With the search word or phrase presented in the middle of the screen and approximately “seven or eight words presented at either side” (O’Keeffe, McCarthy & Carter 2007: 8), this format of presentation is called *Key-Word-in-Context* or short also *KWIC format* (Bernardini 2000: 121), as presented in Figure 1 below.

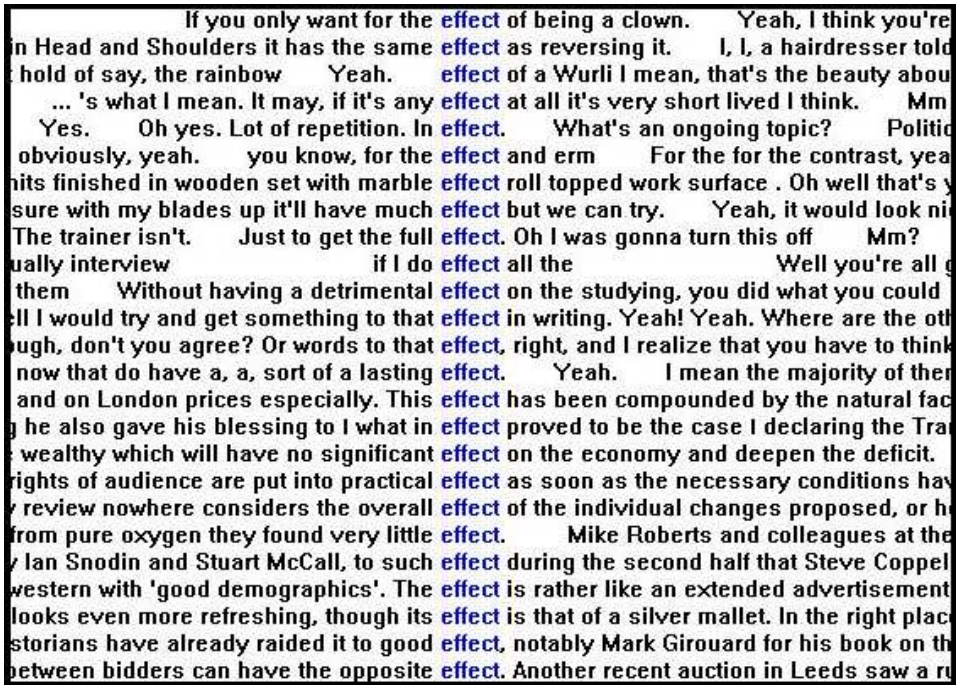

Figure 1: Concordances presented in *Key-Word-in-Context (KWIC)* format (Wynne 2007)

On the vertical axis (see Figure 1), a concordancer provides “information on counts of frequency, i.e. the recurrence of words”, based on which word frequency lists can be generated. On the horizontal axis “the co-text of a particular search word, i.e. the co-occurrence of words” (Kaltenböck & Mehlmauer-Larcher 2005: 70) is presented which allows the researcher to draw conclusions of a qualitative nature.

2.4.3.2. Word frequency lists

One of the most frequently mentioned ways of analysing corpora quantitatively with the aid of computers are word frequency lists which can be calculated for any range of texts (O’Keeffe, McCarthy & Carter 2007: 11). Being able to scan through corpora rapidly, analytic programmes can easily determine the words which appear most or also least frequently in any collection of texts, which are then arranged in a list according to their frequency of occurrence. Wordlists, as Mukherjee (2009: 66) explains, provide a good foundation for the lexical analysis of any kind of corpus and can, for example, illustrate that certain words are used more often in specific domains, like genres, language modalities or registers than in others (cf. Anderson & Corbett 2009: 28). However, the interpretation of word frequency lists needs to be conducted cautiously, as

> by breaking the text into individual word forms it removes the words from their original contexts. One effect of this is that word forms which can have more than one meaning are gathered together and counted as one single word. (Barnbrook 1996: 53)

Therefore, when interpreting the numbers given in word frequency lists, as Mukherjee (2009: 83) argues, one should never draw hasty conclusions, as these may not always be valid from a statistical point of view. To make sure that analysis results are well-founded “when comparing two data sets of unequal size”, it is necessary that normalised frequencies are calculated for both corpora, as these enable the researcher to determine how often specific words occur “per thousand, or sometimes per million words” (Anderson & Corbett 2009: 30). The calculation of normalised frequencies is diagnostically most conclusive, if both corpora under investigation are “maximally representative finite sample[s]” of the

language, language variety, genre or register they mirror (McEnery & Wilson 1996: 61).

2.4.3.3. Key word analysis

The so-called *key word analysis* is another tool available in order to analyse computerised corpora. Key words can be defined as words "whose frequency is unusually high in comparison with some norm" (O'Keeffe, McCarthy & Carter 2007: 12). When investigating a particular, often specialised corpus regarding its key words, a wordlist of the corpus under investigation and of the general corpus representing the aforementioned norm, need to be created so as to be able to check them against each other (O'Keeffe, McCarthy & Carter 2007: 12). It is important to note here that not only positive key words which appear more often than in the reference corpus can be identified in this way (Mukherjee 2009: 71). Negative key words which are lexical items occurring less often in the corpus under investigation than in the reference corpus, can also be of interest to the analyst.

2.4.3.4. Cluster analysis

Another function frequently provided by corpora programmes is the so-called *cluster analysis*. This kind of investigation allows the researcher to find out more about word combinations and chunks of words occurring in a specific corpus. Similar to the creation of word lists containing single words, a cluster analysis does not only arrange individual words based on their overall frequency, but allows for word chunks being made up of selectable numbers of words to be ranked, as O'Keeffe, McCarthy and Carter (2007: 13) outline.

2.5. Corpora as a means of identifying regularities of use

All over the world each language or language variety follows its own rules of application and has its own regularities of use. In order to describe any language in more detail, those concerned with the study of languages need to analyse it. But how do researchers actually gain the information required to write grammar books, dictionaries or practical usage guides of specific languages? Corpora, as the heading of this chapter already reveals, are one means helping to approach

language data. Introspection and elicitation are two other methods (Hölzl 2003: 5). What is special about corpora in this context, as opposed to introspection and elicitation, is that corpora do not only provide the linguist with information about language usage in general, but about how a particular language is used in naturally occurring everyday-life situations. In order to illustrate this point better, the following subsections will deal with each of the three methods mentioned above in more detail, beginning with introspection and elicitation, before elaborating on the use of corpora as an instrument to identify regularities in language use.

2.5.1. Introspection

One of the first and most traditional ways of approaching language data from a linguistic point of view is that of reflecting on actual language usage based on the limited experiences and recollections of individual speakers themselves (Sinclair 1991: 1). As "[i]t became fashionable to look inwards to the mind rather than outwards to society" (Sinclair 1991: 1), intuition and introversion became key terms in the study of languages and *introspection* a popular method of approaching language data. According to Sinclair (1991: 39), two kinds of introspection can be distinguished, as both native-speakers of a specific language as well as linguists, in this context often referred to as 'armchair linguists' (Kaltenböck & Mehlmauer-Larcher 2005: 68), are able to provide language material usable for analysis. To sum up,

> '[i]ntrospective' or 'first person' data consists of sentences obtained from informants, frequently linguists themselves, who reflect on their own language knowledge and use. What introspective data therefore represents is the informant's conscious knowledge or intuitions about their language behaviour. (Hölzl 2003: 5)

Many linguists clearly prefer to approach language data via the means of introspection, as they claim "to get closer to the abstract organization of the language by this method rather than by studying the output of spoken and written language as used in communication" (Sinclair 1997: 28). One major point of criticism with regard to introspection often mentioned in specialist literature, however, is that this method does not really provide the researcher with insights

regarding how language is actually used. This is due to the fact that “[i]n addition to only representing the conscious and those areas of the unconscious knowledge which are accessible, data gained from introspection is also entirely determined by the informants’ individual language uses and preferences” (Hölzl 2003: 6). As a result, introspective data can only provide the researcher with subjective ideas about how language works in one human mind, rather than presenting objective and universal facts applicable to all language users (cf. Sinclair 1991: 39). Hence, introspective data cannot “be considered as representative of the entire language spectrum of which it only presents very limited or biased extracts” (Hölzl 2003: 6). This is especially true if those providing introspective data are both native-speakers as well as linguists, as these people hold “a particular perspective, and this, while giving them insights denied to others, at the same time limits their view”, as Widdowson (2003: 77) claims. Consequently, researchers need to be disqualified as objective language informants in the field of linguistics (Widdowson 2003: 82), which also led to the fact that the exclusive usage of introspection has regularly been criticised in the past, even though it is still a highly esteemed instrument “in evaluating evidence rather than creating it.” (Sinclair 1991: 39)

2.5.2. Elicitation

As mentioned by Sinclair (1997: 27), “[a]nyone who does not have full competence in a language of study has to get the crucial evidence from outside“, as is the case with

> ’[e]licited’ or ‘second person data’ [which] is obtained from informants other than the researchers themselves. Essentially, it is introspection data, but the person reflecting on their language behaviour or knowledge is not the one who has set up the questions in the first place. (Hölzl 2003: 6)

Therefore, the reliability of data gained by means of elicitation can be criticised for the same reasons as that obtained by introspection, even though it can be assumed that language speakers asked about their language experiences “are more reliable informants in that they are innocent of any analytic intent.” (Widdowson 2003: 83) Nevertheless, results originating from both introspection and elicitation need to be regarded critically (Hölzl 2003: 6), as language data

gained in these ways can only be described as abstract rather than concrete, as it can merely “reveal what people know about what they do but not what they actually do.” (Widdowson 1996a: 72)

According to Henry Widdowson (2003: 83), three different kinds of elicitation can be distinguished, namely *co-textual, contextual and conceptual elicitation.* While co-textual elicitation aims at “finding out how far informants could, suitably promoted, provide information about the occurrence and co-occurrence of words in texts”, contextual elicitation focuses on describing specific situations to native-speakers and then asking them about the kind of language they find suitable to use in this context. The third kind called conceptual elicitation concentrates on revealing “how linguistic encodings are mentally organized” (Widdowson 2003: 83). A survey which exemplifies this type of elicitation quite well has been conducted by Rosch years ago. She asked numerous human guinea pigs for “the word which sprang most immediately to mind as an example of a particular category”, which was in her case a superordinate term, like, for instance ‘bird’ or ‘fruit’. The subjects would then normally answer with hyponymous words, like ‘sparrow’ or ‘apple’, which implies that these are more strongly linked to the category than other terms and have thus “some marked mental prominence for them.” (cf. Widdowson 2003: 83).

As can be seen from the descriptions given above, elicitation enables the researcher to gain information regarding language only in a quite limited way, as just “highly specific and narrowly constrained questions” (Kaltenböck & Mehlmauer-Larcher 2005: 68) can lead to those detailed answers aimed at. The subsequent section will now have a closer look at corpora as a means of identifying regularities of use.

2.5.3. Observation and corpora

The third way of approaching language data is by observation, which means that native-speakers of a language function as informants regarding how they actually use language in natural communication (Hölzl 2003: 7). This is also what makes this third type of identifying regularities in language use so special, as neither

introspective nor elicited data can provide the researcher with information regarding the naturally-occurring usage of language in real-life situations. Third-person language data can thus only be provided by speakers in situations in which they are not aware of being observed, because otherwise the produced output may be modified either consciously or unconsciously by informants whose beliefs with regard to how language should be used correctly do not necessarily have to correspond with their actual language usage (Widdowson 2003: 81). Consequently, one of the advantages of using attested data in language research is that its analysis enables researchers to generalise and draw conclusions regarding naturally-occurring language usage and not how people think it should be used, as the following quotation taken from O'Keeffe, McCarthy and Carter (2007: 30) illustrates:

> When we do look at what speakers and listeners do, we may not hear native speakers as we might want to hear them or as how we might have learned to expect to hear them. But we do hear real people interacting with one another, working at full stretch with the language, adjusting millisecond by millisecond to the interactive context they are in, playing with the language, being creative, being affective, being interpersonal and, above all, expressing themselves as they engage with the processes of communication which are most central to our lives.

Thinking back to the definition of corpora provided in section 2.1., the attentive reader may have noticed already that observed language data is exactly the kind of language material collected in order to make up corpora. Barnbrook (1996: 140) even goes as far as comparing a corpus to "a tireless native-speaker informant, with rather greater potential knowledge of the language than the average native speaker". However, although corpora are solely made up of observed data, it cannot be claimed that they are entirely free of the influence of human intuition, as the process of choosing texts to be included in corpora as well as the interpretation of analysis results are "always and inevitably subject to human reasoning." (Hölzl 2003: 8)

Another point of interest to be mentioned is that as corpora only contain evidence of an external nature (McCarthy 2001: 125), they exclusively include language

data which has actually been performed, but do not record what speakers could have said instead, as Kaltenböck and Mehlmauer-Larcher (2005: 68) remark. Therefore, as corpora consist of performance data only, McEnery and Wilson (1996: 5) argue that they "must of necessity be a poor guide to modelling linguistic competence." Widdowson (1996a: 73), however, contradicts their statement by stating that "[i]t is surely better to find out what people actually do than depend on intuitions which are often uncertain and contradictory" (Widdowson 1996a: 73) and sometimes not even able to reveal the kind of data asked for (Widdowson 2000: 6). Yet, also corpora are not able to provide linguists with answers to all questions, since they only include actually performed and thus "attested" data, which does not necessarily have to imply that language usage of this kind is actually "possible" or even "feasible" in the sense of Hymes (cf. Kaltenböck & Mehlmauer-Larcher 2005: 68).

Although numerous researchers argue in favour of approaching real language data in the study of languages by studying third-person data, others claim that the intuition of the linguist can never be fully supplanted by it (cf. Partington 1998: 1). This attitude was particularly wide-spread among linguists shortly after Chomsky had criticised the acknowledgement of observed data and thus corpora as an approved way of approaching real language data in the study of language (Granath 2009: 63), as it was debated that both first-person as well as second-person data are able to reveal details regarding language usage which "cannot be evidenced by a corpus" (Cook 1998: 59). This argument certainly becomes clearer when being informed that "a three hundred million word corpus is equivalent to only around three thousand books, or perhaps the language experience of a teenager" (Cook 1998: 59). However, even though introspective and elicited data may provide the researcher with findings which cannot be supplied by observed data, this is also the case the other way round. Corpora can certainly help researchers to identify regularities in language use going beyond the intuition of a native speaker "in discovering facts about the language which cannot be analysed intuitively by native speakers (Granath 2009: 63), because their "cumulative language experience [...] remains far larger and richer" (Cook 1998: 59). Moreover, corpora definitely have the advantage that they can be

accessed, manipulated and analysed permanently, while introspective data allow for “random and incomplete access” (Cook 1998: 59) only.

As the preceding paragraph illustrates, the linguistic discussion of how to approach language data best is relatively controversial. In a world in which the intuition of the linguist has long been regarded as the only trustworthy way of identifying regularities in language use, corpora and thus also the results to be gained from their analyses have gained importance, as they cannot only be seen as a way of both approaching real language data (Anderson & Corbett 2009: 2), but also as an instrument for detecting myths “arisen from reliance on intuition-based ‘armchair’ linguistics” (Johns 1994: 296). Therefore, real language data should always be approached by observation first, whereas “both the evidence of secondary sources and the evidence of introspection should be brought in at a late stage” (Sinclair 1991: 40). The simple reason underlying this chronological order is that it allows linguists to reflect critically upon analysis results and enables them to perceive the most detailed picture regarding actual language usage.

While this section aimed at illustrating what a corpus actually is and how it can be used in the field of linguistics in general, the subsequent part attempts to give a better overview regarding the role of corpora in the field of English language teaching (ELT). For this reason the next section will be concerned with the most fundamental issues regarding the introduction of corpora into language pedagogy and their development in the field of foreign language teaching and learning.

3. Corpora in (foreign) language teaching and learning

Opening her paper with the line that “[c]orpora seem to have entered the classroom from the backdoor”, Silvia Bernardini (2004: 15) summarises what can regularly be read in specialist literature on the implementation of corpora in language pedagogy. Although the study of corpora has already been regarded as enormously important in linguistic areas like “language research, grammar construction, dictionary making, natural language processing [and] cognitive studies” (Hidalgo, Quereda & Santana 2007: ix) for a long time, it has not found its

way into language teaching and learning until relatively recently. Exactly this development and its implications for language pedagogy in general and English Language Teaching (ELT) more specifically will be addressed in this section.

Corpus-based research already left some marks in the area of language pedagogy between the 1920s and 1950s, when the occurrence frequency of particular lexical items in practical language usage was used to inform language teaching (Kennedy 1998: 282). However, a more deliberate introduction of corpora into the language classroom did not take place until the 1980s with the publication of the first corpus-informed materials for language learners in the context of Sinclair's COBUILD project (O'Keeffe, McCarthy & Carter 2007: xi). But why was there such a long time interval between the dates when corpora were first accepted as research tool in linguistics until they were finally recognised as pedagogical devices? Kennedy's (1992: 364) explanation for this slow development is that "teachers tended to show more interest in the learner and the learning process" for quite some time, before their focus shifted towards the question of what to teach in order to facilitate natural language acquisition and foreign language learning best. From then on, analytic studies of corpora became of increasing importance, informing and influencing teaching materials in language pedagogy (Braun, Kohn & Mukherjee 2006: 1). Afterwards the first publications "devoted to the use of corpora in language teaching" (Chambers 2007: 3) began to appear.

Articles and books concerned with corpora released up until the 1980s dealt almost exclusively with language theory. Sinclair's COBUILD project, however, clearly revolutionised the scene insofar as the influence of corpora on language pedagogy soon became a core theme in the written works of specialists (Gavioli 2005: 17-18). While the first publications dealing with corpora in language teaching and learning were still quite theory-laden and mainly comprised articles in ELT journals and books, later publications directed their attention to practical issues, including teaching materials, practical usage guides, works of reference as well as corpus-informed course books (cf. Johns 1994: 296). With the earliest existing corpus-based language materials primarily focusing on the question of

what to teach in language classrooms, the emphasis slowly but surely shifted towards the issue of potential methods to be used in order to convey selected contents to learners efficiently (Gavioli 2005: 1). This whole development regarding corpus-informed reference materials led to the trend that nowadays there are hardly any dictionaries to be found which are not based on corpora (O'Keeffe, McCarthy & Carter 2007: xi). Yet, the influence of electronic language databases on textbooks for both native speakers as well as foreign language learners is still not as strong as might be expected or even wished for (Römer 2009: 90).

With the beginning of the so-called *TaLC* (Teaching and Language Corpora) *conferences* taking place every two years since 1994, more and more specialists started to recognise the potential underlying the usage of corpora in language pedagogy (Granath 2009: 47). One reason which certainly affected this development positively was and still is the technological progress of computers, as they did not only improve in terms of financial availability over time, but also with regard to accessibility and user-friendliness (Sinclair 2004: 2). However, these developments were not the only ones benefiting the status of corpora in language pedagogy. Hard work has been required in order "to integrate existing corpora, corpus methods and tools into teaching practice", as Braun, Kohn and Mukherjee (2006: 1) stress. These efforts have particularly been made by corpus specialists trying to familiarise other linguists as well as language teachers with their subject of expertise by supplying them with publications addressing and stressing the advantages of the implementation of computer corpora in language pedagogy (Meunier & Gouverneur 2009: 179-180). However, from the compilation of the first computerised corpus in the 1960s "it took almost three decades for the use of corpora to spread beyond the inner circle of corpus linguists" (Granath 2009: 47). Nowadays not only language researchers, but also teachers and even students work with corpora (cf. Barlow 1996: 2). In the teaching contexts practitioners concerned with the subject of 'English for Specific Purposes' (ESP) "were among the first to appreciate the pedagogical potential of corpus work" in the language classroom, as Bernardini (2004: 21) remarks.

