Corpus linguistics and language pedagogy: The state of the art – and beyond

Joybrato Mukherjee

Justus Liebig University, Giessen

Abstract

The present paper provides a selected overview of the state of the art in corpus-informed language pedagogy. Starting off from a general assessment of the impact that the corpus revolution has already had on English language teaching (ELT), the focus of the main part of this paper is on some typical examples of corpus use in three language-pedagogically relevant areas: (1) using corpora for ELT (e.g. producing learner dictionaries); (2) using corpora in the ELT classroom (e.g. in data-driven learning); (3) using learner corpora. With regard to learner corpus research, for example, the paper also sketches out some prospects for future research, e.g. the compilation of local learner corpora.

1 Introduction: the corpus revolution and English language teaching

There is general agreement among empirically-oriented linguists that the advent of large, computerised corpora has revolutionised the linguistic description and analysis of the English language. In modern corpus linguistics, not just any group of texts qualifies as a corpus, but it must be "a collection of texts assumed to be representative of a given language, dialect, or other subset of a language" (Francis 1982: 7). Representativeness is a key issue in corpus design because it captures the attempt to compile a database that provides a statistically viable sample of language use in general (or a relevant subsection thereof). In spite of various problems involved in putting the concept of representativeness into practice (cf. Biber 1993), it is only by assuming some sort of representativeness in corpus design that one can extrapolate from corpus findings general trends in language use (cf. Mukherjee 2004a).

From its very beginnings, modern corpus linguistics has been intricately intertwined with the development of computers and software programs for corpus analysis. As Biber et al. (1998) point out, this has to do with the storage capacities that are needed for large databases and with the reliability of automatic searches (in contradistinction to manual searches):

> Computers make it possible to identify and analyze complex patterns of language use, allowing the storage and analysis of a larger database of natural language than could be dealt with by hand. Furthermore, computers provide consistent, reliable analyses. (Biber et al. 1998: 4)

The computer-assisted analysis of corpus data, especially by means of wordlists, concordancers and other tools that modern corpus-linguistic software packages like *WordSmith Tools* (Scott 2005) offer, opens up entirely new perspectives for linguistic analysis because linear text analysis is no longer necessary: "The corpus is stored in such a way that it can be analysed non-linearly, and both quantitatively and qualitatively" (Hunston 2002: 2).

It is hard to exaggerate the extent to which corpus-linguistic research has influenced the agenda of ELT over the past two decades. In essence, there are two major areas in ELT to which corpora turn out to be relevant:

> The development of corpora has the potential for two major effects upon the professional life of the language teacher. Firstly, corpora lead to new descriptions of a language, so that the content of what the language teacher is teaching is perceived to change in radical ways […]. Secondly, corpora themselves can be exploited to produce language teaching materials, and can form the basis for new approaches to syllabus design and to methodology. (Hunston 2002: 137)

In the following sections, I would like to take stock of the impact of the corpus revolution on the two levels: using corpora for ELT ('content'; cf. section 2), and using corpora in the ELT classroom ('methodology'; cf. section 3). Before reviewing the various fields in which corpora have been of particular importance to ELT, it seems to be in order, however, to also note that there still remains a wide gap between the wide range of corpus-based activities that have been suggested by applied corpus linguists and the relatively limited extent to which corpora are actually used in the ELT classroom (cf. Mukherjee 2004b). As Granger (2004) notes somewhat grudgingly:

> The main fields of application of corpus data are materials and syllabus design and classroom methodology. In all three fields, there is very active work in progress, but, with the exception of ELT dictionaries, the number of concrete corpus-informed achievements is not proportional to the number of publications advocating the use of corpora to inform pedagogical practice. (Granger 2004: 136)

Tribble (2000: 31) also confirms that "[n]ot many teachers seem to be using corpora in their classrooms". It thus seems that there might be a clash between the corpus linguist's enthusiasm about the language-pedagogical use of corpora

on the one hand and the average teacher's reluctance to use corpora in his/her own classroom on the other. It thus appears to be of paramount importance that many more teachers get actively involved in working with – and thus disseminating knowledge about – corpora. As will be discussed in section 4, the compilation and analysis of local learner corpora represent a very fruitful way of initiating teacher-centred corpus-based classroom action research.

