ReCALL **17** (1): 47–64. © 2005 Cambridge University Press
DOI: 10.1017/S0958344005000510 Printed in the United Kingdom

From pedagogically relevant corpora to authentic language learning contents

SABINE BRAUN
Englisches Seminar, Universität Tübingen, Wilhelmstraße 50,
D-72074 Tübingen, Germany
(*email: sabine.braun@uni-tuebingen.de*)

Abstract

The potential of corpora for language learning and teaching has been widely acknowledged and their ready availability on the Web has facilitated access for a broad range of users, including language teachers and learners. However, the integration of corpora into general language learning and teaching practice has so far been disappointing. In this paper, I will argue that the shape of many existing corpora, designed with linguistic research goals in mind, clashes with pedagogic requirements for corpus design and use. Hence, a 'pedagogic mediation of corpora' is required (cf. Widdowson, 2003). I will also show that the realisation of this requirement touches on both the development of appropriate corpora and the ways in which they are exploited by learners and teachers. I will use a small English Interview Corpus (ELISA) to outline possible solutions for a pedagogic mediation. The major aspect of this is the combination of two approaches to the analysis and exploitation of a pedagogically relevant corpus: a corpus-based and a discourse-based approach.

1 Introduction

Language corpora and the methods used for their exploitation have great potential for language learning: corpora provide systematic access to naturally occurring language,[1] and corpus-linguistic methods support exploratory learning and are well-suited to encourage autonomous learning and teaching. It is not surprising therefore that talking about using corpora for language learning purposes has become somewhat trendy. Apart from dedicated conferences such as Teaching and Learning with Corpora (TALC) and Practical Applications in Language Corpora (PALC), many general conferences in the field, including EUROCALL, have established corpus strands or sections which enjoy increasing popularity (cf. Berglund & Chambers, 2004).

[1] The related question as to whether the language contained in corpora is authentic or real has given rise to an ongoing debate, which is well summarised in Seidlhofer (2003: 77–123). This point will be discussed in more detail in section 3.

Moreover, the vast number of recent publications on the use of corpora in language learning and teaching can be taken as an indicator that a lot is happening in this area. Topics range from elaborations of the concordance-based method of data-driven learning originally developed by Tim Johns[2] (e.g. Cobb, 1997; Flowerdew, 1996, 2001; Hadley, 2002; Kettemann, 1995) to the design and exploitation of Languages for Specific Purposes (LSP) and/or genre-specific corpora (e.g. Henry & Roseberry, 2001; Lee, 2002; Tribble, 2001, 2004) and the more recent attempts to integrate spoken corpora into language learning and teaching (e.g. Mauranen, 2004; Simpson, Lucka & Ovens, 2000).

The insights into language use which corpora offer seem to leave no doubt that they hold great potential both as a resource for the creation of rich and interesting learning materials and for direct exploitation by learners. The data they provide is:

- realistic, showing language in real use;
- rich, providing more (and more diversified) information than dictionaries or reference grammars can;
- illustrative, providing actual patterns of use instead of abstract explanations;
- up-to-date, revealing trends in language use and evidence for short-term historical change (as shown, for example, by Mair, Hundt, Leech & Smith, 2002).

However, a careful look around the many different places where languages are learnt and taught makes clear that corpora, while being the ‘buzzword’ in language research departments, are still far from being part of mainstream teaching practice, if not *terra incognita* altogether. This raises the question why the real uptake of potentially so valuable a resource has been rather limited to date. To find an explanation and to initiate a change, we need to first of all take a closer look at the nature of existing corpora.

2 Existing corpora in the light of language learning and teaching

The number of corpora available for different languages varies considerably. As far as English is concerned, over the past decade a number of very large corpora have emerged (cf. Kennedy, 1998 for an overview of English language corpus development). Multi-million-word corpora such as the British National Corpus (BNC, 100 million),[3] the Bank of English (450 million) or the International Corpus of English (ICE, 20 million) are now widely available and can be run on an ordinary PC or accessed via the Web. This potentially opens them up to a broad user group including language learners and teachers.

Indeed, if we compare the present situation with that described by McEnery and Wilson (1997a) in a seminal article in this field, we can note that the overall conditions for corpus use in language learning and teaching have improved: the increased number of English corpora, their accessibility on different platforms (Windows, Unix, Web) and the availability of enhanced corpus retrieval software over the past years have made it easier for learners and teachers of English to exploit corpora. However, as has been

[2] cf. Johns 1986, 1991a, 1991b; the term “data-driven learning” was introduced by Johns (1991a: 2).

[3] References to all corpora mentioned in the text can be found in Appendix 1.

pointed out by McEnery and Wilson and others, we should recall that corpora are a type of data which were originally designed for the purposes of linguistic research. Some of the design principles established by corpus linguists prevail in many of the more recent attempts to create pedagogically relevant corpora. Therefore it comes as no surprise that learners and teachers still face some fundamental problems in using corpora. In the following, I will first discuss some of the 'traditional' design principles and then look at the implications for the use of corpora by learners and teachers.

Size has always been an important issue in corpus design, which shows very clearly that design decisions are driven by the goals of prospective users: the trend towards mega-sized corpora makes a lot of sense in many areas of linguistic research. Lexicography is a good example here: only a large corpus can provide enough instances of each word in a language, enabling the linguist or dictionary maker to study also the less frequently used words. As Aston has pointed out, from a language learning perspective "the virtues of large corpora seem less readily apparent". A corpus size of 20,000 to 200,000 words is regarded to be sufficient or even preferable in the learning context (Aston 1997: 54).[4] I will return to this point below.

Another issue is *content*: corpora are large collections of text that are usually designed to be representative of a language (or of a particular variety). This again is perfectly in line with the needs of many language researchers. Large corpora such as the BNC and the Bank of English – so-called balanced corpora – include texts from a wide range of genres, and the individual texts are taken from as varied a range of topics and sources as possible. In other words: intertextual coherence is not important or may not even be desirable. As a consequence, only statistical counts and 'vertical' reading (e.g. looking at frequency lists and KWIC concordances) make sense for those corpora. No-one – except the corpus designers – would actually read the individual texts.

While many of the larger corpora contain language from both written and spoken genres, the *data format* is usually text only, taken directly from written samples or transcribed from spoken ones. The option to include audiovisual materials in order to represent spoken language has to date only been used in a handful of smaller and more recent corpora such as the MICASE corpus, the Santa Barbara Corpus of Spoken English, the IViE corpus and the VOICE corpus. The alignment between the audiovisual data and the transcripts still presents some methodological questions. From a linguistic or corpus analysis point of view, the transcribed version has the clear advantage that it can be searched with corpus retrieval tools. However, transcribing spoken data usually implies some loss of information. To keep it to a minimum, conversational and/or prosodic mark-up has been introduced in some of the spoken corpora (e.g. in the London-Lund corpus, the IViE corpus and the Freiburg LeaP corpus).

