This is the final draft of a chapter published as: Braun, S. (2008). Designing and exploiting small multimedia corpora for autonomous learning and teaching. In: E. Hidalgo, L. Quereda, J. Santana (Eds), *Corpora in the Foreign Language Classroom. Selected papers from TaLC6*. Amsterdam: Rodopi, 31-46.

Designing and exploiting small multimedia corpora for autonomous learning and teaching

Sabine Braun, University of Tübingen

The use of corpora in the second-language learning context requires the availability of corpora which are pedagogically relevant with regard to choice of discourse, choice of media, annotation and size. I here describe a pedagogically motivated corpus design which supports a direct and efficient exploitation of the corpus by learners and teachers. One of the major guidelines is Widdowson's (2003) claim that the successful use of corpora requires a learner's (and teacher's) ability to 'authenticate' the corpus materials. In line with this, I argue for the development of small and pedagogically annotated corpora which enable us to combine two methods of analysis and exploitation to mutual benefit: a corpus-based approach (i.e. 'vertical reading' of e.g. concordances), which provides patterns of language use, and a discourse-based approach, which focuses on the analysis of the individual texts in the corpus and of linguistic means of expression in relation to their communicative (situational) and cultural embedding. To illustrate my points, I use a small multimedia corpus of spoken English which is currently being developed as a model corpus with pedagogical goals in mind.

1 Introduction

By now it has been well established that corpora can be used in a variety of ways to support language learning and teaching. As a result of improved opportunities to make corpora available to non-specialists (i.e. language learners and teachers), the focus of academic discussion and practical corpus application has clearly shifted from the indirect uses by e.g. publishers or syllabus designers to direct uses by learners (and teachers) themselves. Gavioli & Aston, for example, emphasise "the potential of corpora as tools in the hands of learners" (2001: 238) and show how corpora as a pedagogical resource can considerably enrich the learning and teaching environment and support autonomous language learning and teaching. However, we (still) encounter theoretical, methodological and practical problems in this area of corpus use:

From a methodological and practical point of view, we face the dilemma that the most widely accessible corpora have been developed with linguistic research goals in mind and are not necessarily the corpora with the most obvious pedagogical value.[1] A plethora of smaller and often genre-specific corpora has emerged to remedy some of the shortcomings of 'mainstream' corpora (especially with regard to size and diversification of content) but, as Guy Aston has pointed out at the TALC6 conference, these smaller

'home-made' corpora have not travelled well beyond the institution in which they have been created (c.f. Aston 2004).

From a theoretical point of view, if we look at corpora in terms of the overall goal of language learning, we are faced with another discrepancy: according to the helpful distinction between *text* and *discourse* made by Widdowson (1979) and reiterated with regard to corpora in Widdowson (2003), corpora are a collection of *text*, i.e. *products* of language use isolated from any communicative situation.[2] In contrast, language learning is concerned with *discourse*, i.e. the use of language in concrete communicative situations. Therefore, the acquisition of linguistic means of expression must be embedded in the acquisition of the knowledge and skills required to master the *processes* of producing and understanding discourse in a foreign or second language (for further theoretical discussion of this point c.f. Braun (2005). With particular regard to corpora as containing samples of 'real' language use, Widdowson (2003) has emphasised the following crucial point:

> People make a text real by realizing it as discourse, that is to say by relating it to specific contexts and communal cultural values and attributes. And this reality does not travel with the text. So although this is a real example of actually occurring text, learners will be unable to ratify it as an example of discourse if, as outsiders, they are not privy to the contextual conditions upon which the discourse realization depends. Even if they could track down the meanings of all the unknown words in their dictionaries, they are still unlikely to realize the effect of the pragmatic use of these words, which makes the text real for the discourse community for which it was designed. (Widdowson 2003: 98)

If learners are to use corpora successfully, they need to be (en)able(d) to 'authenticate' the texts for themselves. Or in other words: a corpus is like a 'text museum'. The exhibits are real (as real as e.g. historical artefacts) but if you enter without preparation and appropriate background knowledge, your benefits will be limited. The question which then arises is how the authentication process can be best supported in connection with corpus-based language learning. A number of suggestions have been made in this respect:

Gavioli & Aston (2001) suggest that authentication can be achieved through (corpus) observation. While they firmly rely on concordances as a starting point for corpus observation, they also point out that this has to be complemented by analyses of the wider co-texts behind concordance lines. Moreover, they emphasise that learners have to be directed gradually towards the use of corpora, starting with smaller, less

diversified corpora and mediation through the teacher. They also demand more reflection on the integration of corpora in the overall learning process.