After having a closer look at how corpora gained importance in the field of language teaching and learning from a practical point of view, it is now time to find out more about the scientific links existing between corpus linguistics and language pedagogy. In this context it is necessary to remember that "[t]hose who teach languages depend on those who describe them for [...] basic information" (Sinclair 1997: 29). This fundamental information may, for example, concern the nature of language acquisition and learning as well as the questions what to teach as well as how to do so best. However, in reality the necessity of this cooperation between theorists and practitioners is often challenged, as it is generally expected that teachers themselves, based on their linguistic knowledge and pedagogical education, are able to decide which contents to teach and which methods to apply. Also, the connection existing between theory and practice and thus between linguistics, "communication studies, [...] psychology and sociology" (Yule 2004: 197), on the one hand, and language teaching on the other hand, is not as straightforward and uncomplicated as it might be expected. Therefore, mediation is needed in order to "interpret the results of theoretical and descriptive studies in such a way as to reveal their relevance to the language teacher", as Henry Widdowson (1980: 215) stresses. This mediating field of research is generally referred to as *applied linguistics* and is, among numerous other things, also concerned with the implementation of computer corpora in language pedagogy.

However, even though specialists working in this discipline try to mediate between theoretical corpus linguistics and practical language teaching, "there still seems to be a gap between what applied corpus linguistics has to offer and what teachers actually do (or don't do) with corpora in their teaching practice", as Joybrato Mukherjee (2006: 20) notices. Then the author suggests that

> [t]his gap can only be bridged if, firstly, teachers are involved to a much larger extent in corpus-based classroom action research (for which linguistic assistance and professional help is no doubt needed, e.g. in terms of in-service teacher training programmes) and if, secondly, all corpus-based activities are evaluated under real-time conditions in actual classroom context and both from teachers' and learners' perspectives.

This proposal already points at the most problematic issue when it comes to the application of corpora in the language classroom, namely that “many important developments in the field of corpus linguistics are not always communicated or usefully mediated in terms of their implications for language teaching” (O’Keeffe, McCarthy & Carter 2007: xi). One reason for this lack in information transfer may be that corpus linguists and thus researchers tend to discuss issues among each other rather than share their gained knowledge with practitioners like teachers. But even within the area of applied linguistics a distinction needs to be made between a more theoretical and a more practical approach, to be more precise

> between studies which look at teaching applications of corpora from the linguist’s point of view (giving hints and suggestions about what corpora can do for language teaching) and studies which look at uses of corpora from the teacher’s point of view (starting from a teaching problem and looking at how this problem can be tackled with corpus tools). (Gavioli 2005: 22)

In this context it is important to note that both of these described courses of action are not one-way processes. As language researchers, teachers as well as students may gain insights by analysing corpora, all these discoveries can flow from research into teaching and vice versa (Hidalgo, Quereda & Santana 2007: xiv).

Even though the study of corpora clearly influences linguistic descriptions as well as language pedagogy in a positive way and has been identified as legitimate and useful pedagogical aid in the field of language teaching and learning (cf. Meunier & Gouverneur 2009: 179), corpora are still rarely used in the foreign language classroom and sometimes not even familiar to teachers (Aijmer 2009: 1). But why, one may ask. An answer can be found in Hunston (2002: 192), who argues that although corpus linguists are highly enthusiastic about the usage of this analysis instrument in the field of language pedagogy, they do not seem to be able to convey the potential of this new tool to teachers. Consequently, pedagogues are unable to recognise the value of corpora for the language classroom and unwilling to use them. It is therefore especially important for corpus specialists to involve pedagogues actively “in working with – and thus disseminating knowledge about –

corpora" (Mukherjee 2006: 7), so as to allow them to discover the potential underlying the implementation of corpora in the teaching context themselves. Therefore, "[i]t is not enough to tell teachers that curricula, reference works or teaching materials are based on corpus analysis" (Kennedy 1992: 367), as they should not only be able to work with corpus-based materials provided by applied linguists or publishers. Teachers should also be able to compile corpora and analyse specific language phenomena themselves. This experience then allows them to reflect more critically on curricula, methods and available teaching materials, like, for instance, course books, as well as on the usage of corpora in language teaching in general.

Such a critical analysis of textbooks nowadays would certainly reveal that the majority of language course books available for educational settings are still not corpus-informed, but "based on intuitions about how we use language, rather than actual evidence of use" (O'Keeffe, McCarthy & Carter 2007: 21). As a result, they contain hardly any information concerning the occurrence frequency of specific words, phrases, collocations or even grammatical constructions used by native speakers or speakers of a specific speech community, which may be important for students in order to learn to communicate effectively in a foreign language. With intuition playing a significant role in the compilation of course books, the conclusion can be drawn that the content of these teaching devices may be sometimes inaccurate, especially with regard to semantic as well as grammatical issues (O'Keeffe, McCarthy & Carter 2007: 21).

One instance exemplifying this point is the way in which irregular verbs are frequently introduced to students at grammar school level, where learners are often only provided with a long list of verb forms which they are supposed to learn by heart. A study conducted by Grabowski and Mindt (cf. Kennedy 1998: 283-284), however, came to the conclusion that when analysing both the 'Brown' as well as the 'LOB' corpus regarding the practical application of 160 irregular verbs, "the 20 most frequent irregular verbs account for 83.6% of all irregular verbs in the written corpora" (Kennedy 1998: 284). This outcome illustrates that if teachers really aim at enabling their learners to communicate in English efficiently,

precisely these twenty irregular verbs detected in the survey should be taught first. However, “discrepancies between what is taught in schoolbooks and what is actually used” (Braun, Kohn & Mukherjee 2006: 1), like the one just illustrated, do not always originate from the fact that the actual language usage of specific phenomena has not been investigated or has not yet found its way into language teaching. Specific rules of application can also be victims of intention-based overgeneralisations so as to simplify the contents which students need to learn (Aston 2000: 8).

Another essential problem which teachers instructing foreign or second language learners are confronted with is that their students are not exposed to the target language they are aiming to learn to the same extent as they have been when acquiring their mother tongue (cf. Tsui 2004: 40). Therefore, learners in foreign language classes are

> unlikely to acquire the language efficiently without systematic guidance on linguistic forms. By focusing on words which have a high frequency of occurrence and by concentrating on the usual rather than the exceptional, teachers can help learners acquire the language more efficiently, especially at elementary and intermediate levels. The findings of corpus analysis can be used as a basis for selecting and sequencing linguistic content, as well as for determining relative emphases. (Tsui 2004: 40)

Amy Tsui is not the only specialist emphasising the advantage of finding out about the frequency of specific lexical items via the use of quantitative analysis of corpora – a tool helpful and vital for the foreign language teaching and learning setting. Gregory Hadley (2002: 99), too, highlights that the introduction of corpora as an instrument in language pedagogy certainly influenced the whole area of language teaching. It even caused a major paradigm shift resulting in the fact that the achievement of communicative competence no longer functions as an exclusive objective behind language teaching, as “the role that form and structure [... play] in educating language learners” (Hadley 2002: 99) has begun to gain importance.

While *Communicative Language Teaching* (CLT) as one of the most prominent approaches in ELT methodology focuses “on the development of skills that allow the learner to actively use a language in a given context” (Adolphs 2006: 99), it clearly contrasts with other language learning and teaching models, like, for instance, those focusing on lexical or grammatical language items. Reflecting critically on CLT, Carter (1998: 51), for instance, emphasises that “[i]n spite of numerous pedagogic advantages, communicative teaching has not encouraged in students habits of observation, noticing, or conscious exploration of grammatical forms and functions”. These processes, however, are all included in the study of corpora, as Carter (1998: 51) highlights, which requires “structures, patterns and predictable features [...] to be ‘unlocked’ by the human intelligence.” (Leech 1997: 3) Although it may appear as if CLT and a more corpus-based approach to teaching can never be used in combination, this impression is wrong, as Mukherjee (2002: 88) exemplifies. He distinguishes between two different dimensions of communicative competence which students should achieve when learning a foreign language, namely to be able to identify the most prominent differences existing between spoken and written language usage as well as those language features specific to particular genres. Both of these dimensions clearly aim at advancing the students’ communicative competence and can easily be fulfilled when teaching students a foreign language with the help of corpora. This is due to the fact that the analysis of electronic language databases allows for sophisticated analyses to be made with regard to which characteristics are relevant to both modes of language as well as to different genres. Therefore, the use of corpora in the language classroom certainly helps to raise the learners’ awareness of the language they are learning. Apart from the two dimensions of communicative competence just referred to above, the implementation of corpora as a pedagogical tool also fosters other abilities in students, like, for example, media, computer and corpus literacy, learner autonomy as well as intercultural learning (cf. Mukherjee 2002: 71-116).

A further discovery made in connection with the study of corpora in the field of language pedagogy which may alter general teaching practice considerably is that native speakers use and remember their language mainly based on fixed, semi-

fixed or variable phrasal units (Johansson 2009: 38). Thus, it only appears to be sensible that

> at the elementary and intermediate stages of instruction, high frequency items in the language rather than intrinsically difficult items should receive the main pedagogical focus in the content and sequencing of the curriculum, in the weight of emphasis in the classroom and in the assessment of achievement and proficiency. (Kennedy 1998: 282)

As can be concluded from this quotation, the study of corpora can certainly help to select which language items are worth being emphasised in the foreign or second language classroom. Kennedy (1992: 335) even stresses that "[s]ince pedagogy attempts to reduce the time that would be necessary to learn a language through exposure alone, potential usefulness and likelihood of occurrence have been seen as relevant for deciding what to teach or learn." Nevertheless, it also needs to be stressed that the role which corpora play in the language classroom should not be overvalued, as the analysis of electronic databases of language usage is certainly "no replacement for natural communication" (Johansson 2009: 42) and cannot substitute the teacher. Therefore, even though corpora are useful instruments applicable in the language classroom, they cannot be regarded as the one and only teaching method that can solve all problems learners and teachers are confronted with in the complex process of teaching, acquiring or learning a language.

Since the analysis of corpora can certainly function as "effective learning tool" (Johansson 2009: 42) which complements various other teaching and learning methods or approaches, "[i]t is the task of language teachers to find the right balance, and tailor methodology to the type of learner and the stage of learning." (Johansson 2007: 26) Naturally, in order to be able to decide which methods to use in teaching a language and how to combine different ones, teachers need to be aware of how corpora can be used in the foreign language classroom. Therefore, several well-selected applications will be introduced in the next section.

4. Applications of corpora in language teaching and learning

As already mentioned, the topic of how to apply corpora best in language pedagogy has only gradually gained importance in applied linguistics. Since the usage of corpora in the teaching of languages has mainly been considered from a theoretical position up to the present day, the practical viewpoint, namely that of teachers already working with corpora in the language classroom, has hardly found its way into recent publications (Braun 2006: 25). In order to give the teacher's perspective more priority, the subsequent parts of this section will focus on various ways in which corpora can be applied in the language classroom. Prior to that, however, this introductory part will be concerned with the different types of language data and methods available for applying corpora in the language classroom.

Generally, there are two kinds of language data that can be analysed by language teachers and learners. Firstly, corpora can contain language samples produced by native speakers of a specific language or language variety. Secondly, corpora can include language data created by learners of the target language themselves. While native speaker corpora provide teachers and students with "opportunities for [...] observing regularities" (Aston 1997b: 63) in naturally-occurring language data, learner corpora offer a number of other advantages. They do not only allow students to "be the authors or providers of corpus materials themselves" (Stewart, Bernardini & Aston 2004: 2) and to reflect critically on this language data, learner corpora can also "be helpful in analysing patterns of deviation from native speaker English and [can], as such, highlight particular problem areas for learners." (Adolphs 2006: 98) As both types of data have varying potential, the choice whether to use native-speaker or learner corpora in the language classroom primarily depends on the purpose teachers try to pursue in analysing corpora for themselves, for their students or together with them.

This already leads to another distinction with regard to how corpora can be used in language teaching, as they can be applied in three different ways. Aston (2000: 7), who bases his remarks on the writings of Leech and Fligelstone, refers to them as teaching explicitly about corpora, exploiting corpora to teach or teaching

students to exploit corpora. These methods of using language corpora in pedagogy are of a direct nature and enable participants to gain knowledge and hands-on experience concerning the work with real language data and sometimes even regarding the immediate analysis of corpora. Electronic language databases, however, can also be applied in teaching more indirectly. This can be done, for example, by making use of reference works, textbooks or other language materials based on the outcome of corpus analysis (McEnery, Xiao & Tono 2006: 97). In this paper, however, the indirect application of corpora in language pedagogy will not be dealt with in greater detail, as the major focus lies on the direct implementation of corpora in language teaching and learning. Therefore, the following paragraphs will give a brief overview of the direct uses of corpora in language teaching and learning.

The first method mentioned above of involving corpora in the language classroom is that of teaching explicitly about corpora and using them as research tools in class (Aston 2000: 7). This way of including the study of corpora both theoretically as well as practically in language teaching is especially suitable for students at university level, who are deeply interested in linguistics and will most likely use corpora later themselves – either as researchers or teachers (Römer 2009: 92). While courses like these are at least occasionally incorporated in the study plan of universities, the method of teaching explicitly about corpora is hardly used at secondary school level, as Karin Aijmer (2009: 2) emphasises.

The second way of introducing corpora to language learners, as Aston (2000: 7) notes, is that of teachers exploiting corpora themselves in order to use the results for teaching. In other terms, teachers use corpora to expand their personal knowledge concerning specific phenomena occurring in natural language usage in general or in particular text types more specifically (see section 4.1.1.). The results gained from these analyses can then be presented to the students in varying forms (cf. Partington 1998: 5). In order to do so, teachers have to mediate between the electronic language databases investigated and their students by choosing and even adapting specific “materials derived from [... corpora] before giving them to learners” (Aston 1997a: 205). Materials which are created by

teachers especially for their language learners are often claimed to support the process of acquiring or learning the target language best (see section 4.1.2. and 4.1.3.). Stewart, Bernardini and Aston (2004: 2) even go as far as arguing that learners are “the ultimate beneficiaries of corpus insights” when teachers provide them with naturally-occurring language data in a mediated way.

Finally, the third method of applying corpora in the language classroom is that of letting students compile or at least analyse corpora themselves in search of any prominent findings in general or for answers regarding particular research questions. This process allows students to build up mental concepts regarding the usage of the language they are learning, which may consequently help them to achieve both a higher level of language awareness and competence. In contrast to the first way of ‘teaching about corpora’, this third method focuses on the practical analysis of language data contained in corpora, instead of introducing students’ to corpora and their potential from a more theoretical point of view. As Stewart, Bernardini and Aston (2004: 8) observe quite correctly regarding the origin of the method described as ‘teaching to exploit’:

> Right from the first TaLC conference, there were papers which viewed corpora primarily as tools which learners could use to find out about the language (and the culture behind that language) for themselves, with or without the help of their teachers.

In other terms, the method of teaching students to work with and analyse corpora themselves was already promoted by applied linguists and teachers right at the beginning of the introduction of corpora into the field of language pedagogy (Kaltenböck & Mehlmauer-Larcher 2005: 78). Nowadays this way of applying corpora in the language classroom is frequently referred to as *discovery or exploratory learning* (see section 4.2.2.). Even though these terms may imply that students are given independence when exploring corpus data, this does not necessarily mean that teachers are not allowed to function as mediators between their students and the electronic language database at all. Especially in the students’ early days of working with corpora, they certainly require training in working both with electronic language databases as well as analysis programmes (cf. Partington 1998: 5). But also later on, learners may be guided by their

teachers in their explorations, as pedagogues may, for instance, select, edit or adjust naturally occurring language data for them. Once students are familiar with the most fundamental tools for analysing corpora, however, they can also be left to analyse corpus data on their own, knowing that their teachers are at hand when help is needed. Therefore, "there seems no a priori reason why learners should not be given direct access to corpora for independent use." (Aston 1997a: 205)

When it comes to introducing language learners at secondary school level to corpora, especially the last two methods seem very promising. Therefore, they will be dealt with in more detail in the two subsequent sections focusing on potential methods of how to include corpora into language teaching. While section 4.1. will be concerned with the different forms in which teachers can use corpora inside and outside the language classroom, section 4.2. will discuss several ways in which students can interact with corpora directly.

First, however, another distinction of how to integrate corpora in language teaching, apart from that postulated by Aston (2000: 7), will be discussed. This second differentiation has been made by Michael McCarthy (2001: 129), who distinguishes between three approaches concerning the usage of corpora in language pedagogy, namely the so-called *corpus-based, corpus-driven* and *corpus-informed approach*. It is important to note that all of them can be used both in teacher-corpus as well as in student-corpus interaction, depending on the aim underlying the corpus analysis which can take place both inside as well as outside the language classroom (Gavioli & Aston 2001: 244). When using a corpus-based approach to language teaching, corpora are exploited in order to discover language samples which exemplify what is already known. However, "[o]ne can do the opposite, and go with a completely open mind to a corpus, willing to be guided, illuminated by it in ways one could not dream of", an approach which McCarthy (2001: 129) refers to as being corpus-driven. Finally, the corpus-informed approach can be applied, which demands mediation between the language database and the explorer. In this case the teacher creates a corpus together with his or her students with particular research questions in mind. Later on, the self-compiled corpus can then be analysed jointly in order to find answers

to the already formulated questions (cf. McCarthy 2001: 129). Examples of all three approaches can be found in the following subsections on teacher-corpus and student-corpus interaction (see section 4.1. and 4.2.).

A topic also often addressed in connection with the implementation of corpora in the language classroom is that of mediation, which just means that the students' involvement with computer corpora can be manipulated "to a greater or lesser degree [...] by the teacher for the student's benefit" (Leech 1997: 8). Mediation can, for example, take place by filtering language samples out of corpora which exemplify specific language phenomena to learners quite clearly (see section 4.1.3.). When using corpora in an unmediated way for a similar task, students would be asked to investigate a whole corpus themselves in search for particular constructions. The question which remains unanswered in specialist literature regarding this issue is, to which extent "such mediation is desirable in language pedagogy, or whether learners can profit more from direct access to the corpus" (Aston 2000: 14). Concerning this problem neither teachers nor applied linguists appear to be able to agree on one position, as both the unmediated as well as the mediated use of corpora in the language classroom offer different advantages, which will briefly be discussed in the following paragraphs.

First of all, it has to be mentioned that both the mediated as well as the unmediated application of corpora in language teaching give students the opportunity to explore naturally-occurring language data independently. The mediated usage of corpora, however, allows teachers to guide students in their process of analysing and reflecting on practical language usage to a greater extent, as pedagogues can, for instance, preselect the language samples to be explored by their learners. Another advantage of teachers functioning as mediators between their students and real language data is that they can also encourage their students to authenticate the language contained in corpora by "adding to the reality of the corpus the reality of their own experience of it." (Gavioli & Aston 2001: 244) In other terms, mediation allows teachers to prompt students to contextualise authentic language samples taken from corpora by giving explanatory instructions or raising the learners' awareness with regard to

specific language phenomena. By interpreting samples of naturally-occurring language usage on the basis of their own experience, students may then become able to identify them as genuine instances of language (cf. Widdowson 2003: 93). When asking students to explore corpora directly without any mediation from the side of the teacher, this authentication process hardly takes place at all. This is mainly due to the fact that students do not necessarily establish a connection between language samples taken from corpora and the suitable communicative situations in which these might have been used once or may be used in future. Furthermore, the amount of language data which students are confronted with when analysing whole corpora may just be too high for learners to decide which of the language samples are relevant and should thus be dealt with more profoundly and which are only distracting (Osborne 2000: 165). The unmediated usage of corpora in language teaching can still have its positive aspects though, as it certainly motivates students to conduct their own research independently from the teacher. One requirement for applying corpora in the language classroom in an unmediated way, however, is that students should already be quite proficient in working with corpora. This does not only mean that learners should be able to operate all necessary tools in order to analyse corpora autonomously. Moreover, students also need to be able to estimate the value of the results of their corpus queries correctly.