More generally speaking, many corpora based on written and transcribed spoken texts have been 'enriched' with additional information by various types of *annotation*. In its minimal form this is often just a simple mark-up, similar to html, indicating aspects of document structure such as sentence and paragraph boundaries. Linguistically more interesting types of mark-up are annotations at morpho-syntactic level (part-of-speech tagging, annotation of phrase and clause structures), and at semantic and discourse level (word-sense tagging, anaphoric relationships). In accordance with many linguistic

[4] But see Bernardini (2000) for proposals on successful large-corpus use by learners.

research questions, annotation is usually aimed at complete coverage of *all* relevant instances, e.g. all parts of speech in the corpus. Many efforts have therefore been made to automise annotation processes wherever possible.

In accordance with their primary target group, namely language researchers, corpora seem to fulfil their purposes, and it just seems to be a matter of time until problems relevant to them, e.g. better algorithms for an automatic annotation beyond part-of-speech tagging, or questions relating to multi-layer annotation within richly annotated corpora will be solved.

Having said that, when it comes to using corpora in the learning and teaching context, the parameters described above appear in a different light. First of all, the sheer size of many available corpora and the nature of their content make them difficult to handle for most teachers and learners, even with appropriate hardware and easy-to-use retrieval software. To start with, a large corpus can easily produce too many results, namely hundreds of concordance lines for frequent words. Moreover, as Meunier (2002: 129) points out, corpus results can be 'messy', ambiguous or misleading. The 'messiness', it could be argued, is part of using real-language materials, and there is certainly nothing wrong with removing some unclear or unwelcome lines from a concordance before presenting it to learners. The problem, however, is that the evaluation of corpus search results can be difficult without a study of the wider co-text.[5] This, in turn, proves to be time-consuming when the results are large in number and, in addition, come from a wide variety of entirely different sources which the teacher cannot be expected to be familiar with.

Over the past years, a number of alternatives have been suggested to overcome the problems of size as well as the problems relating to the diversity of content: they include the use of subcorpora derived from large corpora (cf. Aston, 2002), the use of small genre-specific corpora (cf. Henry & Roseberry, 2001; Tribble, 2001, 2004 and others), of LSP corpora (cf. Gavioli, 2002; Lee, 2002 and others) and of corpora created by teachers and learners themselves (cf. Aston, 1997, 2002; Tribble, 1997). Tribble refers to the ad hoc creation of corpora by learners and teachers as "quick-and-dirty solutions" (1997: 109) which may not be ideal by the standards of corpus creation but serve a particular learning purpose. Aston calls them "home-made corpora" and points out that they "may be more appropriate for learning purposes than pre-compiled ones, insofar as they can be specifically targeted to the learner's knowledge and concern." (2002: 9). Gavioli and Aston (2001) suggest that learners start with smaller corpora, which are more limited in variety, and then move on to larger, more varied corpora in order to develop autonomy gradually.

Another point concerns existing corpus annotation schemes. Designed to support complex linguistic queries, they serve the needs of descriptive linguistics and are of little use to most teachers and learners. One simple observation is that common tag sets for part-of-speech tagging include up to 150 different tags such as the UCREL tag set CLAWS 7. They are too specific for the non-linguist. Of course, one could argue that it is always possible to tag any corpus with a simpler tag set, but realistically it has to be acknowledged that learners and teachers are usually not in a position to do this. More

[5] The term 'co-text' is used here to refer to the surrounding text, e.g. a sentence or paragraph. The term 'context', which is sometimes used for this, is in this paper reserved to refer to the communicative situation.

importantly though, even if they were, it is doubtful that the established ways of e.g. morpho-syntactic annotation would be the most helpful basis for pedagogically motivated corpus queries.

Similarly, prosodic mark-up certainly helps to interpret spoken language, which is especially important in language learning and teaching, where the focus has shifted towards spoken language. However, the mark-up may be difficult to decipher for non-linguist language learners.[6] The integration of audio-visual data in corpora will definitely prove to be a very useful extension for language learners and teachers (cf. also McEnery & Wilson, 1997b).

At present, however, the above discussion may be summarised thus: the wide availability of corpora has facilitated access for a broad range of users, including language teachers and even learners. A process of reflection on pedagogic needs in terms of size and content has set in, in part because it has become technologically possible for teachers and learners to create their own corpora. In other areas, especially with regard to data format and annotation, many corpora – large and small – still fail to comply with pedagogic needs.

When we look at the use of corpora in learning and teaching from a wider perspective though, we can see that the above characterisation provides an incomplete picture. What is largely overlooked is that there is more to the pedagogic use of corpora than the ‘right’ design. One very important aspect is the question as to how corpora should actually be used and exploited in the language learning and teaching context.

3 Using corpora in language learning and teaching

The shortcomings of the existing large corpora with regard to pedagogic requirements do not mean that they have no pedagogical value at all. On the one hand, teachers and learners can benefit from the increasingly available corpus-based descriptions of language (e.g. Aijmer & Altenberg, 1991, 2004; Biber, Conrad & Reppen, 1998; Hasselgard & Oksefiell, 1999; Leistyna & Meyer, 2003; Mair & Hundt, 2000; Sinclair, 1991; Thomas & Short, 1996; Wilson, Rayson & McEnery, 2003 and many others). In addition, corpus-based analyses of English have influenced the design of the traditional syllabus (e.g. Mindt, 1996, 1997), the contents of reference works (recently, for example, the Longman Grammar of Written and Spoken English, cf. Biber, Johansson, Leech, Conrad & Finegan, 1999) and the contents of teaching resources (recently, for example, Kennedy, 2003). These are, of course, very indirect uses of corpora, resulting from the general influence of linguistic research on Applied Linguistics. It is not too surprising that the ‘corpus revolution’ in linguistics has helped to shape and inform methods and materials for language learning and teaching.

On the other hand, many fruitful ways of exploiting existing large and small corpora more directly have been suggested. These can range from uses by teachers (e.g. using corpora as a base for material creation, test design, feedback and evaluation references) to direct exploratory uses of corpora by learners themselves (for reviews and systematisation cf. Aston, 1995, 1997, 2001; Fligelstone, 1993; Flowerdew, 1996; Leech 1997;

[6] Different and more differentiated observations on the use of annotated corpora have been reported by Ulrike Gut (personal communication) from her experience with the Freiburg LeaP corpus.

McEnery & Wilson, 1993, 1997a; Mukherjee, 2002). This work has produced a wealth of highly interesting, innovative and relevant language learning activities. However, some caveats are in order: firstly, quite a number of the suggestions have been made on a hypothetical basis and have yet to be tested in practice. Secondly, the empirical/practical work that has been carried out has largely been restricted to the university environment. It has, for example, involved students in modern language departments, who have a linguistic background (e.g. Bernardini, 2000) or students in English for Specific Purposes (ESP) courses, who have specific subject-matter knowledge (e.g. Gavioli, 2002). Thirdly, it has usually focussed on specific aspects of language learning, e.g. vocabulary acquisition or specific aspects thereof (e.g. Cobb, 1997).