While Gavioli & Aston appreciate the fact that concordances put a large number of texts at a learner's fingertips "amongst which learners can choose examples which they find it possible to authenticate in reading" (2001: 243), Mishan (2004) points out that corpora due to their electronic format (in which the texts are usually deprived of their original layout and of complementary materials such as images) and due to their presentation in the form of lists and concordances can easily obscure the communicative intent and wider socio-cultural context of the texts behind the concordance lines. She shows how an exploratory approach to corpora (data driven learning) can yield many effects for the learner – from an understanding of lexical and grammatical phenomena to a better understanding of cultural contexts. She also suggests that learners should create their own corpora to increase familiarity with the texts in the corpus.

Tribble (1997; 2001) as well as Henry & Roseberry (2001) follow a different approach: they use small corpora and show how these enable us to combine two methods of analysis and exploitation to mutual benefit: A corpus-based approach (through frequency and keyword analyses, concordances and 'vertical reading') is complemented by a discourse-based approach, which focuses on the analysis of the individual texts in the corpus and on "whole-corpus reading" (Henry & Roseberry 2001: 99). This serves to study linguistic means of expression in relation to their communicative (situational) and cultural embedding and goes beyond the exploration of texts or paragraphs thereof from the KWIC concordance. In this approach the texts are analysed individually as units in their own right.

I will concentrate here on yet another, complementary aspect: I believe that – beyond the methodological solutions for pedagogical corpus exploitation – the use of corpora in the second-language learning context also requires the availability of (dedicated) corpora which are pedagogically adequate with regard to choice of discourse, choice of media, annotation and size. If we pay more attention to design-related aspects, we can turn corpora into even more efficient resources for language learning and teaching. I will therefore focus on a pedagogically relevant corpus design. I will first outline some pedagogically motivated corpus design considerations (Section 2) and then demonstrate how the proposed design can be used to support autonomous language learning and teaching (Section 3).

2 Designing a pedagogically relevant corpus

To exemplify the discussion in this Section, I will use a small corpus of spoken English (ELISA – English Language Interview Corpus as a Second-Language Application), which is currently being developed with pedagogical goals in mind.[3] At present ELISA contains 15 video-based interviews of 5 to 15 minutes with English native speakers who talk about their professional career. The corpus is being constructed as an experimental corpus to develop and evaluate a conceptual and technical solution for the design proposed here.[4] It is currently being implemented as an XML-based web application.

2.1 Elicitation of the data

As Tribble (1997) has pointed out, *topical relevance* of the corpus materials is one of the key factors supporting discourse authentication. Another is a certain degree of *homogeneity* in the corpus content (c.f. Aston 1997).

The interviews which constitute the ELISA corpus belong to the broad field of professional English. The corpus includes speakers working in education, local politics, tourism, banking, environmental protection, sports and the media and in particular, a number of speakers with 'culture-embedded' careers, e.g. a cattle farmer in the American Southwest and a director of a tropical wildlife centre in the Northeast of Australia. To elicit the data, the speakers received a short briefing with regard to the interview theme. More specifically, they were asked to cover different topic areas, e.g. how they started their career or business, the kind of projects they are working on, challenges for their business, their daily routines and future plans. During the actual interview, the speakers were interrupted as little as possible to ensure a free flow of natural speech. Thus the recordings have a narrative character.

The speech of our samples is best characterised as being spontaneous but thematically focussed (elicited spontaneous speech). This makes the corpus different from a number of other spoken corpora which have been collected for different purposes and (partially) contain unplanned spontaneous conversations (unprompted speech). Understanding them is highly dependent on the immediate situational context and/or on a high degree of familiarity with the conversational interlocutors. They cover situations which are not necessarily characteristic for the situations language learners are most likely to find themselves in. The interviews in the ELISA corpus rely on different types of contexts and knowledge, in particular on the wider cultural and professional context and

related background knowledge, which is highly relevant for most language learners and needs to be explored by them if they want to master the target language successfully.

In other words, topical relevance is achieved through the vocational orientation of the corpus, since with increasing mobility, for instance, there is a growing demand for oral communication in professional contexts. The variety of subject fields which is covered in the interviews make the corpus attractive to a wide variety of learners from different backgrounds (school, university, adult education). At the same time, the common overall theme of the interviews creates a high degree of *intertextual coherence* (i.e. the homogeneity), which makes discourse construction and situational embedding even easier.