However, corpora cannot only be applied in language teaching in a completely unmediated or mediated way. “In between these two extremes, various degrees of mediation are, of course, possible”, as Gavioli and Aston (2001: 244) stress. Therefore, students may not only explore corpora by working with language samples being particularly chosen and sometimes even slightly edited by teachers or by investigating whole corpora in search for an answer concerning a particular research question (Barlow 1996: 30). Learners can also be guided through their analyses only in parts so as to be sure that all students are, for example, able to come to a similar conclusion of how a certain language item can be used in communication. In other terms, the extent to which mediation is needed when introducing corpora into language teaching and also the form in which it is provided may vary according to the needs of the learners (Stewart, Bernardini &

Aston 2004: 6). Overall, it may be advisable for teachers to function as slowly withdrawing mediators between the students and the corpus at first, before letting them investigate real language data without any specific guidance as soon as they have acquired the necessary proficiency.

To conclude, it can be said that no matter in which form corpora are applied in the teaching context, whether mediated or unmediated, they certainly "reduce the learner's dependency on the teacher, and the teacher's on the textbook, allowing teachers to concentrate on their role as learning rather than language experts, " as Aston (1997b: 63) emphasises. After these quite general remarks on the role of both teachers as well as students in the corpus-influenced language classroom, the following section will now address a number of ways in which teachers can interact with corpora for the benefit of their students.

4.1. Teacher – corpus interaction

The most traditional way of applying corpora in the language classroom is their use as sources by teachers (Gavioli & Aston 2001: 244). Pedagogues interacting with corpora can implement electronic language databases in language pedagogy in two different ways, as Susan Hunston (2002: 137) points out. On the one hand, corpora can be used by teachers in order to gain insights into the ways in which native-speakers or learners use language in everyday-life situations and can thus broaden their personal knowledge regarding practical language usage (e.g. for reference purposes – see section 4.1.1.). On the other hand, corpora can be analysed by teachers and introduced to students with the help of mediation of varying degrees (e.g. for material design – see section 4.1.2. and for demonstration – see section 4.1.3.). Both of these uses have already been quite popular in 1994, when the first TaLC conference was held in Lancaster (Stewart, Bernardini & Aston 2004: 6).

Especially the second way of how teachers can apply corpora in the classroom is given much attention in specialist literature (e.g. Hunston 2002: 137). This is due to the fact that teachers mediating between corpora and their students are able to adjust the issues to be dealt with in class exactly to their learners' needs

(Johansson 2009: 42). Another advantage of including corpora into the language classroom via mediation is that students are able to work directly with selected samples of naturally-occurring language rather than with invented examples exemplifying specific language phenomena (Granath 2009: 49). As a resource for this kind of data, corpora

> are invaluable for teachers, in that they can employ them in a number of ways, such as, for example, to create exercises, demonstrate variation in grammar, show how syntactic structures are used to signal differences in meaning and level of style, discuss near-synonyms and collocations, and last (but not least) to give informed answers to student questions. (Granath 2009: 49)

As can be seen from the examples of teacher-corpora interaction provided in the quotation above, corpora are primarily of use for teachers in the preparation of language classes and thus outside the classroom. However, not exclusively, as corpora may also be of help to teachers in certain situations inside the language classroom (Kaltenböck & Mehlmauer-Larcher 2005: 78). Exactly these uses, both inside as well as outside of language classes, will now play an important role in the following two subsections of this chapter. These parts will deal with the two major methods of how teachers can use corpora in the (foreign) language classroom already referred to above by Susan Hunston (2002: 137), namely for reference purposes as well as for material design.

4.1.1. For reference

A well-established method of how to incorporate corpora in the field of language teaching and learning is that of teachers using them as reference materials (Anderson & Corbett 2009: 175). Up to the present day language problems encountered both inside as well as outside the language classroom have regularly been solved by teachers based on their non-native speaker intuition or personal language knowledge. With the help of corpora, however, this field can be revolutionised, as teachers no longer have to depend on their personal evaluation of whether a specific utterance contains acceptable language structures or not. Thanks to the introduction of corpora into language teaching, teachers can now rely on language knowledge going way beyond that of an individual native

speaker and also beyond the information provided in traditional dictionaries or grammar books (Mukherjee 2009: 168). Especially the quantitative and qualitative analysis of corpora can help teachers to verify or modify their subjective evaluation of whether or not certain lexical or syntactic constructions are appropriately chosen in order to convey the intended meaning (Kennedy 1992: 366). Therefore, corpora can be described as precious instruments which teachers can refer to in case of doubt and as tools which are “perhaps more reliable than many of the available teaching materials or (non-corpus-based) usage handbooks.” (Römer 2009: 93)

However, corpora cannot be regarded as the only and thus an exclusive reference tool for teachers. As already discussed, corpora are very helpful when searching for occurrence frequencies of specific words or phrases, for wordlists regarding particular texts as well as for real language samples. However, traditional reference materials, like dictionaries, grammar books or practical usage guides, many of which are based on the results of corpus analysis nowadays anyway, may help teachers to find answers to other questions more quickly and precisely (Leech 1997: 14). This may be the case, for example, when searching for definitions regarding particular words or phrases. In this context, as Partington (2001: 64) remarks, dictionaries or encyclopaedias should be preferably used, as they “are designed to describe conceptual or denotational meanings, arranging the different senses of a word in some kind of order.” In contrast, corpora are less well structured, as they only provide the user with the search word or phrase together with its immediate verbal co-text. Consequently, “it is not always easy to reconstruct the conceptual meaning of a word precisely” from the little information given which encloses it (cf. Partington 2001: 64).

When applying corpora practically inside the language classroom, they can, for example, be used as reference materials so as to answer questions asked by students. Even though teachers may often provide students with spontaneous and thus intention-based answers regarding their questions first, they can conduct a corpus analysis as a follow-up. The analysis results can then be presented to the students in the next lesson. As Granath (2009: 53) points out,

> [t]his has the advantage that the information [provided] is based on what the students are interested in finding out about, and it gives the teacher an additional opportunity to show students how corpora can provide information beyond what we find in reference books.

Moreover, when providing students with answers to their questions by means of authentic language samples, "the language comes alive in the classroom in a way that is not possible by just relying on reference books and made-up examples." (Granath 2009: 55)

An instance of this kind of corpora usage can be found in an article by Granath (2009: 53-54), who provides his readers with an example from his own teaching experience. One of the questions which came up in his language classroom was why some people talk about *the Ganges River,* while others call it *the river Ganges*. In order to clarify whether both options are correct or whether there is any difference in meaning, Granath (2009: 53-54)

> consulted both American and British corpora, and the answer, for once, could be stated very clearly: American speakers will put 'river' after the name (and capitalize it in writing, i.e. *the Ganges River*), whereas in Britain, 'river' precedes the name and can be capitalized, but is more commonly written with a lower-case r: *the river Ganges / the River Ganges.* The most common variant in both varieties, nevertheless, is to use just *the Ganges*, without the apposition *river.*

Students' questions can be manifold, just as corpora. No matter whether lexical, grammatical, stylistic, cultural or other issues are addressed by learners' enquiries, corpora can provide teachers with real language data to investigate. However, as can be seen from the example given above, teachers need to know what they are actually looking for and which features may be determining factors in their search. Sometimes, as in the aforementioned example, it may, for instance, be necessary for teachers to compare corpora of different language varieties. In other situations, however, pedagogues may need to have a closer look at the registers or age-groups in which specific expressions are used. In other terms, depending on the students' questions and thus the individual starting points of particular investigations, the ways in which corpora have to be searched

in order to get satisfactory results may vary. Sometimes it may also seem to be impossible for teachers to answer their students' questions. But also for this problem a solution has been found, at least in China.

In Hong Kong, for example, an internet platform exists which helps English teachers to find answers to difficult language questions asked by students (Römer 2009: 94-95). This online service is called 'TeleNex', which is the abbreviation for "Teachers of English Language Education Nexus" (Tsui 2005: 337). It is hosted and supported by a group of language experts "from the University of Hong Kong who use corpus evidence to respond to questions" (Römer 2009: 94) posted on the platform. As soon as any questions are stated in the teachers' forum, the language specialists hosting the website search through corpora to find suitable answers, summarise the "corpus findings on the use of the items in question and provide selected concordance lines to highlight their central usage patterns." (Römer 2009: 94) The online service just described seems to be a very unique one, nevertheless it is to be hoped that websites of similar structure and content are soon available to "a larger number of teachers in different countries around the world." (Römer 2009: 94)

The direct or indirect usage of corpora in order to answer students' questions, however, is only one way of applying them for reference purposes in the language classroom. Electronic language databases can also be helpful at school when teaching grammar rules or vocabulary in the language classroom. Especially with regard to grammar, textbooks often provide students and teachers with generalised rules of how to apply a certain grammatical phenomenon in practical language usage. More often than not, however, these rules are simplified and thus overgeneralised, which is believed to help students applying them without further difficulty (Tsui 2004: 56). In case of doubt that any grammatical rule may not be completely correct or that exceptions are just not referred to, corpora may be used to supply evidence (e.g. regarding the three traditional conditional forms). In vocabulary teaching, electronic language databases may, for example, be used to find out how specific words collocate (e.g. whether *at short notice* is more frequently used in actual language use than *on short notice* or vice versa*)* or how

often terms included in the vocabulary sections of course books really occur in natural language usage. This may be of interest to teachers, as it certainly does not make sense to let students learn words or phrases by heart which are hardly used in naturally-occurring language anyway.

However, corpora cannot only be used for reference purposes by teachers inside, but also outside the classroom. Especially at home, when teachers correct their students' homework or also their tests, situations may occur in which teachers are not certain whether specific collocations or grammatical constructions can really be used in the way students might have done. In situations like these, teachers may tend to argue that a particular usage is just wrong, because they are not familiar with it or cannot comprehend why students have chosen an adventurous combination of words instead of a construction they should already be familiar with. Often teachers also use external resources instead of intuition in order to be sure whether to classify their students' language usage as correct or incorrect.

A survey conducted by Ute Römer (2009: 86) shows that in case of doubt "whether a certain construction or collocation in a learner's essay [... is] acceptable or not", many teachers prefer to consult dictionaries, preferably monolingual ones (21.84 %), or native speakers (21.84 %). Other teachers, especially those who are not in regular personal contact with native speakers and therefore have to depend on their intuition as non-native speakers, among other things, "also ask colleagues for help (10.14%) or consult the Internet (7.8 %)." (Römer 2009: 87) Even though all of these ways of finding out about the appropriateness of constructions and collocations used by students can help teachers in correcting their learners' work, Römer (2009: 92-93) mentions that many of her informants would want a native-speaker available to them whenever needed. "[T]his is a wish that corpus linguistics can easily fulfil", as the author (Römer 2009: 92-93) points out in her next sentence, as corpora generally consist of language data produced by thousands of native speakers. One clear advantage, however, of using corpora as reference material, as opposed to questioning native speakers, is that corpora "are available 24 hours a day, seven days a week and thus enable teachers to check language points and find

information on common word-combinations or the typical usage of an item whenever they want" (Römer 2009: 93). Furthermore, corpora offer the analyst a more comprehensive picture of actual language use, as they include data from more than just one native speaker.

Coming back to the situations in which teachers sit at home trying to correct their students' homework to the best of their knowledge and belief, corpora can certainly assist them to find informative real language data to base their assessment on. However, corpora cannot help teachers to classify whether the students' language usage is perfectly correct or entirely incorrect. This is due to the fact that the dichotomy between correct and incorrect language usage is just not applicable sometimes, as language does not only have to be correct with regard to lexico-grammatical issues, but also concerning its adaption to the communicative situation it is used in, as Joybrato Mukherjee (2002: 138) stresses. Apart from this so-called 'range of correction', teachers should also consider the 'depth of correction' (cf. Mukherjee 2002: 138) when assessing their students' works based on corpora. This means that the errors and mistakes (e.g. concerning spelling, grammar, punctuation, choice of register or vocabulary) made by students should not all be regarded as being equally serious. Imperfections concerning the "combinability of words and the appropriateness of collocations" (Römer 2009: 94) can be ranked among the mistakes occurring most frequently in students' works. Especially in this area corpora can be of great convenience for teachers, as they can help them to find out how specific words actually collocate.

In order to illustrate the last point made above better, an example will be given. A student uses the phrase *he was forced to make a diet* in his homework (Römer 2009: 93). The teacher might not be certain whether the phrase *to make a diet* actually exists or whether it is an idiomatic expression. As a result, a corpus may be consulted where the teacher looks up the term *diet* "and check[s] the phraseology of the word." (Römer 2009: 93) The teacher would then encounter that native speakers favour to use either the phrase *to go on a diet* or *to be on a diet.* This result can also be confirmed when having a closer look at the 'British

National Corpus' as presented on the internet by Mark Davies from the Brigham Young University (cf. http://corpus.byu.edu/bnc/). When searching the phrase *to make a diet* in this one-hundred-million-word-corpus, not a single occurrence can be found. In contrast, 127 tokens can be found when asking the user-friendly interface for instances regarding the phrase *on a diet.* While 27 occurrences can be found for the collocation *to go on a diet,* the BNC contains only 6 language samples for *to be on a diet.* From these pieces of information the teacher can then conclude that the collocation chosen by the student may be an unusual one and may introduce the learner to one or both alternatives found in the corpus.

For their analytic search, teachers cannot only use corpora which can be bought or those being freely available on the internet, but also the World Wide Web as such. Search engines like, for instance, 'Google' can also be applied in order to find out whether specific word combinations actually occur in English or not (cf. Römer 2009: 93). This is due to the fact that search engines are also able to provide their users with the occurrence frequency of specific search terms and thus allow for conclusions to be drawn whether certain phrases are used more often on the internet than similar ones. However, using the internet to find out more about actual language usage also has its limitations. One of them is that the World Wide Web is neither a principled nor a controllable electronic language compilation. Therefore, "the output of Google and other commercial search engines has to be treated with a lot of caution, in particular with respect to the sources of Web-attested examples", as Ute Römer (2009: 93) points out.

4.1.2. For materials design

Apart from using corpora for reference purposes, "[p]erhaps the most obvious pedagogic use of corpora is to treat them as sources of classroom materials", as Aston (1997b: 52) notes. For this kind of application teachers appear to have a special need, as the materials which they are provided with in language course books nowadays are disappointingly little influenced by analysis results gained from corpora (Römer 2009: 90). The use of corpora in order to design and improve pedagogical language materials by "making [... them] correspond more closely to typical native speaker use", has occasionally also been referred to as

the so-called ‘COBUILD’ or ‘Birmingham’ approach, based on John Sinclair’s ‘Cobuild Project’ (Nesselhauf 2004: 126).

Real language samples gained from electronic language databases can be introduced as teaching materials into the language classroom in two different ways, as Aston (2000: 12) points out:

> The examples may simply serve as illustrations to present the uses in question, or they may be structured in exercises where the learner is asked to classify or complete them, so as to practice the recognition and production of the uses involved.

No matter whether real language samples are used in their natural form or whether they are transformed into exercises, both ways of analysing naturally-occurring language data can be integrated into the process of language teaching and learning inside as well as outside the language classroom if the necessary parameters are given. This means that students either need to have access to corpora (e.g. via the internet) or are provided with printed concordance lines or worksheets by their teachers.

A major advantage of teachers designing language materials based on corpora themselves is that they can create materials concerned with all different kinds of language issues whenever these are required (Römer 2009: 91-92). In other terms, pedagogues do no longer have to rely exclusively on the texts, language examples and topics which are offered in textbooks or in other publications. They are given the opportunity to compose materials based on enormous collections of naturally-occurring language data themselves (Aston 1997b: 52). However, this option is only open to teachers who are willing “to invest time in searching for more interesting and more authentic texts and in creating additional exercises” (Römer 2009: 88-89) for their students.

In order for teachers to be able to design materials based on corpora themselves, however, they first have to learn how to analyse huge electronic language databases (Römer 2009: 91-92). For this purpose cooperation between teachers and researchers is necessary, as teachers need to be trained to become better

acquainted with and improve their skills in handling both corpora as well as available retrieval programmes. This is not the only area in which pedagogues and applied linguists have to work together, as Ute Römer (2009: 92) stresses. More support for teachers from the research community is also needed when it comes to generating teaching materials and thus resources for teachers to use in the language classroom, as this process of creation should “not be left entirely to the teacher.” (Römer 2009: 92) This issue, however, will not be dealt with in more detail in this paper, as it concerns an existing problem between the research community and the teaching profession which cannot be solved in one day. This is due to the fact that it requires willingness to communicate and to collaborate from both theoreticians and practitioners so as to find a good and mutual basis for cooperation.

Coming back to the practical application of corpora in the language classroom, Joybrato Mukherjee (2002: 128) distinguished between three types of language materials, namely ‘informative corpus-based materials’, ‘illustrative corpus-based materials’ and ‘corpus-based exercise materials’. Firstly, he refers to those materials which are created on the basis of dictionaries, grammar books or other reference materials derived from corpora. These materials are called informative corpus-based materials. Secondly, language materials can also be designed on the basis of corpora as such, which Mukherjee (2002: 128) refers to as illustrative corpus-based materials. Finally, corpus-based exercise materials are created in order to help students practise and stabilise specific language forms on their own (cf. Mukherjee 2002: 128).

This distinction made by Mukherjee (2002: 128) is based on the original material used in order to create corpus-based language materials for the classroom. The application of corpora for materials design can, however, also be categorised content-wise, as Geoffrey Leech (1997: 16) illustrates in his article. He distinguishes between three ways in which corpora can contribute to the creation of pedagogical language materials. Firstly, as mentioned on several occasions already, corpora are suitable instruments in order to determine how often specific words or phrases are commonly applied in actual language use. Secondly,

corpora provide those working with them with plentiful naturally-occurring language samples. Thirdly, corpora provide teachers with “computer-delivered learning packages” (cf. Leech 1997: 16), which means that teachers cannot only exploit real language data so as to create language materials themselves, but can also make use of corpora as teaching devices in their entirety. This means, in other words, that corpora per se can function as language materials, which can consequently be analysed together with students in various mediated or unmediated ways. Subsequently, all of these three points will be dealt with in more detail individually so as to illustrate the great potential of self-made language materials created on the basis of corpora.

First of all, corpora can contribute to the creation of language material by supplying those designing the pedagogical devices “in abundance [... with] frequency information” (Leech 1997: 16). However, not all pieces of information of this type are useful for the process of generating language material. This is due to the fact that it is not only important how often certain language items occur in practical language usage, but also how relevant they are for the learners’ ability to communicate efficiently (Widdowson 1996b: 13). In other terms, teachers should not include all kinds of words or phrases into their teaching material just because they occur frequently in real language data, but should also reflect on the relevance of these items regarding the process of learning and becoming proficient in a foreign language.

