ELISA – a pedagogically enriched corpus for language learning purposes
Sabine Braun, University of Tübingen

Abstract

The aim of this paper is to introduce a methodological solution for the design and exploitation of a corpus which is dedicated to pedagogical goals. In particular, I will argue for a pedagogically appropriate corpus annotation and query, and for the enrichment of such a corpus with additional materials (including corpus-based tasks and exercises). The solution will be illustrated with the help of ELISA, a small spoken corpus of English containing video interviews with native speakers. However, the methodology is transferable to the creation of pedagogically relevant corpora with other contents and for other languages.

1 Corpora in language learning: where, what corpora and how

While corpora have made their way into many areas of language pedagogy, for instance to inform teaching materials or learners' dictionaries, their direct exploitation by learners and teachers has to date largely been confined to academic settings. In other learning environments, especially in the school context, the application of corpora has so far remained an exception (cf. Davies & Russell-Pinson 2004, Kaltenböck & Mehlmauer-Larcher 2005) and even the mere awareness of corpora among teachers is low (cf. Mukherjee 2004). One could argue that this is not too surprising because the process of 'percolation' from corpus-based research to corpus use in teacher training and eventually in teaching practice simply takes some time (cf. Chambers 2005, McEnery & Wilson 1997). However, according to Granger (2004) and Tribble (2000) the uptake of corpora in language pedagogy remains generally lower than might be expected in view of the huge amount of corpus-based research carried out over the past decades (see also Mukherjee, this volume), and even in universities corpus-based language courses are mainly run "by researchers and enthusiasts" (Chambers 2005: 1). This suggests that there are more fundamental reasons which prevent a broader uptake of corpus consultation 'in the classroom'. In search of possible explanations it will be helpful to take a look at current practice, especially at the types of corpora used and at the methods of their exploitation.

The corpora used 'in the classroom' range from existing reference corpora (e.g. the British National Corpus, Bernardini 2000) and subsections thereof (e.g. BNC sampler, Gavioli & Aston 2001) to a number of smaller, 'self-made' corpora, especially newspaper
corpora (e.g. Chambers & Ó'Sullivan 2004, Gavioli 1997) or other written corpora including text types familiar to the learners (e.g. Kennedy & Miceli 2001). These corpora have been characterised as more homogeneous and "far more specialised, by topic, by genre, or both" (Aston 1997: 54) than large reference corpora. However, with regard to their overall make-up, especially the range and diversity of texts they include, their annotation and accessibility (query), they follow traditional corpus design criteria. In accordance with this, their main or exclusive way of exploitation – by teachers or by the learners themselves – is through traditional, non-linear methods, especially in the form of concordance-based activities ("Data-Driven Learning", DLL). This, in turn, has limited their use to supporting individual activities such as the exploration of lexico-grammatical patterns and word meanings (e.g. Bernardini 2000, Cobb 1997, Gavioli 1997, Gavioli & Aston 2001, Johns 1991) or the improvement of writing skills (e.g. Chambers & Ó'Sullivan 2004, Kennedy & Miceli 2001, Yoon & Hirvela 2004) and has entailed a reduction of corpora to a sporadically used additional or complementary resource. Moreover, the texts themselves usually remain an 'anonymous mass' to the learners.

Some attempts have been made to use corpora differently, especially in the framework of genre-based learning (cf. Henry & Roseberry 2001, Rohrbach 2003, Tribble 2001). What is interesting to note is that these initiatives rely on very small, self-made corpora and have partially resorted to new ways of corpus annotation and query. Henry & Roseberry and Rohrbach, for example, have used very small corpora (introductions to guest speakers and tourist brochures respectively) to make learners aware of characteristic discourse moves and to let them study the means of expression typically associated with these moves. The corpora used here are so small that the texts can easily be studied in their entirety, while corpus techniques are used to make regular patterns in the individual texts more clearly visible.

Language learning is about the acquisition of communicative competence, i.e. about learning how to use language appropriately in various communicative situations. Corpora contain textual records of real communication, and it makes a lot of sense to use them to support as many aspects of the development of a learner's communicative competence as possible. The above-mentioned examples of 'new' corpus uses suggest that this can be achieved through the development of pedagogically relevant corpora along with pedagogically appropriate methods for their exploration.

Building on such attempts, the aim of this paper is to suggest a methodological solution for the design and exploitation of pedagogically relevant corpora. It will be

---

1 An overview of empirical studies on corpus consultation 'in the classroom' is provided by Chambers (2005).
illustrated with the help of ELISA (English Language Interview Corpus as a Second-language Application), a small corpus of spoken English containing video interviews with native speakers about their professional career. ELISA is a 'by-product' of our research into the development of pedagogical corpora. However, the methodology described here can be transferred to the creation of other pedagogically relevant corpora. One such initiative is indeed underway: the Socrates project SACODEYL aims at the creation of spoken youth language corpora for seven European languages. Before describing the design solution in detail, the theoretical assumptions which gave rise to it will be outlined in section 2.

2 Theoretical background

The two main advantages of DDL and especially of concordance-based activities are often described as being the analysis of 'real language' and the access to a wealth of different materials or 'contexts' through concordances (see e.g. Gavioli & Aston 2001). It seems helpful to explore this in more detail and to distinguish between some key concepts which are relevant here and which have been introduced and elaborated from an Applied Linguistics perspective by Widdowson (1978, 1979, 2003, 2005): text vs. discourse and co-text vs. context. Similar distinctions have been made in the framework of Pragmatics and Discourse Analysis (see e.g. Blakemore 1992, Brown & Yule 1983, Shiffrin 1994, Sperber & Wilson 1995).