2 Using corpora for ELT

The earliest and most significant impact that corpus linguistics had on language teaching can be found in lexicography. The compilation of the Collins Birmingham University International Language Database (COBUILD) in the 1980s, initiated and organised by John Sinclair, led to the first corpus-based dictionary, the first edition of the *Collins COBUILD English Language Dictionary* (1987), which was based on a corpus of 20 million words. In the 1990s, COBUILD was expanded to form the Bank of English, a dynamic corpus to which new texts have been constantly added so that today its size is about 500 million words. COBUILD set new standards in lexicography because the corpus-based description of English and its corpus-based codification has resulted in a new generation of dictionaries that include information that had not been available in traditional pre-corpus dictionaries. The following entry for the noun *assumption*, which is taken from the corpus-based *Macmillan English Dictionary* (2002), exemplifies lexicographical information that can only be derived from corpus data:

> **assumption** /əˈsʌmpʃn/ noun ★★
> **1** [C] something you consider likely to be true even though no one has told you directly or even though you have no proof: *Your argument is based on a completely false assumption.* ◆ **+that** *There is an assumption that all the people who live around here are rich.* ◆ **make an assumption** (=make a decision based on poor evidence) *People tend to make assumptions about you based on your appearance.* ◆ **on the assumption that** *The law works on the assumption that it is preferable for children to be with their mother.*
> **2** [U] a process in which you begin to use your power or status, or begin to be responsible for something: *the assumption of adult responsibilities*
> (*Macmillan English Dictionary* 2002, s.v. *assumption*)

As in virtually all other corpus-based dictionaries, this entry includes information on the general frequency of the word in the English language (cf. the two

asterisks), the frequency of the two main meanings of the word (cf. the ordering of **1** and **2**), frequently co-occurring collocates (e.g. '+ *that*') and frequent lexicogrammatical patterns of the word (e.g. '*on the assumption that*'). By including information on collocations and patterns of *assumption*, the dictionary reflects the general attempt in corpus linguistics to overcome the traditional distinction between lexis and grammar and to establish a unified lexicogrammar (cf. Stubbs 1993). Besides collocations and patterns, corpus-based dictionaries also include information on other frequent routines in language use, e.g. colligations (i.e. the co-occurrence of specific words and word-classes as in 'verb/adjective + preposition + *the* + *naked eye*', cf. Sinclair 1991) and semantic prosodies (i.e. the tendency of a word to occur in positive or negative contexts, e.g. *provide* in positive contexts and *affect* in negative contexts (cf. Stubbs 1995).

While corpus-based dictionaries include relevant grammatical information on the use of individual words, corpus-based grammars complement the grammatical description of specific syntactic structures with lists of words that tend to be used in a given structure. The first corpus-based grammar was also published by the COBUILD project. The *Collins COBUILD English Grammar* (1990) is a learner grammar, the grammatical categories of which are easily accessible and semantically-oriented. The following entry on 'emphasizing adjectives' reveals that learner grammars, too, have profited considerably from the advent of corpora because learners can now learn syntactic structures together with their frequent lexical realisations:

> You can emphasize your feelings about something that you mention by putting an adjective such as 'complete', 'absolute', and 'utter' in front of a noun.
>
> *He made me feel like a complete idiot.*
> *Some of it was absolute rubbish.*
> *...utter despair.*
> *...pure bliss.*
>
> You generally use an adjective of this kind only when the noun indicates your opinion about something. Because they are used to show strong feelings, these adjectives are called **emphasizing adjectives.** Here is a list of emphasizing adjectives:
>
> | | | | |
> |---|---|---|---|
> | absolute | outright | pure | true |
> | complete | perfect | real | utter |
> | entire | positive | total | |
>
> (*Collins COBUILD English Grammar* 1990: 69)

The 1990s saw intense activities not only in corpus-based lexicography, but also in corpus-based grammar writing, culminating in the *Longman Grammar of*

Spoken and Written English (Biber et al. 1999), a very useful reference grammar with a great deal of quantitative information on the distribution of grammatical structures across four registers of present-day English, and *An Empirical Grammar of the English Verb System* (Mindt 2000), a grammar focusing on a corpus-based description of the English verb phrase. There are also various corpus-based on-line grammars available on the World Wide Web, e.g. the *Chemnitz Internet Grammar of English* (<http://www.tu-chemnitz.de/phil/InternetGrammar/>, cf. Schmied 1999), which is geared to the needs of advanced foreign language learners and which focuses on a wide range of sources of frequent learner errors. It should also be mentioned in this context that corpus-based insights into English usage have also exerted a considerable influence on 'ordinary' learner grammars, for example Ungerer's (1999) *Englische Grammatik heute*, which points out in its preface that data from the British National Corpus (BNC) had been used for the grammar at hand.