Even more importantly, the focus lies very strongly on concordance-based materials and activities. Certainly, these have brought back to language learning and teaching a 'healthy' focus on form, which, as Leech (1997: 22) puts it, had "passed into relative oblivion through the influence of the communicative teaching methodology". To a certain extent though, this kind of corpus exploitation is rooted in a tradition of what Widdowson (1980) terms 'linguistics applied', ie of taking linguistic findings more or less directly 'to the classroom'. In line with this, the discussion about the 'right' use of corpora not infrequently just revolves around the question as to whether corpus data should be 'filtered' by the teacher or whether learners should have 'free' access to the corpora themselves.

I would like to take a different perspective on the use of corpora in language learning and teaching here – an 'Applied Linguistics' perspective in Widdowson's sense. If we look at corpora in terms of the overall goal of language learning, we are faced with a well-known discrepancy: according to Widdowson's (1979) helpful distinction between *text* and *discourse*, which he reiterated with regard to corpora in (2003), corpora are a collection of *text*, ie *products* of language use isolated from any communicative situation. In contrast, language learning is concerned with *discourse*, ie the use of language in concrete communicative situations. More specifically, language learning is concerned with the development of the knowledge and skills required to master the *processes* of producing and understanding discourse in a foreign or second language.

What is important here and what has been asserted by Relevance Theory (Blakemore, 1992; Sperber & Wilson, 1995) is that we do not perceive a communicative situation directly but that we construct a context in our mind, drawing on our perceptual abilities, our knowledge about the communicative situation in question, our previous experience with it, our attitudes towards it, our background knowledge as well as textual clues (including co-text) and other factors (cf. Blakemore, 1992). If communication is to be successful, a relevant context has to be constructed by the discourse participants (cf. Sperber & Wilson, 1995).

Against this backdrop it becomes clear why learning and teaching with corpora creates a number of problems that need to be addressed from a pedagogic or applied linguistics perspective: as Widdowson (2003) points out, one of them is that the users of a corpus are not the original addressees of the texts in the corpus, hence they are isolated from the original discourses, but understanding can only be achieved when the discourse which gave rise to a text can be reconstructed. While our ability to create contexts in our mind normally enables us to comprehend many artefacts of language – texts as well as concordances – even if we did not participate in the original discourse, this proves to be much more difficult for a learner. In our native culture and language, we usually manage to extrapolate an appropriate context from the textual clues (bottom-up processing) and our background knowledge (top-down processing, cf. Brown & Yule, 1983). Learners, in contrast, are in quite a different situation: they are less likely to share the cultural background assumptions of the L2 culture. And, of course, they are less well able to extrapolate from textual clues.

Such difficulties have led Widdowson (1990, 2003) to also put forward the following well-known argument, which is important in our context: based on his (1978) distinction between the *genuineness* of texts and the *authenticity* of discourse, he claims that the use of real-language texts ('genuine instances of language use', 1978: 80) in learning and teaching does not yet provide a guarantee for successful *authentication* by the learner. Real-language texts, therefore, are only useful insofar as the learner is able to *authenticate* them, ie to create a relationship to the texts.

Discourse authentication is seen as a key element for learner motivation and eventually for learning success. Isolation from the original discourse is likely to make authentication more difficult but not impossible – as long as learners are able to construct an appropriate context in their mind. Context construction heavily relies on subjective knowledge. Therefore, each learner will construct their own individual context, and the construction process will be greatly facilitated if the topic is familiar to, and interesting for, the learner.

With this in mind, the 'anonymous' mass of *genuine* materials in corpora which have not been collected in accordance with pedagogical considerations does seem to create some problems for authentication (cf. section 1, cf. also Tribble, 1997; Widdowson, 2003).[7] From an applied linguistics perspective, the crucial question then is how we can enable learners and teachers to move from text to discourse when using corpora. In the light of the above discussion, it should be clear that the answer needs to be given at various levels: on the one hand, it concerns the contents of a corpus as well as its design in terms of size, data format and annotation. On the other hand, it also relates to the overall approach which is taken to analyse and exploit the data or, more specifically, the question of how corpus techniques should be embedded and complemented by other methods. Furthermore, it raises the question as to how the corpus materials can be 'enriched' by additional, pedagogically relevant materials. In the following I will briefly discuss each of these points.

A first requirement concerns the *content* of a pedagogically relevant corpus. The content should support the authentication process in the following way: the choice of texts should make it easy to prepare and motivate learners and teachers for the kind of language and communicative situations they can expect, before they start exploring individual texts. On the one hand, this implies that the materials should be relevant for the needs of the target group. On the other hand, it requires a 'coherent' corpus, ie more homogeneity than the approaches underlying the creation of genre-specific corpora or LSP corpora mentioned in section 2. The kind of coherence we have in mind here is one that is brought about by a common overall theme around which all the texts in the corpus revolve. It will enable learners and teachers to establish a relationship to the

[7] Gavioli and Aston (2001) argue that the multitude of texts in a corpus will facilitate the authentication process, giving learners a choice of texts to choose from.

materials in the corpus that would hardly be achievable with any of the 'traditional' corpora.

Another point concerns the overall *methodological approach* taken to the creation of a pedagogically relevant corpus and to its exploitation by learners and teachers. Due to the vital role of discourse authentication, it will be fruitful to adopt an approach which combines two methods of analysis and exploitation to mutual benefit: The *corpus-based approach* (through frequency and KWIC analyses, concordances, ie 'vertical reading') clearly provides patterns of language use which do not reveal themselves easily when reading 'horizontally'. However, this approach needs to be complemented by a *discourse-based approach* (ie 'whole-corpus reading') which focuses on the analysis of linguistic means of expression in relation to their communicative (situational) and cultural embedding. A similar claim has been made by Henry and Roseberry, reflecting on the advantages of small corpora for language learning: "Small corpus analysis, like large corpus analysis, relies heavily on computer assistance to manipulate data and capture patterns that would otherwise be difficult or impossible to spot. But unlike large corpus analysis, it also relies heavily on expert judgements, based on whole-corpus reading and study" (Henry & Roseberry 2001: 99). This approach will be greatly facilitated by the type of coherence in the corpus content described above.