With regard to the claim that corpora contain 'real language', it is worthwhile noting that they contain texts, i.e. (decontextualised) products of language use. What makes them attractive for language learning purposes is that they are genuine, i.e. attested instances of language use (Widdowson 1978: 80). Text, however – whether genuine or invented – has to be distinguished from discourse, which is a meaningful unit including the implicit meanings that a reader/hearer has established in a particular communicative situation, on the basis of the textual clues along with linguistic and (cultural) background knowledge.

2 The development of ELISA was funded by a grant from the University of Tübingen from 2003-2004. Currently the corpus contains 26 interviews with an average length of 10 minutes, amounting to about 60,000 words in total. The video interviews were transcribed to text. The corpus contains both the transcripts and the video clips. It is accessible online on http://www.uni-tuebingen.de/elisa.

3 The SACODEYL project (System-Aided Compilation and Open Distribution of European Youth Language), in which the University of Tübingen is involved, is funded under the Minerva action "Promotion of open and distance learning" in the EU Socrates programme from 2005-2008 (Ref: 225836-CP-1-2005-1-ES-MINERVA-M co-ordinator: University of Murcia, Spain).
If we come across, for example, a heading such as *England dig deep in hunt for draw*,4 this can be interpreted variably as referring to football, possibly the world cup draw as well as to a cricket match (or it may not make much sense at all to a non-native speaker). Similarly, concordances contain *co-texts*, i.e. (usually small) chunks of text surrounding the examined word or phrase. *Co-text* has to be distinguished from *context*,5 which is not given in a text or a concordance but constructed by the hearer/reader in the process of, and as a basis for, discourse comprehension (Blakemore 1992: 87).

With regard to the claim that concordances give access to a wealth of materials it follows from these distinctions that there are at least two difficulties: The first is that exploring texts from corpora usually means dealing with texts outside the communicative situation in which they were originally produced.6 To understand these texts, we have to be able to construct a context in which they could have been produced. This processes has been termed *authentication* by Widdowson (1979). Its success hinges on a variety of factors, especially on knowledge about the communicative situation, appropriate linguistic and (cultural) background knowledge to interpret textual clues. Whether we come across the textual record of a ten-year old parliamentary debate, of a comment on England's latest cricket match or of the account of somebody's life, it will be difficult to make sense of it unless we remember, or know about, the events referred to or can relate to them on the basis of general world knowledge. Furthermore, this process is controlled by the relevance of topic and the reader's interest in the text.

The second problem arises from the specific non-linear or 'vertical' ways in which corpora are usually analysed, especially through concordances and word lists. The point I would like to make here is not whether it is possible to authenticate a concordance as such. Rather, the question is how easy or difficult it is for learners to authenticate a text which they 'access' through a concordance line – without knowing to which part of the text this takes them and without pedagogically appropriate information about the text and the communicative situation in which it was produced.

The ability to exploit the full potential of concordances will vary among learners depending on their level of competence, autonomy and experience with concordances. Also, teachers can, of course, customise concordances, e.g. by deleting unwanted lines (cf. Johns 1991), or tell the learners to ignore lines which make no sense to them (cf.

---

5 For an in-depth discussion of this distinction and its relevance in connection with corpus analysis, see Widdowson (2005: 58-73).
6 This is not restricted to texts included in corpora but it is prominent with corpus texts.
Gavioli & Aston 2001). So to be clear, none of the above is to invalidate the usefulness of concordances and especially their role in revealing lexico-grammatical patterns, aspects of word meaning and supporting form-focussed learning activities. My concern is not with the usefulness of concordances (and other corpus techniques) as such, but with their standing in 'corpus-based' approaches to language learning. It seems that the exploitation of corpora through concordances (and other traditional corpus techniques such as word lists and frequency counts) alone restricts their potential for the language learning context.

The above considerations suggest that learners who work with corpora need support in the authentication of corpus materials. It seems useful to give them access to entire texts or well-defined parts thereof along with appropriate information about the texts (as well as to word lists and concordances). Therefore, the methodological solution introduced here approaches the use of corpora from a different angle. The starting point is a text-based exploration of the corpus content, focussing on the wider social and cultural context of the materials. The analysis with corpus techniques is seen as complementary, for the in-depth study of linguistic means of expression on the basis of familiar texts. Such an approach requires that corpus contents be chosen in accordance with pedagogical needs (rather than the needs of linguistic description or lexicography), and it has further implications for the size, annotation and query options of the corpus. In section 3 and 4, I will take up the design issues. In section 5, I will discuss how this solution can be further enhanced by enriching the corpus with additional, pedagogically relevant materials.

3 Pedagogically motivated corpus contents

The need for pedagogically motivated corpus contents is well reflected in the many initiatives to create 'home-made', 'ad-hoc' or 'DIY' corpora for language learning or translator training purposes (cf. Aston 2002, Ghadessy, Henry & Roseberry 2001, Varantola 2003, Zanettin 2001, Tribble 1997, Rohrbach 2003). Tribble emphasises the motivating effect which tailor-made corpora have on the students who work with them, as a result of successful authentication.

