Deliverable no.1 (Arbeitspaket AP1 / Work Package WP1)

Brief review of existing resources for representing knowledge about languages (English and German)

Kerstin Kunz, Karin Maksymski, Erich Steiner, 8 June 2009

The following overview will discuss existing resources for the description and comparison of the language pair English-German. The resources outlined in Figure 1 are aligned along the dimensions of

- modelling (system-based vs. text-based)
- language-orientation (monolingual vs. comparative)
- level (lexis vs. grammar vs. text / discourse).

This alignment yields 12 types of (potential) resources **for representing knowledge about languages**; of these, it turns out, our project (*GECo*) understands cells C2 and C4 as its major areas of work.

Methodology / Level		1 **system-based model monolingual**	2 **system-based model comparative**	3 **text-based model monolingual**	4 **text-based model comparative**
A	**Lexis**	✓ dictionaries, lexicons	✓ bi-lingual dictionaries	✓ text-based dictionaries (monolingual), e.g. COBUILD	✓ text-based terminologies (multilingual)
B	**Grammar (clause / sentence)**	✓ grammars	✓ contrastive grammars	✓ text-based grammars (monolingual), e.g. Longman Grammar	Ø
C	**Grammar (text / discourse)**	✓ text- / discourse grammars	Ø	Ø	Ø

Figure 1: Existing resources for representing knowledge about languages

Below, we briefly discuss the different approaches and identify the gap we see as a field of research for our current project.

1. Monolingual, system-based works

In this area substantial coverage has already been achieved on all three levels, and we do not see any major gap here.

A1 Lexis

On the level of lexis (cell A1 of Figure 1), works that employ a system-based model are monolingual dictionaries and lexica (thesauri, ontologies). They concentrate on one single language and mainly draw on earlier lexical resources as their knowledge source for compilation. Additionally, they use texts *as sources for samples*. The information they give is on lemmata and their use; these lemmata are sorted either alphabetically, or else by semantic relations such as synonymy, hyponymy, semantic oppositions, or other semantic relations. Examples for extensive monolingual dictionaries in English are the *Oxford English Dictionary* or the *Webster's Dictionary*, in German *Duden Wörterbuch der deutschen Sprache* or *Wahrig Deutsches Wörterbuch*.

B1 Clause / Sentence grammar

Concerning grammar with an emphasis on clauses and sentences we find different monolingual grammars for both languages, e.g. Quirk et al. (1985) or Huddleston and Pullum (2002) for English. The *Duden Grammatik der deutschen Gegenwartssprache* by Eisenberg et al. (1998) or Eisenberg (^{3}1994), Heidolf (1981), Engel (2004) and Helbig / Buscha (2001) are examples for German clause / sentence grammars (Zifonun et al. (1997) will be mentioned in B3). They constitute repositories of the grammatical system of a language that are usually structured by rank, class, function, or other grammatical criteria. They lay down rules on the use of a language and are thus often used, together with bi-lingual dictionaries, for didactic purposes. For their compilation they draw on earlier grammars and linguistic theories as a knowledge source and use texts as sources for samples as well.

Other monolingual clause grammars for English are, for example, Halliday and Matthiessen (2004). Since there is a consideration of text and discourse in this work, too (e.g., in a chapter on cohesion), this grammar somehow belongs both to B1 as well as C1. Treatments of German in a similar model include Steiner and Ramm (1995), Steiner and Teich (2004) and Steiner (2006).

C1 Text / Discourse grammar

Other grammatical descriptions are located on a third level, that of grammar with a focus on text and discourse. These works do not constitute a comprehensive grammar of a language, but deal with specific questions of text and discourse. Here Halliday and Hasan (1976), Brown and Yule (1983) and in fact de Beaugrande and Dressler (1981) have to be mentioned (see also Schubert 2008). Halliday and Hasan focus on cohesion, i.e. linguistic elements establishing ties to other linguistic elements above the clause and sentence level. De Beaugrande and Dressler include more phenomena than Halliday and Hasan by discussing seven standards of textuality (cohesion, coherence, intentionality, acceptability, informativity, situationality, intertextuality). Brown and Yule offer a discourse analysis of English. All three are rough parallels, though more as instructions for analysis than as overviews of “a grammar”.