Word frequency lists, however, cannot only be integrated into language pedagogy in order to tell teachers which lexical items are worth teaching and which are not, they can also be used directly as language materials in the classroom, as the following example taken from Laura Gavioli (1997: 88-89) attempts to illustrate. She describes that a wordlist containing a number of frequently occurring words regarding a particular text (see Figure 2) can, for instance, be handed out to students in order to let these reconstruct the content of this text.

the	14	in	6	are	4
of	10	they	6	on	4
drivers	8	a	5	their	4
to	7	for	5	young	4

Figure 2: Twelve most frequently occurring words in a text (Gavioli 1997: 89)

After having had the opportunity to read through the provided word frequency list, learners may be asked to guess the topic of the text underlying it. As ten out of twelve are functional words, only two words will be of importance for the learners to base their hypotheses on, namely *drivers* and *young* (Gavioli 1997: 88). The students' individual ideas regarding the content of the text can then be used as a basis for a writing task, as teachers can, for instance, ask their students to write short texts which need to contain all words from the list in their correct quantity. After this exercise, students may be given a number of other words occurring quite frequently in the source text (see Figure 3) and can then be asked to revise their personal hypotheses regarding the topic of the text.

accidents	3	cent	2	or	2
been	3	certain	2	other	2
driving	3	considered	2	per	2
from	3	deaths	2	plates	2
is	3	has	2	said	2
move	3	have	2	such	2
after	2	involving	2	them	2
an	2	Ireland	2	there	2
be	2	more	2	two	2
being	2	new	2	were	2
by	2	Northern	2	with	2
Carlisle	2	number	2	would	2
cars	2	options	2	years	2

Figure 3: Words occurring in the source text three or two times (Gavioli 1997: 97)

Finally, students may be invited to form small groups in order to discuss the influence of this new word frequency list to their initial ideas regarding the issue dealt with in the source text and may be asked to merge these to a group-version. Each group can then present their own hypothesis regarding the content of the source text in class, before the original text (cf. Gavioli 1997: 99) is handed out and discussed.

The second contribution of corpora to materials design mentioned by Leech (1997: 16) is that electronic language databases can provide teachers with authentic samples of actual language use. Here it is necessary to distinguish between language samples and examples, as "corpora can be used to teach students to interpret instances as *samples* rather than *examples"*, as Bernardini (2004: 21) stresses. While samples are natural instances of language use, which are collected rather than created, examples are language models specifically designed in order to illustrate a particular language phenomenon. Therefore,

> unlike the examples provided by textbooks and dictionaries, the samples of language provided by corpus data do not immediately illustrate particular linguistic patterns. A concordance does not make sense in itself: sense has to be attributed to it by the reader, who must infer patterns which will as far as possible account for the data. (Gavioli & Aston 2001: 241)

Unfortunately, hardly any textbooks or other teaching devices available on the educational market are nowadays based on naturally-occurring language data. Most of them are based "almost exclusively on made-up examples", as Adolphs (2006: 99) notes. This fact is regarded particularly critically by John Sinclair (1991: 5), who argues that someday "it will be realized that there is just no reason or motivation to invent an example when one is knee-deep in actual instances". However, up to the present day language course books do not even include conversations and dialogues which are characterised by the most prominent features of spoken language, "such as ellipsis, turn overlaps, false starts and repetition" (Adolphs 2006: 107). Nevertheless, "[i]t is to be hoped that materials [...] which include genuine instead of invented language and take corpus findings into account, will soon also be available" (Römer 2009: 90) to teachers around the

world. Teaching materials based on real language data would certainly assist teachers to find suitable language material easier, just as they would help students to learn how to communicate efficiently in a motivating way (Aston 2000: 12).

One exercise which may in future be part of a corpus-based textbook, could, for instance, be a fill-in exercise like the following one illustrated in Figure 4 (cf. Römer 2009: 91). This task requires students to decide which of two near-synonyms, namely *speak* and *talk*, is the correct contextual choice for each sentence. What needs to be mentioned with regard to this task, however, is that students cannot simply be asked to fill in the correct forms without any further explanation. This is due to the fact that the limited number of sample sentences does not really provide the learners with an opportunity to deduce any rule of application regarding the distinction between *speak* and *talk.* Therefore, the corpus-based exercise illustrated in Figure 4 can mainly be used as a follow-up activity after students have already been introduced to some guidelines of application.

What is the missing word in each of the following sentences – 'speak' or 'talk'?

I'm not here to _____ on behalf of the theatre at all.
Are you able to _____ English fluently?
I'd like to _____ about something with you.
I managed to put her off that idea, managed to _____ her out of that.
So you're free to _____ your mind.
Excuse me could you _____ up a little bit? Yes yes er thank you.
I will _____ to David about it as well.
Mothers and fathers _____ differently to sons and daughters.
Men tend to _____ like that, don't they?
You're not allowed to _____ for the rest of the week.

Figure 4: Corpus-based exercise on the near-synonyms *speak* and *talk* (Römer 2009: 91)

Exercises of similar structure can easily be created by teachers regarding different kinds of lexical (e.g. near-synonyms: *small* or *little)*, grammatical (e.g. pronouns / adjectives: *some* or *any)* or also stylistic (e.g. conjunctions: *if* or *whether)* issues

(cf. Römer 2009: 91). Moreover, corpus-based exercises like the one presented above cannot only be used in order to help students practise language features which are new to them. They can also be given to students for reasons of refreshment, especially if teachers observe that some students in their class have problems with certain language phenomena. These can then be focused on from a theoretical point of view once more, before becoming a field of attention in practice again.

Apart from the fill-in exercise presented in Figure 4, naturally-occurring language data can also be introduced to the students in different exercise formats. Learners can, for example, not only be provided with fill-in exercises in which they have to decide which of two language items is the correct choice for each sentence. They can also be asked to fill the gaps occurring in a text with various different words given in a bar at the top or bottom of the exercise or may have to tick whether a specific language sample is correct or incorrect with regard to a specific grammar rule which they have learned recently. Depending on the purpose teachers pursue in providing their students' with exercises based on real language data, also multiple choice questions, word sequencing tasks, short answer questions, editing exercises, structure identification tasks, sentence transformation exercises and matching tasks between sentences and definitions or between sentence halves can be used as exercise formats when designing classroom materials (Harris & McCann 1994: 36-38).

It is now time to have a closer look at the third way in which corpora can contribute to the creation of language materials (Leech 1997: 16), as even whole corpora can be used as language material: they can be analysed together with or individually by the learners in an unmediated or mediated way. One specific way of integrating corpora into language teaching is, for example, to use them "for reading work, focussing on their meaning rather than on their linguistic properties", as Aston (2000: 14) notes. However, as the direct exploitation of whole corpora by students falls into the category of student-corpora interaction (see section 4.2.) and more precisely into section 4.2.2. on discovery and

exploratory learning, it will be dealt with in more detail in the following section of this paper.

What has been said so far with regard to the usage of corpora in language teaching and learning in order to design pedagogical materials has mainly focused on the teacher's interaction with native-speaker corpora. However, as already mentioned in the introduction to this section, learner corpora can also be used for the construction of language materials (O'Keeffe, McCarthy & Carter 2007: 23). Pedagogical devices based on this kind of language data may help students "to study features of interlanguage (often in comparison with the language produced by native speakers) and to analyse 'errors' ", as Stewart, Bernardini and Aston (2004: 3) remark. The analysis and identification of characteristics of learner language may consequently help students to reflect more intensely on the process of language learning. Furthermore, the explicit exploitation of language data produced by students in order to create language materials may also assist teachers to adapt "teaching methods and contents more precisely so as to speed acquisition." (Stewart, Bernardini & Aston 2004: 3)

4.1.3. For demonstration

Apart from applying corpora as reference materials and for materials design, electronic language databases can also be used in the field of language teaching and learning for demonstrative purposes, namely in order to illustrate practical language use and thus to provide learners with samples of naturally-occurring language. This demonstrative use of corpora will now be dealt with in more detail subsequently.

Often when teachers try to illustrate or clarify specific language phenomena in the classroom, students are first provided with theoretical explanations or rules regarding the application of these features. Only then examples of how specific items can be integrated in actual language use are shown to learners. When using corpora for demonstrative purposes in language pedagogy, the basic procedure is that "the learner is provided with generalised explanations of particular linguistic uses, and then works on corpus data which exemplifies these"

(Aston 2000: 12). Tim Johns (1991: 3) generally refers to teaching situations in which explanations precede the practical exploitation of the language phenomena under consideration as instances of *deductive learning*.

According to Gavioli (2005: 25), the use of corpora for demonstrative purposes is particularly suitable for "those areas which are traditionally considered 'difficult to deal with' and where descriptions provided by grammars and/or dictionaries seem inadequate." This is due to the fact that the use of corpora in order to demonstrate how particularly difficult items are incorporated in actual language use may assist students to create mental concepts regarding the construction and application of these features. With regard to the difficulty of the language samples and exercises used for demonstrative purposes, Granath (2009: 63) notes that especially if the levels of difficulty of these are geared to the language competence of the students, "corpus work can help raise their awareness of structures." Therefore, teachers should always try to estimate their students' level of competence in order to pick naturally-occurring language samples which are neither too easy nor to challenging for them. Teachers putting this advice into practice may even use corpus samples with beginners, as

> [c]orpus examples are important in language learning as they expose students at an early stage in the learning process to the kinds of sentences and vocabulary which they will encounter in reading genuine texts in the language or in using the language in real communicative situations. (McEnery & Wilson 1996: 104)

The use of authentic language samples for demonstrative purposes is an often discussed and quite important topic in the field of language teaching and learning (e.g. Leech 1997: 16). This is due to the fact that teachers are regularly required to "come up with examples of a particular expression or construction" (Aston 1997b: 53) in order to illustrate to their learners how specific items are used in context. In these situations corpora can be quite helpful, as they allow teachers to retrieve all instances of a search word or phrase together with their immediate verbal surrounding from a large amount of real language data. After having filtered out all samples of a particular word or collocation, these instances "can be viewed, selected and sorted in a variety of ways before being printed or saved,

giving the teacher a range of [... samples] with which to illustrate a particular usage." (Aston 1997b: 53) There are two major advantages of finding authentic language samples with the help of corpora, as Anderson and Corbett (2009: 176) note. Firstly, corpora may include instances of practical language usage in specific contexts unfamiliar to both teachers and learners and secondly, corpora provide those analysing them with transcriptions and "recordings of speakers of English in many contexts that are not usually found in textbooks and their accompanying audio-visual resources".

However, as Sinclair (1997: 31) stresses, the fact that only language samples taken from corpora can provide teachers with actually-occurring written and spoken language data does not mean that pedagogues are not allowed to provide their students with invented language examples at all. In some situations there is just no time left for the teacher to search through a corpus so as to find suitable samples of specific language phenomena. Therefore, they have to come up with made-up examples spontaneously. However, if teachers have time to prepare language instances for presentation in class, they are certainly better off time-wise and effort-wise to analyse corpora in search of particular language samples themselves, instead of inventing examples which are supposed to sound natural (Sinclair 1997: 31).

Nevertheless, even when searching for samples of particular language phenomena in large corpora, sometimes none can be found. However, if this really happens, teachers should always reflect critically on the term or collocation they have been looking for. This is due to the fact that numerous words or phrases are just not as common in actual language use as one might assume (Sinclair 1997: 31). Consequently, searchers who are not able to find evidence for a particular language phenomenon in large corpora will have to come to the conclusion that this item is rarely used in naturally-occurring language data and possibly not even worth teaching to students.

However, there are also some problematic points which need to be considered when selecting language samples from corpora. First of all, the process of

choosing language samples from corpora should always be conducted attentively and carefully. Otherwise, the selected samples may provide learners with “a chaotic picture of the language” (Kennedy 1992: 366). Another problem which teachers may come across when searching for language samples in corpora is that even though they find instances of the phenomenon they are looking for their search item may be “mixed up with other material which is unsuitable for the job in hand.” (Sinclair 1997: 31) However, when using large corpora in search for useful material, this should not be the case, as at least one of the language samples to be found should be suitable for the language classroom. A further point which teachers, according to John Sinclair (1997: 31), should be aware of when using language samples taken from corpora is that “other patterns which co-occur with a cited word or phrase are actually independent of it.” Therefore, teachers should not only present students with single phrases contained in corpora without any co-text, but with complete units of meaning in order “to avoid the risk of distorting real language patterns” (Aston 2000: 12). In this way teachers can prevent setting their students on the wrong track regarding the meaning, construction and practical application of the language items under discussion.

In order to illustrate the principles stated above regarding the demonstrative use of corpora in the language classroom an example will be given now, which deals with the corpus-based introduction of the three traditional conditional forms into the language classroom (cf. Gavioli 2005: 26). In the teaching situation students will first be confronted with some rules regarding the construction and usage of the three traditional types of if-constructions. As a next step, students will be exposed to language samples from corpora including different conditional forms (Gavioli 2005: 26), which can then be assigned to the three types of conditional clauses introduced previously. After that they can be asked to identify which combinations of tenses are used in the remaining language samples. Finally, the teacher can then point out that “the three-conditional model is a useful one for beginners” (Gavioli 2005: 26), even though the actual usage of the if-clauses is a more complex one. According to Partington (1998: 80-86), even only 40% of the if-constructions occurring in the language material investigated by him belong to one of the three traditional types of conditional forms. The remaining 60% consist

mainly of “other mixed conditional or non-conditional uses”, as Gavioli (2005: 26) points out. When demonstrating the variety of conditional types which exist in practical language use to students, teachers should, however, try to convey the idea to their learners that it is satisfactory if they are able to apply the three basic types at their stage of language learning.

4.2. Student – corpus interaction

After having dealt with several options of how teachers can integrate corpora into language teaching in the previous section, this part of the paper will now focus on a number of ways in which students can interact with corpora both inside as well as outside the language classroom (Kaltenböck & Mehlmauer-Larcher 2005: 78). The active involvement of learners with corpora is a brilliant tool in order to support students in their learning process. This claim is based on the fact that both in the complex process of learning a language as well as when dealing directly with corpora, students are constantly “engaged in hypothesis formation and hypothesis testing” (Johansson 2009: 37). In the process of creating, verifying or falsifying assumptions students function as researchers (Johansson 2009: 37), who are able to get involved with corpora in different ways, as Bernardini (2000: 139) points out:

> [I]n some cases corpora are accessed directly by the learners, in others learners only interpret the results obtained, and possibly sorted and thinned, by the teacher; in some cases learners work with general language corpora, in others with specialised corpora.

What needs to be taken into account, however, when students interact with corpora for the first time is that they cannot just be confronted with enormous amounts of language samples and asked to analyse them regarding specific language phenomena. This is due to the fact that corpora can only unfold their unrestrained potential to those who are “able to exploit them effectively” (Aston 1997a: 206-207). Therefore, students need to be introduced to corpora in a step-by-step approach in which theoretical explanations go hand-in-hand with practical exercises. Teachers function as guides, advisors and facilitators in this process, rather than as an authorities or language experts (Leech 1997: 8). Owing to the

teachers' guidance, students will progressively learn to use corpora as direct resources and will start "to problematize language, to explore texts, and to authenticate discourse" (Gavioli & Aston 2001: 244). The strategy of letting students approach corpora slowly, but gradually certainly assists learners in deepening their understanding of how language works, how it can be studied and how questions concerning language usage "can be usefully asked and answered by reference to a corpus of data", as Leech (1997: 9) notes.

As a result, both the students' language awareness as well as their language competence will most likely improve (Leech 1997: 8) and especially if students get feedback on guided corpus tasks, they will learn how to use this new instrument of analysing real language data to "explore corpora for their own purposes." (Leech 1997: 8) If used correctly and with the required proficiency, corpora can finally become "powerful learning resources" (Aston 1997a: 206), as their exploitation helps to enhance and focus

> the input to the student. They provide authentic data. They encourage reflection. They are well suited for consciousness-raising activities and for the training of inferencing. They stimulate the student to work actively and independently, and in this way they probably increase both the motivation of the student and the learning effect. (Johansson 2009: 38)

As can be seen from this description, "corpora seem able to significantly enrich the learning environment" (Aston 1997b: 63), as they do not only provide language learners with instances of naturally-occurring language, but also require them to use the target language and identify regularities in it themselves. Consequently, the idea of many people that real language data can primarily be used in the language classroom for students "as language models to either reproduce or imitate" (Gavioli 2005: 3) is a rather restricted one. People holding this opinion clearly fail to see that there is a lot of potential in "using corpora and interpreting the data interactively" (Gavioli 2005: 3), as students learn to reflect critically on language samples contained in corpora. As a result, learners become skilled in conducting their own corpus searches which are possibly even grounded in research questions of their own devising.

The following subsections will now deal with two different applications of corpora in the language classroom which are based on student-corpus interaction. After having had a closer look at the use of corpora as reference materials for students, the incorporation of electronic language databases for discovery learning activities will be addressed.

4.2.1. For reference

Whereas section 4.1.1. has already illustrated in which ways corpora can be used as reference materials by teachers, this sections now deals with the interaction between students and corpora for reference purposes. The potential of this application “was readily perceived”, as Aston (1997a: 207) remarks and is still a very prominent issue in specialist literature on the implementation of corpora in the language classroom until the present day (Braun, Kohn, Mukherjee 2006: 1). Since the basic facts regarding the use of corpora as reference tools have already been referred to in the section on teacher-corpus interaction, this part of the paper will try to focus especially on the learner’s perspective of the use of corpora for reference purposes.

Similar to the application of corpora described in section 4.1.1., students can use corpora as reference tools to look up naturally-occurring and thus authentic language samples. These can especially be of help to learners for clarifying “doubts on particular problems which had arisen in other tasks.” (Aston 1997a: 207) Put differently, the use of corpora for reference purposes by students is usually motivated by problems encountered in various communicative contexts (Aston 1997a: 209). Some of them may be solvable quite easily by using traditional monolingual or bilingual dictionaries, grammars or practical usage guides, while others may require corpus-based reference materials or corpora as such in order to be sorted out. This is due to the fact that in traditional dictionaries, for example, specific terms may only be provided together with their definition, however, without any further information on their actual use and thus typical phrases or collocations which the search term can be part of. This lack of information in traditional reference works is quite unfavourable for students (Mukherjee 2002: 161).

In order to exemplify this argumentation, we will now have a closer look at an example provided by Jessica Hölzl (2003: i-ii), who questions whether “the preposition *for* [... is] indeed idiomatic in combination with *relevance*.” While a consulted traditional dictionary did not provide the author with any information regarding which preposition goes best with the noun under consideration, Sinclair’s corpus-based ‘Collins COBUILD English Dictionary’ provided Hölzl with a quite detailed explanation. “According to COBUILD, *relevance* is an uncountable noun which is accompanied by supplementary information, and often occurs in the pattern ‘noun followed by the preposition *to* followed by another noun’.” (Hölzl 2003: ii) In cases like these, corpus-based reference materials as well as corpora per se are able to provide those searching with various instances of real language data including the search item. While relevant concordance lines taken from corpora still need to be analysed and interpreted by the searcher, corpus-based dictionaries, as can be seen above, may already provide ready-made answers for the encountered problem with regard to how a specific word or phrase is applied in practical language usage. Especially in situations like the one just described above, corpora or reference materials based on them can help students on more individual terms “by giving them access to a ‘native-speaker consultant’ who would be at their beck and call.” (Granath 2009: 48)

Corpora can be particularly useful for reference purposes outside the language classroom, for example, when writing essays for homework assignments or preparing oral presentations to be held in class. Inside the school building, however, corpora may be applied less often as reference materials by students. This is due to the fact that an application of this kind would not only require technical equipment in the form of computers in each classroom, but also enough time for students to look up specific language items before having to present them in class. Spontaneous communication and language tasks which are to be fulfilled at short notice, however, do not allow for long periods of preparation. As a result, students may want to use corpora for reference purposes outside rather than inside the language classroom.