In the *exploitation* of a pedagogically relevant corpus, learners and teachers will best be able to authenticate the corpus materials when they start by studying a general corpus description and some of the texts in their entirety. This kind of familiarisation process will later help them to recognise and interpret corpus search results. This is in line with Gavioli's (1997) claim that it is much easier to interpret concordances when they come from known texts. Then learners and teachers can move on to analysing the language in the corpus in more detail, e.g. through studying characteristic means of expression for a particular topic in the corpus. On the basis of this, they can then use corpus techniques, e.g. to focus on means of expression or structures which they encountered in the text and find worth exploring. By so embedding the corpus techniques in a discourse-based approach, the pedagogic advantages of these techniques can be exploited more effectively.

In the *creation* of a pedagogically relevant corpus, a discourse-based approach will support the corpus developers in identifying not only relevant formal units such as individual means of expression, grammatical structures, larger text passages, but also the discourse functions they fulfil, the context-specific meanings they carry and the topics they relate to. At the same time, the discourse-based approach will help to capture the knowledge and background information which is required for understanding. Together, this will support the identification of potential problem areas for specific learner groups.

Moreover, this type of corpus analysis will be the main source for *pedagogic annotation*, which will, to a large extent, have to be manual and will often focus on units beyond concordance lines or sentences. The major task of the annotation is to support *pedagogically motivated* corpus queries. Relevant options include, for instance, a keyword-based search for topics in the corpus as a starting point for a detailed study of the characteristic means of expression related to it. This involves the identification and annotation of larger passages in the corpus, beyond a concordance line or sentence – a task that will hardly be feasible on the basis of corpus techniques alone.

Both the content-related requirements and the methodological approach to corpus development and exploitation suggested here provide arguments for the restriction of corpus *size* which go beyond the suggestions outlined in section 2. It is clear that the coherence requirement and especially the discourse-based approach, which involves whole-corpus reading, can only be realised with a small corpus.

With regard to the *format of the data* in the corpus, the inclusion of audiovisual materials for corpora of spoken language, mentioned in section 2, will be helpful because it will give learners and teachers an idea of the overall communicative situation in which the material was produced. Moreover, features of spoken language such as intonation have a direct impact on meaning and interpretation. As pointed out in section 2, it will be easier and 'livelier' for learners and teachers to work with sound and video files instead of relying on prosodic annotation alone.

In their study of the corpus, learners will need additional information, e.g. to bridge knowledge gaps, to learn about cultural, political and other aspects referred to in the texts and thus situate the producers of a text with regard to the positions they take vis-à-vis the issues they discuss. Moreover, learners will of course need facilities to look up word meanings and other issues. Learners using a corpus autonomously will also need guidance on how to make best use of it as well as exploratory tasks and opportunities to practise and test their knowledge. Teachers using the corpus as a resource for the creation of learning materials may wish to have pre-fabricated tasks and exercises.

One way to cover these issues is to integrate into the corpus complementary materials which are relevant for learners and/or teachers. Such a *pedagogic enrichment* is crucial in supporting the authentication process, ie in helping learners make the move from the text materials presented in the corpus to a real discourse situation. The materials should comprise comments and explanations, exploratory tasks and exercises, study aids and didactic hints for learners and teachers, to name just a few options. Ideally there should also be room for learners' or teachers' own enrichments (e.g. a learner's personal notes or a teacher's exploratory tasks for a specific learner group).

The points discussed in this section make it clear that a kind of *pedagogic mediation* of corpora (Widdowson, 2003) is required to guide the learners and teachers in the authentication process. As shown in this section, this mediation should be realised at the level of corpus content and design as well as at the level of corpus exploitation by learners and teachers. In the following section, I will introduce a small corpus which is currently being developed with these issues in mind.

4 ELISA: An English language interview corpus as a second-language application

The ELISA corpus is a collection of video-based interviews with English native speakers. It is currently being developed by the Applied English Linguistics group at the University of Tübingen.[8] It is intended as a resource for the creation of learning materials as well as for autonomous exploitation by learners. A demo of the corpus is available at www.corpora4learning.net, showing some of the audiovisual materials and some of the features outlined below.

The speakers in the ELISA corpus represent different varieties of English (currently US, British, Australian and Irish). The overall theme, which serves to create coherence,

[8] This work has partially been financed by a research grant from the University of Tübingen.

is their professional career (e.g. in banking, local politics, tourism, sport, the media, agriculture or environmental technology). We asked the speakers to talk about different topics, e.g. how they started their career or business, the kind of projects they are working on, their daily routines and future plans. This is why there is a high degree of intertextual coherence between the interviews.

The corpus currently contains fifteen interviews of five to fifteen minutes and amounts to about 60,000 words in total. It will be further expanded but will remain small. The interviews were recorded on video and transcribed to text. A method for the alignment of transcripts and videos was developed (see below).

The materials in the ELISA corpus support culture-embedded language learning. Moreover, they are particularly well-suited to fulfil the demand for communicatively relevant contents, because learning a language, and in particular the English language, is often done to acquire professional, vocational or somewhat 'technical' language. Hence the materials have an appeal for a broad range of learners from different backgrounds and in different environments, including university, school and other institutions.

The interviews have a narrative character. The speakers received a short briefing with regard to relevant interview topics beforehand. During the actual interview, they were interrupted as little as possible to ensure a free flow of natural speech.

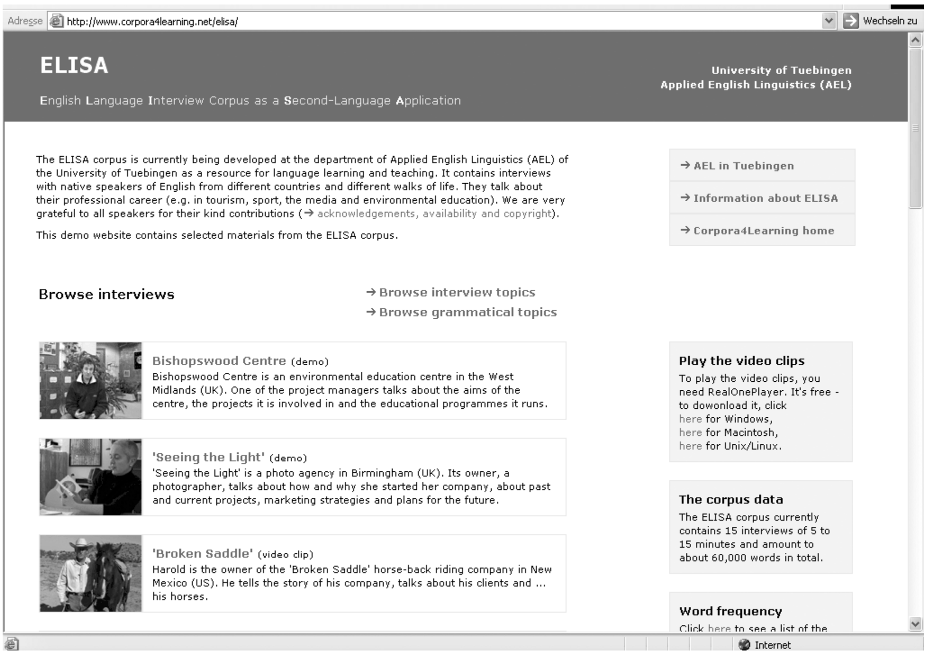
The corpus is currently analysed thematically, linguistically and functionally (e.g. with regard to relevant topics, means of expressions, structures, discourse and grammatical functions) to see what is most relevant in the language learning context.