2.2 Transcription and encoding

Transcribing spoken language always raises methodological questions, and transcription conventions are dependent on the anticipated purpose. With regard to pedagogical use conventions still have to be developed. In ELISA, we have decided to keep the transcripts as closely as possible to written conventions (with regard to spelling and punctuation) without compromising the crucial features of spoken language, which have to play an important role in language learning and teaching (c.f. Carter & McCarthy 1995). We have however, excluded hesitation and filler words ("uhm", "er"), word repetitions and length of pauses. While they do, of course, contribute to a correct understanding of an utterance or entire situation, their exclusion from the transcripts here makes working with the transcripts much easier for the non-linguist, and the corpus users are strongly encouraged to read the transcripts in conjunction with watching the video from the interview (the integration of the videos will be described below, c.f. Section 2.4).[5] In contrast, we have indicated syntactic breaks and points where the video has been cut to 'warn' the transcript readers of structural or – in the case of video edits – (minor) pragmatic disruptions.[6]

While topical relevance and coherence in connection with manageable corpus size and adequate transcripts are important ingredients for a pedagogically relevant corpus, these features alone do not seem to suffice to encourage a more wide-spread use of the smaller 'home-made' or 'dirty' corpora (c.f. Aston 1998 and Tribble 1997 respectively) which are now in existence in so many institutions. One crucial point, I believe, is that at their origin these corpora fulfil an immediate demand for particular contents, i.e. they match an institutional context which makes them immediately relevant for their users (and makes authentication very easy). This effect is even increased where learners themselves have been involved in the creation of the corpus. A more wide-spread use of

these corpora requires a more explicit description and documentation as bridge-building elements to support prospective autonomous uses and authentication processes.

In the ELISA corpus, each interview file has received a header which is specifically geared towards information relevant for learners and teachers. Apart from the rather common classification of the interviews with regard to language variety, speakers, duration, number of word tokens, speech rate (words per minute), they have also been given a short non-linguistic description of their content, which specifically addresses learners' curiosity. I.e. they have an appetizer function but at the same time they give a first clue towards the interview situation. Some examples of these descriptions are shown in Figure 1.

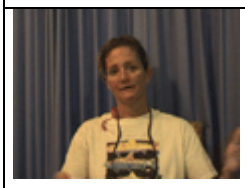
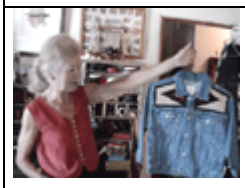
	**Bishopswood Centre** Bishopswood Centre is an environmental education centre in the West Midlands (UK). One of the project managers talks about the aims of the centre, the projects it is involved in and the educational programmes it runs.
	**Horse Caravanning in Ireland** Dieter lives in county Wicklow (Ireland) where he rents out horse-drawn caravans to people on holiday. He talks about how his business has grown and how tourism has changed since he started about 30 years ago.
	**A tour guide from Uluru (Ayers Rock)** Chris works in Yulara resort in the Northern Territory (Australia). He is a tour guide at Uluru. He talks about his life in Yulara, his daily routines, the people he meets and the people who are the traditional owners of Uluru.
	**The life of a travel publicity person** Heather is in the travel business. Originally from America, she now lives in Sydney (Australia). She has worked in publishing and advertising, covering e.g. Bermuda and the Caribbean, South and Central America and the Pacific.
	**Making clothes out of old Navaho rugs** Barbara from Santa Fe (New Mexico, US) uses old Navaho rugs to turn ordinary denim wear into ethnic-style clothing. She explains how she started her business and what is involved in making and selling these clothes.

Figure 1: Interview descriptions

Moreover, each interview has received a short linguistic characterisation including information about the language used in the interview, the learning activities it is useful for, the skills which can be trained with the interview (or different parts of it) and the knowledge which is required to understand the interview and/or can be acquired when studying it. This information serves as a didactic help for learners and teachers. With the help of XML technology, the different pieces of information in the document header can be displayed or hidden as required in different views of the corpus materials (c.f. Ward 2002).

2.3 Annotation

To facilitate the detailed study of the materials, an annotation which accounts for pedagogically motivated corpus queries is necessary. Existing annotation schemes mainly take into account linguistic research goals and are not necessarily suitable for pedagogical queries. Besides that, they are often too sophisticated for teachers and learners without linguistic background.