According to learners' needs, the contents of these corpora can be very specific. Thus, Rohrbach (2003), building on Henry & Roseberry's (2001) approach, worked with a small corpus of tourist brochures, which was analysed by his students as a model for their own production of a local tourist brochure. However, a general orientation for relevant areas of content is provided by the Common European Framework of Reference (CEF). It situates language use in 4 broad domains – personal, public, occupational and
educational (CEF: 45). Thus, Chambers & Ó'Sullivan (2004) and others have used newspaper corpora because they include texts on topical issues in the public domain.

A similar, but more focussed approach has been taken in ELISA. It contains a series of narrative interviews with native speakers of different varieties of English and from different walks of life. The informants talk about their professional career. All interviews follow a general pattern, covering a similar range of topics, e.g. the what the speakers do, their educational background, how they started their career or business, the type of projects they are involved in, their daily routines and future plans. While some of the speakers engage in unusual professional occupations (among them are e.g. a tour guide at Ayers Rock, a guitar teacher, a travel journalist and an arts therapist) and thus make for part of the attraction of the materials, they all describe issues of general interest in professional contexts, e.g. their motivations for doing their respective jobs, how they go about working on their projects or the kind of work they do in different fields of education and training.

Due to their different national, educational and professional backgrounds the speakers in ELISA talk about the above topics from a multitude of perspectives, using a good variety of means of expression. The corpus includes, for example, a range of interviews with speakers from the American Southwest, which give insight into the region's history, lifestyle and problems. Thus, three political players of the area – the Mayor of Santa Fe, a city councillor and the president of the local Chamber of Commerce – express their views on the problems with water supply, one of the greatest challenges this very dry region faces. Similarly, some speakers interviewed in Scotland voice different opinions on Scotland's political independence and its (new) own parliament in Edinburgh. Another interesting focus is on environmental issues, coming up in interviews with different orientations. They play a role in the Santa Fe water debate as well as in interviews with speakers from an environmental education agency in England, a tropical wildlife centre in Queensland and a water company in Scotland.

The main point about ELISA is that it covers a variety of communicatively relevant topics from the broad area of professional, social and cultural life. According to the CEF it can be said to reflect relevant uses of language in the occupational and personal domains. Apart from that, it contains spoken narrative, which may constitute a genre that can be analysed in its own right and used to raise learners' awareness for spoken language.

Spoken discourse is becoming increasingly important in international communication and hence in the learning and teaching context. The importance of spoken language and the usefulness of spoken language corpora in the pedagogical context has
been emphasized by many researchers (e.g. Mauranen 2004, McCarthy 1998). However, suggestions for teaching spoken language have often focussed on conversation (e.g. Carter & McCarthy 1995). While spontaneous conversation, as the most basic type of communication at all, certainly has a role to play in language learning, relevant professional scenarios also call for other speaking skills including presentation skills (presenting oneself or one's institution), addressing an audience or defending one's standpoint. The CEF mentions 'sustained monologue', e.g. to 'describe experience', as an important form of spoken discourse. At level B2, for example, it is required that the language user "can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest" (CEF: 59). The topic-related narrative of others as it can be found in ELISA is a useful source of reference in this context. 

It should be noted, however, that native speakers are not necessarily the only 'model'. Tribble (1997: 109) believes that "the most useful corpus for learners of English is the one which offers a collection of expert performances" and emphasises that a non-native but proficient expert in a particular field of interest can be a better 'linguistic model' that a native non-expert. Moreover, an opportunity to compare native and non-native 'expert performances', as would be possible with two comparable corpora may help learners to better and more actively identify appropriate communication strategies in the target language (see also Widmann & Braun, in this volume). 

Following Tribble, we can say in conclusion that topical relevance is the most important criterion for the content of a pedagogically relevant corpus. In the SACODEYL project, youth language corpora will be collected since it will be interesting for students in school to listen to 'peer voices' and to young people's topics. As Mukherjee (2002) points out, small corpora can also be used (or created) to cover areas or genres which are less well presented in course books or in traditional corpora but of interest to learners (e.g. from song texts for school students). A good source are the growing text and media archives (such as Project Gutenberg7 or the Oxford Text Archive8) which provide a wealth of interesting materials for many languages. 

Any small corpus is, of course, limited with regard to the means of expression it provides and would not be suitable for lexicographically motivated research on e.g. word meaning or collocation. However, in the learning and teaching context small corpora, and especially small and homogeneous corpora, have a number of advantages. They provide a more systematic range of material than any individual text or sample of spoken language

7 http://www.gutenberg.org
8 http://ota.ahds.ac.uk
(such as the occasional recording of a TV show) or other 'scattered' material collection (such as paper copies of newspaper articles etc.) which are often used in teaching contexts. Contributions by different speakers/writers to similar topics make even a small corpus less dependent on the idiosyncrasies of an individual speaker/writer and provide a greater range of expressions. The next question with regard to the corpus design then is how to make the contents accessible for teachers and learners, i.e. for corpus-linguistic non-experts.