In order to exemplify how language learners may apply corpora as reference materials both outside as well as inside the classroom an example will now be used to illustrate the potential of this kind of application. Sometimes, when writing an essay in class or as a homework assignment, students may have doubts how to combine certain words correctly. From time to time learners may, for example, be uncertain which prepositions can be combined with specific verbs (e.g. *to bargain*) or nouns (e.g. *information)* (cf. O'Keeffe, McCarthy & Carter 2007: 3). Problems like these, however, may not only be encountered by language learners at elementary or intermediate levels, but also by advanced language users. In the process of writing this thesis, for instance, I made regular use of Mark Davies's BYU-BNC (cf. http://corpus.byu.edu/bnc/). Once, for instance, I questioned myself whether to use the preposition *on* or *at* prior to the phrase *short notice.* As a consequence, I looked up both phrases in the above-mentioned corpus. While the phrase *on short notice*, my personal favourite, only occurred three times in the one-hundred million words corpus, *at short notice* had 141 hits. In case that the result would have been less clear, the concordance lines provided by the corpus would have given me the opportunity to analyse single language samples regarding the actual application of the search terms (O'Keeffe, McCarthy & Carter 2007: 3).

However, apart from the potential of corpus use for reference purposes by students both inside and outside the language classroom, an application of this kind can also be problematic. Students may draw wrong conclusions when analysing real language samples from corpora. They can, for instance, assume that the combination of several words in one or more concordance lines implies that these lexical items usually co-occur, instead of the fact that they may just sometimes co-occur in practical usage (Aston 1997a: 208). Moreover, students should be warned that the language usage mirrored in corpora does not necessarily have to be entirely correct. This is due to the fact, as already mentioned on several occasions in this paper, that the language data contained in corpora is taken from real-life situations. Therefore, samples included in corpora represent language how it is actually used and not how it should be used

according to prescriptive grammar books or other traditional reference materials and this is precisely where the value of corpora lies.

4.2.2. For discovery learning

This section will now focus on *discovery learning* in more detail. This approach to integrate students actively in the analysis of corpora is sometimes also referred to as *exploratory* or *data-driven learning (DDL)* in specialist literature. Data-driven learning, a concept introduced into the language teaching and learning context by Tim Johns, seems to be a more recent term for discovery learning (cf. Hunston 2002: 171). This assumption is based on the fact that both approaches to language teaching, namely discovery as well as data-driven learning, rest upon the same principles. Therefore, the terms *discovery learning*, *exploratory learning* and *data-driven learning* will be used interchangeably in the following paragraphs.

Discovery learning is a method in which real rather than invented language samples are directly integrated "in the classroom, by having students either analyse the corpus itself or examples from the corpus prepared by the teacher." (Nesselhauf 2004: 126) What is so special about this way of 'Computer Assisted Language Learning' (CALL) (O'Keeffe, McCarthy & Carter 2007: 24) is that students are not immediately taught how to apply specific language phenomena correctly in actual language use. In contrast, learners are only presented with evidence of naturally-occurring language use which consequently allows them "to make hypotheses and draw conclusions" (Hunston 2002: 184) about the application of particular language features in communicative situations themselves. Odlin (1994: 319-320) even describes discovery learning as "an approach to language teaching that gives central importance to developing the learner's ability to 'puzzle out' how the target language operates from examples of authentic usage". Thus, it can be argued that data-driven learning is quite similar to Ellis's (1994: 643-645) concept of *consciousness raising* "in the way that it allows the learner to explore language data and thus to derive patterns of language use." (Adolphs 2006: 109) Data-driven learning, as suggested by Johns (1991: 4), is mainly based on the idea that students analyse real language data in a three step approach. First students have to identify recurring language features

in the language samples under consideration, then these phenomena need to be classified and finally generalisations are to be found by the learners in order to create usage rules on the basis of practical exploitation (cf. Johns 1991: 4).

As can be deduced from this description, discovery learning is mainly based on a inductively-oriented rather than on a deductively-based approach towards language teaching (Hadley 2002: 108). In order to allow students to become more independent and interested in research it is necessary that

> [a] supportive, non-authoritarian environment is created: the teacher is not artificially setting up tasks requiring learners to provide information that she already has: rather, everyone in the classroom is actively trying to find the solution to a problem, discussing a solution proposed by one of the participants, guessing at the meaning of an expression, and so forth." (Bernardini 2004: 28)

Thus, pedagogues can basically be described as "director[s] and coordinator[s] of student-initiated research" (Johns 1991: 3) rather than as authorities regarding linguistic or pedagogical issues (Bernardini 2004: 21). Furthermore, they also function as guides, who do not explicitly tell students what they should do, but advise "them on how to pursue their own interests, suggesting alternative ways to proceed, other interpretations of the data or possible ways forward." (Bernardini 2002: 166) What makes data-driven learning such a unique approach is that it tries to omit teachers functioning as middlemen between the real language data and the learners "as far as possible and to give direct access to the data so that the learner can take part in building up his or her *own* profiles of meaning and uses." (Johns 1994: 297) Nevertheless, as Osborne (2000: 166) points out, teachers still need to assist their students in certain situations throughout the process of generating their own rules of application, as

> a certain amount of sign-posting will probably be necessary if they are to make fruitful connections between what they notice about the data, what they already know about the grammar, and what we might like them to develop in terms of overall understanding. (Osborne 2000: 166)

However, before learners are even able to analyse corpora with regard to specific language items, teachers first have to show them in which ways the data can be read and how “findings can be critically discussed and evaluated.” (Johansson 2007: 25) Students also need to be made aware of the fact that corpora do not provide them with instant answers to their research questions, but only with real language data which can be manipulated, analysed and interpreted for diverse purposes (cf. Gavioli 1997: 84). In this context it is important to note that corpus data never provide information about the “characteristics of ‘the language’ in general” (Gavioli 1997: 84), but only regarding the particular language, language variety, register or genre contained in the corpus under investigation. Consequently, a corpus simply remains a passive and silent informant presenting students with real language samples, as Johns (1991: 1) notes. This does not imply, however, that learners are able to acquire language features directly from corpora, as language data can only be processed efficiently by students if they are shown how to interact with the presented language samples by their teachers. Therefore, language data included in corpora cannot be

> expected to provide right answers, and often does not, but constantly presents new challenges and stimulates new questions, renewing the user’s curiosity and offering ample opportunity for researching aspects of language and culture, which may easily become a subject for research projects, reports and discussion. (Bernardini 2002: 166)

Put differently, corpora “simply provide the evidence needed to answer the learner's questions, and rely on the learner's intelligence to find answers.” (Johns 1991: 2) Therefore, activities which support the notion of discovery learning are “designed to favour learner-centred, open-ended, tailored learning” (Bernardini 2004: 27) so as to enable learners “to explore, to investigate, to generalize [and ...] to test hypotheses” (Leech 1997: 5). As can be seen from this description, exploratory learning tasks do not attempt to “initiate or direct the path of learning” (Leech 1997: 5). They are themselves only a means by which students are encouraged “to become more autonomous in their studies, taking responsibility for their own learning” (Bernardini 2004: 27), which may consequently lead to an increase in their motivation to learn the target language (Adolphs 2006: 109).

The role of the learners working on discovery learning tasks is often described as that of language researchers (Adolphs 2006: 109). Bernardini (2000: 143), however, does not appreciate this comparison of students and researchers very much. She argues that it would be better to refer to students carrying out discovery learning tasks as ‘travellers’. Bernardini (2000: 143) argues her choice by noting that students do not only have to develop strategies and test hypotheses when analysing real language data, but also need to plan their encounters and reflect on them afterwards (cf. Bernardini 2000: 143). “Besides, the researcher metaphor does not seem to do justice to the entertaining side of corpus browsing”, as Bernardini (2000: 143) remarks. Furthermore, when regarding students as travellers, “it is the journey itself that matters, not its endpoint: the traveller metaphor views learning experiences as rituals of change, which imply the learner’s displacement to a more advanced level of competence, capacity and awareness.” (Bernardini 2000: 143) Tim Johns (2002: 108) does not go for the researcher or the traveller metaphor when discussing data-driven learning activities in his articles, but refers to his learners as detectives in the sense of Sherlock Holmes. Johns (2002: 108) explains the choice of this metaphor by stating that discovery learning exercises demand students “to recognise and draw conclusions from clues in the data”, just as detectives have to do when solving a murder or other mystery.

This process just described in the last quotation is referred to as *inductive learning* by Tim Johns (1991: 3), which is based on three steps, namely “observation, classification and generalization” (Aston 2001: 19). Data-driven learning “provides opportunities for the learner to develop inductive discovery strategies, or to ‘learn how to learn’ along with the opportunities to increase her/his own competence”, as Bernardini (2000: 135) notes. Furthermore, discovery learning activities being grounded in inductive learning strategies seem to be “particularly effective for the acquisition of grammar and vocabulary [...], leading learners to notice patterns in the input [...], stimulating deeper processing [...], and improving subsequent retention”, as Guy Aston (2001: 19) argues. Activities based on data-driven learning may, for example, require students to make generalisations about

> whether an observed pattern is used more in speech than in writing or vice versa, whether it can be said to belong to a more formal or rather to a more informal kind of register [... or] whether any specialised use is discernible [...]. (Bernardini 2002: 168)

Generalisations and rules derived from discovery learning tasks often help students to understand the application of specific language features better than if they are only confronted with descriptions provided by either teachers or textbooks (Aston 2001: 19).

However, despite all these reasons arguing for the use of discovery learning activities in the language classroom, only few teachers do so (Götz & Mukherjee 2006: 49). One cause for this may be that teachers do not want to spend time on the quite laborious creation of exploratory learning activities. A possible solution for this problem may be that all kinds of discovery learning exercises on various issues suitable for students of different levels of proficiency are created which are then accessible via platforms or websites on the internet, as Ute Römer (2009: 91) suggests. A response to this request can be found in Mukherjee (2006: 14), who points out that

> an impressive array of DDL activities are [already] on offer, including both more teacher-guided and more learner-autonomous methods, many of which are also used in freely accessible DDL websites (cf. T. Cobb's *Compleat Lexical Tutor* and T. Sripicharn's *DDL Materials*).

According to Götz and Mukherjee (2006: 49), however, data-driven learning does not only support inductive learning, as it "also enhances deductive learning". This is due to the fact that learners do not always approach corpora for discovery learning activities without having any idea in mind as to what they are actually looking for. Students often "apply previously-acquired generalizations in order to classify concordance data, testing the 'rules' they have learned, and thereby consolidating and/or refining their knowledge." (Aston 2001: 19-21)

As already mentioned when explaining the principle underlying discovery learning above, there is not only one way of applying discovery learning in language teaching, as these can vary depending on the language proficiency and level of

autonomy of the students. Therefore, the application of exploratory learning in the language classroom can

> range from very controlled learning tasks, where the teacher may select a few specific concordance lines that illustrate a grammatical point or usage of a particular lexical item, to more complex tasks for advanced learners. The latter might consist of a task where the learners are asked to consult raw corpus data in order to determine the use and associated patterns of a set of lexical or grammatical items, or it might involve the learner setting up their own line of linguistic enquiry and using raw corpus data to address this task. (Adolphs 2006: 109)

This description already illustrates that corpora cannot only be used in their raw form, but also in the shape of teacher-made materials. Learners can, for instance, be provided with discovery learning exercises "based on concordance lines which the teacher has selected" (Hunston 2002: 177) which are accompanied by questions guiding students towards relevant pieces of information contained in these samples. While the results to be gained from raw corpora can hardly be predicted by teachers prior to their students' analysis of the data, the selection of particular concordance lines for discovery learning activities allows teachers to determine the analysis outcome considerably (Hunston 2002: 170-171). The use of selected instances of naturally-occurring language data is especially advantageous for language learners "who are not advanced enough to benefit from 'raw' concordance data" (Hunston 2002: 177), as language samples containing difficult or uncommon expressions may just be eliminated from the material by teachers in advance.

In one of his papers Tim Johns (1994: 298) describes the way how data-driven learning materials can be created in more detail. First of all, teachers have to decide on the language feature they want their students to study and search raw corpora for all instances of this phenomenon. As a next step, teachers need to select suitable examples for the language classroom. When conducting this procedure,

> [t]he most important principle that has to be borne in mind [...] is that the inevitable process of selection should not distort the evidence – that is to say, the concordance extracts chosen should represent as far as

> possible the full range of linguistic and communicative features of the raw data. (Johns 194: 298)

Thus, teachers should not only select those samples of a specific language phenomenon which are "in themselves [...] perfectly justifiable (for example, that preference should be given to citations that are relatively self-contained and self-explanatory)" (Johns 1994: 298), as this would certainly bias the outcome of the analysis conducted later on by the students.

The following example of a discovery learning task is taken from Hunston (2002: 172) and is based on the fact that a student wrote "the following phrase in an essay: ... *in their efforts to prevent such incidents to ever happen again.*" The teacher who did not approve of this use of the verb *prevent* decided to give the student the opportunity to discover "the conventional usage of this verb" (Hunston 2002: 172) independently on the basis of real language data. For this purpose the student was sent to a corpus with the task to analyse a particular number of randomly chosen concordance lines. "What the teacher 'wants' the learner to see in these lines are the presence of the patterns *prevent something, prevent something happening* and *prevent something from happening* and the absence of the pattern *prevent something to happen.*" (Hunston 2002: 173) As can be seen from this example, discovery learning activities cannot only be conducted with all students of a class at the same time, but can also be given to single students for personal investigation. However, if data-driven learning tasks are set, for example, as homework tasks for students which are to be conducted individually, it is necessary that learners are already familiar with the process they have to undergo when analysing real language data in order to come to a satisfactory conclusion. Furthermore, teachers need to be aware of their students' level of competence when confronting them with tasks like the one just described above.

Teachers' decisions concerning the level of difficulty of discovery learning activities "require a good deal of judgement and sensitivity with regard to the level of the learner, the corpus data and the benefit which may be derived from the activity." (Adolphs 2006: 109) Even though exploratory corpus activities can already be introduced to language learners who are just beginning to study a

specific language as soon as they are able to appreciate them (Hadley 2002: 119), discovery learning “is most suitable for very advanced learners who are filling gaps in their knowledge rather than laying down the foundations.” (Hunston 2002: 171)

The general idea underlying the application of exploratory learning activities in the classroom is that they assist learners’ in improving their language awareness and competence as “effective language learning is itself a form of linguistic research” (Johns 1994: 297). The question of “what is to be taught and learned” (Johns 2002: 110) by means of data-driven learning, however, is still an open one. This is due to the fact that the aim underlying exploratory learning activities can either be determined by the teacher who preselects materials or by the learners browsing through corpora in search of any prominent language features.

The fact that teachers do not necessarily have to know “in advance exactly what rules or patterns the learners will discover” (Johns 1991: 3) is quite a distinctive feature of discovery learning. Sometimes teachers may not even able to predict analysis results beforehand, as things might be discovered “which have probably never been brought to notice before, even in the most detailed dictionaries and grammars of a language.” (Leech 1997: 3) Bernardini (2004: 28) even goes as far as claiming that the teachers’ limited knowledge regarding the practical usage of certain language features may even “facilitate a process of democratization of the learning setting”. This is due to the fact that “[t]he ‘puzzle’ that the data create puts teachers and learners on essentially ‘the same ground’ ” (Gavioli 2005: 128). This situation then allows teachers and students to negotiate the analysis results together and exactly this process of negotiation is what Peter Voller (1997: 109) regards as a central aspect in the students’ development of becoming more autonomous. Moreover, it also benefits the development of the students’ “metalinguistic and metacognitive awareness” (Aston 2001: 23).

The fact that neither teachers nor students may be able to say which outcome can be expected from a corpus analysis is exactly what makes this way of approaching real language data in the classroom so motivating and interesting for

students (Hunston 2002: 170). As Johns (1991: 3) notes, it is exactly “this element of challenge and of discovery that gives DDL its special flavour and stimulus.” Sometimes students “may not […] be able to reach indisputable conclusions from the available evidence” (Bernardini 2000: 136) at all. Nevertheless, even in these instances students “can exercise their inferring and generalising skills” (Bernardini 2000: 136). These competencies and inductive learning strategies can also be of help to language learners outside the school building in the context of life-long learning (Johns 1994: 297).

On a more theoretical basis it can be said that discovery learning is a way in which to approach real language data which is process-oriented rather than product-oriented. This is due to the fact that it is not only the outcome of the analysis which counts, but the process of deducing rules of application from actual language usage (cf. Bernardini 2000: 136). Nevertheless, data-driven learning “also draws from product teaching in that it provides authentic material for study.” (Hadley 2002: 107) Consequently, it allows students to internalise discovered regularities in language usage in a way which is more effective than when just providing students with fewer or invented language examples for analysis (Mukherjee 2002: 67-68). Another advantage of using discovery learning activities based on corpora in the language classroom is that students may notice that rules of practical language usage may sometimes be more complex than illustrated in course books (Mukherjee 2002: 67-68).

However, using discovery learning activities in language teaching can also be regarded from a more critical perspective. The danger underlying this kind of teaching method pointed out most frequently in specialist literature is that students might feel overwhelmed by the huge amount and complexity of real language data they are confronted with and thus feel left out in the rain by their teachers (Gavioli 1997: 83-84). However, it may not only be the students who experience a complete lack of power to act, as teachers may share the same feelings. Teachers may, for instance, “feel that a loss of expertise has occurred” (Hunston 2002: 171), when difficult questions are raised which they cannot answer immediately without the consultation of reference materials. Moreover, teachers

may find it difficult to make their students see more in corpus data than just instances of naturally-occurring language samples (Gavioli 2005: 127). Another critical point with regard to data-driven learning is that they "rely heavily on the learners' unflagging curiosity and interest." (Bernardini 2002: 167). Unfortunately, many students lose interest in working with corpora "after the initial enthusiasm" (Bernardini 2002: 167) and favour using more traditional reference materials, like, for instance dictionaries or grammar books. "A more basic problem is that not every teaching situation allows the luxury of one-to-one consultations, or sufficient computer access for students to undertake investigations on their own", as Hunston (2002: 171) points out. Consequently, students may have to work together which, however, also restricts their scope of action, as all steps to be done during the process of analysis need to be coordinated between the partners. As an alternative, teachers may print out the language samples required for discovery learning activities, however, this clearly limits the students' autonomy in approaching the data under consideration (cf. Hunston 2002: 171).

Besides Tim Johns' concept of data-driven learning, which regards "*learning* as *research*" (Bernardini 2004: 22), discovery learning can also take place in an even more autonomous form, called *serendipitous corpus browsing.* This concept created by Silvia Bernardini (2004: 22) considers learning as a process of discovery and thus as "an approach to learning from corpora in which learners are guided to browse large and varied text collection in open-ended, exploratory ways". When browsing corpora in this way, students do not to have any particular research question in mind, "but are expected to note any form and structure that they may find interesting, to analyse the form and structure at hand and to move from here to other interesting forms and structures." (Mukherjee 2006: 14) By exploring one language phenomenon after the other and thus following their personal interests, language learners are provided with various "opportunities to develop their capacities and competences so that their searches become better focused, their interpretation of results more precise, their understanding of corpus use and their language awareness sharper." (Bernardini 2004: 23) Nevertheless, serendipitous corpus browsing is often regarded critically both by teachers as well as by experts, as the linguistic search conducted by students in this approach to

language teaching can neither be predetermined by teachers nor does it allow for any common outcome which can later be discussed jointly in the classroom. Furthermore, with all students working on different language samples and thus discovering varying language features it is no longer possible for teachers to supervise all their students at the same time (cf. Mukherjee 2006: 14). Moreover, teachers may quickly lose track of their students' projects, just as they will no longer be able to keep them at an approximately equal level of competence.