Aston (2002: 10) summarises the major uses of corpora by learners: apart from "form-focussed activity" (study of lexico-grammatical patterns of use, based on concordances and the like) he also lists "meaning-focused activity" (to study and distinguish different meanings of a word), "skill-focused activity" (to practice e.g. reading comprehension skills), "reference activity" (corpus as look-up facility to support reading, writing, translating) and "browsing activity" (curiosity-led exploration, c.f. also Bernardini 2000). To support these activities, the following ways of working with a corpus such as ELISA are particularly relevant:

- working with a particular interview (or more generally with a particular text in any corpus);
- studying of a particular section of an interview;
- comparing similar sections across interviews;
- studying linguistic means of expression.

A combination of these will be necessary if a corpus is to be a comprehensive 'tool' supporting the acquisition of relevant skills and knowledge. In fulfilment of the first of these requirements, the ELISA corpus gives immediate access to each individual interview – via an index page where all interviews are listed together with relevant information from the document headers, in particular the summary descriptions. This makes the corpus different from most other corpora, in which texts can be viewed but

where their presentation is less 'immediate' and where they are not necessarily complemented by a description.

While the access to the interviews is achieved through the overall file and header structure of the corpus, the fulfilment of other requirements listed above is achieved with the help of a topic-based annotation structure.

To this end, the interviews have been divided into sections. This was done a) on the basis of the topics which the speakers were asked to cover and b) on the basis of an analysis of characteristic means of expression or features such as tense, aspect and modality. An overview of the topic areas together with a brief explanation is given in Figure 2. The sections which fall into the category "challenges", for instance, are marked by a frequent occurrence of the modal verbs of obligation *have to* and *must*. The segmentation was carried out by three different analysts with linguistic and discourse-analytic background and familiar with the whole corpus. A similar segmentation has been suggested by Tribble (2001) for a small genre-specific corpus.

TOPIC	**EXPLANATION**
location	descriptions of where the speakers work and live
the past and getting started	including how the speaker grew up, their education, their past occupations and how they started their current business or job
current work	descriptions of what the speakers do, their business, their aims and values and what makes their business or institution special
projects	examples of project-based work, e.g. partnership projects
education and training	examples of work in education and training, e.g. providing management training
business issues	e.g. the organisational structure of the business, marketing strategies, business plans
challenges	what makes business difficult …
future plans	e.g. planned career changes, planned projects, goals for the speaker's business
job routines	e.g. daily routines and routines in the office

Figure 2: List of topics in the ELISA corpus

In addition to their (discourse-oriented) topic annotation, the sections have also been annotated with two other keys: firstly they have received one or several subject keys where appropriate. These can be used to retrieve characteristic means of expression for an

area a learner is particularly interested in. As the subject-specific differences between the interviews become more apparent in some of the topics (e.g. "current work" or "projects") and less so in others (e.g. "job routines"), the subject keys are a more effective filter for targeted queries than the classification of an entire interview into subject areas.

Secondly, the sections have been annotated with one or several grammatical keys to indicate prominent grammatical structures or to mark a section as an example for the linguistic realisation of a grammatical function. Examples are the above-mentioned modal verbs of obligation in the topic "challenges" or the forms for expressing habitual past in the topic "the past and getting started".

2.4 Retrieval and presentation of the data

While designing our model for a pedagogically relevant corpus, the importance of topic segmentation (as outlined in Section 2.3) has become very clear to us. It can serve a multitude of functions in the retrieval and presentation of the data.

First of all, the topic segmentation facilitates the exploration of each individual interview. Whole-text reading, which has been identified in Section 1 as one important step towards discourse authentication, is supported by this segmentation because the topic headings can be displayed in the interview texts, which gives them some sort of structure and thus facilitates reading (and listening) comprehension.

Secondly, the segmentation is also used to give direct access to each interview section individually as well as to comparable sections across interviews. The former is important because the sections, which have an average length of 1 to 3 minutes, are appropriate units for closer inspection by the learner. I will return to this point later (c.f. Section 3). The comparison across interview sections will enable the corpus users to study e.g. the means of expression which are typically related to a particular topic.

One topic can have different realisations across interviews. The topic "current work", for instance, can be realised as "What we do" or "What I do", "Our aims and values", "What makes us special" and "What makes us different from our competitors". This two-foldedness takes account of the fact that a topic can have different aspects. Talking about their current work, for example, most speakers cover more than one of the aspects named above. Furthermore, it accounts for 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 free lancers and just talk about themselves. Yet others have an educational background and, for instance, do not talk in

terms of "competitors". Figure 3 shows the topic structure of an interview with a representative from a Chamber of Commerce.

