Christian Fandrych, Cordula Meißner & Adriana Slavcheva (2012):

The GeWiss Corpus: Comparing Spoken Academic German, English and Polish

Schmidt, T. / Wörner, K. (Hg.): Multilingual corpora and multilingual corpus analysis (= Hamburg Studies in Multilingualism (HSM)), Amsterdam: Benjamins, S. 319-337.

The *GeWiss* corpus

Comparing spoken academic German, English and Polish

Christian Fandrych, Cordula Meißner and Adriana Slavcheva

Research on academic language has flourished in recent years, including academic German. The corpus resources available for larger, empirically based research projects remain, however, limited, even with regard to written academic language, and they are practically non-extant for spoken academic language. A detailed, empirical analysis of linguistic conventions and formulaic language used in (oral) academic communication is, however, all the more important in a day and age where our academic landscapes are becoming ever more internationalised. *GeWiss* aims to lay a foundation for such research: With *GeWiss* we are currently constructing a parallel corpus consisting of spoken academic language data from German, English, and Polish. Our contribution outlines the corpus design and the methodological procedures used for the construction of *GeWiss*. The project focuses, at least initially, on two key genres: Academic papers / student presentations and oral examinations. These are recorded, transcribed and stored in a searchable database. The corpus will comprise native speaker data from Polish, English, and German academics and students, as well as German as a Foreign Language (GFL) data of non-native speakers of German. These are recorded at all three partner institutions. *GeWiss* will therefore be the first corpus comprising GFL learner data. The paper focuses on corpus design as well as data collection and transcription. It also discusses selected research questions for which *GeWiss* can serve as an empirical basis.

1. Putting *GeWiss*[1] into context: Motivation, aims and applications

Creating a new corpus, and in particular one comprising (recorded and transcribed) spoken data, is a labour- and cost-intensive project. It requires clear ideas about the kind of research questions that it might help to answer as well as about the kind of applications it may be used for. It also needs strong support from various participants and institutions if the intellectual investment and financial expenditure is to be justified. In the case of *GeWiss*, there were three main types of motivation that have lead us to embark on this project: First, we were interested in creating a corpus that would be useful for comparing various academic languages in a comprehensive way, including genre-specific features. A comparison of genres across languages is an interesting but challenging proposition since it requires the creation of sub-corpora containing instances of discourse that fulfil similar functions in their respective institutional settings, cover broadly comparable topics in similar (sub-) disciplines, are delivered by speakers with a comparable degree of expertise, and are recorded and transcribed using the same standards. The advantages of such parallel corpora are, however, great: They allow not only for the comparison of linguistic features on a morpho-syntactic, phonological or lexical level, but also on the level of discourse, including the contrastive analysis of text architecture, of various strategies in constructing and linking arguments, of self-positioning in a scholarly argument, of addressing the audience and the handling of data, as well as of the linguistic devices and strategies that are typically employed for such purposes. This may unveil different scholarly conventions, traditions, styles, but also different choices in the conceptualisation of key phenomena and of 'doing science'. A second motivating factor for the building of our corpus was that it should also allow us to find out more about the differences between various degrees of expertise and language proficiency with regard to different academic genres: How do students compare with experts when it comes to scholarly presentations? How do L2 speakers of German from different backgrounds and with different L1s compare with German L1 speakers, and how homogeneous or varied are the L2 data? Our aim is to analyse selected linguistic aspects in such a comparative way, e.g. the use of connectors and modal particles for the construction of arguments; the use of metadiscourse; and lexical means employed to distance oneself from, or align oneself with, previous research. The third motivating factor was of a pedagogical nature: since all the project partners are eminently interested in issues to do with German as a Foreign Language, an

[1] *GeWiss* is funded by the Volkswagen-Stiftung as part of its funding scheme *Deutsch Plus – Wissenschaft ist mehrsprachig*, see http://www.volkswagenstiftung.de/foerderung/herausforderungen/deutsch-plus-wissenschaft-ist-mehrsprachig.html.

important aim of the *GeWiss* project was to create a database which could serve as a sound resource tool for creating appropriate teaching materials. A genre-oriented approach is of great importance for such pedagogical purposes, too, since genre-specific competences and a broader understanding of the social practices of which these genres are a part are required for a language learner to be able to act successfully in a foreign-language academic environment.