4 Corpus annotation and query for pedagogical purposes

The remarks in the previous sections will have made clear that the ways in which corpora are used in a pedagogical context differ from linguistic research contexts. Consequently, a pedagogically relevant corpus design should differ from traditional corpus design in at least the following dimensions (for a more detailed discussion, see Braun 2005). With regard to the contents, homogeneity and topical relevance are more important than representativeness in the traditional sense. Regarding the methods of data analysis, a combination of a text-based approach with corpus-linguistic methods is more appropriate to support the contextualisation of corpus data than relying too strongly or exclusively on non-linear and quantitative methods. In addition, the inclusion of audiovisual data can further facilitate the authentication process, as opposed as using corpora which contain only textual data allowing searchability and statistical analysis as is often required in linguistic research. As for the size of pedagogically relevant corpora, small corpora are more appropriate since they enable whole-text reading and hence support the desirable combination of text-based and corpus-based exploration. Moreover, since learners and teachers are likely to rely more on qualitative corpus analysis than on quantitative analyses, small corpora provide a more appropriate (i.e. manageable) amount of data than the multi-million word corpora which are available today. Last but not least there are specific requirements for the search in corpora in the pedagogic context because firstly, learners and teachers have their own 'research' questions (which differ from those in linguistic research or lexicography) and secondly, they are no experts in corpus or computational linguistics. This in turn has an impact on corpus query and annotation.

With regard to what was said about authentication in section 2, the most relevant starting point for corpus consultation by learners and teachers may not be a concordance,

---

9 The usefulness of integrating audiovisual materials in the corpus will be further discussed in section 5.
as often suggested in DLL, but one or more texts from a small corpus, e.g. texts which match a unit in the syllabus. Teachers and (autonomous) learners will therefore need some support in identifying and selecting appropriate texts in the corpus. In ELISA, pedagogically relevant metadata have been included to this end. Thus each interview was given a title and an easy-to-read summary description. Both are tailored to learners' needs, serve as an introduction and appetiser (to increase motivation). The descriptions also provide a useful quick reference for those users who do access a text through a concordance and seek further information on the text. Secondly, a linguistic characterisation was added to each interview. It provides an orientation for a teacher (or an autonomous learner) e.g. on the level of difficulty and abstractness of the text and on the speech rate.

ELISA is accessible on the web, with an index page (cf. Figure 1) giving access to the titles and summary descriptions of the interviews and, via the titles, direct access to each interview. Moreover, it offers a thematic index of the topics covered in the interviews and other indices (see below) as well as further options which will be explained in section 5.
The core of ELISA is the topic index. It is based on a thematic annotation and supports a) the detailed exploration of an interview and (b) a comparison of sections across interviews. As mentioned earlier, the interviews all follow a similar pattern. The speakers covered some or all of the following topics:

01 Our location 07 Economic issues
02 What we do 08 Business issues
03 Personal history 09 Job routines
04 Getting started 10 Challenges
05 Project examples 11 Future plans
06 Education and training

On average the speakers devoted one to three minutes to a topic. This way, every interview can be divided into a number of sections. Each section was annotated with a topic key (e.g. "What we do") and a section title. Figures 2 and 3 show the thematic
structures of two interviews, one with the arts therapist and one with the tour guide at Ayers Rock.

<table>
<thead>
<tr>
<th>Topic key</th>
<th>Section title</th>
<th>Length</th>
<th>Word count</th>
</tr>
</thead>
<tbody>
<tr>
<td>02 What we do</td>
<td>What I do</td>
<td>1:14</td>
<td>133</td>
</tr>
<tr>
<td>03 Personal history</td>
<td>What I used to do</td>
<td>1:24</td>
<td>158</td>
</tr>
<tr>
<td>02 What we do</td>
<td>My current work</td>
<td>2:51</td>
<td>323</td>
</tr>
<tr>
<td>06 Education and training</td>
<td>Teaching at the university</td>
<td>2:10</td>
<td>267</td>
</tr>
<tr>
<td>04 Getting started</td>
<td>Living and working abroad</td>
<td>2:11</td>
<td>263</td>
</tr>
<tr>
<td>03 Personal history</td>
<td>Growing up and studying</td>
<td>1:14</td>
<td>198</td>
</tr>
</tbody>
</table>

Figure 2: Structure of the interview "Working as an arts therapist"

<table>
<thead>
<tr>
<th>Topic key</th>
<th>Section title</th>
<th>Length</th>
<th>Word count</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 Our location</td>
<td>Where I am based</td>
<td>0:30</td>
<td>91</td>
</tr>
<tr>
<td>02 What we do</td>
<td>What I do</td>
<td>0:43</td>
<td>121</td>
</tr>
<tr>
<td>04 Getting started</td>
<td>How I got my current job</td>
<td>0:56</td>
<td>176</td>
</tr>
<tr>
<td>03 Personal history</td>
<td>What I used to do</td>
<td>1:11</td>
<td>215</td>
</tr>
<tr>
<td>02 What we do</td>
<td>The company I work for</td>
<td>1:36</td>
<td>278</td>
</tr>
<tr>
<td>02 What we do</td>
<td>What makes us special</td>
<td>1:37</td>
<td>279</td>
</tr>
<tr>
<td>10 Job routines</td>
<td>My daily routines</td>
<td>1:27</td>
<td>229</td>
</tr>
<tr>
<td>02 What we do</td>
<td>Our clients</td>
<td>1:03</td>
<td>181</td>
</tr>
</tbody>
</table>

Figure 3: Structure of the interview "A tour guide from Ayers Rock"