While corpus data have been widely accepted as relevant input for learner dictionaries and learner grammars, the language of ELT textbooks is still very often not in line with what corpus analyses have revealed about the way the English language is used in reality. In many regards, therefore, the language of ELT textbooks still needs to be refined so that it becomes more natural and native-like. A good case in point is the use of discourse markers. Müller (2004) shows in her corpus-based study of advanced learner language that in three widely used ELT textbooks in Germany, i.e. *Learning English Green Line*, *English G* and *Notting Hill Gate*, the discourse markers *well* and *so* are used very frequently (up to 24 occurrences and 20 occurrences per volume, respectively), while the discourse markers *you know* and *like* are used only rarely in ELT textbooks (up to 6 instances and 3 instances per volume, respectively). Thus, Müller (2004) draws the following conclusion, which shows the need for further refinement of ELT textbook texts in the light of corpus data:

> Given this representation of the four discourse markers in German textbooks of English, it is not surprising that the German speakers in the GLBCC [Giessen Long Beach Chaplin Corpus] did not have much difficulty using *well*, but apparently were not used to employing *you know* and *like* as much as native speakers did. (Müller 2004: 257)

Since in most curricula of English studies at German universities corpus linguistics is a topic that can easily be replaced with other specialties, many teachers are not at all familiar with corpora, corpus-linguistic resources and corpus-linguistic methods. As shown in a recent survey among 248 grammar school teachers in North-Rhine Westphalia, Germany's most populous state, nearly 80% of the participants do not know anything about corpus linguistics, while 10% have already heard of corpus linguistics and another 10% are familiar

with corpus studies (cf. Mukherjee 2004b: 241). It is necessary, therefore, to develop and offer many more in-service teacher training programmes for English language teachers that introduce them to key issues in corpus linguistics, basic functions of standard corpus software like *WordSmith Tools* and major applications of corpora in the teaching practice.[1] Various studies show that it is both very important and useful to make corpus data and/or corpus-linguistic expertise available to English language teachers. Tsui (1996, 2004), for example, reports on the TeleNex project in Hong Kong, which contains a website for English language teachers (<http://www.telenex.hku.hk>). On this website, a conference area with various discussion corners is offered, including a language corner.[2] Here language specialists with access to a wide range of large corpora of English (including the Bank of English, the BNC and Tele Corpora with Hong Kong English) answer linguistic questions that are put forward by English language teachers. The following two questions refer to the choice of the verb form after coordinated subject noun phrases:

> Teacher 4:
> Hello! Which one is correct?
> There is a man and a woman outside.
> Or
> There are a man and a woman outside.
> Please give some comments, any one.
>
> Teacher 5:
> Hi,
> What should we use in the following sentences? Is or are?
> 1. There _____ an apple and some oranges on the table.
> 2. There _____ some oranges and an apple on the table.
> Thanks. It seems to me that 'are' is okay in both. Is there any rule here?
> (Tsui 2004: 50)

These questions exemplify the usefulness of the language corner because very often English language teachers as non-native speakers feel insecure about

---

1 See also Breyer's (this volume) software programme *My Concordancer*, which is geared to the needs of language teachers and learners.

2 Note in this context that this language corner is of particular relevance to English language teachers in Hong Kong because the status of the English language in Hong Kong – and, accordingly, the question of what the target model in ELT should be – is a controversial issue (cf. Bolton 2000): is English a foreign language in Hong Kong that is entirely dependent on an exonormative British English model or is it an institutionalised second-language variety with its own norm-developing potential?

specific aspects of English usage. The following reply was given by TELEC staff members:

> TELEC staff responded to the teachers' questions by pointing out that usually the singular form of 'be' is used when the first noun that follows is singular and the plural form of 'be' is used when the noun group after it is plural (see also *Collins Cobuild English Grammar*, p. 416). However, a search through the corpus does show an instance of the following:
> According to PACE, suspects can only be detained at designated police stations where **there are a custody and a reviewing officer**.
> (Tsui 2004: 51)

Apart from the fact that teachers are provided with a general rule of thumb, it is also highly significant that the grammatical rule which is also included in the *Collins COBUILD English Grammar* does not cover every case. In fact, the analysis of corpus data reveals that the scope of virtually all grammatical rules is limited and that there is a remainder of instances which deviate from the rules.[3] It is very important to make English language teachers fully aware of the fact that with regard to many forms and structures "the question is not about possibility but about probability of usage" (Tsui 2004: 51).

While the present section has focused on a selection of areas in which corpora are used for language-pedagogical purposes outside the classroom, the following section will deal with some major fields of corpus use in the classroom setting itself.

3 Using corpora in the ELT classroom

For a long time already, corpus linguists have suggested various ways of how to make students work with corpora themselves. Picking up on inductive and leaner-centred autonomous approaches to language learning such as Widdowson's (1990) 'learning as discovery', Johns and King (1991) were in the vanguard of developing concordance-based learning procedures, for which they introduced the term *Data-driven Learning* (DDL):

> [Data-driven learning is] the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language, and the development of activities and exercises based on concordance output. (Johns and King 1991: iii)

---

3 This remainder is called "the compost of language" by Mindt (2002: 211).

DDL activities can be plotted on a cline of learner autonomy, ranging from teacher-led and relatively closed concordance-based exercises to entirely learner-centred corpus-browsing projects. The DDL material in Table 1 exemplifies concordance-based exercises. This material is intended to make German learners of English use synonyms of the adjective *important*, which is often overused even by advanced learners, more frequently in appropriate adjective-noun collocations.