The topic-title dichotomy allows to select titles which match the content of the relevant interview section as closely as possible. So, while the classification of sections into topics supports and enables the exploration of similar sections across interviews, the individuality in the topic *titles* helps the learners to explore individual interviews as described above.

TOPIC	**TOPIC TITLE**	**LENGTH**	**WORDCOUNT**
current work	What we do	2:35	359
projects	Controlling the growth rate	2:40	414
projects	Managing natural resources	2:07	338
business issues	Wage policy	1:40	264
challenges	The current job situation	1:05	143
challenges	Challenges and perspectives for the economy	2:53	427
job routines	Networking	2:23	396
education and training	Offering seminars for business people	0:44	98
current work	Reasons for joining us	1:31	216
job routines	Lobbying	1:46	278
challenges	Language barriers	1:40	238

Figure 3: Structure of the Chamber of Commerce interview

Another important function of the segmentation is that the sections are the basic units for the alignment of the transcripts with the video. The links to the video clips are integrated in the XML files on the basis of SMIL technology[7] in such a way that the video can either be watched as a whole or called up for each section.

Furthermore, as pointed out in Section 2.3, the segmentation will also allow queries by subject key and grammatical key, e.g. to elicit subject-related vocabulary or means of expression used to realise a particular grammatical function. Finally, the sections will be the most important unit for the integration of complementary materials ('pedagogic enrichments') into the corpus (e.g. exploration tasks and exercises).

The best way to present a corpus like the one outlined here is a Web interface. Access to the corpus materials can then be enabled via index pages for a) the interviews, b) the topic classes and the individual sections, c) the subject keys and d) the grammatical functions. The interview index, which can also be used to display the summary descriptions and other meta-information, can be customised for different user groups (at

the simplest level, just distinguishing between different types of meta-information relevant for learners and teachers).

3 Working with a pedagogically relevant corpus

After outlining a pedagogically driven corpus design and some of the retrieval options it holds in store, I will give some specific examples of how learners and teachers can exploit such a corpus. I will use an investigation of differences in word meanings and corresponding lexico-grammatical patterns (Section 3.1) and the exploration of grammatical functions (expressing habitual activity in the past, Section 3.2) to illustrate the overall idea of topic-based corpus exploitation and to demonstrate the benefits of combining discourse-based and corpus-based methods of analysis. The examples are taken from the ELISA corpus.

3.1 Exploring word meanings and corresponding lexico-grammatical patterns

The basis for the discussion which follows is a passage which represents the first section of an interview with a (former) president of the Chamber of Commerce of Santa Fe (New Mexico, US). The outline structure of this interview was shown in Section 2.4, Figure 3. Anyone who chooses to analyse this interview is likely to notice the frequent occurrence of the word *business/es* right in this first section.

> **What we do**
> The chamber in Santa Fe is a traditional chamber of commerce in a sense of a <break/> traditional chambers of commerce in the United States are supported almost entirely by their membership, by the **businesses** within the community. They get very little assistance of any kind from any other source other than from their members. In New Mexico there are only three chambers of commerce that are the traditional chambers of commerce. One is in Albuquerque, the Albuquerque chamber, and they're the largest chamber of commerce in the state. The second is here in Santa Fe, we're the second largest, and then the third is in Las Cruces. And the difference is quite simply: in all other communities in New Mexico, the chambers of commerce receive a significant amount of money either from their city or from their county government to perform tourism functions, to act as the marketer for the community. And in Santa Fe, government has done that for themselves, as they do in Albuquerque and Las Cruces. So, it makes us dependent upon our members, which for chambers of commerce is really the best kind of organisation in

this country. And the reason it's the best kind of organisation is it allows us to take exception with things that the city may do. And in Santa Fe <break/> this is a very difficult community in which to do **business** because there is so much anti-**business** sentiment. And that comes from a lot of different sources, but the primary source that we get it from is from city government, where there is not a **business**-friendly atmosphere at city government. They would just as soon make things more difficult for **business** as make it easy for **business**. So it allows us as a member-driven organisation to take exception. When they want to do things that are detrimental to **business**, we then take exception with those. That assists us in establishing our reputation and our standing with our membership. So, hopefully what happens is that all that energy makes it easier for us to get support from our members and we don't have to depend on money from any other source. [chamber_of_comm_us 0:00]

In the British National Corpus, *business/es* occurs 394 times and ranks 41 among all nouns in the BNC. In the ELISA corpus *business/es* occurs in most of the interviews and 71 times altogether, which makes it a very prominent and relevant content word, taking rank 6 of all nouns and offering multiple study opportunities in the corpus.