In short, the broader aim of our corpus is to enable research along various comparative dimensions with the increasingly globalised and mobile academic world in mind. Being able to sensitise scholars and students to differences and similarities in academic discourse across languages, to build their communicative competence in a foreign academic language based on realistic and authentic material becomes ever more important in a world in which academic exchange at every level is on the rise, yet the success of study time or research periods abroad are crucially dependent on appropriate communicative and social skills. Yet it is still very difficult to determine exactly what kind of linguistic and communicative skills and competences are required of somebody in order for him / her to be able to participate successfully in a foreign-language academic setting. A prerequisite for such questions to be answered is a sound empirical basis, ideally an accessible and fully searchable electronic corpus which allows for linguistic analyses of various kinds, from the phonetic level right through to the analysis of discourse practices in the target language community. *GeWiss* aims to be a first step towards such a corpus, initially focusing on only two spoken academic genres, but with the long-term aim of growing into one of the major reference corpora for academic German in comparison to other languages.

2. The design of the *GeWiss* corpus

2.1 Research rationale

At present, publicly available and searchable corpora on (spoken) academic German do not exist. Previous studies in the field (e.g. Meer 1998; Jasny 2001; or Guckelsberger 2005) are based on rather small-scale data collections which were compiled according to individual research questions and which have not been made available to the larger research community. The *GeWiss* project aims at making a first step towards remedying this situation by creating an electronically accessible multilingual corpus of academic discourse with academic German at its core. It is to be hoped that such a publicly accessible corpus of spoken data will encourage further empirical work from a variety of angles and perspectives, including the fields of contrastive and applied linguistics. The range of spoken academic data will allow for comparative studies in the following areas:

- contrastive analyses of German and other L1 data
- contrastive analyses of students' and experts' academic language in their respective L1s and in German as a Foreign Language
- contrastive analyses regarding differences in the discourse practices and conventions of academic German in different academic settings, i.e. in the context of the German academic tradition in comparison to Polish and British academic tradition.

In order to obtain a suitable data base, the *GeWiss* corpus is carefully designed. In a nutshell, it can be described as a balanced composition of two prototypical academic discourse genres, a monologic (i.e. academic talk) and a dialogic one (i.e. oral examinations), recorded in comparable disciplines (German, Polish, and English philologies) in a German, a British and a Polish academic context. It comprises L1 and L2 data of German as well as L1 data of English and Polish. As can be seen, English and Polish and their respective academic settings are chosen as points of comparison to German in the first instance. However, the corpus is designed in such a way that it is easily expandable in various ways – more languages and academic traditions can be added, more genres can be included, and further learner data could be added (e.g. data from secondary school settings).[2]

2.2 Parameters of the corpus design

There are two key parameters determining the structure of the *GeWiss* corpus: the type of discourse genre chosen, and the language constellation of the speakers recorded.

The two discourse genres that are included in the *GeWiss* corpus were selected because they were considered to be of great relevance in many academic settings (in various disciplines, academic traditions, and linguistic communities), though, of course, they are far from being universal: Conference papers / student presentations were selected as a key monologic genre, as they are of prime importance for scholarly success and successful academic graduation, respectively (cf. Ventola 2002: 14; Guckelsberger 2005: 11). We decided to include oral presentations / conference papers held by expert scholars as well as presentations held by students in seminars to allow for the comparison of different levels of academic

[2] There is already a further partner who is using the *GeWiss*-design to create a comparable parallel corpus: Sofia University 'St. Kliment Ochridski', Bulgaria.

literacy. The recordings also include the respective follow-up discussions as we regard them as integral parts of the communicative events as a whole. Oral examinations were chosen as a dialogic genre of prime importance for student success because in many disciplines and countries they are a prerequisite for graduation.

With regard to the parameter 'language constellation of the speakers', we took recordings of L1 productions of German, English and Polish native speakers on the one hand, and L2 productions in German on the other hand.[3] The L2 productions are recorded in three different academic settings (German, British, and Polish), possibly also reflecting influences of different discourse traditions in the sense proposed by Koch and Österreicher (cf. Koch & Österreicher 2008: 208ff.).

2.3 Comparability of the recordings

As has already been stated above, *GeWiss* is a multilingual corpus which strives to collect data that are broadly comparable in terms of their disciplinary nature (modern philology in a wider sense, comprising literary, linguistic, cultural and pedagogical themes) as well as broadly comparable in terms of genre and function. Genres such as student presentations or oral examinations may, of course, differ in their structure, length, degree of preparedness, role of participants etc.; here, we adopted an ethnographic approach in the selection of appropriate communicative events: our project partners in Poland and UK were consulted as experts familiar with the conventions of the academic genres in their countries. What they regarded as a reasonably 'typical' realisation of a given genre was included in the corpus. This will allow for a comparison, at least to a certain degree, of different discourse traditions and genre-related conventions. As mentioned above, we also strive to reflect the breadth of the respective philological disciplines recorded without prescribing too narrowly the contents of each individual recording. In general, the recordings were made in areas that typically make up philology departments in many countries: The recordings made in the German academic context, for example, contain comparable quantities of data on topics ranging from linguistics, language teaching pedagogy / language acquisition and