This analysis will provide the basis for the development of annotation categories which will account for the needs of a pedagogic corpus use. Complementary to this bottom-up approach we will also develop top-down categories: the starting point for those is what learners may be interested in, e.g. depending on their native language, their personal requirements and profiles. This is in line with the factors which have been identified in section 3 as supporting the authentication process. The ELISA corpus will be annotated with regard to:

- content-related categories (topics, keywords),
- L2-related categories (ie lexical, grammatical, pragmatic and discourse properties),
- learner-related categories (level of proficiency, relevant knowledge requirements, skills which can be practised, challenges and difficulties).

The annotation will be used in two ways: on the one hand, it will support the presentation and retrieval of the corpus data. On the other hand, it will provide the basis for attaching the enrichments envisaged for the corpus, including the audiovisual materials.

The corpus will be available via the Web. Access to the corpus materials will be possible in the following ways: the start page (see Figure 1) provides an overview including a short overall description of the corpus, summaries of each interview and an index of all interviews, corpus topics such as 'how did you start your business', 'current projects', 'daily routines', and linguistic functions which are worth exploring, for example 'tenses in contrast', 'habitual past', showing by which means of expression they are realised. From these indices, the user has direct access to each interview, topic and function. The interviews are displayed with a division into topic-based subsections (see Figure 2)



http://www.corpora4learning.net/elisa/

ELISA

English Language Interview Corpus as a Second-Language Application

University of Tuebingen
Applied English Linguistics (AEL)

The ELISA corpus is currently being developed at the department of Applied English Linguistics (AEL) of the University of Tuebingen as a resource for language learning and teaching. It contains interviews with native speakers of English from different countries and different walks of life. They talk about their professional career (e.g. in tourism, sport, the media and environmental education). We are very grateful to all speakers for their kind contributions (→ acknowledgements, availability and copyright).

This demo website contains selected materials from the ELISA corpus.

→ AEL in Tuebingen
→ Information about ELISA
→ Corpora4Learning home

Browse interviews

→ Browse interview topics
→ Browse grammatical topics

Bishopswood Centre (demo)
Bishopswood Centre is an environmental education centre in the West Midlands (UK). One of the project managers talks about the aims of the centre, the projects it is involved in and the educational programmes it runs.

'Seeing the Light' (demo)
'Seeing the Light' is a photo agency in Birmingham (UK). Its owner, a photographer, talks about how and why she started her company, about past and current projects, marketing strategies and plans for the future.

'Broken Saddle' (video clip)
Harold is the owner of the 'Broken Saddle' horse-back riding company in New Mexico (US). He tells the story of his company, talks about his clients and ... his horses.

Play the video clips
To play the video clips, you need RealOnePlayer. It's free - to dowonload it, click
here for Windows,
here for Macintosh,
here for Unix/Linux.

The corpus data
The ELISA corpus currently contains 15 interviews of 5 to 15 minutes and amount to about 60,000 words in total.

Word frequency
Click here to see a list of the

Fig. 1. ELISA corpus: access to interviews.

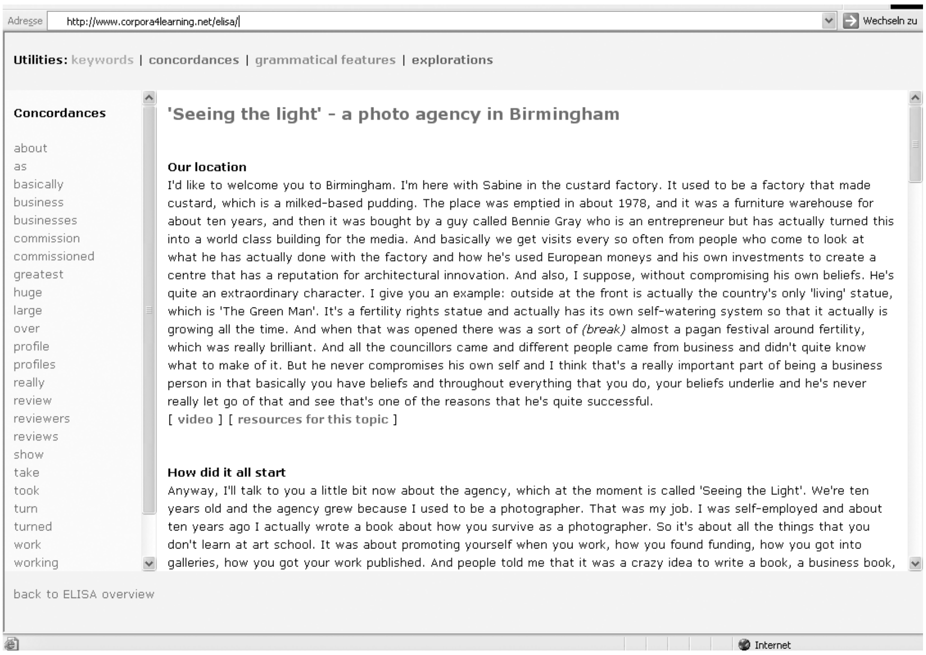
based on the topic annotation. In the topic and function views, all relevant interview sections which fit into the category in question are displayed. One of the main options is the detailed exploration of individual sections of each interview.

The division of the interviews into topic-based sections is also used to align the transcripts and the videos,[9] which allows the users to watch the video and read the transcript section by section (see Figure 2).

Most of the enrichment materials will be attached to the sections for which they are relevant and can be accessed by following the links 'video' and 'resources' at the end of each passage (see Figure 2). Other enrichment materials will be attached directly to individual means of expression in the corpus and can be reached from within the text. The enrichment materials for the ELISA corpus fall into four main categories:

- audiovisual materials
- information and explanations
  - – lexical, grammatical, cultural and other comments
  - – usage notes
  - – word frequency lists
  - – pre-fabricated concordances, e.g. to study similar words, collocations and word families
- tasks and exercises

[9] XML is used to structure the texts (cf. http://www.w3.org/XML/). The video alignment is done on the basis of the SMIL technology (cf. http://www.w3.org/AudioVideo/) used e.g. by the RealPlayer.