As already mentioned in the course of this section, discovery learning activities can be based both on raw corpora as well as on selected concordance lines picked out of electronic language databases by teachers. However, exploratory learning tasks can also be grounded on a third type of language material. Especially "[i]n recent years it has been suggested that it may be both useful and motivating for teachers and learners to construct their own corpora to analyse with appropriate interrogation software." (Aston 2002: 9) When basing discovery learning activities on self-made corpora, naturally learners also need to be trained in compiling representative databases of language as well as using them properly (Aston 2000: 14). The huge advantage of self-compiled corpora is that all real language samples to be added can be selected based on certain criteria, for example, only texts belonging to a specific genre can be taken into account. Consequently, the texts to be included in a corpus "can be specifically targeted to the learner's knowledge and concerns" (Aston 2002: 9) and may thus suit already formulated or potential research questions best. This is especially important as not all types of corpora, like, for instance, general ones, are suitable to answer quite specific questions. The process of selecting and compiling a corpus may also help learners to acquire "skills and knowledge which may be of value to them in the future." (Aston 2002: 10) Nevertheless, even though the construction of home-made corpora may have numerous advantages, "the costs of corpus construction" need to be balanced, as Guy Aston (2002: 10) emphasises. It certainly takes longer to compile a whole corpus than to use an already existing one and furthermore, professionally designed corpora are of higher quality and thus more representative than self-made ones (Aston 2002: 10).

Nevertheless, in specific circumstances home-made corpora are more useful than ready-compiled ones, as they represent, for example, a specific genre, register or dialect. One approach to language teaching which appreciates self-made corpora very much is the so-called *genre-based approach* (Mukherjee 2006: 17). “The overall aim of a genre approach to language teaching is to make learners aware of the relationship between the communicative purpose of a genre, the context, and language chosen to achieve the purpose”, as Alex Henry and Robert Roseberry (2001: 94) notice. Even though a great number of corpora exist on all different kinds of genre, not all existing text types are covered already. A detailed description of how the genre-based approach to language teaching works can be found in Mukherjee (2006: 14-15), who explains the genre approach to language teaching developed by Henry and Roseberry (2001). Their approach consists of two steps with the first being quite analytical, while the second one is more productive:

> (1) In the first step, corpus texts of a particular genre are analysed by the learners with regard to the basic textual moves that are typical of the genre (e.g. scientific papers). For each of the typical moves (e.g. the conclusion), learners then find out which linguistic patterns are preferred for the verbalisation of the textual move (e.g. *to conclude..., in conclusion,...)*. (2) In the second step, learners write new texts of the genre at hand, albeit with a different thematic focus, by sticking to the overall move structure and by using the preferred linguistic patterns. (Mukherjee 2006: 15)

As can be seen from this example of how corpora can be integrated into the field of language teaching and learning in an exploratory way, they are not used in order to find proof for the particular use of a language item. In contrast, discovery learning activities are mainly applied in the language classroom to give students the opportunity to deduce rules of actual application from naturally-occurring language samples comprised either in already established or in self-made corpora.

5. Potential and limitations of corpora in language teaching

The topic of how to integrate corpora into the field of language teaching is frequently examined in specialist literature, as the preceding sections of this paper

have tried to illustrate. However, “[w]hen such classroom uses are being discussed, it is often in highly positive and uncritical terms, with computer corpora being presented as the new revolution in language teaching”, as Kaltenböck and Mehlmauer-Larcher (2005: 66) point out. Therefore, it is especially important for practising teachers to be aware of the challenges coming along with the application of corpora in the language classroom, which can be of “technical, linguistic, logistic, pedagogical and philosophical” (Johns 2002: 107) nature. Kaltenböck and Mehlmauer-Larcher (2005: 66) even go as far as claiming that “[k]nowing about these limitations [...] is as important for successful integration of computer corpora into language teaching as knowing about their assets and potential.” Unfortunately, especially the limitations are often only mentioned in the literature while other points of relevance regarding the implementation of corpora into language pedagogy are being discussed.

Advantages and restrictions regarding the application of corpora in the language classroom cannot only be found in a teaching context, but are also intrinsic to corpus use in general. Apart from numerous advantages offered by corpora, which have already been discussed in more depth elsewhere in this paper, like, for example, that they “provide information not easily available from other sources” (Kaltenböck & Mehlmauer-Larcher 2005: 81), they also have some limitations. One of these is, for example, that corpora can “only record what people have said (externalized language), not what they can say (internalized language)”, as Kaltenböck and Mehlmauer-Larcher (2005: 81) stress. Furthermore, computer software can only be used proficiently by those who are able to manipulate data according to their own needs and purposes as well as to interpret the gained language data correctly (cf. Kaltenböck & Mehlmauer-Larcher 2005: 81-82).

The two most important issues regarding the potential and limitations of corpora, however, concern the authenticity (see section 5.1.) as well as the representativity (see section 5.2.) of the language data contained in electronic language databases (cf. Kaltenböck & Mehlmauer-Larcher 2005: 81-82). The named subsection will thus focus on these two points in more detail, as they are most

frequently and controversially discussed in the subject literature and have not yet been paid sufficient attention in the course of this paper.

The subsequent sections will then focus on the availability of both computers and corpora in educational settings (see section 5.3.), before some financial aspects of the integration of direct corpus exploitation into the language classroom will be considered (see section 5.4.). Afterwards, the skills required both from teachers as well as from students in order to work with and analyse corpora appropriately will be dealt with (see section 5.5.). Finally, the last part of this section will address the issue of learner autonomy (see section 5.6.) when exploiting corpora in the language classroom.

5.1. Authenticity

Corpora have certain advantages as well as restrictions when it comes to the language data they comprise. As already mentioned when defining the term corpora, electronic language databases contain naturally-occurring language samples, which are frequently referred to as authentic instances of language use in specialist literature. Whether this description is really appropriate or not will be discussed in more detail in the following paragraphs.

One of the most frequently debated issues regarding the authenticity of language data contained in corpora concerns the way in which the naturally-occurring language samples are displayed to the researcher. “Language items retrieved from corpora are not isolated linguistic items” (Kaltenböck & Mehlmauer-Larcher 2005: 81), they are presented together with their co-text, which can also be described as linguistic context (Yule 2004: 129). The co-text of a particular search item consists of “the words, grammatical constructions, phonological or discourse features” (Anderson & Corbett 2009: 155-156) in the immediate surroundings of the search term. However, while providing analysts with the immediate verbal surroundings of the chosen search terms, corpora do not supply their users with information regarding the context in which particular words or phrases have actually been applied. The context, also referred to as physical context by Yule (2004: 129) is the “non-linguistic environment of any language activity” (Sinclair

1991: 171). It is especially relevant in communication, as it provides the researcher with information concerning the underlying communicative purpose, as Henry Widdowson (2007: 5) notes.

While the co-text presented along with specific search terms in corpora "has a strong effect on what we think the word means" (Yule 2004: 129), only the context, "e.g. the situation paralinguistic communication, cultural knowledge, other texts, or other parts of the same text" (Cook 2003: 126) can provide the analyst with a clear idea concerning an appropriate interpretation of the language sample under consideration. Consequently, only if both co-text and context of a specific lexical item are given analysts become able to assess "the totality of the features of its background which may affect the language we use." (Anderson & Corbett 2009: 158)

"Context, however, is precisely what is missing in a corpus", as Kaltenböck and Mehlmauer-Larcher (2005: 69) emphasise, since text passages included in corpora are "extracted from the context of larger communicative units and presented in detachment" (Widdowson 1996b: 79-80). This is particularly unfortunate when it comes to the implementation of corpora in the language classroom, as the missing situational and communicative context (cf. Kaltenböck & Mehlmauer-Larcher 2005: 69) reduces the naturalness of the language data contained in electronic language databases. Consequently, texts comprised in corpora can only be referred to as containing decontextualised language samples, which are merely "real because of the presupposed reality of the discourses of which they are a trace", as Widdowson (2000: 7) points out. As a consequence, texts need to be "distinguished from *discourse*, which is a meaningful unit including the implicit meanings that a reader/hearer has established in a particular communicative situation, on the basis of the textual clues along with linguistic and (cultural) background knowledge" (Braun 2006: 28). Partington (1998: 145) even goes as far as claiming that language data contained in corpora are "as decontextualised as any linguistic information could possible be and therefore cannot count as communication."

At this point it is necessary to mention an often quoted and popular distinction made by Widdowson between *authentic* and *genuine* instances of language use. When providing learners with language data taken from real-life communication, they are presented with genuine language samples (Widdowson 2003: 93). These genuine samples, however, need not necessarily be authentic as well. While genuineness "is a characteristic of the passage itself and is an absolute quality [..., a]uthenticity is a characteristic of the relationship between the passage and the reader and it has to do with appropriate response." (Widdowson 1996b: 80) Therefore, authenticity "depends on a congruence of the language producer's intentions and language receiver's interpretation" with regard to the meaning underlying a certain language samples. This meaning can only be discovered when the speaker and the analyst share a common ground regarding their "knowledge of conventions" (Widdowson 1980: 166). Thus, the prevailing problem when introducing students simply to extracts of naturally-occurring language samples is that the authenticity of these samples as such is quite low (Widdowson 1996b: 80). As a solution, Widdowson (1996b: 80) thus suggests that in order to enable students "to acquire communicative abilities [..., they] must ultimately be induced to treat reading passages as discourse, to adopt the same attitude to them as [... they] would to written discourse in [... their] own language."

As a result, naturally-occurring language samples comprised in corpora need to be interpreted by both teachers and students in order to authenticate and contextualise them (Gavioli & Aston 2001: 244). In this process of re-contextualisation language instances taken from corpora "can only be made real as discourse to the extent that [... they] can be appropriately related to context." (Widdowson 2003: 104) The authentication of language samples taken from corpora is particularly important, as it allows students to interact with naturally-occurring language data, which helps them to make "language patterns (recurrent or 'deviant' ones) real or memorable to them." (Gavioli 2005: 130) This process, however, cannot only take place when students interact with language data in a direct way, as an interaction between students and teachers can also assist learners to authenticate language samples, as Gavioli (2005: 130) points out. This is due to the fact that the active discussion of language data in the classroom

helps to make these instances more authentic by becoming real-life issues when being reviewed. This, however, does not necessarily mean that all students contextualise specific language samples in the same way, as individual learners "'authenticate', or give authenticity to a text from [... their] own state of knowledge and frame of reference." (Breen 1985: 64) In order to be able to do so, though, teachers need to introduce their students to "a text-based exploration of the corpus content, focussing on the wider social and cultural context of the materials." (Braun 2006: 29) In this process of authentication, which is based on the personal interpretation of texts by the students, "we may regard texts as potential *means* for the learner towards authentic communication in the target language." (Breen 1985: 64)

Even though Widdowson's concept regarding the genuineness and authenticity of language data is a frequently discussed one, corpora are still often referred to as containing 'real', 'authentic' or 'contextualised' language samples. According to Sinclair (1991: 5), however, "no example is ever complete unless it is a whole text" and thus neither language samples taken from corpora, nor invented examples can be regarded as being authentic. While real language samples can be interpreted based on their natural situational and communicative contexts, which are, however, often not immediately tangible, invented language examples "appeal for their authenticity to a non-existent context, which would eventually be evaluated by someone's intuition, with all the misleading consequences of that" (Sinclair 1991: 5). In some cases it may even be easier for students to contextualise unreal language examples rather than real language samples (Widdowson 2003: 104). This is due to the fact that even though invented language instances are "unlikely to be reproduced in actual contexts of use" (Carter 1998: 47), learners are sometimes able to understand them more easily than naturally-occurring samples. Invented examples also present the linguistic patterns they illustrate more directly (Gavioli & Aston 2001: 241) than corpus samples, which may consequently be harder to comprehend and contextualise for students and may thus be regarded as unrealistic by them (Carter 1998: 50).

As authentic language data do not exist as such, as the preceding paragraphs have attempted to illustrate, it is the task of the language teacher to design “a methodology which will establish the conditions whereby this authenticity can ultimately be achieved.” (Widdowson 1980: 172) Therefore, teachers have to fulfill several tasks in the language classroom when it comes to establishing authenticity, as Michael Breen (1985: 61) notes. Pedagogues, for instance, need to keep an eye on the ‘authenticity’ of the texts used in the language classroom as well as on the ‘authenticity’ of their “learners’ own interpretations of such texts” (Breen 1985: 61). In order to enable students to authenticate concordances, it is the teacher’s task to provide them with “pedagogically appropriate information about the text and the communicative situation in which it was produced.” (Braun 2006: 28)

In conclusion, it can thus be said that regarding the potential and limitations of corpora in the language classroom, the issue of authenticity can be classified as belonging to both categories. The fact that corpus data is not authentic, but ‘merely’ genuine may at first glance appear to be a limitation. However, exactly this characteristic of corpus data allows both teachers as well as students to reprocess language samples taken from corpora and thus to deal with them in more depth.

5.2. Representativity

A further issue which is frequently discussed in connection with the potential and limitations of corpora is their representativity. As already mentioned several times throughout this paper, the investigation of data contained in computer corpora allows researchers to gain manifold pieces of information regarding actual language use. Researchers, however, “should not accept corpus evidence uncritically, but should appraise it in the light of other sources of information about language such as introspection and elicitation”, as Susan Hunston (2002: 193) emphasises. In other terms, any outcomes of corpus analyses should never be treated as ultimate truths, as the following remarks attempt to illustrate.

The first point which needs to be made with regard to the representativeness of corpora concerns the fact that a “corpus, no matter how large and varied, is only ever representative of itself” (Partington 1998: 146) and thus not of language in general, as Gavioli (2005: 18) points out. Put differently, corpora “are neither infallible nor omnipotent” (Partington 1998: 146), as all regularities derived from a particular language database “hold true only for the portion of language contained in that corpus.” (Partington 1998: 146) Therefore, if a certain item or construction cannot be found in a specific corpus, this does not necessarily have to imply that it does not occur in actual language use at all, as corpora only register those utterances actually made in real-life situations, while “they cannot tell us what is possible or not possible.” (McEnery, Xiao & Tono 2006: 121) Another reason why specific phrases cannot be found in a particular corpus may be that the electronic language database under investigation contains language data which is just too specific, for example, with regard to the text types or registers included (cf. Partington 1998: 146). Consequently, it is important to keep in mind that “descriptive facts such as frequency findings cannot automatically be taken as basis for pedagogic prescription” (Kaltenböck & Mehlmauer-Larcher 2005: 77), since frequency, according to Widdowson (2003: 87), says nothing about the prototypicality of a language feature.

Nevertheless, and this is what Widdowson (2007: 78) clearly points out as one potential of the representativity of corpora, electronic language databases are able to inform us about the distribution of particular lexical items or grammatical constructions “in different domains of use”. Moreover, a corpus also allows researchers access to vast amounts

> of quantitative information [...] about the frequency of linguistic tokens [... and] the recurrence and co-occurrence of words. It provides a detailed profile of what people do with the language. It is a fascinating revelation and its importance for linguistic description can hardly be exaggerated. (Widdowson 2003: 80-81)

Still, “frequency should not be the only factor in deciding what to teach” in the language classroom, as Hunston (2002: 193) notes, since both saliency as well as relevance of particular language features should be considered as well. In other

words, teachers should not believe that quantity can be equated with relevance or communicative value when it comes to integrating language features taken from corpora into the field of language teaching and learning.

The second point to be made within this section on representativity of corpora regards the issue of generalisability. As corpora can only "yield findings but rarely provide explanations for what is observed" (McEnery, Xiao & Tono 2006: 121), it is the task of the analyst to come up with explanations concerning the questions why and how particular language phenomena occur in naturally-occurring language use. An often unseen danger in this process of deducing usage guidelines from corpora is that generalisations may be based on only small numbers of language samples. According to Sinclair (1991: 27), this may lead to the fact that in the majority of cases derived rules are oversimplified and influenced by the researcher's intuition. In other terms,

> it is important to keep in mind that the findings based on a particular corpus only tell us what is true in that corpus, though a representative corpus allows us to make reasonable generalizations about the population from which the corpus was sampled. Nevertheless, unwarranted generalizations can be misleading. (McEnery, Xiao & Tono 2006: 121)

5.3. Availability of computers and corpora

An issue which is of major importance with regard to the implementation of computer corpora into the field of language teaching concerns the availability of both computers as well as corpora in the educational setting. In this context a number of requirements need to be considered by teachers when planning to use corpora in the classroom, as the following remarks will try to illustrate.

Nowadays almost every school provides both teachers as well as students with modern technological devices like computers, which can be used more or less freely inside the school building. The ever-increasing number of computers available for both learners and teachers inside as well as outside the language classroom can mainly be based on the facts that they have become smaller, cheaper and better with regard to their storage capacity. However, the fact that

computers have become more easily affordable does not necessarily have to imply that they are also more easily available in schools. In some schools each classroom may only be equipped with one or two computers, while in others computer rooms or labs may be available which have to be shared equally among all classes. Only in rare instances portable computers may be provided in class strength.

Together with the improvement of computers, electronically stored corpora have gained importance in linguistics in general as well as in applied linguistics and language pedagogy more specifically, as Leech (1997: 2) points out:

> Inevitably, while computers were limited to large mainframes available to the initiated few, computer corpora were largely restricted to research use. But, as computers have grown smaller, cheaper, and massively more powerful, their use in teaching has grown immeasurably.

Hunston (2002: 1) even goes as far as claiming that the "improved accessibility of computers has changed corpus study from a subject for specialists only to something that is open to all." But is this really true? Is everyone able to access corpora easily without being confronted with restrictions?

Leech and Fligelstone (1992: 121) tried to answer this last question already approximately twenty years ago. Unfortunately, their response still had to be a negative one: "Regrettably, there is a huge gulf between the amount of computer corpus material in existence, and the amount which is available for use in any realistic sense." The fact that not all corpora were simply available to everybody interested at that time was mainly due to "copyright law and the law of confidentiality [... which] effectively bar[red] the use of the vast majority of corpora from the vast majority of potential users." (Leech & Fligelstone 1992: 121) While researchers might have been able to use certain corpora with the help of password systems or by paying annual fees (cf. Anderson & Corbett 2009: 9), schools did neither have the authority nor the money to do the same. However, this was mainly the state of the art until approximately ten years ago.

The good news for teachers is that nowadays “more and more corpora can be accessed immediately and freely on the Internet.” (Anderson & Corbett 2009: 9) These corpora are consequently immediately-usable in the language classroom, provided that teachers know where to find them, know how to use them and “know how to maximise the potential of language corpora, with all their idiosyncrasies and differences.” (Anderson & Corbett 2009: 2) With user-friendly software and interfaces being available in order to analyse these corpora (Sinclair 2004: 2), “it is no longer impracticable to use, and teach learners to use” electronic language databases, as Bernardini (2000: 120) notes. Since the convenient availability and accessibility of corpora via the internet is still a quite recent advancement, however, it is rarely mentioned in publications which are older than five years.

Numerous corpora which are publicly and also freely available are listed in Wendy Anderson’s and John Corbett’s (2009: 183-187) book called ‘Exploring English with Online Corpora’. All of the following electronic language databases can, for example, be accessed online:

- the ‘British National Corpus’ (BNC)
  (cf. http://corpus.byu.edu/bnc/)
- the ‘TIME Corpus of American English’
  (cf. http://corpus.byu.edu/time/)
- the ‘Michigan Corpus of Academic Spoken English’ (MICASE)
  (cf. http:// quod.lib.umich.edu/m/micase/)
- ‘The Corpus of Contemporary American English’ (COCA)
  (cf. http://www. Americancorpus. org)
- ‘The Corpus of Historical American English’ (COHA)
  (cf. http://corpus.byu. edu/coha).