[3] We are, of course, aware of the fact that determining the 'native language' of a speaker can be rather difficult and that there is a wide variety of individual language constellations and biographies out there. Since our main research interest lies in the study of academic discourse and academic discourse conventions, we broadly define the L1 of our participants as the language in which they (predominantly) received their school education. This still means that in certain cases, more than one L1 can be identified. Since aspects of the participants' language biography are documented as part of the metadata questionnaire, such more complex language constellations are made transparent in the metadata data base (see section *Metadata*, below).

literary / cultural studies. Data at the other partner institutions are also selected in such a way that they cover these typical sub-divisions within the discipline. The topics of the presentations and examinations recorded within these areas were kept as varied as possible to reflect the diversity of the sub-divisions with regard to range of topics and theoretical approaches.

2.4 Corpus size

The first version of the *GeWiss* corpus will comprise a total of about 120 hours of recording, i.e. 60 hours per genre and 40 hours of data originating from German, English, and Polish academic settings respectively. The Table 1 gives a more detailed breakdown of the recording time for each parameter combination.

Table 1. The *GeWiss* corpus: structure and size in hours of recorded speech events (hrs.)

Language and location of recording	Academic presentation	Oral examination
German in a German academic context (40 hrs.)	10 hrs. academics (L1 German) 5 hrs. students (L1 German) 5 hrs. students (L2 German)	10 hrs. students (L1 German) 10 hrs. students (L2 German)
German in a British academic context (20 hrs.)	5 hrs. academics (L1 English) 5 hrs. students (L1 English) (talks held in German)	10 hrs. students (L1 English) (examination in German)
English in a British academic context (20 hrs.)	5 hrs. academics (L1 English) 5 hrs. students (L1 English)	10 hrs. students (L1 English)
German in a Polish academic context (20 hrs.)	5 hrs. academics (L1 Polish) 5 hrs. students (L1 Polish) (talks held in German)	10 hrs. students (L1 Polish) (examination in German)
Polish in a Polish academic context (20 hrs.)	5 hrs. academics (L1 Polish) 5 hrs. students (L1 Polish)	10 hrs. students (L1 Polish)

2.5 *GeWiss* in comparison to existing corpora of spoken academic language

How does the *GeWiss* corpus relate and compare to other corpora of native and non-native spoken academic language? As mentioned above, there are currently no other corpora of spoken academic German available. The situation is much better with regard to English where there are three corpora of spoken academic English to which the *GeWiss* corpus may be compared: MICASE (Michigan

Corpus of Academic Spoken English),[4] BASE (British Academic Spoken English),[5] and ELFA (English as a Lingua Franca in Academic Settings).[6] The first two contain mainly L1 data while ELFA comprises only L2 recordings.

Apart from the obvious fact that MICASE, BASE and ELFA are centred around English, while the *GeWiss* corpus focuses on academic German in contrast to other languages, there are also further remarkable differences in the corpus design.

The *GeWiss* corpus combines the different perspectives taken by MICASE and ELFA with regard to the usage contexts of a foreign language. MICASE contains English L2 data produced only in a native (American) English context while ELFA contains English L2 data recorded at Finnish universities, i.e. in one specific non-English academic environment. As can be seen from the corpus design summarised in Table 1, the *GeWiss* corpus comprises both types of L2 data, thus allowing for a comparison of academic German used by L2 speakers of German both in a German context and in two non-German academic contexts.

With regard to the range of disciplines and discourse genres covered, the *GeWiss* corpus is much more focussed (or restricted) than both MICASE and ELFA. The latter two both contain recordings of a wide range of spoken academic genres (or 'event types', as they are called in the ELFA-Corpus): MICASE for example contains lectures, colloquia, discussion sections, student presentations, seminars, office hours, consultations, study groups and even campus/museum tours and service encounters.[7] In *GeWiss*, in contrast, the number of genres covered is confined to two (expert conference papers / student presentations and oral examinations). In this regard, *GeWiss* is comparable to BASE which contains only lectures and seminars.[8]

There are similar differences between *GeWiss* and MICASE / ELFA with regard to the range of disciplines covered. While the latter contain recordings from several different subject groups including social sciences, humanities, and natural sciences, the disciplinary range of *GeWiss* is very narrowly limited to philology. This may be seen as a shortcoming in some respects, but it is a gain when it comes to research interests that go beyond the general lexical and morpho-syntactic

[4] For further information on MICASE see http://micase.elicorpora.info/about-micase.

