Compiling a Multilingual Spoken Corpus

Introduction The present paper describes the compilation of the spoken part of an English-German corpus, which has been created for the investigation of cohesion. The corpus is one of the few existing resources supporting contrastive studies of cohesion and, to our knowledge, the only one permitting a contrastive analysis of spoken registers in the two languages. In addition, our corpus data offer further research potentials for contrastive linguistics and translation studies as well as for numerous NLP research areas.

Background and Motivation Comprehensive accounts of cohesion are only existent from a largely systemic and monolingual perspective, see e.g. (Halliday and Hasan 1976), (Brown and Yule 1983), (Schubert 2008) and (Esser 2009) for English, and (De Beaugrande and Dressler 1981), (Vater 2005), (Brinker 2005) for German. Empirical analyses (both monolingual and contrastive) in the area of cohesion mainly deal with indiviual cohesive devices. To our knowledge, empirical analyses of spoken discourse only exist for German, e.g. in (Ahrenholz 2007), or for German and French, e.g. in (Schreiber 1992), from a contrastive perspective.

Particular cohesive devices [1] are expected to occur either in registers of spoken language only or with a much higher frequency than in written discourse, see e.g. (Schreiber 1992), (Ahrenholz 2007). Indeed, preliminary corpus linguistic analyses in registers of written language [2] have shown that occurrences of the German demonstrative pronouns *der, die, das* as well as instantiations of cohesive ellipsis and substitution are mainly traced in registers that approximate spoken language, such as fiction or written speech [3].

Therefore, we design a corpus to establish a comprehensive model of cohesion in English and German that integrates differences between written and spoken registers.

The Data Collection Our multilingual spoken corpus contains two registers: interview and academic speech. These registers are added to the eight registers of written language (popular-scientific texts, tourism leaflets, prepared speeches, political essays, fictional texts, corporal communication, instruction manuals and websites) of the already existing corpus, cf. (Amoia *et al. submitted*).

To create the German-English spoken corpus, we use parts of already existing speech corpora and collect our own data, cf. table 1.

The GECCo multilingual corpus		
	German subcorpora	**English subcorpora**
comparable spoken	**original**	**original**
		ELISA
	BACKBONE-DE	BACKBONE-EN
	GECCo spoken collection	MICASE

Table 1. The structure of the GECCo corpus

For English, we take the MICASE corpus data [4], the English part of the BACKBONE corpus [5] and the

[1] We take the classification by (Halliday and Hasan 1976) as a starting point, according to which cohesion includes five categories: reference, substitution, ellipsis, conjunctive relations, lexical cohesion.

[2] cf. (Kunz *et al.* 2009), (Klein 2007) and (Birster 2007)

[3] The extractions were done on the CroCo corpus, cf. (Neumann 2005)

[4] The Michigan Corpus of Academic Spoken English (MICASE) is a collection of nearly 1.8 million words of transcribed speech (almost 200 hours of recordings) from the University of Michigan and includes lectures, classroom discussions, lab sections, seminars, and advising sessions, cf. (Simpson *et al.* 2002)).

[5] The BACKBONE pedagogic corpus contains corpora of video-recorded spoken interviews with native speakers of various European languages, cf. (Kohn 2011).

ELISA corpus [6]. The data from the corpora were extracted according to criteria such as nationality of speaker, type of speech event, degree of speaker interaction.

For German, we use the German part of the BACKBONE corpus, which contains interviews with German native speakers (including variants of German). This subset is comparable to the interviews in ELISA and the English part of the BACKBONE corpus. In addition, we compile our own corpus of spoken academic discourse consisting of transcribed recordings [7] of lectures from all departments of the Saarland University.

Problems in Spoken Corpus Compiling In the process of data collection for the German part of spoken academic discourse, we have encountered a number of practical problems. For instance, we initially planned to include recordings of seminars for analysing dialogues. However, the seminars turned out to be less interactive and dialogic than assumed. Moreover, the collected student presentations constitute prepared speech and thus lack the authentic character of spontaneous speech. Therefore, our German academic corpus currently consists of lecture recodings only.

The recorded data contain too much noise to permit an automatic transcription (speech recognition). Hence, we decide for manual transcription, which requires the formulation of transparent transcription guidelines. Since the English data was transcribed according to differing guidelines we elaborate a consistent scheme for both languages to annotate breaks, linguistic variants and extralinguistic information, as marked bold in example (1).

(1) Transcription example from MICASE:

S1: **ugh**, i hate this board.

S2: well it's **cuz** somebody, pulled it all the way down and it says, not to.

S1: yeah.

S2: turkeys.

<SU-f: LAUGH>

In order to guarantee comparability in frequency and function of cohesive devices between the written and spoken registers we had to restrict each register to 10-14 texts with around 34 thousand tokens each.

The existing registers of written language contain both comparable and parallel texts of English and German. However, for the spoken registers, only comparable texts are available, cf. table 1. One possible solution for obtaining aligned texts would be to create interpretations for the existing originals. Interpreted texts however are produced under very specific conditions and are affected by various contraints such as time pressure, limited short-term memory capacity, linearity, etc. (see e.g. (Gumul 2010), (Pöchhacker 2001)). They are not considered as reflecting spontaneous speech and differ considerably from translations. We will thus integrate transcriptions of films and their synchronysations in our corpus, although these are subject to other well-known limitations (see e.g. (Herbst 1994), (Döhring 2006)).