Table 1: *Concordance-based DDL material (Flowerdew 2001: 368)*

Alternative word list: critical / crucial / major / serious / significant / vital		
1. nd imagine. This is the first and most	.....	advantage of science and tech
2.		

WORKSHEET

handsome good attractive pretty tall successful

1. In pairs, choose three of these adjectives and have a look at their concordances.
   - Which words do they most frequently occur with?
   - Do they occur as part of a series of adjectives?

2. Discuss your findings with other pairs, then with the rest of the class.

Figure 1: *DDL material initiating learner-centred corpus analysis (Gavioli 2001: 120)*

into a particular linguistic form

structures by using corpus data. It is quite clear, however, that such language-related *Facharbeiten*, which require autonomous language learners, can only be successful if students are imparted with some basic 'corpus literacy' (cf. Mukherjee 2002: 179), including, for example, a basic understanding of what a corpus is, what you can (and cannot) do with a corpus, how concordances can be analysed, how one may (or may not) extrapolate from corpus data general trends in language use.

At the very end of the autonomy cline of DDL activities, Bernardini's (2004: 22) concept of 'serendipitous corpus browsing' can be plotted, which she describes as a "an approach to learning from corpora in which learners are guided to browse large and varied text collections in open-ended, exploratory ways". The idea here is that students no longer browse the corpus with any specific *a priori* topic or assignment in mind, but are expected to note any form and structure that they may find interesting, to analyse the form or structure at hand and to move from here to other interesting forms and structures. It is doubtful, however, whether this extremely autonomous corpus-based activity can be fruitfully put into practice in the reality of ELT classrooms. For one, it remains entirely unclear what the linguistic syllabus of this corpus-browsing activity could be. Secondly, and more importantly, the teacher could easily face major difficulties because he/she can no longer exert any control over what students are using the corpus for. As Hunston (2002) notes:

> A possible disadvantage for the teacher is that they have very little control over what happens. If the corpus is consulted and no answer is apparent to student or teacher, or if further difficult questions are raised, the teacher may feel that a loss of expertise has occurred.

genre approach to language teaching includes an analytic first step and a productive second step: (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 textual 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. Rohrbach (2003) has shown how this approach can, *mutatis mutandis*, already be fruitfully applied to a class 9 in a German grammar school. He compiled a small corpus of travel brochures extracted from the freely accessible website of the Yorkshire Tourist Board. He then guided his students in analysing the move structure of the travel brochures and the preferred patterns for each typical move. The following list provides an overview of some of the typical moves of a travel brochure text; for the fifth move, 'recommendation of night life and amusements', the preferred patterns that the students identified are also given:

Move 1: Promotion of the general character of an area
Move 2: Presentation of fascinating history
Move 3: Praise of beautiful landscape and natural sights
Move 4: Recommendation of cultural assets
  a) Museums
  b) Exhibitions and events
Move 5: Recommendation of night life and amusements
  a) Clubs, Pubs etc.
    - *If clubbing means anything to you, X means everything to you.*
    - *Don't miss out on X - the chance to enjoy...*
    - *The more adventurous might want to try …*
  b) Shopping
    - *Not to mention serious shopping ...*
    - *…including the only XY shop outside London.*
    - *The town has a… and is host to…*
    - *…all found in a pedestrianised city centre*
    - *For shoppers XY offers many choices*
    - *Of course there is…*
    - *For those who love to shop, XY offers a first-class opportunity to browse in unrivalled shopping facilities, including…*
    - *a weekly market is held every…*

> Move 6: Famous places or persons from movies and books
> […]
> (cf. Rohrbach 2003)

The following text, exemplifying the potential outcome of the second productive phase of the genre approach, was written by one of the students. The task was to write a travel brochure text on the students' own hometown, Göttingen, by picking up on the move analysis and by making use of the preferred patterns for each move.