[5] For further information on BASE see http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/.

[6] For further information on ELFA see http://www.helsinki.fi/englanti/elfa/elfacorpus.html.

[7] Cf. http://micase.elicorpora.info/about-micase.

[8] But these were recorded '…in a variety of departments', see http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/.

levels and include pragmatic and textual dimensions. It is also a key feature for any cross-language comparison.

It has already been pointed out that, as a comparative corpus, *GeWiss* differs from MICASE, BASE, and ELFA, in that it contains L1 data from three different languages (recorded in all the selected discourse genres). Taken together, the focus on comparable disciplines, topics, genres, and data size allows for methodologically sound contrastive analyses of various kinds, including an investigation of variations of academic German in different contexts and settings and possible factors determining such variation.

When comparing the *GeWiss* corpus size to that of MICASE (200 hours / 1.8 million tokens), BASE (1.6 million tokens) ELFA (131 hours / 1 million tokens), it is noticeable that *GeWiss* is somewhat smaller (a total of 120 hours / approx. 0.8 million tokens[9]). But if one compares the ratio of data per genre the differences are not generally all that big. MICASE, for example, contains a total of 143.369 tokens originating from student presentations covering different disciplines[10] while *GeWiss* will contain an estimated amount of 30 hours / 210.000 tokens of this genre from a single discipline (but in three different languages), or about 20 hours / 140.000 tokens in German (but including both native and non-native speakers of German).

In sum, although the first version of *GeWiss* is somewhat smaller than other publicly available corpora of spoken academic discourse, its specific design offers a valuable database for comparative investigations of various kinds in the areas mentioned above.

2.6 Accessibility and possibilities for further usage

The potential applications and uses of *GeWiss* go, however, far beyond comparative linguistic investigations. Since the transcription convention is transparent and easily readable for non-specialists (see below), the data may be re-used by scholars from other (sub-) disciplines in the Humanities or Social Sciences for various purposes, including pedagogical aims (e.g. using the data as a basis for

[9] As the transcription of the *GeWiss* corpus is still ongoing, its size in tokens can only be estimated. If we take the experiences of MICASE (1.8 million tokens/200 hours, equalling a ratio of 1 hour: 9000 tokens) and ELFA (1 million tokens/131 hours, equalling a ratio of 1 hour:7600 tokens) as examples, a slightly more conservative assumed ratio of 1 hour:7000 tokens would give us an estimate of 0.8 million tokens for *GeWiss*.

[10] The amount of data for different genres contained in MICASE is given on http://micase.elicorpora.info/researchers/micase-statistics-and-transcription-conventions/statistical-overview-of-speakers-and-spe.

designing teaching materials for courses in L1 and L2 academic discourse or for the familiarisation with different discourse traditions).

The data of the *GeWiss* corpus may also be re-used by scholars with linguistic research interests which require further annotation of the transcriptions. As the recordings of *GeWiss* are transcribed in EXMARaLDA (Schmidt & Wörner 2009; see below), such further annotations may be added simply by inserting further annotation tiers in the EXMARaLDA transcription files.

Once the transcription will be completed and all metadata will have been entered into the database it is planned to make *GeWiss* publicly available through a web interface after free registration. The available data will comprise transcriptions, audio recordings and additional material such as presentation slides, handouts or manuscripts. The *GeWiss* team is also currently considering a possible integration of the corpus and the tools developed in the process of the construction of the corpus into the CLARIN infrastructure.[11]

3. Data acquisition

In order to assemble a balanced corpus of spoken academic discourse in the various philological disciplines (German, Polish, and English), a variety of speakers and communicative events was chosen. Care was taken to include a variety of philological topics and sub-disciplines (e.g. linguistics / applied linguistics, literary and cultural studies, and language pedagogy). Apart from these content criteria, the collection of the speech events was opportunistic. The recruitment of participants, especially for the subcorpus containing oral examinations, turned out to be a substantial challenge.

The procedure for the recordings of student presentations and oral examinations was as follows: first, selected lecturers of specific (sub-) disciplines were contacted and asked for their consent to the recordings. Once the lecturers had agreed to the recordings, the *GeWiss* project was presented to the students in order to recruit student participants. For the recordings of conference papers, the speakers were contacted several weeks in advance and asked for their consent. The contact with potential participants was facilitated by the organising committees of the various conferences. Once the speakers had agreed to the recording they were asked to sign a consent form allowing for the anonymised recorded data and the pseudonymised transcriptions to be used for research and teaching

[11] CLARIN stands for 'Common Language Resources and Technology Infrastructure'. It is a pan-European project which develops infrastructures that will enable the long term accessibility of language corpora, cf. http://www.clarin.eu.

purposes as well as for papers and conference papers. The consent forms adhere to the various data protection regulations in the respective countries in which the data and the consent forms were collected and archived.