Topic keys and section titles have complementary functions. The topic keys are used to generate a topic index which allows the corpus users to access, read and compare sections from different interviews with similar content. Thus, selecting the topic "What we do" from the index would provide five section from the two interviews shown above (and many more from other interviews). The section titles, in contrast, do not have a classificatory function but are used as a reading and comprehension aid. They are displayed as subheadings in the interview transcripts (cf. Figure 4). The topic key/section title dichotomy accounts for the fact that one topic can have different realisations across interviews. The topic "What we do", for instance, can be realised as "What we do, "What I do" or "What my company does" or even "What makes us special" and "What makes us different from our competitors". The titles reflect the different situations and backgrounds
of the speakers. Thus, some of the interviewees run a business and describe what the business does as a whole, while others are freelancers and just talk about themselves. So, while the classification of sections into topic keys supports the exploration of similar sections across interviews, the individuality in the section titles helps the learners to explore individual interviews.

While the actual content of sections which fall into the same topic can vary, they usually show a number of linguistic and communicative similarities. With regard to the topic "Getting started", for instance, the arts therapist describes how her work experience in different countries led her to become an arts therapist, whereas the tour guide describes the how he got his job after meeting people working in the field. What makes all "Getting started" sections similar linguistically and communicatively is, for example, the frequent use of different past tense forms as well as similar vocabulary. Another topic in which past tense forms occur frequently is "Personal history". Here the forms of habitual past are particularly prominent (used to and would).
The segmentation of the interviews is also used to link the videos to the transcripts. This makes it possible to call up the corresponding video sequence for each interview section. Figure 5 shows the underlying XML-based annotation scheme for the topics and the video alignment (the start and end time of the corresponding video sequence), using the first section of the interview "Working as an arts therapist" as an example. The links to the video clips are integrated on the basis of the SMIL technology.10

```
<event start="0m0" end="1m14" video="arts_therapist_us" duration="1m14"
    wordcount="133">
    <topic>
        <topic_title>What I do</topic_title>
        <topic_key>02 What we do</topic_key>
        <subject_key>arts and photography</subject_key>
        <subject_key>university</subject_key>
        <function_key>introducing yourself</function_key>
        <grammar_key>tenses simple present</grammar_key>
    </topic>
    <speaker name="Judith">I'm on vacation in Alice Springs and taking a break from work and responsibilities in Melbourne. I flew up here a couple of days ago, and tomorrow I'll be visiting a colleague who is <break/> facilitates an aboriginal women's arts co-op, and in this arts co-op the aboriginal women produce art work that is sold nation-wide. They produce paintings, musical instruments and they've been doing that for a couple of <break/> several decades. <cut/> In Melbourne I work as a <break/> an arts therapist. I have a private practice working with adults, children, young people using the arts for health, well-being. <cut/> And I also teach at RMIT University, which offers a programme in creative arts therapy. It's a postgraduate programme, and the students come to that field of arts therapy from education, arts, nursing, sociology, psychology backgrounds.
    </speaker>
</event>
```

Figure 5: XML-based thematic annotation

As can be seen in Figure 5, each interview section represents an 'event' (within an xml file) which includes one topic key and section title as well as (optional) other keys which are used to build additional indices, in particular, subject and grammatical indices as well

---

10 SMIL stands for Synchronized Multimedia Integration Language (cf. http://www.w3.org/AudioVideo/) and is used e.g. by the RealPlayer.
as an index of communicative functions. The rationale behind all of them is that a slightly more extended co-text (such as an interview section) is a better basis for the analysis and understanding of many relevant aspects of language (subject-specific vocabulary, grammatical structures as well as appropriate means of expression for a communicative function) than the one-sentence examples frequently found in traditional materials or the cut-off lines in corpus concordances. This is why the additional keys are also given to entire interview sections rather than to smaller units such as individual words or sentences.

The passage shown in Figure 5, for example, has been given the function key "introducing yourself", since it illustrates how people introduce themselves to strangers when they start a conversation. The grammar key for this passage is "tenses simple present" since it provides a range of examples of how the simple present (in contrast to the present progressive) is used to describe habitual actions. Interview sections which were classified with the topic key "Getting started" frequently have the grammar key "tenses past" whereas "Personal history" sections are usually given the grammar key "habitual past". Finally, the subject keys can help to find subject-specific words and phrases. They are especially relevant in interviews which cover a broader range of subject areas. While the interview with the arts therapist, for example, evolves around both the arts (because of the occupation of the speaker) as well as around university life (because of her work environment), another interview with the owner of a private photo agency partially also relates to the arts but at the same time to running a small business.

The keyword-based annotation and the resulting indices provide useful starting points for a corpus query by learners and teachers. They represent a pedagogically viable alternative to the traditional forms of annotation such as part-of-speech tagging or syntactic parsing. As Kaltenböck & Mehlmaurer-Larcher (2005: 71) have pointed out, corpus queries in the pedagogical context usually "focus primarily on lexical queries, especially collocations". In other words, teachers do not make use of the potential offered by traditionally annotated corpora because they are simply not familiar with building complex linguistic queries.