'Seeing the light' - a photo agency in Birmingham

Our location

I'd like to welcome you to Birmingham. I'm here with Sabine in the custard factory. It used to be a factory that made custard, which is a milked-based pudding. The place was emptied in about 1978, and it was a furniture warehouse for about ten years, and then it was bought by a guy called Bennie Gray who is an entrepreneur but has actually turned this into a world class building for the media. And basically we get visits every so often from people who come to look at what he has actually done with the factory and how he's used European moneys and his own investments to create a centre that has a reputation for architectural innovation. And also, I suppose, without compromising his own beliefs. He's quite an extraordinary character. I give you an example: outside at the front is actually the country's only 'living' statue, which is 'The Green Man'. It's a fertility rights statue and actually has its own self-watering system so that it actually is growing all the time. And when that was opened there was a sort of *(break)* almost a pagan festival around fertility, which was really brilliant. And all the councillors came and different people came from business and didn't quite know what to make of it. But he never compromises his own self and I think that's a really important part of being a business person in that basically you have beliefs and throughout everything that you do, your beliefs underlie and he's never really let go of that and see that's one of the reasons that he's quite successful.

[**video**] [**resources for this topic**]

How did it all start

Anyway, I'll talk to you a little bit now about the agency, which at the moment is called 'Seeing the Light'. We're ten years old and the agency grew because I used to be a photographer. That was my job. I was self-employed and about ten years ago I actually wrote a book about how you survive as a photographer. So it's about all the things that you don't learn at art school. It was about promoting yourself when you work, how you found funding, how you got into galleries, how you got your work published. And people told me that it was a crazy idea to write a book, a business book,

Fig. 2. ELISA corpus overview page.

  - comprehension questions relating to a particular passage
  - exploratory tasks
  - various lexical and grammatical, comprehension and production exercises
- study aids and didactic hints
  - simplified transcript versions, e.g. to provide help for learners at lower proficiency levels
  - translation aids, e.g. for users with different native languages
  - best-practice guidelines for teachers and learners.

In the following I would like to present two examples from the ELISA corpus to illustrate some of the features described above. The first one is a passage from an interview with the owner of a photo agency in Birmingham. In the passage, the speaker describes one of the projects on which the agency worked some years ago. The transcript reads as follows:

> So at any one time we might be working on a huge project like ‘The People and the City’ exhibition, which basically is in this book, which profiles some of the people who actually live and work in the city and also profiles what the city actually looks like from the air and that in an extraordinarily creative way. This was commissioned to support the Capital of Culture bid for 2008. We were one of the six cities in the running. And that was an extraordinarily well paid commission, which I ‘stole’ from an agency in London because when I saw what they were actually going to do, we just knew that we could do it better. And luckily we actually got the commission.

> We had six weeks to produce the show. We commissioned two people, Tom and Brian, both who have their roots here in Birmingham. Brian made portraits of people and Tom took the pictures from the air. And then we opened the show in London in the architect Richard Roger's new building, on the sixth floor of his building in Soho. And we invited the world and the world came. And it was a very successful opening and it made the local and national news and Sunday magazines. And gave Birmingham a really brilliant profile. And although we didn't actually win the Capital of Culture 2008 award, I think that opportunity allowed us to both profile ourselves, our photographers and the city – nationally and also internationally.

In the first sentence the non-restrictive relative clause *which basically is in this book* stands out. It is a good example of why video clips should be included in the corpus. When the video is viewed, the clause makes immediate sense: the speaker holds up a booklet in which the project is documented. As mentioned above, in the ELISA corpus video material is treated as one kind of enrichment. Users can either watch the complete interview or a part of it to look at a particular section in detail or to compare similar sections across interviews, where they deal with the same topic.

Three relatively frequent words in this passage are *commission, profile* and *actually*. If we take a German learner of English, a number of interesting issues arise. For example, *commission* is used in this passage as a noun and a verb. Moreover, as a noun it is used in a 'technical' sense, which is less likely to be familiar to many learners than some other frequent meanings of this word such as 'committee' and 'fee paid to an agent'. In the above passage, the two instances where it is used as a noun would translate into German as *Auftrag* and *Zuschlag* respectively but due to its rather low overall frequency it is one of those words which may not immediately spring to mind when searching for an English equivalent for *Auftrag* or *Zuschlag*.

Various tasks could arise from this. One simple task is, of course, to make learners discover that the word is used as a verb and a noun. A more challenging task would be to identify and compare the meaning(s) of the word in this particular passage, in other passages relating to the topic 'projects' and in the entire corpus. What would be helpful here is a function to call up concordances with different scopes, namely featuring the instances of *commission* (a) in the whole interview, i.e. showing how the word is used by this particular speaker, (b) in the whole corpus or (c) in all the passages relating to the topic 'projects'. However, as can be seen from the above passage, a detailed study of co-texts beyond individual concordance lines will be more helpful to get all the different shades of meaning. Furthermore, a learner's interest in this word is most likely to arise in connection with passages like the one above. This implies that the starting point for calling up the different kinds of concordances should be the passage in question.

Another interesting point in our example passage is the relationship between *commission*, in this 'technical' sense, and *bid*. A relevant task would be to identify meaning differences and overlaps. Examples like this show that it will in many cases be appropriate to attach exploratory tasks to entire passages rather than just to individual lexical items in the corpus. Hence, the idea of topic-based access to most of the enrichment materials.

With regard to *profile,* learners could be again challenged to discover the twofold use of the word as verb and noun. Moreover, it would be interesting, at least for German learners of English, that the verb can be used reflexively and transitively in English. In German, the similar verb, and false friend, *profilieren* can only be used reflexively (*sich profilieren*). Also, it has a negative connotation. It may be interesting then to study the uses and connotations in English. A corresponding task would most suitably be integrated as a topic-related enrichment to cover all (verb and noun) uses of *profile* within one task. Moreover, concordances of *profile* will again only provide the starting point for this task. They will surely help to discover the different grammatical uses (noun, verb, transitive, reflexive). However, to study the connotations of the word, it may well be necessary to consider the whole passage or even the entire interview in order to get an impression of the speaker and the way she talks about her business in general.

The adverb *actually* occurs frequently in the ELISA corpus but the speakers vary with regard to the frequency of using it. The photo agency owner, who produced the above passage, uses it particularly often. On the basis of concordances for *actually* learners could compare different speakers with regard to their use of *actually*. Here again, it would be helpful to have access to different concordances, in this case from each interview and from the whole corpus.

Another interesting point in this particular passage is the clause *which I 'stole' from an agency in London*. From the intonation and the facial expression of the speaker it becomes clear that she says it 'with a twinkle in her eye'. This is why we decided to put it in quotation marks in the transcript. If annotated appropriately, the clause would provide a good starting point for exploring the intonation. As mentioned earlier, the video clip will be crucial in enabling learners to do this. Moreover, the clause could be enriched with a comment introducing the idiom *with a twinkle in one's eye* to a learner. This would be an example of an enrichment which actually relates to an individual means of expression rather than to an entire passage.