> Situated in the south of Niedersachsen, Göttingen is famous for its university and is often called: university town Göttingen. Famous poets like Theodor Heuss (sic!) or Heinrich Heine described this "small city" as a small global city with an international atmosphere. In the town centre is the old marketplace with the old town hall and the famous "Gänseliesel" fountain.
>
> Built over 1,000 years ago as the village "Gutingi", Göttingen is home to a lot of historic churches and the romantic ruin of the Plesse Castle. A small piece of the city wall and a defencetower can also be visited.
>
> Don't forget to enjoy the beautiful countryside around Göttingen, with its wooded hills and footpaths. The floral beauty and also the small villages in the region with their farms are breathtaking.
>
> Discover the "Städtische Museum" with a great collection of musicinstruments. In addition there is the "Archäologisches Institut", which shows the artcollection of the university, the museum for Geologie and Paläontologie, the Völkerkundliche Sammlung and the Zoologisches Museum, presenting an exhibition of nature and animals with changing special exhibitions.
>
> Don't miss out on Göttingen's nightlife. The club- and discoscene is very famous in the region. Many people from other towns come here to enjoy the entertainment. A variety of bars, restaurants, cafés, discos, and nightclubs offers something for everyone. From business class to primitive pubs [sic!] for students.
>
> If you want to stay here you've got wide range of accommodation to choose from: a cheap youth hostel or a nobel hotel. If you want to have a wellness/sport hotel, the "Freizeit Inn" is a must. It’s the third best wellness hotel in Germany.
>
> To sum it up, Göttingen is worth a trip.
> (Rohrbach 2003: 388)

In spite of the remaining errors, the text shows a remarkably high degree of appropriate word choices and idiomatic phrases for a class 9 student. It also represents the typical structure of a travel brochure text. The sample text thus illustrates that students gain from the genre approach to language teaching both at the level of lexis and grammar (in terms of idiomaticity) and at the level of text design. The genre-based approach is a good example of a corpus-based method which is not just a playful add-on to existing and traditional teaching methods at a fairly advanced level of language learning but an innovative way of increasing learners' language and genre competence at a much earlier level: corpus-based methods of language learning thus help to achieve a wide range of language-related goals that are specified in ELT curricula. One can only hope that future curricula will take the full potential that DDL and other corpus-based activities provide into account to a much larger extent than has been the case so far.

4 Using learner corpora

A fairly recent research area in corpus linguistics is the compilation and analysis of learner corpora, which Granger (2002) defines as follows:

> Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance. (Granger 2002: 7)

Among others, Granger (2004), Meunier (2002) and Nesselhauf (2004) show that learner corpus research has a lot to offer to ELT materials designers and language teaching professionals, for example because learner corpora provide an empirical basis for the identification of frequently occurring mistakes at various stages of the language learning process. The largest learner corpus of English which has already been completed is the International Corpus of Learner English (ICLE, cf. Granger et al. 2002, Nesselhauf, this volume). It includes 2 million words of learner essays produced by advanced learners of English with different mother-tongue backgrounds (e.g. German, French). The spoken counterpart is the Louvain International Database of Spoken English Interlanguage (LINDSEI, cf. de Cock et al. 2003), which is being compiled at present and which will include spoken English produced by advanced learners of

English with different mother tongues in standardised interview situations.[5] Both ICLE and LINDSEI have been coordinated by Sylviane Granger and her staff at the Centre for English Corpus Linguistics of the University of Louvain-la-Neuve, which has been in the vanguard of learner corpus research world-wide. Another spoken learner corpus, which is similar in size, is the Giessen Long Beach Chaplin Corpus (GLBCC, cf. Jucker et al. 2003), a corpus of oral narratives and conversations between two students at a time on the silent Chaplin movie *The Immigrant*. The EFL component of this corpus includes 95,555 words of spoken English produced by advanced learners of English.

The aforementioned corpora can be labelled 'reference learner corpora' because they have been designed in such a way that they are representative of what Granger (1998: 7) calls the 'archetypal learner', i.e. an abstract learner with an average level of language competence. The analysis of reference learner corpora and its comparison with comparable native corpora offer important quantitative and qualitative insights into the extent to which learners of English at a certain stage of the learning process have already approximated to the native-speaker norm and where they still deviate from the target model.[6] For example, Lorenz (1999) analyses adjective intensification in learner English by comparing the German component of ICLE and comparable native data. He groups the deviances from the native norm that he finds in learner English into four categories:

1. Overuse: Learners use linguistic forms significantly more frequently than native speakers (e.g. the pattern *'really* + adjective' as in *really big*, *really important*, *really interesting*).
2. Underuse: Learners use linguistic forms significantly less frequently than native speakers (e.g. the pattern *'particularly* + adjective' as in *particularly difficult*, *particularly useful*).
3. Misuse: Learners use English forms wrongly, e.g. due to interference (e.g. the overextension of the pattern 'maximizer + adjective' to non-gradable adjectives as in ?*absolutely silly*, ?**totally damaged*).
4. Learner-idiosyncratic forms: Learners use linguistic forms and structures that do not exist in English (e.g. the non-existent pattern '*a* + *too* + adjective + noun' as in **a too complex problem*).

---

[5] The German component of LINDSEI, including spoken language produced by advanced German learners of English, has already been completed and will be published together with other national components in 2006 (cf. Brand and Kämmerer, this volume).