What is important to notice when subjecting learners to working with sections from different interviews (on the basis of the indices-based text retrieval) is that this may present a challenge to the learners depending on their level of competence and autonomy. The way ELISA is designed, especially due to its small size, it is possible for a teacher to carefully select the sections to start with. In a first step, learners can work with only a small number of sections - preferably those with which they are familiar from reading the entire interviews – to formulate a hypothesis on a particular lexical, grammatical or other
issue. In a second step the range of interview sections can be extended in order to have a more substantial amount of data to study the phenomenon in question and to further explore the hypothesis. Thus learners can gradually be led to benefit from the advantage of a corpus (compared to just working with one text).

Furthermore, apart from the index-based access to the corpus, ELISA also allows for complementary types of queries, especially through its online concordancer which provides extended search options (regular expressions, wild cards, search for phrases), allows the user to select the texts to be included in the search and offers a choice between different concordance formats (KWIC of variable length, sentence concordance).

What has been described so far is how the corpus design suggested in this paper supports learners and teachers indirectly – through appropriate corpus size, contents, annotation and pedagogically motivated query options. Another part of the design solution will be introduced in section 5. It is a more direct type of support through the enrichment of the corpus with complementary materials.

5 Pedagogical enrichment

Enriching a corpus with additional materials which are relevant for language learning and teaching can be seen as a major step towards the 'pedagogic mediation' of corpora requested by Widdowson (2003). Moreover, it is in line with constructivist requests for a rich learning environment (cf. Wolff 1994). Broadly speaking, such materials should help learners to bridge the gap between the textual records in the corpus and the discourse situations they have to (re)construct in order to exploit the corpus materials efficiently for learning purposes. In more detail this means that learners need support for text comprehension and knowledge construction as well as opportunities to practise and test their knowledge. In addition, teachers (but also learners using the corpus autonomously) need to be supported in selecting and preparing appropriate materials from the corpus.

One type of enrichment which fulfils a multitude of purposes is the inclusion of audiovisual materials into a (pedagogically relevant) corpus. ELISA contains video clips of the interviews in addition to the transcripts. One advantage is that the visual, gestural and intonational clues in the video clips greatly help to contextualise and clarify problematic utterances. A second, equally important advantage is that audiovisual corpora such as ELISA open up entirely new ways of exploitation. In particular, they can be used for listening comprehension activities. In ELISA the possibility to call up the video clips directly from the index page and to watch them before reading the transcripts supports listening-for-gist activities. On the other hand, the division of the interviews into sections
and the possibility to call up the corresponding video sequence for each section facilitates more detailed comprehension tasks. Moreover, the integration of comprehension questions and exercises into the corpus further increases the efficiency of comprehension training with the help of the videos. Finally, since ELISA includes interviews with speakers of different national origin, it can also be used to train learners in understanding different varieties as required e.g. by the German teaching curricula (syllabus) for English in secondary schools. A suitable complement in this connection are explorative tasks which help learners to compare and work out characteristic features of different accents.

In section 1 the importance and communicative relevance of spoken language (including 'sustained monologue') was emphasised. It is certainly not exaggerated to say that the full potential of spoken language corpora can only be exploited in connection with the audiovisual records of the data. However, written corpora used for pedagogical purposes may also be enhanced by integrating audiovisual materials such as the original layout of a text and related images (cf. Mishan 2004).

A second type of enrichment are informative and illustrative materials which can range from simplified transcript versions as well as translations (for comprehension support) to comments and explanations (e.g. lexical, grammatical, cultural) and to ready-made corpus analysis results such as word lists or concordances. Such ready-made data help learners and teachers who are not familiar with corpus techniques or do not have the time, tools or occasion to apply them to nevertheless benefit from those techniques. In ELISA four types of frequency-based lists are available for the entire corpus and for each interview: a plain and a lemmatised list of all words as well as a plain and a lemmatised list of content words only. A future extension will be the inclusion of ready-made word lists for all sections with similar content (i.e. with the same topic key).

The various lists provide different 'views' on the corpus and the texts. As Gavioli (1997) and Tribble (2001) have shown, word lists of small corpora or even of individual texts can – in connection with whole-text reading – provide a useful basis for a number of activities. Content word lists can, for example, be used by teachers in the preparation phase to get a quick overview of the lexical difficulty of a text and to select the vocabulary items which may have to be introduced or explored in class before working with the text. Alternatively these lists can also be given to the learners to have them guess the overall topic of a text before studying it in detail or to give an account of it afterwards. A comparison of a plain and lemmatised content word list can raise the learners' awareness for word forms and families. Figure 6 shows a non-lemmatised and a

---

11 Tasks and exercises form another type of enrichment and will be discussed below.
lemmatised list of the most frequent content words in the arts therapist interview. What is potentially interesting for further exploration by learners here is the range of forms of ‘LIVE, LIFE, WORK, ENJOY and LEARN’. At the same time these examples – especially the forms work and lives – also demonstrate the limitations of looking at a word list alone. I will return to this point below but will first make some remarks on the ready-made concordances included in ELISA.