The first task at hand could be an analysis of the different meanings of *business* in its singular and plural forms in the above section and the formation of an adequate hypothesis.[8] This task can be followed by the investigation of a KWIC concordance for *business/es* from the entire corpus to confirm, extend (or correct) the hypothesis. Alternatively learners could be advised to compare concordances from the individual interviews. This would be feasible here due to the restricted size of the corpus. At the same time it would reveal good results due to the prominence of the word (in many interviews). A concordance from the whole Chamber of Commerce interview is shown in Figure 4.

1.	s, making new contacts. And we do a	**business**	After Hours which is not <br
2.	ity government, where there is not a	**business**	-friendly atmosphere at city
3.	omething that's very big in American	**business**	- business people getting to
4.	siness because there is so much anti-	**business**	sentiment. And that comes fr
5.	at's very big in American business -	**business**	people getting together and
6.	re going to face in their day-to-day	**business**	. <cut/> </speaker> <speake
7.	y difficult community in which to do	**business**	because there is so much ant
8.	ople getting together and exchanging	**business**	cards, making new contacts.
9.	soon make things more difficult for	**business**	as make it easy for business
10.	ult for business as make it easy for	**business**	. So it allows us as a member
11.	that are advocating growth and good	**business**	practices. And that's what w

12.	nd to interact with them to gain new	**business**	contacts, to either sell som
13.	eption, along with most of the other	**business**	organisations in the state.
14.	o network, the ability to meet other	**business**	people and to interact with
15.	embers to get together to meet other	**business**	people. We also offer <break
16.	re donating their time to help other	**business**	people. And we do a series o
17.	ve high-paying jobs where people own	**businesses**	, and then we have people w
18.	d Executives. And they're successful	**business**	people who have retired, but
19.	ces issues, just various things that	**business**	people are going to face in
20.	then we have people who work for the	**businesses**	that feed tourism, which a
21.	entirely by their membership, by the	**businesses**	within the community. They
22.	e then, someone can click onto their	**business**	and go directly to their web
23.	condary reason is exposure for their	**business**	. And we provide that through
24.	m almost anywhere prefer to do their	**business**	from an area that has a very
25.	olved in technology who can do their	**business**	from almost anywhere prefer
26.	to do things that are detrimental to	**business**	, we then take exception with

Figure 4: concordance of *business/es* in the Chamber of Commerce interview

The advantage of a comparison across interviews is that it reveals differences in the use of the word in relation to the overall interview contexts. Due to the easy access to other interviews and individual interview sections in which the word figures prominently (via the subject key) the different uses can easily be correlated with the interview situations. This will help to underpin the learners' hypotheses and contribute to a deeper understanding of the differences.

The formation (or confirmation) of meaning hypotheses should be complemented by an investigation of characteristic lexico-grammatical patterns for each of the meanings including article use, use of singular and plural form, compound options (e.g. *anti-business, business-friendly, business people / cards / contacts / practices*), collocations (e.g. *to do business with somebody*) as well as a closer look at the interesting phrase *Business After Hours*. Some research on this phrase, e.g. on websites of Chambers of Commerce (in the US at least) will allow learners to familiarise themselves with an interesting part of (American) business culture. In addition, the study of business can be extended to a comparative analysis of similar words (e.g. *business – industry – economy, business – enterprise – company – firm, business – trade*).

Beyond the explorations outlined above, the interview section also provides a range of opportunities for other types of investigations. Learners could, for example, be asked to analyse the means of expression which the speaker uses here to put up a contrast between the aims of the Chamber of Commerce and those of the local government. The independence of the Chamber from the government is clearly put in a positive light (*it*

allows us to, that assists us in) whereas the local government appears in a negative light *(anti-business sentiment, not a business-friendly atmosphere, detrimental to business, make things more difficult*). Such investigations open up the way for the study of some argumentation strategies, which – once they have been noticed – can again be confirmed with the combined methods of interview-based, topic-based and concordance-based analyses.

3.2 Exploring a grammatical function

The interview section which I will use to illustrate how a grammatical function can be explored is taken from an interview with two founders of small a Credit Union (a membership-based type of bank). In the excerpt given below they talk about their first board meetings after founding the bank.