The second example – this time for a function-based approach – will conclude this section. According to our interview questions, many of the speakers in the ELISA corpus outlined earlier stages in their professional career. In those passages the use of different tenses in contrast is particularly interesting. The following passage is one example of this. It is the beginning of an interview with the owner of a horse riding company:

> I'm the owner of the Broken Saddle Riding Company, have been for the last eleven years. I used to be in the horse racing industry back in New Jersey, worked around horse racing. I wasn't making any money, and the woman I was seeing at the time decided that she wanted to come to Santa Fe. So I decided to come along with her, because I was very much in love. …I started this eleven years ago. Came up with the name Broken Saddle, because when I started the business, my saddle broke. And I've been doing it now full-time for the last nine years. It took me about two years to get it going and it's been just a lot of fun. For the last nine years we've been riding in the hills. Silver, turquoise mines, the canyons.

As Granger (1999) has pointed out, tenses should not be studied and taught on the basis of concordance formats but at discourse level. An appropriate annotation of entire passages such as the one above will enable learners and teachers to find relevant materials for studying and practising tenses. Relevant enrichment materials for this passage can range from grammatical explanations to awareness-raising exercises (e.g. in gap-filling format) as well as guided production tasks. In connection with the tenses in the above passage the relevant prepositions (*for, since*) which trigger the use of the present perfect-could also be practised and studied with the help of the passage. This implies that some passages in the corpus may receive multiple annotations to allow access from different perspectives and with different questions in mind.

5 Summary and outlook

In this paper, I have argued that the successful use of corpora for learning and teaching hinges to a great extent on a successful 'pedagogic mediation' between the corpus materials and the corpus users. On the basis of the arguments put forward by Widdowson (1978, 1979, 1990, 2003) I have outlined why this mediation is necessary: its aim is to support learners and teachers in reconstructing the discourses which gave rise to the texts in the corpus. Moreover, I have demonstrated possible solutions for a pedagogic mediation. I have shown that it has to be approached at different levels. Firstly, the mediation is supported by a coherent and relevant content, a restricted size, a multimedia format and a pedagogic annotation of the corpus. Secondly, it will be greatly facilitated by a combination of corpus techniques with a discourse-based analysis both for the creation and exploitation of pedagogically relevant corpora. Thirdly, the integration of pedagogically relevant enrichment materials can be regarded as another direct 'response' to the mediation requirement.

Furthermore I have introduced the ELISA corpus, which is currently being created with pedagogic requirements in mind. The potential of the ELISA corpus lies in the multidimensional access to the materials (e.g. entire interviews, individual corpus topics and relevant linguistic functions). Thus it will be possible to account for different user perspectives, interests and profiles.

One aim related to the construction of the ELISA corpus is to develop and evaluate a *methodology* for the creation and exploitation of a pedagogically relevant corpus which can be transferred to other corpus projects or for other languages and/or themes.

One critical point is the necessary trade-off between the manual work involved in the creation of the corpus, especially in encoding and enriching it on the one hand and a desirable easy extension of the corpus on the other. A solution for this problem may come from linguistic research where work on (semi-)automatic text pattern recognition and similar analyses is well underway.

References

Aijmer, K. and Altenberg, B. (eds) (1991) *English Corpus Linguistics*: *Studies in honour of Jan Svartvik*. London: Longman.

Aijmer, K. and Altenberg, B. (eds) (2004) *Advances in Corpus Lingusitics*. Amsterdam: Rodopi.

Aston, G. (1995) Corpora in language pedagogy: matching theory and practice. In: Cook, G. and Seidlhofer, B. (eds), *Principle & Practice in Applied Linguistics*. Oxford: OUP, 257–270.

Aston, G. (1997) Small and large corpora in language learning. In: Lewandowska-Tomaszczyk, B. and Melia, P. (eds), *op. cit.*, 51–62.

Aston, G. (2001) Learning with corpora: an overview. In: Aston, G. (ed), *Learning with Corpora*.

Houston/Texas: Athelstan, 4–45.

Aston, G. (2002) The learner as corpus designer. In: Kettemann, B. and Marko, G. (eds), *op. cit.*, 9–25.

Berglund, Y. and Chambers, A. (2004) Trends in corpora in language learning: *EUROCALL 2004. TEL & Cal* 4/04: 81–82.

Bernardini, S. (2000) Systematising serendipity: proposals for concordancing large corpora with language learners. In: Burnard, L. and McEnery, T. (eds), *op. cit.*, 225–234.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Investigating language structure and use*. Cambridge: CUP.

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

Blakemore, D. (1992) *Understanding Utterances*. Oxford & Cambridge, MA: Blackwell.

Brown, G. and Yule, G. (1983) *Discourse Analysis*. Cambridge: CUP.

Burnard, L. and McEnery, T. (eds) (2000) *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the third international conference on teaching and language corpora*. Frankfurt: Lang.

Cobb, T. (1997) Is there any measurable learning from hands-on concordancing? *System* **25**(3): 301–315.

Fligelstone, S. (1993) Some reflections on the question of teaching, from a corpus linguistics perspective. *ICAME Journal* **17**, 97–109.

Flowerdew, J. (1996) Concordancing in language learning. In: Pennington, M. (ed), *The Power of CALL*. Houston/Texas: Athelstan, 97–113.

Flowerdew, J. (2001) Concordancing as a tool in course design. In: Ghadessy, M., Henry, A. and Roseberry, R.(eds), *op. cit.*, 71–92.

Gavioli, L. (1997) Exploring texts through the concordancer: guiding the learner. In: Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds), *op. cit.*, 83–99.

Gavioli, L. (2002) Some thoughts on the problem of representing ESP through small corpora. In: Kettemann, B. and Marko, G. (eds) *op. cit.*, 293–303.

Gavioli, L. and Aston, G. (2001) Enriching reality: language corpora in language pedagogy. *ELT Journal* **55**(3): 238–246.

Ghadessy, M., Henry, A. and Roseberry, R. (eds) (2001) *Small Corpus studies and ELT: Theory and practice*. Amsterdam: Benjamins.

Granger, S. (1999) Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus. In: Hasselgard, H. and Oksefiell, S. (eds), *op. cit.*, 191–202.

Hadley, G. (2002) Sensing the winds of change: an introduction to data-driven learning. *RELC Journal* **33**(2): 99–124.

Hasselgard, H. and Oksefiell, S. (eds) (1999) *Out of Corpora. Studies in honour of Stig Johansson*. Amsterdam: Rodopi.

Henry, A. and Roseberry, R. (2001) Using a small corpus to obtain data for teaching a genre. In: Ghadessy, M., Henry, A. and Roseberry, R. (eds), *op. cit.*, 93–133.

Mair, C., Hundt, M., Leech, G. and Smith, N. (2002) Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* **7**(2): 245–264.

Johns, T. (1986) Microconcord: a language-learner's research tool. *System* **14**(2): 151–162.