<table>
<thead>
<tr>
<th>N</th>
<th>Content words</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ARTS</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>WORK</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>PEOPLE</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>STUDENTS</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>WORKED</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>MELBOURNE</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>SWEDEN</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>ADULTS</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>AUSTRALIA</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>MUSIC</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>THERAPY</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>CULTURE</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>TIME</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>WORKING</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>AREA</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>CREATIVE</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>ENJOYED</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>EUROPE</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>HEALTH</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>LEARNING</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>LIVE</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>LIVES</td>
<td>4</td>
</tr>
<tr>
<td>23</td>
<td>LOT</td>
<td>4</td>
</tr>
<tr>
<td>24</td>
<td>MOVEMENT</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>PROGRAMME</td>
<td>4</td>
</tr>
</tbody>
</table>

shortened:
<table>
<thead>
<tr>
<th>N</th>
<th>Content words</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>ENJOYING</td>
<td>3</td>
</tr>
<tr>
<td>45</td>
<td>LEARN</td>
<td>3</td>
</tr>
<tr>
<td>46</td>
<td>LIFE</td>
<td>3</td>
</tr>
<tr>
<td>47</td>
<td>LIVED</td>
<td>3</td>
</tr>
<tr>
<td>48</td>
<td>LIVING</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Content words</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ART</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>WORK</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>PEOPLE</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>LIVE</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>STUDENTS</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>LEARN</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>LIFE</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>ENJOY</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>MELBOURNE</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>SWEDEN</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>ADULTS</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>AUSTRALIA</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>FIND</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>MUSIC</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>THERAPY</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>CULTURE</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>LOT</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>STORY</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>TIME</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>TRAVEL</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>AREA</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>CREATIVE</td>
<td>4</td>
</tr>
<tr>
<td>23</td>
<td>EUROPE</td>
<td>4</td>
</tr>
<tr>
<td>24</td>
<td>HEALTH</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>MOVEMENT</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 6: Content word lists from the interview "Working as an arts therapist"

12 The lemmatisation of lives was done manually in this list. It turns out that in this interview the form is only used as the plural of life.
The ready-made concordances for ELISA were created from a complete, lemmatised word list of a) the entire corpus and b) each individual interview. The resulting, web-based concordances offer an alphabetical word list in which each word can clicked to see a concordance. The sources of the concordance lines are also provided, and the user can jump directly to the relevant text passage. Figure 7 shows an extract from the web concordance for the arts therapist interview, focussing on the different forms of *live/life*.

Figure 7: Web concordance for the arts therapist interview

The text versions of the transcripts which were used to create the web concordances received paragraph and line numbers to optimise orientation. Thus the example in Figure 7 shows that the third and fifth section of the arts therapist interview include many of the

13 created with Concordance (http://www.concordancesoftware.co.uk/)
14 The alphabetical word list is lemmatised. The form *lives* (NPI) is, after manual lemmatisation, listed under its base form *life* (NSg).
different forms of *LIVE* and *LIFE*. This can, for instance, help teachers (in the preparation phase) to find text passages for which it is particularly relevant to introduce or discuss a particular word and/or its forms.\(^{15}\) At the same time this information can be useful for anyone wanting to create tasks and exercises.

As was pointed out above, word lists have a limited value for learners if looked at in isolation. The limited value of concordances without proper 'embedding' was discussed in section 2. It does not mean that this data is entirely irrelevant in the learning and teaching context. What it does mean is that concordances and word lists need to be integrated into appropriate learning activities in which their potential can be exploited. Therefore, the inclusion of **ready-made tasks and exercises** (comprehension questions, lexical, grammatical and other exercises and explorative tasks) in a corpus dedicated to pedagogical uses can further increase its value. In particular they can support:

- the "warm-up" phase preceding the actual work with the corpus (e.g. quiz-type activities, awareness-raising exercises);
- the detailed work with one text or with similar passages across texts, to explore words and phrases, grammatical issues, topics or communicative functions (e.g. lexical, grammatical and other exercises);
- the more 'global' work with several texts or the entire corpus (e.g. explorative tasks, project work).

Most of the tasks and exercises included in ELISA were created with the learning and authoring tool *Telos Language Partner*\(^ {16}\) and are accessible from the ELISA index page or the transcript pages as appropriate (cf. Figure 1 and 4 in section 4). One small set of examples is given in Figure 8-10 to demonstrate the support of the detailed work with one interview. The set is taken from an exercise module created for the interview with arts therapist (Judith) and focuses on the use of the grammatical forms of the of most frequent lexical verbs in this interview (*LIVE, WORK, LEARN, ENJOY*) as well as on the noun *LIFE*. To be clear, these exercises are meant to be done after a phase of in-depth study of this interview (e.g. through listening comprehension, discussion of its contents in class and a number of awareness-raising exercises).\(^ {17}\)

---

\(^{15}\) An alternative would be to have the words highlighted in the source text.

\(^{16}\) cf. [http://www.sprachlernmedien.de](http://www.sprachlernmedien.de)

\(^{17}\) The full exercise module can be found on the Elisa website, cf. note 2.
Words and their forms (1) – live / life

How does Judith use the words life and live? Use the "Browse words" function and look at her uses of life, live, lives, lived, living. Then fill in the correct word forms.

Judith works with adults who have faced life's challenges.
Arts therapy can enable people to find something away from the challenges of their lives.
It can help people to remember things they wouldn't ordinarily think about in their lives.
When she was a student, Judith lived in Germany for one year.
Judith has relatives who live in the south of Sweden.
She fell in love with the countryside and the way of living.
When she lived in Sweden, she created a film about people's lives there.
She enjoyed living and working in Sweden.
She ended up living in Sweden for 12 years.
But then she went to Australia because she wanted to live in a warmer country.
She enjoys the pace of life in Australia. It's close to the pace of life the people live in Sweden.