[6] The underlying assumption here is that native-like usage is the relevant target norm in the ELT classroom (cf. Mukherjee 2005). However, it should be noted that a specific line of learner-corpus research, the aim of which is the compilation and analysis of corpora of English as a lingua franca in non-native settings, is based on the assumption that the native norm no longer provides the only possible target model in ELT (cf. e.g. Jenkins 2004).

Lorenz's (1999) study is representative of many learner-corpus studies in which the focus is on the description of the gap between the native target norm and representative learners' actual language use, e.g. Nesselhauf's (2005) analysis of collocations in learner English. Other studies take into account the language that is used in pedagogical material, e.g. ELT textbooks, and discuss to what extent they reflect native-like usage, cf. e.g. Römer's study (2005) of the English progressive.

While reference corpora are no doubt valuable resources for the overall and supra-individual comparison of interlanguage and native usage, actual teachers in real ELT classrooms might also be interested in their own students' output and, thus, in the compilation and analysis of 'local' learner corpora, including their own students' interlanguage. As Seidlhofer (2002: 220) notes, "FL pedagogy, and presumably any pedagogy, has to be local, designed for specific learners and settings." By compiling a local learner corpus, teachers are provided with a powerful resource for systematic error analysis. Mukherjee and Rohrbach (2006) report on a pilot project in which a small local learner corpus consisting of 32,000 words of written examinations was compiled at Hainberg-Gymnasium Göttingen. This corpus, the Giessen-Göttingen Local Learner Corpus of English (GGLLC), consists of two subcorpora: (1) the complete set of papers of a written examination in class 12, which took place in 2003; (2) the complete set of papers of a written examination in the same group one year later, i.e. in class 13. As Mukherjee and Rohrbach (2006) show, such a corpus can be compiled in a "quick and dirty" way, easily stored in three different formats (plain texts as produced by the students, texts with the teacher's correction marks, texts with correction marks and the teacher's corrected versions), and used for various language-pedagogical ends:

- By using corpus-linguistic software like *WordSmith Tools*, the teacher can analyse the range of general and topic-related vocabulary that students in general or individual students have used.
- The corpus makes it possible to focus on individual learners' interlanguage and to provide them with tailor-made feedback.
- The corpus allows for longitudinal studies of learner language progression across time (both for the entire class and for individual learners). Thus, this local learner corpus has a monitoring quality.
- The corpus can be analysed both by teachers and by students. For example, individual learners can use a concordance display of their own mistakes as a starting point for data-driven learning activities.

The compilation and analysis of truly local learner corpora is still in its infancy. Besides opening new ways of describing and evaluating students' interlanguage

output, the localisation and individualisation of learner corpus research as envisaged here will have positive effects:

> [F]irstly, the focus on their own students' output will involve many more teachers in corpus-based activities and […], secondly, the exploration of learner data by the learners themselves will motivate many more learners to reflect on their language use and thus raise their foreign language awareness. (Mukherjee and Rohrbach 2006: in press)

Learner corpus research is an area which is immediately relevant to the teaching and learning of languages but its full potential has not yet been exploited in language pedagogy. The compilation and analysis of local learner corpora is but one promising and enticing avenue for future learner corpus research. Another area is the empirical comparison of interlanguages of learners with different mother tongues, which is made possible by corpora like ICLE and LINDSEI – this field of research has recently been labelled 'Contrastive Interlanguage Analysis' (CIA, cf. Granger 2002).

5 Concluding remarks

By way of exemplification, the present paper has sketched out the wide range of language-pedagogical applications of corpus-based research. While it is beyond the scope of the present article to give an exhaustive overview of the corpus-linguistic influence on language teaching and the full extent of corpus-based activities, there is no doubt that corpora can be used for various ends in language pedagogy: (1) for dictionaries and other materials for the ELT classroom; (2) as a database and tool in the ELT classroom itself; (3) as representative samples of learner language. Learner corpus research in particular has a much greater potential to offer for the corpus-informed classroom of the future than has been realised so far.

In spite of the undeniably large number of corpus-based activities that have been suggested by researchers in applied corpus linguistics, it cannot be ignored that 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. This 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 contexts and both from teachers' and learners' perspectives. Götz and Mukherjee (this volume) report on a language-pedagogical case study carried out in the context of a linguistic

seminar at the University of Giessen in which advanced learners evaluated the advantages and disadvantages of various corpus-based activities and DDL methods. While the results do not provide any conclusive answers, they clearly show that once learners become familiarised with corpora, they tend to find corpus work and corpus-based DDL activities both interesting and beneficial to their own learner language. Many more case studies are needed in order to get a more comprehensive and realistic picture – from the consumer end, as it were – about the benefits of corpus-based language learning.