> **Organising meetings**
> **Nora:** And we **used to** have board meetings. Now we got a little bit better handle on it, but some board meetings **would** last for three or four hours. It was just intense. Many things that needed to be accomplished and voted on, and they had to get done.
> **Philip:** And sometimes <break/> and I was trying to keep things down to an hour or two at the most. My philosophy was anything over an hour is no longer a meeting, it's a party. And whoever has <break/> whoever called the party should supply beer. But we never had anything to drink up there. All we had was
> **Nora:** water
> **Philip:** water.
> **Nora:** It was probably just as well.
> **Philip:** Bottled water. We **would** bring in bottled water by the case. And people **would** drink bottled water.
> **Nora:** It was just as well. We had to get a lot done.
> **Philip:** I remember once in a while somebody **would** bring in cookies and that **would** sustain us through a lengthy meeting.

The interesting point here is the use of different expressions to describe habitual activities in the past (*used to, would*). In the entire corpus, both forms occur with nearly identical frequency (*used to*: 20, *would* in this function: 24). A concordance-based analysis will, of course, reveal that the modal auxiliary *would* has a range of other functions or meanings (which in itself would be a topic of investigation). It will also show the interesting interrogative construction *what did we used to do* (in the Credit Union interview) and the *do*-construction *didn't used to*.

More interestingly in our context though, it also reveals that the two forms are distributed in the corpus very unevenly. Whereas *used to* occurs in many interviews, *would* is restricted to the Credit Union interview with just some scattered uses in very few other interviews. The ensuing question to be explored is why Nora and Philip use this form so frequently. Once again, reading the above passage and even the entire Credit Union interview carefully (and watching the speakers in the video!) will be more helpful than just studying the concordances to answer this question. In the *would*-concordances it is even difficult to decide which meaning was intended. Reading the whole interview, in contrast, will reveal the atmosphere (i.e. the discourse situation): the founders of the Credit Union are very proud of what they have achieved over the years, and it is with some amusement but also with a lot of emotion that they remember the very beginnings of their business. While they are telling the story of how they got started, they reminisce and take an insider perspective again. Hence the use of the form *would*, which seems somewhat less formal and less matter-of-fact than the form *used to* and creates the impression of a 'live report'.

4 Conclusion

The approach described here combines discourse-based and corpus-based forms of analysis. The crucial point is that learners use the same resource – a pedagogically motivated corpus – a) to discover interesting items in the discourse-based study of the materials and b) to explore them further, e.g. with corpus-based methods of investigation.

The corpus *design* outlined here certainly involves a degree of manual work which restricts the possible size of such corpora, but such size restrictions also have a pedagogical motivation as has been acknowledged by others researching the field (e.g. Ghadessy, Henry & Roseberry 2001).

Something similar can be said about the topic-based annotation. At first this might just appear to be a good compromise between the really 'quick and dirty' ways of creating corpora and the time-consuming creation of exhaustively annotated corpora. The former are very useful but offer only limited exploration and query possibilities, the latter are difficult to produce, because in order to be fruitful for pedagogical purposes they would have to include annotations at the semantic and pragmatic level, which are hard to achieve. On closer examination, however, the topic-based annotation turns out to be much more than a good compromise: the topics as they are exemplified by the ELISA corpus form a highly relevant unit of investigation in the pedagogical context.

The whole approach as described here is intended as a model which can be transferred to meet the specific needs of educational and training institutions. Despite the resources required to develop a corpus of this type, it is a feasible and rewarding task.

The exploitation opportunities offered by such a corpus support learning activities which are likely to create long-term consolidation as they lead to a deep understanding and help to build up comprehensive (strategic) knowledge of how to use the means of expression investigated. This is in line with constructivist models of learning, where learning is viewed as a process of knowledge construction and therefore is best supported by a rich, interesting and motivating learning environment (c.f. Wolff 1994). In the end, these considerations also support the requirement of discourse authentication put forward by Widdowson (2003) and outlined in Section 1.

Learner (and teacher) requirements must be at the top of the pedagogical agenda for corpus design and exploitation. Only this way it will be possible for learners and teachers to turn the 'text museum' into a lively and motivating 'educational experience'.

Notes

[1] But c.f. Bernardini (2000) and Aston (1997; 2000) for suggestions on how to use larger corpora with language learners.

[2] C.f. also Brown & Yule (1983) for the text/discourse distinction.

[3] This development has been partially funded by a research grant from the University of Tübingen. A demo of the corpus is available at www.corpora4learning.net/elisa, showing some of the audiovisual materials and some of the features outlined below. The copyright of the ELISA is held by the group of Applied English Linguistics at the University of Tübingen.