Figure 8: Gap-fill exercise on the different forms of live/life

Words and their forms (2) – the pronunciation of lives

Look at the following two sentences. Do you know how the form lives is pronounced there? If you don't, go back to the video and listen again.

1. Arts therapy can enable people to find something away from the challenges of their lives.
   Pronunciation: [laɪvz][lɪvz]
2. Judith lives in Australia.
   Pronunciation: [laɪvz][lɪvz]

Justify your choice.

Figure 9: Select exercise on the pronunciation of lives

---

18 “Browse words” function is the name for the web concordances in ELISA.
Words and their forms (3) – other words with more than one form

Now go to the word lists of the interview and search for other words which have different forms. Explore the grammatical functions of the words you have found (i.e. find out why they appear in different forms). Then fill in the correct word form:

I grew up in the Los Angeles area but I spent quite a lot of time in Northern California as well. And as I mentioned earlier I lived and worked in Sweden. I taught English to Swedish business people who needed English in their work.

I ended up living in Stockholm for twelve years. I learned a lot about the culture there. But whilst I had enjoyed being able to travel and work throughout Europe, I wanted to settle in a warmer country and migrated to Australia.

At the moment, I'm on vacation in Alice Springs. I'm taking a break from work and responsibilities in Melbourne. I haven't travelled around Australia very much. So, I'm really enjoying this. I'm learning a lot about the aboriginal culture.

Figure 10: Gap-fill exercise on the pronunciation of lives

As our own pilot study of ELISA in a secondary school has shown, one of the major problems learners face when working with corpus data such as word lists or concordances is that they do not necessarily have the analytical skills required for the interpretation of this type of data. It is, therefore, crucial to give learners very focussed and manageable study tasks and to design tasks which are communicatively relevant (or authentic). So, while learners are encouraged in the above exercises to make use of the ready-made corpus resources (word lists, concordances as well as the video clip to listen for pronunciation), the exercises as a whole are discourse-oriented and aim to help the learners apply word forms in utterances taken from the interview (sometimes slightly adapted).

Other exercises which fulfil the 'integration requirements' are, for example, explorative tasks around one individual interview such as giving a summary of the arts therapist's life or – since the arts therapist devotes part of her interview to her travel experience – making learners talk about their own travel experience and plans as well as writing a CV of the arts therapist's life and their own CV. In all of these activities, the interview word lists together with the concordances can serve as a starting point. The exercises can be applied according to the learners' level of competence, keeping in mind that giving a summary of the interview text is a simpler exercise than 'transforming' it into a CV or using it as a model to talk about one's own life (see also Widdowson 1979).
Examples of explorative tasks involving more than one interview are the exploration of different accents as mentioned above as well as the study of different viewpoints and related means of expression, e.g. on the water debate mentioned in section 3. A project task related to the water debate would be a learner (group) research project on the water policy of their home town authorities and the preparation of a role play or debate of different points of view. A whole range of such tasks has been developed for ELISA, e.g. in pre-service teacher training seminars at the University of Tübingen19 and in the preparation of the pilot study on the use of ELISA in a secondary school. These can serve as a resource pool for the further use of ELISA and as models for the creation of other corpora and related materials.

6 Conclusion

The corpus design solution suggested here caters for a range of learner needs and interests, and supports a variety of teaching methods. It enables learners and teachers to access a corpus from different but pedagogically appropriate 'entry points' and explicitly supports the possibility to choose and combine different methods of exploration. A corpus designed according to this method is a pedagogical resource which can be exploited more comprehensively than has traditionally been possible with corpora in language learning and teaching. Furthermore, the enrichment of a corpus with relevant materials prevents reiterated ad-hoc creation and re-invention of these by teachers and hence further contributes to a better usability of corpus materials in the classroom.

A small and dedicated corpus as outlined in this paper has an obvious advantage over working with individual texts, which usually do not provide a sufficient and systematic range of material on the relevant point of learning and overemphasise an individual speaker's/writer's idiosyncrasies. At the same time, such a corpus may be preferable to working with larger corpora, which make access to individual texts usually more difficult, offer a wide range of texts that is too diversified for a learner and generally provide amounts of data that cannot be handled well by learners.

It is the possibility to study and compare a manageable number of speakers/writers with regard to their different but similar ways of talking/writing about a particular topic that constitutes the attraction and advantage of such a corpus. The method which will eventually be used to analyse the speech/text samples – corpus techniques, text-based

19 The students in these seminars made numerous creative suggestions for the pedagogical use of ELISA.
exploration or a combination of both – depends on the study question, on the analytical skills and level of autonomy of the learners.

At present, much of the corpus processing, especially the keyword-based annotation, has to be done manually. This calls for support tools which are tailored to pedagogical needs of corpus creation. Such tools will be developed in the SACODEYL project (cf. section 1). Moreover, as Müller & Strube (this volume) show, more sophisticated, (semi-) automatic methods of text analysis and annotation are also in reach and if we follow the 'percolation' argument mentioned in section 1, these methods should eventually find their way into Applied Linguistics contexts.

7 References

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents, ReCALL 17 (1), 47-64.
Chambers, A. & Í. O'Sullivan (2004): Corpus consultation and advanced learners' writing skills in French, ReCALL 16 (1), 158-172.