All audio recordings were made with a digital audio tape recorder (mostly Olympus LS10) as PCM encoded audio data in a WAV container. Consent forms permitting, the events were additionally recorded with a digital video recorder (mostly Canon HV40) as DV tapes, which are archived separately, and converted into HDV files. That applied to all conference papers and student presentations; for oral examinations, however, it proved impossible to convince participants to give their consent to the video recordings. In order to ensure sustainability of the recorded data, we also plan to convert all recorded files into open formats (e.g. WebM and/or Ogg with Theora and Vorbis as codecs).

The research assistants who conducted the recordings were also present as a participant observer in most of the recorded speech events in order to identify speakers and to collect metadata about the event, the setting and the speakers. However, in oral examinations where an observer's presence could be legally problematic and disturbing, the research assistants left the room after the equipment was set up.

4. Metadata

The *GeWiss* corpus contains detailed metadata describing the event, the recordings themselves, the transcriptions associated with the speech events, as well as the main participants. The metadata are stored and administered using the EXMARaLDA Corpus Manager (COMA) (cf. Schmidt & Wörner 2009 and Wörner in this volume). The *GeWiss* metadata for the speech events and the participants are described in more detail below, with a focus on selected data types. The full list of both types of data can be found in the appendices.

Metadata about the speech event are collected by the researchers conducting the recordings during the event and stored according to the COMA metadata model in the specific section ('container') ***Communications*** (see Appendix 1).

Information relating to the nativeness / non-nativeness of the language use of a specific speaker in a given communicative event is entered in the key ***03 L1 Communication***[12] in order to allow for various comparisons between speakers of different native languages (this is one of the design features of the corpus). There are three possible values of that key – *yes*, i.e. the communication took place in

[12] For the difficulty in distinguishing between L1 and L2, see our remarks in the section *Parameters of the Corpus Design* above.

the native language of the main speaker, *no*, i.e. the communication did not take place in the native language of the main speaker, or *mixed*, e.g. in student presentations held by a group of native and non-native speakers of German. The second relevant design parameter of the GeWiss corpus – the type of speech event – is represented in the data type ***04 Genre***, again with three possible values (*conference paper*, *student presentation,* and *oral examination*). The speech event is further characterised with regard to the main topic covered (***08 Topic***, and ***13 Summary***) and any additional material (such as *power point slides*, *notes* or *handouts*) that may have been used during the event (***10 Additional material***).

An overview of the languages used in a particular communication can be found in the section ***Language(s)***. We distinguish between the ***Main language of interaction*** of a given communicative event and any further languages that may have been used (***Language alternation***). In addition, we characterise the event according to the ***Degree of orality*** (*freely spoken*, *read aloud* or *learnt by heart*) based on the evaluation of the participant observer conducting the recording and the additional materials to the speech event like scripts, power point slides or notes. However, there is a gradual transition in the degree of orality.

Finally, the section ***Setting*** describes some general conditions of the communicative event (partly in their relevance for the recording). In addition to the number of participants, we specify the relationship between the main speaker and the audience, i.e. whether speakers are familiar with their audience or not, whether we are dealing with a public event or a presentation in a classroom environment etc. (***03 Relation between the speaker(s) and the audience***). Furthermore, all additional materials used in the event, irrespective of their availability, are listed in ***02 Media used***.

As for the metadata about the speakers, speaker information is collected with a short form which has to be filled in by all main speakers. The full list of this metadata set can be found in Appendix 2. Apart from some basic socio-demographic information, the form includes questions about the education as well as the languages spoken by the speakers. According to the COMA metadata model, particular stages in the education of the speakers are stored as different ***Locations***. There are, however, three different types of locations in the metadata set of the *GeWiss* corpus: ***Education*** is used as a general heading for both primary and secondary education of the speaker which is assumed to be significant for the socialisation of the speaker in the educational system of a particular language community and thus might be relevant for the specific and general academic language skills of the speaker. (First) university degrees in the country of the speakers' school education are not listed as an extra category in the metadata set of *GeWiss*; any study-abroad period, however, is entered into the metadata set under the heading of ***Study abroad***. The third location type in *GeWiss* is ***Stay abroad*** which lists (longer) periods abroad for non-academic purposes, e.g. for a job, an internship or a language course. A further set of questions concerns speakers' language competences. Since the *GeWiss* corpus aims to provide a comparison of the academic style of speakers of three different language communities, with a particular emphasis on the distinction between native and non-native speakers of German, metadata on both the ***L1*** – defined as the language(s) of the educational socialisation (see above) – as well as ***L2*** – defined as any additional language – were collected. Note that for all cases where German is the L2 there is an additional item ***Evaluation of the language competence.*** This should allow us to compare the specific academic language skills, as represented in the recordings of the speech events, with the general language competence.