No matter which type of corpus teachers may want to use in their classrooms (e.g. corpora of spoken or written language, corpora of contemporary, historical, British or American English), numerous professionally designed electronic language databases can nowadays be found in full size on the internet. This development,

however, is not only advantageous for teachers planning to introduce corpora into the area of language teaching, but also for students, as it allows them to access and search language databases individually whenever they have questions regarding the actual use of specific language features. Being a frequent user of online corpora myself, I am sure that both the easier availability as well as accessibility of corpora containing between one-hundred million (e.g. BNC, TIME corpus) and four-hundred million words (e.g. COHA, COCA) open up new possibilities for the implementation of computer corpora into the field of language pedagogy. While the limited access to corpora may once have been a major limitation for applying corpora in language teaching, it is my own opinion that their easier availability certainly has a positive influence on their application by both teachers as well as students being familiar with their use and potential nowadays. This is due to the fact that neither teachers nor learners have to rely on electronic language databases and retrieval programmes which first need to be bought and installed on computers any longer, but are able to use online corpora for various purposes inside, but also outside the language classroom as soon as they have access to the World Wide Web.

5.4. Financial aspects

Another issue which needs to be dealt with regarding the potential and limitations of introducing corpora into the field of language teaching and learning concerns the financial aspects. In this context, unfortunately, the restrictions of corpus use in the language classroom may sometimes prevail, as computers, retrieval software and at times even special corpora as such may be quite expensive resources (Leech 1997: 23). One of the most basic problems when planning to integrate corpora into the language classroom is that “[u]nfortunately in many countries, schools do not have the money to purchase the software and equipment needed” (Hadley 2002: 110).

Apart from the acquisition of the compulsory hardware, however, expenses may also need to be calculated for the purchase of computer programmes, including corpora and retrieval programmes – at least if special software is needed. As already mentioned in the preceding section on the availability of computer and

corpora (see section 5.2.1.), numerous electronic language databases can nowadays already be easily accessed for free via the internet. Some of them, as, for instance, the 'British National Corpus' (cf. http://corpus.byu.edu/bnc) or the 'TIME corpus of American English' (cf. http://corpus.byu.edu/time) are even equipped with user-friendly interfaces, which offer "the facility of identifying collocates, comparing words across registers, and viewing all hits for search terms in the corpus." (Anderson & Corbett 2009: 183-184) Even if special analysis programmes are required which facilitate the work with self-made corpora, teachers can make a find on the internet, where, for instance, the 'AntConc software' can be downloaded for free. Therefore, at least with regard to corpora and retrieval programmes money can be saved by teachers and schools, except if teachers want their students to work with quite specific corpora which are not publicly available (cf. Granath 2009: 55).

Regrettably, as Hunston (2002: 171) remarks with regard to the use of computer corpora in the language classroom, "not every teaching situation allows the luxury of [...] sufficient computer access for students to undertake investigations on their own." Teachers should thus always think of alternatives of how to include corpora in their language classrooms if not enough or even no computers are available. As already suggested elsewhere, if only a limited number of computers are at hand, students may be asked to work with partners and thus to share computers. If no computers are available for direct exploitation of corpora at all, concordance lines can, for example, "be printed on to paper to be used with a whole class." (Hunston 2002: 171) However, the disadvantage of this application is that students will be quite restricted in their analysis of the material provided, as the topic of the investigation and thus to some extent also the outcome have been predetermined and influenced by the teacher. This may consequently lead to the fact that "students are potentially less motivated to search for or remember the target information." (Hunston 2002: 171)

5.5. Skills required

Although access to corpora "has become fairly easy on standard small computers [... and] user-friendly software is available for most normal tasks" (Sinclair 2004:

2), this does not necessarily mean that teachers and students can start exploring naturally-occurring language data as soon as these basic requirements are given. A further requirement which has to be fulfilled when corpora are to be used in the language classroom are the skills needed by students in order to work with corpora proficiently. Mukherjee (2002: 179) refers to two abilities which need to be acquired by learners so as to manage the analysis of computer corpora in the field of language teaching and learning efficiently. These are called *computer literacy* and *corpus literacy* and will be dealt with in more detail subsequently.

As modern corpora are entirely computer-based, it is necessary for learners to be competent in handling computers before being introduced to the study of corpora. This is what Mukherjee (2002: 179) refers to as computer literacy. Nowadays, most students are already familiarised with the handling of computers before they even start attending grammar schools. In case learners are not yet skilled in dealing with them, they will most likely learn how to do so in their computer classes at school. As corpora are in the majority of cases used in higher-level language classes, almost all students should have gained computer literacy by then.

The term corpus literacy, as used by Mukherjee (2002: 179), refers to the students‘ competence to work with computers in order to manipulate and search corpora. Furthermore, it includes those skills needed to analyse language databases which are not necessarily connected with computer software. What can be understood as corpus literacy has also been circumscribed by Laura Gavioli (1997: 96), who notes that

> [u]sing the computer as a source of information about the foreign language in the classroom requires more than simply giving students direct access to data. Learners need also to be introduced to the process by which the data can be analysed. If students are to learn to edit and classify data, they must be shown how to read frequency tables, and concordances. This can lead students to formulate hypotheses, which can then be confirmed or falsified by consulting more text. This process of ‘navigating’ through the corpora leads to extensive reading of texts. Students thus learn to use the concordancer as an instrument of research and learning.

This whole process of analysing language samples contained in corpora often irritates and unnerves both teachers as well as students, as numerous sub-skills are required by both parties to work with electronic language databases proficiently and thus effectively. “The initial reactions of most teachers and of many learners to using corpora, particularly where multiple texts and contexts are retrieved, often stress their linguistic difficulty”, as Aston (1997b: 61) emphasises. This is due to the fact that both teachers and students need some basic linguistic knowledge in order to analyse texts compiled in corpora. Linguistic background knowledge is, for example, necessary, when analysing electronic language databases regarding particular collocations so as to find out whether a specific verb is commonly used in connection with certain prepositions, prepositional phrases or noun phrases. Granath (2009: 49-50) even goes into more detail with regard to the linguistic knowledge which students need to acquire before being able to analyse corpora proficiently. The author argues that first of all, “students need to know how to identify word classes, since in English, for many lexical items, it is the use of a word in context rather than inflectional morphology that signals the word class.” (Granath 2009: 49) Secondly, students need to learn how to distinguish the relevant results gained from corpus analysis from those which are less helpful in order to answer their research questions. Finally, “students need to know how they can calculate variation based on a limited number of sample sentences.” (Granath 2009: 50)

What is important when providing students with corpora for the first time is that the required linguistic competences in order to analyse and deduce regularities from language data are taught to and demanded from learners in a step-by-step approach, as Aston (1997b: 61-63) points out. Therefore, Aston (1997b: 62) suggests that teachers who want their students to learn how to interact with corpus data efficiently should choose language samples which are easily understandable, limited with regard to their number and relatively predictable regarding the analysis results for the first few investigations. Apart from selecting and manipulating language samples for students, teachers can also help their learners to acquire analysis skills more easily by confronting them with corpus tasks of gradually increasing difficulty (Aston 1997b: 63). To sum up, it can

therefore be argued that it is the task of the teachers to provide their students slowly but surely with both theoretical explanations and “the necessary ‘hands on’ know-how”, as Geoffrey Leech (1997: 8) calls it.

One area in which students may, for instance, need both theoretical as well as practical guidance in order to become proficient corpus analysts concerns the interpretation of occurrences of particular language samples. This is due to the fact that the occurrence of a particular lexical item only “tells learners how the word has been used on one occasion, but they do not know how representative this occurrence is.” (Widdowson 2003: 102) Furthermore, when students are introduced to corpora first, it is also important that they are trained in working with potential regularities occurring in corpora. As learners are used to “maximally generalized rules” (Aston 1997b: 60) in language pedagogy in general and grammar teaching more specifically, the fact that corpora do not offer them any rules per se may take some getting used to. Learners may consequently need “practice in identifying patterns of collocation, colligation, connotation and discourse structuring” (Aston 1997b: 60).

Students, however, do not only shy away from linguistic analyses of corpora, as Granath (2009: 61) notes, but also from handling corpora and analysis programmes as such. However, this problem can be solved quite easily, when teachers select the kind of corpus and retrieval software to be used in the language classroom carefully. Granath (2009: 55), for instance, uses the so-called ‘MicroConcord’ software, “because it is so simple that students can learn both simple and complex queries in a matter of minutes”. Owing to the rapid improvement in technology in the recent past, “the difficulties involved in the use of general corpora in language pedagogy are nowadays not as great as they used to be”, as Bernardini (2000: 120) already pointed out ten years ago. Nevertheless, the use of corpora in the language classroom still requires “a certain amount of expertise both for retrieval of corpus data and for their correct analysis and interpretation” (Kaltenböck & Mehlmauer-Larcher 2005: 82).

5.6. Learner autonomy

As a last point in this section, one particular advantage of the use of corpora in the language classroom needs to be mentioned, namely its positive influence on the autonomy of the learners. But what exactly does *learner autonomy* refer to in this context? Benson and Voller (1997: 1-2) note that the term has at least five major meanings in the educational framework, as it is equally used

1. for *situations* in which learners study entirely on their own;
2. for a set of *skills* which can be learned and applied in self-directed learning;
3. for an inborn *capacity* which is suppressed by institutional education;
4. for the exercise of *learners' responsibility* for their own learning;
5. for the *right* of learners to determine the direction of their own learning.

For reasons of simplicity, however, the remarks in this section will not be based on this quite complex definition, but on a broader one taken from Henri Holec (1981: 3). He refers to learner autonomy as "the ability to take care of one's own learning" and thus comprises all of the previously mentioned interpretations by Phil Benson and Peter Voller (1997: 1-2).

With regard to the role of learner autonomy when using corpora in the language classroom, it needs to be said that students are not able to experience how a language can be learned efficiently and independently as soon as they get in direct contact with electronic language databases for the first time. For this reason students need to be guided towards an independent application of this instrument by their teachers gradually. Corpus-based language teaching certainly has the potential to support learner autonomy by engaging pupils in interesting and motivating activities providing them with opportunities for interaction with both teachers as well as fellow students in a relatively stress-free atmosphere (cf. Bernardini 2004: 31-32). This can, for instance, be the case when students are allowed to function as language researchers being encouraged to form and test hypotheses themselves (Adolphs 2006: 109).

Before trusting their students to analyse naturally-occurring language samples quite autonomously, however, it is necessary that teachers protect their students to be "overwhelmed by too much information", as Gavioli (2005: 127) notes. This is due to the fact that an overload of language data may

> lead to frustration on the part of the learners. The learners' age, their general level of language competence, levels of expert knowledge and the learners' attitude towards increasing their learner autonomy all have to be taken into consideration when deciding on how corpora can be used in a foreign language learning context. (Kaltenböck & Mehlmauer-Larcher 2005: 80-81)

The degree of learner autonomy demanded of students when working with corpora, can be regulated by the teacher from highly autonomous corpus searches towards less autonomous corpus activities. The "development of learner autonomy is therefore best understood as a gradual process on a continuum", as Kaltenböck and Mehlmauer-Larcher (2005: 80) point out, whereby full learner autonomy can hardly be reached in educational settings. A quite autonomous way for students to explore corpora is, for example, Bernardini's (2004: 22) approach called 'serendipitous corpus browsing' (see section 4.2.2.). Another activity which requires students to work quite autonomously is when teachers create corpus tasks in which students have to analyse all concordance lines regarding a particular search item on their own (Hunston 1995: 18). Moving one step further away from total autonomy, learners may also be asked to investigate only a number of randomly selected language samples taken from corpora. A fairly determined and hardly autonomous activity with regard to the usage of corpora in the language classroom, however, would be to confront students with only a number of well-chosen concordance lines which clearly exemplify specific lexical or grammatical constructions and thus do not require much autonomy on the part of the learners (cf. Hunston 1995: 18).

Even though learner autonomy is generally referred to as positive characteristic which teachers want their students to achieve in the process of getting older, some problems may occur when asking students to work with or analyse corpus data independently. As already mentioned above, some learners may feel unable

to cope with the amount of language data they are confronted with when exploring corpora. This problem, however, can be solved easily by introducing students only to selected concordance lines or chosen language features. Another difficulty which may occur when students are asked to analyse corpora is that they may not always experience autonomous learning tasks as favourable for their personal learning success. Put differently, learners should be given the opportunity and autonomy to find out and decide themselves how they learn and remember specific language items best. “In many cases, this will be via grammatical rules and lists of lexical items,” as Susan Hunston (2002: 193) notes. Even though some students may feel more comfortable when learning grammar rules or collocations by heart, others may still prefer to analyse and exploit naturally-occurring language samples themselves. Thus, although learners “should not be forced to approach English” (Hunston 2002: 193) exclusively via the analysis of corpora, they should at least be introduced to different learning strategies, since students can only find out which strategies assist them best, if they are provided with alternatives to choose from.

6. Conclusion

My first confrontation with corpora took place in an introductory lecture at university, before they were briefly mentioned again only about two years later in a linguistics lecture. All I knew about corpora from these two courses was that they actually exist and can be used as a means of approaching language data. At that time I was not in the least aware that corpora can actually be applied as pedagogical instruments in the field of language teaching. I only realised this potential when I was first introduced to corpora on a more practical basis, namely, when I decided to attend both an ‘Introduction to Corpus Linguistics’ course as well as a seminar in the field of ‘English for specific purposes’ (ESP) course called ‘Approaching ESP texts’. While the first of these two courses aimed at familiarising students with the structure, analysis and potential of already existing corpora both for linguistic study as well as for the language classroom, the second course attempted to enable students to create and analyse self-made corpora in order to derive teaching materials from them.

Retrospectively I have to admit that I was lucky to be able to attend both courses during my studies, as the first one is only offered every three to four semesters, while the second one is part of a freely selectable additional training programme, which is not an obligatory part of the curriculum. The point which I would like to make here is that even though I had the opportunity to get to know and gain personal experience in working with corpora, this is still exceptional for university students. Unfortunately, neither students of linguistics nor student teachers are provided with the most basic pieces of information regarding the potential of corpora in their field of expertise as a compulsory part of their studies.

Being a student teacher myself, I am especially concerned about the fact that both experienced teachers as well as teachers in training are hardly introduced to the potential of the application of corpora in the field of language pedagogy in their basic or also further education. This feeling is particularly strong, when reading that the results of field investigations “indicate that many of the problems teachers have could be solved, at least partially, if they were introduced to some basic corpus resources and received more support from corpus researchers.” (Römer 2009: 95) Therefore, since teachers are often not even aware that corpora exist and can assist them in solving some of their problems, most pedagogues may actually “need persuading that corpus linguistics can make a contribution to their professional activity.” (Kennedy 1992: 368) This can, to my mind, only be done by introducing teachers to corpora as powerful instruments in the areas of linguistics and pedagogy both from a theoretical as well as a practical perspective.

This is also the idea underlying the organisation of this thesis, which has first attempted to introduce already experienced as well as future language teachers to corpora by providing them with facts concerning the definition, the historical development and the general potential of corpora. Only then more practical issue have been discussed, like the question of how corpora have gained importance in the area of language teaching and in which ways electronic language databases can be applied by teachers (e.g. for reference, for materials design, for demonstration) and students (e.g. for reference, for discovery learning) both inside as well as outside the language classroom. Finally, the last section has

been concerned with the potential and limitations of using corpora in language pedagogy, as evaluated based on existing specialist literature and to a certain extent also on my personal practical experience in working with corpora both as a foreign language learner and future teacher.

On the grounds of the information provided in this thesis individual teachers, however, still have to decide themselves whether the direct use of corpora or corpus-based language materials in language teaching is really worth the extra effort by balancing the presented potential against the limitations of corpora in the field of language pedagogy very carefully. The results gained from these personal evaluations may then assist language teachers “to make informed choices” (Kaltenböck & Mehlmauer-Larcher 2005: 82) regarding the questions if, when and how corpora can be applied in their classes.

Even though, I am aware of the fact that the application of corpora in the language classroom certainly also entails a number of limitations (see section 5.), I am still convinced that “the potential clearly outweighs possible problems”, just as Gunther Kaltenböck and Barbara Mehlmauer-Larcher (2005: 82) note. Therefore, I hope that this thesis will have a share in “spread[ing] the word about corpora” (Römer 2009: 95) and in informing teachers and future teachers about the potential of their application in the field of language teaching and learning so that perhaps one day corpora will be given the attention they deserve and will become an obligatory part of the university curriculum.

List of references

Adolphs, Svenja. 2006. *Introducing electronic text analysis: a practical guide for language and library studies.* London: Routledge.

Aimjer, Karin. 2009. “Introduction: Corpora and language teaching.” In Aijmer, Karin (ed.). *Corpora and Language Teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 1-10.

Anderson, Wendy; Corbett, John. 2009. *Exploring English with Online Corpora.* New York: Palgrave Macmillan.

Aston, Guy. 1997a. “Involving learners in developing methods: exploiting text corpora in self-access.” In Benson, Phil; Voller, Peter (eds.). *Autonomy and Independence in Language Learning.* London: Longman, 204-214.

Aston, Guy. 1997b. “Enriching the Learning Environment: Corpora in ELT.” In Wichmann, Anne; Fligelstone, Steven; McEnery, Tony; Knowles, Gerry (eds.). *Teaching and Language Corpora.* London: Addison Wesley Longman, 51-64.

Aston, Guy. 2000. “Corpora and language teaching.” In Burnard, Lou; McEnery, Tony (eds.). *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang GmbH, 7-17.

Aston, Guy. 2001. “Learning with corpora: an overview.” In Aston, Guy (ed.). *Learning with Corpora.* Bologna: Clueb, 7-45.

Aston, Guy. 2002. “The Learner as Corpus Designer.“ In Kettemann, Bernard; Marko, Georg (eds.). *Teaching and Learning by Doing Corpus Analysis.* Amsterdam / New York: Rodopi B.V., 9-25.

Barlow, Michael. 1996. “Corpora for theory and practice”. *International Journal of Corpus Linguistics 1(1)*, 1-37.

Barnbrook, Geoff. 1996. *Language and Computers: A practical introduction to the computer analysis of language.* Edinburgh: Edinburgh University Press.

Benson, Phil; Voller, Peter. 1997. “Introduction: autonomy and independence in language learning.” In Benson, Phil; Voller, Peter (eds.). *Autonomy and Independence in Language Learning.* London: Longman, 1-17.

Bernardini, Silvia. 2000. *Competence, Capacity, Corpora.* Bologna: Cooperativa Libraria Universitaria Editrice.

Bernardini, Silvia. 2002. “Exploring New Directions for Discovery Learning.” In Kettemann, Bernard; Marko, Georg (eds.). *Teaching and Learning by Doing Corpus Analysis.* Amsterdam / New York: Rodopi B.V., 165-182.

Bernardini, Silvia. 2004. “Corpora in the Classroom: An overview and some reflections on future developments.” In Sinclair, John (ed.). *How to use corpora in language teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 15-36.

Braun, Sabine. 2006. “ELISA: A pedagogically enriched corpus for language learning purposes”. In Braun, Sabine; Kohn, Kurt; Mukherjee, Joybrato (eds.). *Corpus Technology and Language Pedagogy.* Frankfurt am Main: Peter Lang GmbH, 25-47.

Braun, Sabine; Kohn, Kurt; Mukherjee, Joybrato. 2006. “Introduction”. In Braun, Sabine; Kohn, Kurt; Mukherjee, Joybrato (eds.). *Corpus Technology and Language Pedagogy.* Frankfurt am Main: Peter Lang GmbH, 1-4.

Breen, Michael. 1985. “Authenticity in the Language Classroom”. *Applied Linguistics* 6(1), 60-70.

Carter, Ronald. 1998. “Orders of reality: CANCODE, communication, and culture.” *ELT Journal 52(1),* 43-56.

Chambers, Angela. 2007. “Popularising corpus consultation by language learners and teachers.” In Hidalgo, Encarnación; Quereda, Luis; Santana, Juán (eds.). *Corpora in the Foreign Language Classroom.* Amsterdam / New York: Rodopi B.V., 3-16.

Cook, Guy. 1998. “The uses of reality: a reply to Ronald Carter.” *ELT Journal* 52*(1):* 57-64.