Johns, T. (1991a) Should you be persuaded: two examples of data-driven learning. In: Johns, T. and King, P. (eds), *op. cit.*, 1–13.

Johns, T. (1991b) From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In: Johns, T. and King, P. (eds), *op. cit.*, 27–45.

Johns, T. and King, P. (eds) (1991) Classroom concordancing. Birmingham University: *English Language Research Journal* **4**.

Kennedy, G. (1998) *Introduction to Corpus Linguistics*. London: Longman.

Kennedy, G. (2003) *Structure and Meaning in English*. Harlow: Longman/Pearson.

Kettemann, B. (1995) On the use of concordancing in *ELT. TELL&CALL* **4**: 4–15.

Kettemann, B. and Marko, G. (eds) (2002) *Teaching and Learning by doing Corpus Analysis*. Amsterdam: Rodopi.

Lee, D.(2001) Defining core vocabulary and tracking its distribution across spoken and written genres. *Journal of English Linguistics* **29**(3): 250–278.

Lee, D. (2002) Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through The BNC jungle. In: Kettemann, B. and Marko, G. (eds), *op. cit.*, 247–292.

Leech, G. (1997) Teaching and language corpora: a convergence. In: Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds), *op. cit.*, 1–23.

Leistyna, P. and Meyer, C. (eds) (2003) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi.

Lewandowska-Tomaszczyk, B. and Melia, P. (eds) (1997) *Proceedings of PALC 97*, Lodz: Lodz University Press.

Mair, C. and Hundt, M. (eds) (2000) *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.

Mauranen, A. (2004) Spoken – general: spoken corpus for an ordinary learner. In: Sinclair, J. M. (ed), *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins, 89–105.

McEnery, T. and Wilson, A.(1993) The role of corpora in computer-assisted language learning. *Computer Assisted Language Learning* **6**(3): 233–48.

McEnery, T. and Wilson, A. (1997a) Teaching and language corpora. *ReCALL* **9**(1): 5–14.

McEnery, T. and Wilson, A. (1997b) Multi-media corpora. In: Lewandowska-Tomaszczyk, B. and Melia, P. (eds), *op. cit.*, 24–33.

Meunier, F. (1999) The pedagogical value of native and learner corpora in EFL grammar teaching. In: Granger, S., Hung, J. and Petch-Tyson, S. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 119–141.

Mindt, D. (1996) English corpus linguistics and the foreign language teaching syllabus. In: Thomas, J. and Short, M. (eds), *op. cit.*, 232–247.

Mindt, D. (1997) Corpora and the teaching of English in Germany. In: Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds), *op. cit.*, 40–50.

Mukherjee, J. (2002) *Korpuslinguistik und Englischunterricht: Eine Einführung (Sprache im Kontext series 14*) Frankfurt: Lang.

Seidlhofer, B. (2003) *Controversies in Applied Linguistics*. Oxford: OUP.

Simpson, R., Lucka, B. and Ovens, J. (2000) Methodological challenges of planning a spoken corpus with pedagogical outcomes. In: Burnard, L. and McEnery, T. (eds), *op. cit.*, 43–49.

Sinclair, J. (1991) *Corpus, Concordance and Collocation*. Oxford: OUP.

Sperber, D. and Wilson, D. (1995) *Relevance: Communication and cognition*. Oxford & Cambridge, MA: Blackwell.

Thomas, J. and Short, M. (eds) (1996) *Using Corpora for Language Research: Studies in honour of Geoffrey Leech*. London: Longman.

Tribble, C. (1997) Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. In: Lewandowska-Tomaszczyk, B. and Melia, P. (eds), *op. cit.*, 106–117.

Tribble, C. (2001) Small corpora and teaching writing: towards a corpus-informed pedagogy of writing. In: Ghadessy, M., Henry, A. and Roseberry, R. (eds), *op. cit.*, 381–408.

Tribble, C. (2004) The text, the whole text. In: Lewandowska-Tomaszczyk, B. (ed.), *Practical Applications in Language and Computers (PALC 2003)*. Frankfurt: Peter Lang, 303–318.

UCREL tag set CLAWS 7 – http://www.comp.lancs.ac.uk/ucrel/claws7tags.html

Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds) (1997) *Teaching and*

Language Corpora. London: Longman.

Widdowson, H. G. (1978) *Teaching Language as Communication*. Oxford: OUP.

Widdowson, H. G. (1979) *Explorations in applied linguistics*. Oxford: OUP.

Widdowson, H. G. (1980) *Models and fictions. Applied Linguistics* **1**(2): 165–170.

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

Widdowson, H. G. (2003) *Defining issues in English language teaching*. Oxford: OUP.

Wilson, A., Rayson, P. and McEnery, T. (eds) (2003) *Corpus Linguistics by the Lune*. Frankfurt: Lang.

Appendix 1

Bank of English (University of Birmingham and HarperCollins, since 1980): 450 million words (written and spoken British and American English), http://www.titania.bham.ac.uk and http://www.collinswordbanks.co.uk.

BNC British National Corpus (Oxford University Press, Longman, British Library, 1991–1995): 100 million words (written and spoken British English), http://www.natcorp.ox.ac.uk.

ELISA English Language Interview Corpus as a Second-Language Learning Application (University of Tuebingen, since 2003): ca 60.000 words (spoken English, different national varieties), http://www.corpora4learning.net/elisa/.

ICE International Corpus of English (co-ordinated at University College London, since 1991): 20 x 1 million words (subcorpora with different national varieties of English, some of them completed), http://www.ucl.ac.uk/english-usage/ice/.

IViE Intonational Variation in English (University of Oxford, 1997-2002): 36 hrs of speech (from nine urban varieties of British English), http://www.phon.ox.ac.uk/~esther/ivyweb/.

LeaP Learning the prosody of a Foreign Language (University of Freiburg, 2001–2003): 359 recordings of 2 to 20 minutes (spoken learner English), http://www.phonetik.uni-freiburg.de/leap/.

London-Lund Corpus of Spoken English (University College London and Lund University, 1960s-1980s): 500.000 words (British English), http://khnt.hit.uib.no/icame/manuals/LOND-LUND/.

MICASE Michigan Corpus of Academic Spoken English (University of Michigan, since 1997): 1.7 million words (spoken American English in various academic settings), http://www.lsa.umich.edu/eli/micase/.

Santa Barbara Corpus of Spoken English (University of California, Santa Barbara, since ca. 1990): 300.000 words (spoken American English), http://www.ldc.upenn.edu/Projects/SBCSAE/.

VOICE Vienna Oxford International Corpus of English (University of Vienna, since 2001): under construction, 250.000 words to date, to be extended to 1 million within the next 3 years (English as a lingua franca, spoken English of competent non-native speakers), http://www.univie.ac.at/Anglistik/voice/.