After this overview over the key metadata types of the *GeWiss* corpus the following section describes the transcription conventions and the software tool used for the transcription of our primary data.

5. Transcription

All recordings were transcribed using the *EXMARaLDA Partitur-Editor* (cf. Schmidt & Wörner 2009) and the minimal transcription level of the GAT2 transcription conventions (cf. Selting et al. 2009). GAT was developed more than 10 years ago by German linguists with a mainly conversation analytical background for the specific purposes of the analysis of spoken language data. The current version of GAT2 aims at further improving the transcription conventions (and the 'minimal transcription' level in particular) in such a way that digital processing and archiving of electronic transcription corpora is enhanced. GAT is also supported and used by one of the main research institutions in German linguistics today, the *Institut für Deutsche Sprache* (Mannheim). The minimal transcription level is especially suited for corpus projects (such as ours) who aim to make larger quantities of data available in a relative short period of time. The fact that the minimal transcript can later be expanded into the 'basic transcript level' of GAT2 (which, amongst other things, provides more prosodic information) without too much difficulty was another important reason for our choice of this convention.

We now summarise briefly some of the fundamental principles of the transcription conventions used in the *GeWiss* corpus.

According to the GAT2 conventions on the minimal transcription level (cf. Selting et al. 2009: 359–369), words and phrases are transcribed orthographically in the speaker's tier, without any punctuation or capitalisation and using the standard pronunciation as reference norm. This means that spoken realisations which adhere to standard pronunciation (such as terminal devoicing or terminal reduction syllables in German) are transcribed using standard orthography (e.g. [*tha:k*] = *tag*, [*habn*] = *haben*). Strong deviations from standard pronunciation, however, such as dialectisms and regionalisms as well as idiosyncratic forms are transcribed using a 'literary transcription', e.g. *mitnanner* as an idiosyncratic form of *miteinander*. Furthermore, no attempt is made to represent foreign accents as part of the orthographic transcription.[13]

Some of the GAT2 conventions were expanded and developed further to improve the searchability of the *GeWiss* corpus, in particular in cases of clitisation and idiosyncratic forms resulting from phonetic processes within the word boundaries. In cases where one element is incorporated into another one, thus loosing its original form, the underscore character is used to indicate the word boundary, e.g. *hab_s*, *s_gibt*. This should enable the automatic search for this particular phenomenon. However, since clitisation of this kind is quite common in spoken German, we have stipulated as a general guideline for our transcriptions to be rather conservative when it comes to transcribing clitisation – only clearly perceptible instances are transcribed in such a way. The reduction of the German indefinite article (another rather common phenomenon in colloquial German, e.g. *n auto*) is regarded as a different phenomenon and is therefore not treated as clitisation (not least because the degree of lexical and syllabic independence of the reduced element can vary widely). Grammaticalised contractions of preposition plus article, e.g. *im* (< *in dem*) or *aufs* (< auf das) are not treated as clitisation, either. In cases where the word boundaries are completely fused, the underscore character is also omitted so that the two words appear as one complex, e.g. *simma* for *sind wir* or *auffa* for *auf der*. Since the lexical and morphological basis of these realisations can be rather opaque, an expanded (standard) version is usually given in a separate comment tier. This also applies to idiosyncratic forms resulting from assimilation, reduction, deletion or contraction within the word boundaries, e.g. *mitnanner* as an idiosyncratic form of *miteinander* or *wern* for *werden*.

In addition, all clitisations and idiosyncratic items are listed in a separate document together with their expanded (standardised) equivalents to enable automatic searches and to homogenise transcription practises across the *GeWiss* project.

As for the transcription of the Polish and English spoken data, the GAT2 conventions were adapted by the Polish respectively English project group according to the specific spoken language phenomena of each of the two languages.

[13] Distinct accents may, however, be noted in the head of the transcription and additional comments may be added in a commentary tier.

All new transcribers undergo extensive training courses and are then gradually initiated to more complex data, with extensive feedback sessions and corrections of their first independent transcription attempts. This is to ensure transcription standards are maintained across the project and to familiarise all transcribers with the complex transcription software. Any new transcription issues, problem cases and other matters are discussed at transcription meetings conducted regularly by each research group. In order to ensure consistency of the transcribed data within the English, German and Polish subcorpora of *GeWiss*, issues of greater importance are discussed at coordination meetings of the whole project and summarised in the transcription documentation.