List of References

Aston, G. (ed.) (2001): *Learning with Corpora*. Houston, TX: Athelstan.

Aston, G., S. Bernardini & D. Stewart (eds.) (2004): *Corpora and Language Learners*. Amsterdam: John Benjamins.

Bernardini, S. (2004): "Corpora in the classroom: an overview and some reflections on future developments", *How to Use Corpora in Language Teaching*, ed. J. Sinclair. Amsterdam: John Benjamins. 15-36.

Biber, D. (1993): "Representativeness in corpus design", *Literary and Linguistic Computing* 8, 243-257.

Biber, D., S. Conrad & R. Reppen (1998): *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999): *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Bolton, K. (2000): "The sociolinguistics of Hong Kong and the space for Hong Kong English", *World Englishes* 19(3), 265-285.

Collins COBUILD English Language Dictionary (1987). London: Collins.

Collins COBUILD English Grammar (1990). London: Collins.

De Cock, S., S. Granger & S. Petch-Tyson (2003): "The Louvain International Database of Spoken English Interlanguage – LINDSEI", available at <http: //www.lftr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/lindsei.htm>, accessed 13 Jul 2003.

Flowerdew, L. (2001): "The exploitation of small learner corpora in EAP materials design", *Small Corpus Studies and ELT: Theory and Practice*, ed. M. Ghadessy, A. Henry & R.L. Roseberry. Amsterdam: John Benjamins. 363-379.

Francis, W.N. (1982): "Problems of assembling and computerizing large corpora", *Computer Corpora in English Language Research*, ed. S. Johansson. Bergen: Norwegian Computing Centre for the Humanities. 7-24.

Gavioli, L. (2001): "The learner as researcher: introducing corpus concordancing in the classroom", *Learning with Corpora*, ed. G. Aston. Houston, TX: Athelstan. 108-137.

Ghadessy, M., A. Henry & R.L. Roseberry (eds.) (2001): *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins.

Granger, S. (1998): "The computer learner corpus: a versatile new source of data for SLA research", *Learner English on Computer*, ed. S. Granger. London: Longman. 3-18.

Granger, S. (2002): "A bird's eye view of learner corpus research", *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, ed. S. Granger, J. Hung & S. Petch-Tyson. Amsterdam: John Benjamins. 3-33.

Granger, S., E. Dagneaux & F. Meunier (eds.) (2002): *International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. (2004): "Computer learner corpus research: current state and future prospects", *Applied Corpus Linguistics: A Multidimensional Perspective*, ed. U. Connor & T. Upton. Amsterdam: Rodopi. 123-145.

Henry, A. & R.L. Roseberry (2001): "Using a small corpus to obtain data for teaching a genre", *Small Corpus Studies and ELT: Theory and Practice*, ed. M. Ghadessy, A. Henry & R.L. Roseberry. Amsterdam: John Benjamins. 93-133.

Hunston, S. (2002): *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jenkins, J. (2004): "ELF at the gate: the position of English as a lingua franca", *The European English Messenger* 13(2), 63-69.

Johns, T. & P. King (eds.) (1991): *Classroom Concordancing*. Birmingham: University of Birmingham.

Jucker, A.H., S.W. Smith & T. Lüdge (2003): "Interactive aspects of vagueness in conversation", *Journal of Pragmatics* 35, 1737-1769.

Lorenz, G. (1999): *Adjective Intensification - Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Macmillan English Dictionary: School Edition for Advanced Learners (2002). Hannover: Schroedel.

Meunier, F. (2002): "The pedagogical value of native and learner corpora in EFL grammar teaching", *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, ed. S. Granger, J. Hung & S. Petch-Tyson. Amsterdam: John Benjamins. 119-141.

Mindt, D. (2000): *An Enpirical Grammar of the English Verb System*. Berlin: Cornelsen.

Mindt, D. (2002): "What is a grammatical rule?", *From the COLT's Mouth ... and others': Language Corpora Studies – in Honour of Anna-Brita Stenström*, ed. L.E. Breivik & A. Hasselgren. Amsterdam: Rodopi. 197-212.

Müller, S. (2004): *Discourse Markers in Native and Non-native English Discourse*. University of Giessen: PhD thesis.

Mukherjee, J. (2002): *Korpuslinguistik und Englischunterricht: Eine Einführung*. Frankfurt am Main: Peter Lang.

Mukherjee, J. (2004a): "The state of the art in corpus linguistics: three book-length perspectives", *English Language and Linguistics* 8(1), 103-119.

Mukherjee, J. (2004): "Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany", *Applied Corpus Linguistics: A Multidimensional Perspective*, ed. U. Connor & T. Upton. Amsterdam: Rodopi. 239-250.