Cook, Guy. 2003. *Applied Linguistics.* Oxford: OUP.

Ellis, Rod. 1994. *The study of second language acquisition.* Oxford: OUP.

Francis, Nelson. 1992. “Language corpora B.C.” In Svartvik, Jan (ed.). *Directions in Corpus Linguistics.* Berlin: Mouton de Gruyter, 17-32.

Gavioli, Laura. 1997. “Exploring Texts through the Concordancer: Guiding the Learner.” In Wichmann, Anne; Fligelstone, Steven; McEnery, Tony; Knowles, Gerry (eds.). *Teaching and Language Corpora.* London: Addison Wesley Longman, 83-99.

Gavioli, Laura; Aston, Guy. 2001. “Enriching reality: language corpora in language pedagogy.” *ELT Journal* 55(3), 238-246.

Gavioli, Laura. 2005. *Exploring Corpora for ESP Learning.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Granath, Solveig. 2009. “Who benefits from learning how to use corpora?” In Aijmer, Karin (ed.). *Corpora and Language Teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 47-65.

Götz, Sandra; Mukherjee, Joybrato. 2006. “Evaluation of Data-Driven Learning in university teaching: A project report”. In Braun, Sabine; Kohn, Kurt; Mukherjee, Joybrato (eds.). *Corpus Technology and Language Pedagogy.* Frankfurt am Main: Peter Lang GmbH, 49-67.

Hadley, Gregory. 2002. “An introduction to data-driven learning.” *RELC Journal* 33(2), 99-124.

Harris, Michael. McCann, Paul. 1994. *Assessment.* Oxford: Macmillan Education.

Henry, Alex; Roseberry, Robert. 2001. “Using a small corpus to obtain data for teaching a genre”. In Ghadessy, Mohsen (ed.). *Small Corpus Studies and ELT.* Amsterdam: Benjamins, 93-134.

Hidalgo, Encarnación; Quereda, Luis; Santana, Juán. 2007. “Foreword.” In Hidalgo, Encarnación; Quereda, Luis; Santana, Juán (eds.). *Corpora in the Foreign Language Classroom.* Amsterdam / New York: Rodopi B.V., ix-xiv.

Holec, Henri. 1981. *Autonomy in Foreign Language Learning.* Oxford: Pergamon.

Hölzl, Jessica. 2003. “Corpora and Patterns: Their Relevance to ELT”. Diplomarbeit: University of Vienna.

Hunston, Susan. 1995. “Grammar in teacher education: the role of a corpus.” *Language Awareness* 4(1), 15-31.

Hunston, Susan. 2002. *Corpora in Applied Linguistics.* Cambridge: CUP.

Johansson, Stig. 2007. “Using corpora: from learning to research.” In Hidalgo, Encarnación; Quereda, Luis; Santana, Juán (eds.). *Corpora in the Foreign Language Classroom.* Amsterdam / New York: Rodopi B.V., 17-28.

Johansson, Stig. 2009. “Some thoughts on corpora and second-language acquisition.” In Aijmer, Karin (ed.). *Corpora and Language Teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 33-44.

Johns, Tim. 1991. “Should you be persuaded: two samples of data-driven learning materials.” In Johns, Tim; King, Philip (eds.). *Classroom Concordancing (ELR Journal* 4). Birmingham: University of Birmingham, 1-16.

Johns, Tim. 1994. “From printout to handout: Grammar and vocabulary teaching in the context of Data-driven Learning.” In Odlin, Terence (ed.). *Perspectives on pedagogical grammar.* Cambridge: CUP.

Johns, Tim. 2002. “Data-driven Learning: The Perpetual Challenge.“ In Kettemann, Bernard; Marko, Georg (eds.). *Teaching and Learning by Doing Corpus Analysis.* Amsterdam / New York: Rodopi B.V., 193-202.

Kaltenböck, Gunther; Mehlmauer-Larcher, Barbara. 2005. “Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching”. *ReCall* 17(1), 65-84.

Kennedy, Graeme. 1992. “Preferred ways of putting things with implications for language teaching.” In Svartvik, Jan (ed.). *Directions in Corpus Linguistics.* Berlin: Mouton de Gruyter, 335-373.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics.* London: Longman.

Leech, Geoffrey; Fligelstone, Steven. 1992. “Computers and corpus analysis”. In: Butler, Christopher (ed.). *Computers and written texts.* Oxford: Blackwell, 115-140.

Leech, Geoffrey. 1997. “Teaching and Language Corpora: a Convergence.” In Wichmann, Anne; Fligelstone, Steven; McEnery, Tony; Knowles, Gerry (eds.). *Teaching and Language Corpora.* London: Addison Wesley Longman, 1-23.

McCarthy, Michael. 2001. *Issues in Applied Linguistics.* Cambridge: CUP.

McEnery, Tony; Wilson, Andrew. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McEnery, Tony; Xiao, Richard; Tono, Yukio. 2006. *Corpus-based language studies: an advanced resource book.* London: Routledge.

Meunier, Fanny; Gouverneur, Céline. 2009. “New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material.” In Aijmer, Karin (ed.). *Corpora and Language Teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 179-201.

Meyer, Charles. 2002. *English Corpus Linguistics: An Introduction.* Cambridge: CUP.

Mindt, Dieter. 1975. *Moderne Linguistik.* Düsseldorf: August Bagel Verlag.

Mindt, Dieter (ed.). 1988. *EDV in der Linguistik: Ziele – Methoden – Ergebnisse.* Frankfurt am Main: Diesterweg.

Mukherjee, Joybrato. 2002. *Korpuslinguistik und Englischunterricht: Eine Einführung.* Frankfurt am Main: Peter Lang GmbH.

Mukherjee, Joybrato. 2006. “Corpus linguistics and language pedagogy: the state of the art – and beyond”. In Braun, Sabine; Kohn, Kurt; Mukherjee, Joybrato (eds.). *Corpus Technology and Language Pedagogy.* Frankfurt am Main: Peter Lang GmbH, 5-24.

Mukherjee, Joybrato. 2009. *Anglistische Korpuslinguistik: Eine Einführung.* Berlin: Erich Schmidt Verlag.

Nesselhauf, Nadja. 2004. “Learner corpora and their potential for language teaching.” In Sinclair, John (ed.). *How to use corpora in language teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 125-152.

Odlin, Terence. 1994. “Glossary.” In Odlin, Terence (ed.). *Perspectives on pedagogical grammar.* Cambridge: CUP.

O’Keeffe, Anne; McCarthy, Michael; Carter, Ronald. 2007. *From Corpus to Classroom.* Cambridge: CUP.

Osborne, John. 2000. “What can students learn from a corpus?: building bridges between data and explanation.” In Burnard, Lou; McEnery, Tony (eds.). *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang GmbH, 165-172.

Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Partington, Alan. 2001. “Corpus-based description in teaching and learning.” In Aston, Guy (ed.). *Learning with Corpora.* Bologna: Clueb, 63-84.

Renouf, Antoinette. 1997. “Teaching Corpus Linguistics to Teachers of English.” In Wichmann, Anne; Fligelstone, Steven; McEnery, Tony; Knowles, Gerry (eds.). *Teaching and Language Corpora.* London: Addison Wesley Longman, 255-266.

Römer, Ute. 2009. “Corpus research and practice: What help do teachers need and what can we offer?” In Aijmer, Karin (ed.). *Corpora and Language Teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 83-98.

Schmitt, Norbert. 2002. *Vocabulary in Language Teaching.* Cambridge: CUP.

Sinclair, John. 1991. *Corpus, Concordance, Collocation.* Oxford: OUP.

Sinclair, John. 1997. “Corpus Evidence in Language Description.” In Wichmann, Anne; Fligelstone, Steven; McEnery, Tony; Knowles, Gerry (eds.). *Teaching and Language Corpora.* London: Addison Wesley Longman, 27-39.

Sinclair, John. 2004. “Introduction.” In Sinclair, John (ed.). *How to use corpora in language teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 1-10.

Stewart, Dominic; Bernardini, Silvia; Aston, Guy. 2004. “Introduction: Ten years of TaLC.” In Aston, Guy; Bernardini, Silvia; Stewart, Dominic (eds.). *Corpora and Language Learners.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 1-18.

The Compleat Lexical Tutor. For Data-driven learning on the Web. http://www.lextutor.ca/ (8 November 2010)

TeleNex. A Resource for English Teachers in Hong Kong Schools. Developed by TELEC, The University of Hong Kong. http://www.telenex.hku.hk/telec/pmain/opening.htm (8 November 2010)

Tsui, Amy. 2004. “What teachers have always wanted to know – and how corpora can help.” In Sinclair, John (ed.). *How to use corpora in language teaching.* Amsterdam / Philadelphia: John Benjamins Publishing Company, 39-61.

Tsui, Amy. 2005. “ESL teachers’ questions and corpus evidence.” *International Journal of Corpus Linguistics* 10(3), 335-356.

Voller, Peter. 1997. “Does the teacher have a role in autonomous language learning?” In Benson, Phil; Voller, Peter (eds.). *Autonomy and Independence in Language Learning.* London: Longman, 98-113.

Widdowson, Henry. 1980. *Explorations in Applied Linguistics 1.* Oxford: OUP.

Widdowson, Henry. 1996a. *Linguistics.* Oxford: OUP.

Widdowson, Henry. 1996b. *Teaching Language as Communication.* Oxford: OUP.

Widdowson, Henry. 2000. “On the limitations of linguistics applied.” *Applied Linguistics* 21(1): 3-25.

Widdowson, Henry. 2003. *Defining Issues in English Language Teaching.* Oxford: OUP.

Widdowson, Henry. 2007. *Discourse Analysis.* Oxford: OUP.

Wynne, Martin. 2007. “Searching and Concordancing.” *Pala: Poetics and Linguistics Association.* http://www.pala.ac.uk/resources/sigs/corpus-style/searching/handbook.html (19 October 2010).

Yule, George. 2004. *The Study of Language.* (2nd edition; 11th printing). Cambridge: CUP.

List of corpora and retrieval software

AntConc software (antconc). Laurence Anthony's Homepage. http://www.antlab.sci. waseda.ac.jp/software.html (8 November 2010).

A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). 1964, 1971, 1979. Compiled by W. N. Francis and H. Kučera. Brown University. Providence, Rhode Island. Further information: http://icame.uib.no/brown/bcm.html (9 November 2010).

Cambridge and Nottingham Corpus of Discourse in English (CANCODE). University of Cambridge Press: English Language Teaching. Further information: http://www.cambridge.org/at/elt/catalogue/subject/custom/item3637700/Cambridge-International-Corpus-Cambridge-International-Corpus/?site_locale=de_AT (9 November 2010).

Collins Birmingham University International Language Database (COBUILD). Based within the School of English at Birmingham University. Further information: http://www.mycobuild.com/about-collins-corpus.aspx (9 November 2010).

Davies, Mark. (2004-) BYU-BNC: *The British National Corpus*. Available online at http://corpus.byu.edu/bnc (8 November 2010).

Davies, Mark. (2007-) *TIME Magazine Corpus* (100 million words, 1920s-2000s). Available online at http://corpus.byu.edu/time (8 November 2010).

London-Lund Corpus of Spoken English. (1975-81 and 1985-88). Compiled by Jan Svartvik. Further information: http://khnt.hit.uib.no/icame/manuals/londlund/index.htm (9 November 2010).

MicroConcord Software. (1993). Oxford: Oxford University Press. http://www.lexically.net/software/index.htm (9 November 2010).

R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales. (2002*) The Michigan Corpus of Academic Spoken English* (MICASE). Ann Arbor, MI: The Regents of the University of Michigan. Available online at http://quod.lib.umich.edu/m/micase/ (9 November 2010).

The Bank of English (BoE): An International Language Corpus. Jointly owned by HarperCollins Publishers and the University of Birmingham. Further information: http://www.mycobuild.com/about-collins-corpus.aspx (8 November 2010).

The Freiburg-Brown Corpus ('Frown') (POS-tagged version) compiled by Christian Mair, Albert Ludwigs-Universität Freiburg, and Geoffrey Leech. Further information: http://khnt.hit.uib.no/icame/manuals/frown/ INDEX.HTM (9 November 2010).

The Freiburg-LOB Corpus ('F-LOB') (original version) compiled by Christian Mair, Albert-Ludwigs-Universität Freiburg. Further information: http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/ (9 November 2010).

The LOB Corpus, original version (1970-1978), compiled by Geoffrey Leech, Lancaster University, Stig Johansson, University of Oslo (project leaders), and Knut Hofland, University of Bergen (head of computing). University of Lancaster. Further information: http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM (9 November 2010).

The International Corpus of English – The British Component. (ICE-GB). Copyright: Survey of English Usage. University College of London, 1998. Further information: http://ice-corpora.net/ice/icegb.htm (9 November 2010).

WordSmith Tools. Mike Scott's Web. http://www.lexically.net/wordsmith/index.html (8 November 2010).

List of figures

Summary

Although the integration of computer corpora into the field of language teaching and learning has frequently been discussed from a theoretical viewpoint in specialist literature throughout the last two to three decades, only few teachers appear to apply them in their actual teaching practice. One reason for this fact may be that corpus linguists and applied linguists tend to focus more on their own research, rather than on communicating the existence and potential of corpora to practitioners in the field of language pedagogy.

Therefore, this study attempts to introduce already experienced as well as future language teachers to corpora as such, by defining the term, providing a brief historical overview regarding their development and by discussing their potential. After having addressed the issue of how corpora have gained importance in the area of language teaching, various ways in which corpora can be applied by teachers (e.g. for reference, for materials design, for demonstration) and students (e.g. for reference, for discovery learning) in the language classroom are discussed both from a theoretical as well as from a practical point of view. The final section is then concerned with the potential and limitations of using corpora in language pedagogy, as evaluated based on existing specialist literature and to a certain extent also on personal practical experience.

The major objective underlying this survey is to introduce and make already practising and future teachers aware of the existence as well as the potential of corpora in the field of language teaching. Based on the information and insights provided, language pedagogues can get an idea of the ways in which corpora can be applied in language teaching so as to decide for themselves whether and how they may want to integrate corpora into the language classroom.

Zusammenfassung

Obwohl die Integration von computerisierten Textkorpora in den Sprachunterricht in den vergangenen zwei bis drei Jahrzehnten bereits vielfach vom linguistischen Standpunkt aus in der Fachliteratur diskutiert wurde, scheinen nur wenige LehrerInnen auch im Unterrichtsalltag auf dieses Medium zurückzugreifen. Ein Grund für diese Tatsache liegt sicherlich unter anderem darin, dass sich sowohl Linguisten als auch angewandte Linguisten stark auf die wissenschaftliche Forschung konzentrieren, anstatt ihre theoretischen Ausführungen praktisch aufzuarbeiten und mit Lehrpersonen zu teilen.

Diese Arbeit verfolgt nun das Ziel diese Lücke zu schließen und bereits erfahrene, sowie auch zukünftige LehrerInnen näher mit Textkorpora vertraut zu machen, indem der Begriff an sich definiert, ein kurzer historischer Überblick über die Entwicklung von Korpora gegeben und das generelle Potential von Korpusanalysen erörtert wird. Nachdem erklärt wurde wie Korpora außerhalb der Linguistik auch im Gebiet der Sprachendidaktik Fuß gefasst haben, werden verschiedene Wege sowohl theoretisch als auch praktisch vorgestellt, wie Korpora von Lehrpersonen (z.B. zur Referenz, zur Herstellung von Unterrichtsmaterialien, zu Demonstrationszwecken) und auch von SchülerInnen (z.B. zur Referenz, zum entdeckenden Lernen) im Unterricht verwendet werden können. Das letzte Kapitel beschäftigt sich schließlich mit dem Potential, aber auch den Einschränkungen, die mit der Verwendung von Korpora im Bereich des Sprachunterrichts einhergehen. Diese beiden Aspekte wurden auf Basis existierender Fachliteratur und – in einem gewissen Ausmaß – auch auf persönlichen praktischen Erfahrungen in der Auseinandersetzung mit Korpora evaluiert.

Das Hauptziel dieser Arbeit besteht darin, LehramtsstudentInnen und bereits praktizierende LehrerInnen mit Korpora und ihrem Potential für den Sprachunterricht vertraut zu machen. Basierend auf den bereitgestellten und aufgearbeiteten Informationen wird diesen Personengruppen die Möglichkeit geboten, selbständig zu entscheiden, ob und wie sie Textkorpora in den Sprachunterricht einbeziehen möchten.

Alexandra Hable

Jägerstraße 24/9
A-1200 Wien

alex.hable@gmx.at

Lebenslauf

AUSBILDUNG:

1992 – 1996	Volksschule in Wels (OÖ)
1996 – 2000	BG / BRG Schauerstraße in Wels (OÖ)
2000 – 2005	BBA für Kindergartenpädagogik in Linz (OÖ) Ausbildung zur Kindergartenpädagogin und Horterzieherin mit Zusatzqualifikation „Englisch im Kindergarten“
Juni 2005	Reife- und Diplomprüfung mit ausgezeichnetem Erfolg
Seit Oktober 2005	Lehramtsstudium Englisch und Psychologie & Philosophie an der Universität Wien
Seit Oktober 2007	Diplomstudium Anglistik und Amerikanistik mit Wahlfachmodul „Deutsch als Fremdsprache“ an der Universität Wien

AUSBILDUNGSBEZOGENE ARBEITSVERHÄLTNISSE UND PRAKTIKA:

2000 – 2005	laufende sowie geblockte Praktika in Kindergärten und Horten im Zuge der Ausbildung
2001 – 2003	Gruppenleiterin der kath. Jungschar der Stadtpfarre Wels (OÖ)
2002 – 2003	Betreuerin einer wöchentlich stattfindenden „Mutter-Kind-Gruppe“ im Eltern-Kind-Zentrum Wels (OÖ)
2003 – 2005	Pfarrleiterin der kath. Jungschar und Jugend der Stadtpfarre Wels (OÖ) und Jungscharlagerleiterin
2001 – 2008	Betreuerin der jährlich stattfindenden Ferienlager der kath. Jungschar der Stadtpfarre Wels (OÖ)
September 2003	dreiwöchiger Auslandsaufenthalt in Dublin (Irland) im Rahmen eines Praktikums an der „Little Acres Montessori School“ (Kindergarten)

Seit 2005	Betreuerin von Anfängerkursen beim Skiklub ESKA Wels (OÖ)
Juli/August 2006	zweimonatiger Au-Pair-Aufenthalt in East Bergholt / Suffolk / England (GB)
2007 – 2009	jeweils dreiwöchige Ferialanstellung als Hortpädagogin beim Magistrat Wels (OÖ)
März – Juni 2008	Schulpraktikum für das Unterrichtsfach „Englisch“ im Zuge des Lehramtsstudiums
Jänner – März 2009	Schulpraktikum für das Unterrichtsfach „Psychologie und Philosophie“ im Zuge des Lehramtsstudiums
März – Juni 2009	Hospitations- und Unterrichtspraktikum im Rahmen der Lehrveranstaltung „Methodik“ für das Modul „Deutsch als Fremd- und Zweitsprache“ an der VHS Brigittenau (Wien)

BESONDERE KENNTNISSE UND FÄHIGKEITEN:

Sprachen:	Deutsch:	Muttersprache
	Englisch:	Europäischer Referenzrahmen C1/C2
	Französisch:	Grundkenntnisse
Computerkenntnisse:	Microsoft Office, WordSmith Tools, ICE-CUP software	
Führerschein:	Klasse B	

PERSÖNLICHE DATEN:

Geboren am:	20.08.1986 in Wels (OÖ)
Staatsangehörigkeit:	Österreich
Eltern:	Susanne Hable Richard Hable
Geschwister:	Nina Hable

Wien, am 19. Oktober 2010