The same may be true for German descriptions, such as Linke / Nussbaumer / Portmann (42001), Brinker (2005) or Vater (^{3}2001). Weinrich (1993) offers a text grammar for German (and French) that takes several approaches (such as, for example, textual, dialogical, instructional or cultural ones) as a basis. He sees grammar as a rule set, which gives indications as to the meaning of a linguistic sign ("Sprachzeichen", 1993: 21). While the literature in English mainly focuses on linguistic phenomena available to establish textuality, German literature takes as its starting point general pragmatic, cognitive and semantic principles of coherence, which are reflected in linguistic phenomena. These differences in methodological orientation lead to noticeable differences in the range of phenomena considered, as well as in and the granularity of the descriptions.

In general, monolingual text- / discourse grammars inform about the coherence-building systems of a language and are structured by type and / or function of the system (e.g. (co-) reference, conjunctive relation, lexical / semantic relations, etc.). Their knowledge sources are earlier grammars, linguistic theories and sometimes writings on "stylistics" as well. Again, they use texts as sources for examples.

2. Comparative, system-based works

A second group of methodologies start from system-based models as well, but take on a comparative approach instead of concentrating on one language.

A2 Lexis

On the level of lexis, bi-lingual dictionaries such as the *PONS dictionary*, the *Langenscheidt Collins dictionary* or the *Oxford Duden dictionary* have to be mentioned for the language pair English-German. In very rare cases lexica belong to this category, too, e.g. thesauri and ontologies such as EuroWordNet, a bilingual lexicon for semantic relations, or language-specific instantiations of *word net* (e.g. GermaNet for German). Multilingual terminologies in terminology-engineering and translation studies are another comparative, system-based approach on this level.

As above, these works rely on earlier lexical resources and take texts as sources for samples. Their information concerns lemmata and is sorted either alphabetically, or else by different semantic relations. They can also be sorted by some relations as "cross-lingual correspondence" (loosely speaking "translation").

B2 Clause / Sentence grammar

In the field of grammar there are some contrastive grammars focussing on clause / sentence aspects for a comparison of English and German. Already in the 17th century the *grammar of Port-Royal* tries to establish some common features between different languages (although concentrating on French). Later on, in the 19th century, comparative grammars analyse characteristics of indo-Germanic languages. Nowadays, there are overviews for English and German such as Hawkins (1986), König and Gast (2007) and Königs (2001). However, the latter in particular mainly addresses "translational difficulties" and does not offer a comprehensive contrastive grammar of the two languages.

The structure of these grammars and the information they offer is mainly the same as in the first methodological group (cf. section 1); the same is true concerning the knowledge source for their compilation, i.e., they use earlier grammars and linguistic theories for the modelling, and texts as sources for samples.

C2 Text / Discourse grammar

Concerning grammar with a focus on text and discourse we can establish a first gap in language research. For the language pair English-German, we are currently not aware of any resources dealing with questions of text and discourse (e.g. cohesion) from a comparative, system-based and comprehensive perspective. Nevertheless, we would like to point to episodic comparisons between English and German in de Beaugrande and Dressler (1981). Numerous passages in works by Doherty (e.g. Doherty 2006) deal with relevant questions. These latter are example-based studies analysing contrasts in information structure and information distribution by comparing English originals and German translations. Fabricius-Hansen takes a similar approach for the analysis of German originals and English translation in her works on information distribution (cf. 1996, 1999, 2005). For the language pair English-French, the *Stylistique Comparée* by Vinay and Darbelnet (1958) approaches an early comparative attempt. For English and German, this is a gap we would like to work towards filling. On the one hand, we would start from a system-based approach, since we would take Halliday and Hasan (1976) (cf. C1) as a starting point and model for the analysis of cohesion in German – in comparison to English. On the other hand, we would like to combine this with a text-based approach (more on this in C4).

As to the information and structure of such an approach, it should constitute a repository of the coherence-building systems of a language, thus moving beyond the clause as the basic unit of grammaticalization. It can be structured by type and / or function of the system. As a knowledge source the above mentioned earlier grammars can be employed as well as linguistic theories. In some of Doherty's works, she also takes writings on a comparative stylistics into consideration. Again, texts are used as sources for samples; in our project, we can also make use of aligned data from the *CroCo* project (http://fr46.uni-saarland.de/croco/index_en.html) as examples, without being restricted to this usage.

3. Monolingual, text-based works

We now turn to the methodologies that work with a text-based model and have a look at the monolingual approaches first.

A3 Lexis

Starting from the level of lexis, resources are monolingual dictionaries, lexica (only in very few cases), e.g. thesauri and ontologies such as language-specific instantiations of *WordNet* (for example, *GermaNet* for German), and multilingual terminologies in terminology-engineering and translation studies. The difference to works from the first group is that they were compiled using texts as sources for examples as well as for frequencies of collocations. Earlier lexical resources are taken as a knowledge source, too.