Mukherjee, J. (2005): "The native speaker is alive and kicking: linguistic and language-pedagogical perspectives", *Anglistik* 16(2), 7-23.

Mukherjee, J. & J. Rohrbach (2006): "Rethinking applied corpus linguistics from a language-pedagogical perspective: new departures in learner corpus research", *Planing and Gluing Corpora: Inside the Applied Corpus Linguist's Workshop*, ed. B. Kettemann & G. Marko. Frankfurt am Main: Peter Lang.

Nesselhauf, N. (2004): "Learner corpora and their potential for language teaching", *How to Use Corpora in Language Teaching*, ed. J. Sinclair. Amsterdam: John Benjamins. 125-152.

Nesselhauf, N. (2005): *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Römer, U. (2005): *Progressives, Patterns, Pedagogy: A Corpus-driven Approach to English Forms, Functions, Contexts and Didactics*. Amsterdam: John Benjamins.

Rohrbach, J. (2003): "'Don't miss out on Göttingen's nightlife': Genreproduktion im Englischunterricht", *Praxis des Neusprachlichen Unterrichts* 50, 381-389.

Schmied, J. (1999): "Applying contrastive corpora in modern contrastive grammars: the Chemnitz Internet Grammar of English", *Out of Corpora: Studies in Honour of Stig Johansson*, ed. H. Hasselgård & S. Oksefjell. Amsterdam: Rodopi. 21-30.

Scott, M. (2005): *WordSmith Tools: Version 4.0*. Oxford: Oxford University Press.

Seidlhofer, B. (2002): "Pedagogy and local learner corpora: working with learning-driven data", *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, ed. S. Granger, J. Hung & S. Petch-Tyson. Amsterdam: John Benjamins. 213-234.

Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (ed.) (2004): *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Stubbs, M. (1993): "British traditions in text analysis: from Firth to Sinclair", *Text and Technology: In Honour of John Sinclair*, ed. M. Baker, G. Francis & E. Tognini-Bonelli. Amsterdam: John Benjamins. 1-33.

Stubbs, M. (1995): "Collocations and semantic profiles: on the cause of the trouble with quantitative studies", *Functions of Language* 2(1), 23-55.

Tribble, C. (2000): "Practical uses for language corpora in ELT", *A Special Interest in Computers: Learning and Teaching with Information and Communications Technologies*, ed. P. Brett & G. Motteram. Whitstable, Kent: IATEFL. 31-41.

Tsui, A.B.M. (1996): "The participant structures of TeleNex: a computer network for ESL teachers", *International Journal of Educational Telecommunications* 2(2/3), 171-197.

Tsui, A.B.M. (2004): "What teachers have always wanted to know – and how corpora can help", *How to Use Corpora in Language Teaching*, ed. J. Sinclair. Amsterdam: John Benjamins. 39-61.

Ungerer, F. (1999): *Englische Grammatik heute*. Stuttgart: Ernst Klett.

Widdowson, H.G. (1990): *Aspects of Language Teaching*. Oxford: Oxford University Press.

The present paper provides a selected overview of the state of the art in corpus-informed language pedagogy. Starting off from a general assessment of the im-pact that the corpus revolution has already had on English language teaching (ELT), the focus of the main part of this paper is on some typical examples of corpus use in three language-pedagogically relevant areas: (1) using corpora for ELT (e.g. producing learner dictionaries); (2) using corpora in. the ELT class-room (e.g. in data-driven learning); (3) using learner corpora.Â From its very beginnings, modern corpus linguistics has been intricately intertwined with the development of computers and software programs for corpus analysis. As Biber et al. 112 I. Origin and history of corpus linguistics Ïª corpus linguistics vis-a`-vis other disciplines. 7. Corpora and language teaching. 1. Introduction 2. Indirect applications of corpora in language teaching 3. Direct applications of corpora in language teaching 4. Tasks for the future 5. Concluding remarks 6. Literature. 1. Introduction. 1.1. Corpus linguistics and language teaching.Â Indicative of the popularity of pedagogical corpora use and the need for research in this area is the con-siderable number of books and edited collections Ïª some of which are the result of the successful â€œTeaching and Language Corporaâ€ (TaLC) conference series Ïª that have recently been published on the topic of this article or which bear a close relationship to it. Corpus linguistics is a new branch of linguistics but its status is still debatable - either as a theory or a methodology. This article aims to give an overview of the different approaches and perspectives of corpus linguistics. The neo-Firthians contend that corpus linguistics is a method, while other prominent corpus linguists claim that it is a theory. Other corpus linguists believe that corpus linguistics can be both a methodology as well as a theory depending on the extent and purposes it is used for. The applicability of corpus linguistics as a methodology is observed in English Language Teaching and Learning (ELT).