[4] While we have tried to achieve a balance between different varieties of English (currently American, British, Australian and Irish), gender, age, educational background of the speakers and the subject area they are working in, considerations of sampling and representativeness will have to play a bigger role, if the corpus is to be expanded.

[5] Having said that, the transcripts provide a basis for other uses of the materials for which they could simply be 'expanded' or encoded with additional information.

[6] Main reasons for editing the video materials were changes of place within an interview (e.g. between a speaker's office and an outdoor situation) and meta comments by the interviewees (e.g.. 'Do you want me to go in more detail?' or 'I don't know how much detail you want…'), followed by clarification sequences between interviewee and interviewer.

[7] SMIL stands for Synchronized Multimedia Integration Language (c.f. http://www.w3.org/AudioVideo/) and is used e.g. by the RealPlayer.

[8] It could be argued that learners who are able to understand the passage from the Chamber of Commerce interview are likely to be familiar with the meanings of *buiness/es*. However, they may not necessarily be aware of the correlation between the different meanings and their characteristic patterns of use. In this sense a detailed study of the passage would surely have an awareness-raising and/or confirmation effect even for more advanced learners.

Works cited

Aston, G. 1997. 'Small and large corpora in language learning.' in Lewandowska-Tomaszczyk, B. and P. Melia (eds.), 51-62.

Aston, G. 1998. 'What corpora for ESP?' in Pavesi, M. and G. Bernini (eds.) *L'apprendimento linguistico all'università: le lingue speciali.* Roma: Bulzoni, 205-226.

Aston, G. 2000. 'Learning English with the British National Corpus.' in Battaner, M.P. and C. López (eds.) *VI jornada de corpus lingüístics.* Barcelona: Institut universitari de lingüística aplicada, Universitat Pompeu Fabra, 15-40.

Aston, G. 2002. 'The learner as corpus designer.' in Kettemann, B. and G. Marko (eds.) *Teaching and learning by doing corpus analysis.* Amsterdam: Rodopi, 9-25.

Aston, G. 2004. 'Corpus upon corpus: a bout of indigestion?' Paper presented at TALC6 in Granada, Spain.

Bernardini, S. 2000. 'Systematising serendipity: proposals for concordaning large corpora with language learners.' in Burnard, L. and T. McEnery (eds.) *Rethinking language pedagogy from a corpus perspective.* Frankfurt: Lang, 225-234.

Braun, S. 2005. 'From pedagogically relevant corpora to authentic language learning contents.' *ReCALL*, 17/1: 47-64.

Brown, G. and G. Yule. 1983. *Discourse analysis.* Cambridge: CUP.

Carter, R. and M. McCarthy. 1995. 'Grammar and the spoken language.' *Applied Linguistics* 16/2: 141-158.

Edwards, J. and M. Lampert (eds.). 1993. *Talking data: transcription and coding in discourse research.* Hillsdale, NJ: Lawrence Erlbaum.

Gavioli, L. and G. Aston. 2001. 'Enriching reality: language corpora in language pedagogy.' *ELT Journal* 55/3: 238-246.

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

Henry, A. and R. Roseberry. 2001. 'Using a small corpus to obtain data for teaching a genre.' in Ghadessy, M., Henry, A. and R. Roseberry (eds.), 93-133.

Lewandowska-Tomaszczyk, B. and P. Melia (eds.). 1997. *PALC'97: Practical Applications In Language Corpora.*. Lodz: Lodz University Press.

Mishan, F. 2004. 'Authenticating corpora for language learning: a problem and its solution.' *ELT Journal* 58/3: 219-227.
Tribble, C. 1997. 'Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching.' in Lewandowska-Tomaszczyk, B. and P. Melia (eds.), 106-117.
Tribble, C. 2001. 'Small corpora and teaching writing: towards a corpus-informed pedagogy of writing.' in Ghadessy, M., Henry, A. and R. Roseberry (eds.), 381-408.
Ward, M. 2002. 'Reusable XML technologies and the development of language learning materials.' *ReCall* 14/2: 283-292.
Widdowson, H. 1979. *Explorations in applied linguistics*. Oxford: OUP.
Widdowson, H. 2003. *Defining issues in English language teaching*. Oxford: OUP.
Wolff, D. 1994. 'Der Konstruktivismus: Ein neues Paradigma für die Fremdsprachendidaktik?' *Die Neueren Sprachen* 93: 407-429.