For English, the *Collins COBUILD English Language Dictionary* is an important example; it is based on the COBUILD (Collins Birmingham University International Language Database) corpus. A relevant example for German may be the DWDS (*Digitales Wörterbuch der Deutschen Sprache*) that builds on text corpora as well. Here, the lemmata are sorted by semantic relations; structuring can also be done alphabetically or else by differences in collocations. Additionally, KWIC (Key Word In Context) indices can be presented.

B3 Clause / Sentence grammar

On the level of grammar with a clause / sentence perspective, one (and probably the only one, as far as we know) monolingual grammar that is text-based rather than system-based is the *Longman Grammar of Written and Spoken English* (1999) by Biber et al. It was compiled using texts as sources for samples, but also as sources for variety-specific frequencies. Earlier grammars are part of the basis of such grammars, too, as are linguistic theories, although to a lesser extent. For German, the *Grammatik der deutschen Sprache* by Zifonun et al. (1997) could be listed within this category. In contrast to Biber et al. (1999), this German grammar does not give information on frequency, but nevertheless is text-based in the sense that examples from corpora are used to illustrate grammatical questions.

A text-based monolingual grammar offers information on the grammatical system(s) of a language, structured by rank, class, function, or other grammatical criteria. Additionally, based on the data of a text corpus, it can include information about frequencies in different varieties and / or an extended coverage of weakly-grammaticalized units. The *Longman Grammar*, for example, does this for non-clausal material (cf. Biber et al. 1999: 224ff). Interesting observations about individual aspects of English (and to a limited extent other languages) can be found in e.g. Nevalainen et al. (eds.) (2008); Chambers et al. (eds.) (2004); Nair (2006a,b).

C3 Text / Discourse grammar

A second gap becomes clear when looking at grammar from a text / discourse perspective. Currently there is no resource known to us that, on the one hand, offers a (structured) repository of the coherence-building systems of a language beyond the clause as the basic unit of grammaticalization (as mentioned already in 2.3), and, on the other hand, is enriched by extensive frequency information and by tying that information to linguistic variation, varieties or registers of the language concerned. Knowledge sources that could be used for compilation in this field would be earlier text / discourse grammars (cf. section 1.3), linguistic theories, and characterizations of registers as in Neumann (2008) and similar work. Possibly there is also some work in Biber et al. (1999) that can be used to this end, as well as in Ghadessy (1988, 1993, 1999), but only where that work starts to go beyond the purely lexicogrammatical realm, and where it becomes amenable to the study of frequencies.

With our current project we cannot frontally address this gap and will rather concentrate on work under the comparative aspect. Nevertheless, we would like to make at least some preliminary contributions in this field of research.

4. Comparative, text-based works

The last methodological group we would like to address is that of text-based, comparative models. Here, we can spot even two gaps, with one being of special interest for our current project.

A4 Lexis

To the field of lexis, multilingual terminologies in terminology-engineering and translation studies belong, but only in those cases, where they are text-based. This is true only for very few of them at this stage. Texts should be used as sources for examples and frequencies of collocations in both or all of the languages concerned; at the same time, earlier lexical resources can be included as well. The alphabetical sorting or sorting by semantic relations

should, above all, be supplemented by some sort of “translation / correspondence relation”, or else by differences in collocations. Maybe a presentation of “KWIC” indices would be a desirable additional feature, too.

B4 Clause / Sentence grammar

Concerning grammar the next gaps can be made out, on the level of clause and sentence as well as on the level of text and discourse (cf. section 4.3). A resource in that first area would be a system-based comparative grammar, but complemented by information about frequencies, especially for the differentiation of different varieties across languages. In a certain sense, this would be a fusing of Hawkins (1986) or König and Gast (2007) with a methodology as in Biber et al. (1999). This means that for the presentation of the grammatical systems of two or more languages texts would be used as sources for samples and for variety-specific frequencies, besides earlier grammars and (to a lesser extent) linguistic theories. As in Biber et al. (1999), an extended coverage of weakly-grammaticalized units would also be possible. The sources would have to be available in a multi-layer representation, like can be found, for example, in the *CroCo* corpus (cf. Vela et al. 2007) or tree-banks in general. A very important point would be the inclusion of aligned data from source-target pairs.