6. Annotation

The current version of the *GeWiss* corpus contains mainly orthographic transcriptions of the linguistic actions of the speakers which are entered in a ***verbal*** tier. In addition, for every speaker there is a corresponding ***comment*** tier for describing of non-verbal phenomena of the speakers affecting the communication; it is also used for the standardised versions of abbreviations, dialectisms, regionalisms, idiosyncratic realisations of words and phrases (cf. section **Transcriptions**).

Since the *GeWiss* corpus comprises non-native spoken data of German, too, which may contain instances of code switching and code mixing, we have included an additional annotation layer for language alternation (cf. the data types for ***Languages*** in section **Metadata** for a definition of the term). The language alternation is annotated as *Wechsel* in a separate ***annotation*** tier. In addition, the translation of the passage is given in the comment tier of the particular speaker.

7. Perspectives

As has been outlined above, the design of *GeWiss* can be used to investigate a host of research questions and can also form a useful resource for pedagogical applications. There are, however, a number of ways in which the usability of *GeWiss* could even be enhanced further, of which we would like to discuss a few in this final section.

First of all, *GeWiss* needs to grow. As we know from written corpora, data-driven research projects need a sufficiently large database, and if we take the genre-specific design of *GeWiss* seriously, we need larger sub-corpora for any statistically relevant analyses of frequency and relevance of specific linguistic forms to be made. This means we need more data per genre, as well as per type of speaker and setting. *GeWiss* could and should also grow by including more spoken academic genres, including, e.g., seminar discussions, poster presentations, consultations, to name but a few. Another way in which *GeWiss* could grow is by including more languages and disciplines which would multiply the comparative possibilities.

Second, we need to contribute to the development of tools and annotation conventions designed for spoken language (in collaboration with our partners, including the *Institut für Deutsche Sprache*, Mannheim). The fact that the various existing taggers have all been developed for written data and are not easily adaptable to spoken language is one key problem that needs to be solved. Another challenge lies in the development of annotation conventions for pragmatic and discourse features of spoken (academic) language, enabling corpus searches for categories such as 'stance', 'text-commenting speech action', 'support', 'introducing an argument', and the like. They would enable us to compare languages, disciplines, genres, and speakers with regard to specific pragmatic features, as well as give us a better idea of the frequency and distribution of such pragmatic features in general.

Third, we would like to develop the L2-dimension of the corpus further in a variety of ways. One way of enhancing the existing (and any future) data would be to include error annotation in a similar way as is done in *Falko* for written data.[14] Another way of developing *GeWiss* into a research corpus for language acquisition studies would be to include data originating from non-academic (standardised) oral proficiency tests (taken from the same L2-speakers for which academic language data were recorded), enabling us to compare spoken academic and non-academic language proficiency. Furthermore, spoken L2 data from pre-university academic contexts (secondary schools, intermediate language courses etc.) could be added (using comparable genres) to allow for an exploration of key stages and phases of oral proficiency.

Finally, one pre-eminent aim of *GeWiss* is, of course, to develop and maintain the usability and stability of the corpus on all levels, and to document as accurately as possible all decisions taken and conventions stipulated to enable full transparency of the corpus and its design. As fellow corpus linguists know, this is far from trivial; it becomes even less trivial in a multilateral project such as ours.

[14] See http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko.

References

Guckelsberger, S. 2005. *Mündliche Referate in universitären Lehrveranstaltungen Diskursanalytische Untersuchungen im Hinblick auf eine wissenschaftsbezogene Qualifizierung von Studierenden*. München: Iudicum.

Koch, P. & Österreicher, W. 2008. Mündlichkeit und Schriftlichkeit von Texten. In *Textlinguistik*, N. Janich (ed.), 199–215. Tübingen: Narr.

Jasny, S. 2001. *Trennbare Verben in der gesprochenen Wissenschaftssprache und die Konsequenzen für ihre Behandlung im Unterricht für Deutsch als fremde Wissenschaftssprache* [Materialien Deutsch als Fremdsprache 64]. Regensburg: FaDaF.

Meer, D. 1998. *Der Prüfer ist nicht der König: Mündliche Abschlussprüfungen in der Hochschule*. Tübingen: Niemeyer.

Schmidt, T. & Wörner, K. 2009. EXMARaLDA – Creating, analyzing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19: 565–582.

Selting, M. et al. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10: 353–402.

Ventola, E. (ed.). 2002. *The language of conferencing*. Frankfurt: Peter Lang.