Annotation Layers The spoken registers of the multilingual corpus will be annotated with the same annotations as the written part i.e. lemma, morphology, pos, lexical chains on the word level, sentences, grammatical functions, predicate-argument structures on the chunk level, registers and further metadata information (language variation, speaker age, etc) on the text level. For further research on cohesion we elaborate other layers of annotation such as coreference, lexical chaining and cohesion disambiguation based on the analyses in (Kunz and Steiner *in progress*)'s and (Kunz 2010). The written part of the corpus additionally contains clause-based alignment of originals and translations.

[6] The ELISA corpus contains interviews with native speakers of English talking about their professional career (e.g. in tourism, politics, the media or environmental education), cf. (Braun 2006).

[7] collected by VISU=Virtuelle Saar Universität (Virtual University of Saarland) for Microsoft.

Conclusion We build a spoken corpus for English and German that is enhanced with annotations on several linguistic levels. Our corpus architecture not only allows a text-based contrastive analysis of cohesion in German and English but also permits a comparison of various spoken and written registers. Therefore, our findings will not only complement the existing research gaps in cohesion but also enrich contrastive grammars with a systematic account of discourse phenomena in written vs. spoken mode. Moreover, both the developed resources as well as our findings on cohesion will provide valuable insights for language teaching and translator training and will open up new research options for various fields of linguistic disciplines.

References

Ahrenholz, B. (2007). Verweise mit Demonstrativa im gesprochenen Deutsch. Grammatik, Zweitspracherwerb und Deutsch als Fremdsprache. Berlin u. New York: de Gruyter.

Amoia, M., K. Kunz and E. Lapshinova-Koltunski (*submitted for GSCL-2011*). GECCo: Multlingual Corpus for Multidisciplinary Analysis.

De Beaugrande, R.A. and W.U. Dressler. (1981). Einführung in die Textlinguistik. Tübingen: Niemeyer.

Birster, L. (2007). Kohäsionsmittel im Englischen und Deutschen – ein Vergleich anhand ausgewählter Phänomene. Diploma thesis. Universität des Saarlandes, Fachrichtung 4.6. Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen.

Brinker, K. (2005). Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden. Berlin: Schmidt.

Braun, S. (2006). ELISA – a pedagogically enriched corpus for language learning purposes. In: S. Braun, K. Kohn & J. Mukherjee (eds.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods. Frankfurt am Main: Peter Lang, 25-47.

Brown, G. and G.Yule (1983). Discourse Analysis. Cambridge: Cambridge University Press.

Döhring, S. 2006. Kulturspezifika im Film: Probleme ihrer Translation. Berlin: Frank & Timme.

Esser, J. 2009. Introduction to English Text-linguistics. Frankfurt a.M. u.a.: Peter Lang.

Gumul, E. and A. Lyda. Disambiguating Grammatical Metaphor in Simultaneous Interpreting. In: J. Maliszewski (ed.) Discourse and terminology in Specialist Translation and Interpreting. Frankfurt am Main: Peter Lang, 87-100.

Halliday, M.A.K. and R. Hasan (1976). Cohesion in English. London, New York: Longman.

Herbst, T. (1994). Linguistische Aspekte der Synchroniasation von Fernsehserien. Phonetik, Textlinguistik, Übersetzungstheorie. Tübingen: Niemeyer

Klein, Y. (2007). Übersetzungsspezifische Eigenschaften – eine korpusbasierte Studie am Beispiel der Kohäsion. Diploma thesis. Universität des Saarlandes, Fachrichtung 4.6. Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen.

Kohn, K. Final report of the LLP BACKBONE project. Public part. University of Tübingen, Germany.

Kerstin Kunz and Erich Steiner. Towards a comparison of cohesion in English and German – contrasts and contact. Submitted for *Functional Linguistics*. London: Equinox Publishing Ltd.

Kerstin Kunz, Karin Maksymski and E. Steiner (2009). Suggestions for a corpuslinguistic analysis of cohesion. Deliverable No. 3 of the GECo Project (`http://fr46.uni-saarland.de/uploads/media/GECo_AP3.pdf`).

Kunz, K.(2010). Variation in English and German Nominal Coreference. A Study of Political Essays. Frankfurt am Main: Peter Lang

Pöchhacker, F. (2001). Dolmetschen. Konzeptuelle Grundlagen und deskriptive Untersuchungen. Tübingen: Stauffenburg.

Neumann, S. (2005). Corpus Design. Deliverable No. 1 of the CroCo Project (`http://fr46.uni-saarland.de/croco/corpus_design.pdf`).

Schreiber, Michael (1992). *Textgrammatik – Gesprochene Sprache – Sprachvergleich. Proformen im gesprochenen Französischen und Deutschen.* Frankfurt a.M.: Lang.

Schubert, C. 2008. Englische Textlinguistik. Eine Einführung. Berlin: Schmidt.

Simpson, R. C., S. L. Briggs, J. Ovens, and J. M. Swales (2002). The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan.

Vater, H. (2005). Referenz-Linguistik. München: Fink.