Wörner, K. In preparation. Finding the balance between strict defaults and total openness. Collecting and managing metadata for spoken language corpora with the EXMARAaLDA Corpus Manager. In *Multilingual corpora and multilingual corpus analysis* [Hamburg Studies on Multilingualism], T. Schmidt & K. Wörner (eds). Amsterdam: John Benjamins.

Corpora

BASE: <http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/>

ELFA: <http://www.helsinki.fi/englanti/elfa/elfacorpus.html>

Simpson, R. C., Briggs, S. L., Ovens, J. & Swales, J. M. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor MI: The Regents of the University of Michigan. <http://micase.elicorpora.info/>

Appendix 1. Metadata set describing the speech events in the *GeWiss* corpus

Category	Example	Comments
Description (Communication)		
01 Project Name	GeWiss	
02 Location	Germany	Location of the subcorpus the recording is part of
03 L1 Communication	mixed	Yes = the communication took place in the native language of the main speaker; no = the communication did not took place in the native language of the main speaker; mixed = for some of the main speakers the communication took place in their native language, for another it did not
04 Genre	student presentation	conference paper / student presentation / oral examination
05 Number of the recording	EV_DE_005	[abbr. of the genre]_[abbr. of the location]_[serial number]
06 Short title	Grammar, EuroComGerm	Keyword describing the topic of the event
07 Type of communicative event	student presentation in a M.A. program	More detailed information on the type of communicative event
08 Topic	The project Euro Com(prehension) using the example of EuroComGerm	Full title of the communicative event
09 Data protection	data protection exists	Information on possible restrictions made by the speaker in the written consent
10 Additional material	handout, presentation	An overview of the available additional materials belonging to the speech event
11 List	A	A = best quality; B = middle quality; C = bad quality
12 Classification into the ABC list	all criteria for list A fulfilled	Reasons for the classification of the recordings in the quality list
13 Summary		Short summary of the event
Location		
City	Leipzig	
Country	Germany	
PeriodStart	29.01.2010 15:05:00	
PeriodDuration	00:20:00	
Description (Location)		
01 Institution	University of Leipzig	Institution where the recording took place

Appendix 1. (*continued*)

Category	Example	Comments
02 Room	lecture room	Description of the room where the recording took place
03 Event	master course "Grammaticographie"	Description of the umbrella event where the recording took place
Languages		
Main language of interaction		The main language in which the interaction took place
LanguageCode	DEU	Language code according to ISO 639-3
Description (Language)		
Degree of orality	freely spoken	Description as freely spoken, read aloud or learnt by heart
Language alternation		Languages other than the main language of interaction involved in the interaction
LanguageCode	NONE	If there was no language alternation the value "NONE" is displayed instead of the language code
Setting		
Description (Setting)		
01 Number of participants	3 students giving a presentation, 1 lecturer, audience of approx. 22 students	
02 Media used	handout, presentation	An overview of all additional materials used in the event, irrespective if available or not.
03 Relation between the speaker(s) and the audience	the students giving the presentation are fellow students of the audience and thus familiar to them	Description of the communicative relevant relation of the main speaker to the audience
04 Identification	JP_0208 begins with "…", KF_0205 begins with "…"	The first words of each main speaker
05 Project staff involved in recording	Daisy Lange	Name of the project assistant(s) responsible for the recording.
06 Degree of involvement of project staff	Karsten Feiler involved; social role: lecturer	Pseudonym of the member(s) of the project staff involved in the speech event: non = no project staff involved in the speech event, present = observer, involved + social role

Appendix 2. Metadata set describing the speakers in the *GeWiss* corpus

Category	Example	Comments
Description (Speaker)		
Age	21	
Sex	f	
Name	Wen Zhao	Pseudonym of the speaker
Roles	student giving a presentation	Specification of all communicative roles the speaker occupies in the whole corpus
Locations		For specifying of the education, studies abroad and further stays abroad.
Education		General heading for both primary and secondary education. Academic studies and Ph.D. in the country of the education do not build extra category.
City	n/a	
Country	China	
PeriodDuration	11 years	
Study abroad		E.g. study or Ph.D. in a country other then the country of the education, studies abroad of one or more semesters etc.
Country	Germany	
PeriodDuration	1 year	
Stay abroad		Stays abroad for non academic purposes, e.g. internships, language courses, job etc.
Country	n/a	
Languages		For specifying the language competence of the speaker.
L1		L1 defined as the language of the country of the education
LanguageCode	CHI	Language code according to ISO 639-3
L2		For languages other then the L1
LanguageCode	DEU	Language code according to ISO 639-3
Description (Language)		
Evaluation of the language competence	TestDaF, TDN 4	Information about the language proficiency in the L2