To a limited extent, Teich (2003) can be considered an approach in this area; here we find the verification of system-based comparative grammars by a corpus-linguistic study. However, the analysis is limited to one register and to specific linguistic phenomena.

We do not see the filling of this gap as one of our main goals in the current project, but would rather focus on the question of cohesion. This leads us to the next level, that of text / discourse grammar.

C4 Text / Discourse grammar

In this area of grammar, we are again not aware of any resources. For attempts at creating such a resource, there would be basically the same knowledge sources as mentioned in 3.3 (earlier grammars, linguistic theories, register characterizations, etc., all including aspects beyond the purely lexicogrammatical realm and information on frequencies). In contrast to 3.3, though, a comparative perspective has to be adopted, i.e. sources for two languages (in our case, English and German) would have to be considered. As a result, there would be a model that combines several aspects mentioned before: Information would be given on the coherence-building systems of the languages concerned, but on a textual basis, not only considering the clause as the basic unit of grammaticalization (cf. 2.3) but also including linguistic phenomena of textuality that go beyond the grammar of the clause. The model would be structured along the dimensions of type and / or function of the system. Extensive frequency information would have to be added as well, tied to linguistic variation, varieties or registers of the languages concerned (cf. 3.3). Finally, a tree-bank type multi-layer representation would be necessary, including aligned data for English and German (cf. 4.2).

This is a gap we are planning to address within the current project. We cannot offer a complete text- / discourse grammar for English and German, but would focus on differences in terms of cohesive devices. This limitation is motivated by the fact that the analysis will still be of a manageable size (compared to an analysis of all criteria of textuality). Furthermore, Halliday and Hasan (1976) offers a good model to take as a starting point and as a basis for the comparison German-English.

By using a text corpus including aligned data for English and German we can benefit from several advantages of a text-based model: first, the perspective becomes a broader one, in the sense that more phenomena can be discovered than would come to mind otherwise. Second, frequencies of the actual use of these phenomena can be identified and interpreted

(e.g., as indicator for registers). A third advantage is the possibility to discover the different functions of the phenomena in different contexts. Finally, a text-based approach can also yield answers to questions of language contact, such as to differences between originals and translations and to the influence of one language on the other (for a relevant discussion on the nature of linguistic data cf. Haspelmath 2009 and other contributions in the same volume).

In our project we would thus try to compile a resource that combines a system-based and a text-based approach under a comparative perspective focussing on questions of text and discourse.

5. References

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

Brinker, K. 2005. *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden.* 6th edition. Berlin: Erich Schmidt.

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

Chambers, J.K., Trudgill, P., and N. Schilling-Estes, eds. 2004. *The Handbook of Language Variation and Change.* Oxford. Blackwell.

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

Doherty, M. 2006. *Structural Propensities. Translating nominal word groups from English into German.* Amsterdam/ Philadelphia: Benjamins.

Eisenberg, P. 1994. *Grundriss der deutschen Grammatik.* 3rd edition. Stuttgart, Weimar: Metzler.

Eisenberg, P., Gelhaus, H., Henne, H., Sitta, H., and H. Wellmann. 1998. *Duden - Grammatik der deutschen Gegenwartssprache.* Mannheim: Duden-Verlag.

Engel, U. 2004. *Deutsche Grammatik.* Neubearbeitung. München: iudicium.

Fabricius-Hansen, C. 1996. Information density: A Problem for Translation and Translation Theory. In: Linguistics 34: 521-565.

---. 1999. Information packaging and translation: Aspects of translational sentence splitting (German – English/ Norwegian). Studia Grammatica, 47:175-214.

---. 2005. Elusive connectives. A case study on the explicitness dimension of discourse coherence. In: Linguistics 43-1: 17-48.

Ghadessy, M., ed. 1988. *Registers of Written English.* London: Pinter.

---., ed. 1993. *Register Analysis. Theory and Practice.* London: Pinter.

---., ed. 1999. *Text and Context in Functional Linguistics.* Amsterdam: Benjamins.

Halliday, M.A.K. and R. Hasan. 1976. *Cohesion in English.* London: Longman.

Halliday, M.A.K. and C.M.I.M. Matthiessen. 1994. *An Introduction to Functional Grammar.* London: Arnold.

Haspelmath. Martin. 2009. “Welche Fragen können wir mit herkömmlichen Daten beantworten?“ in: Forum Zeitschrift für Sprachwissenschaft ZfS. 2009. 28.1..: 157ff.

Hawkins, J. A. 1986. *A Comparative Typology of English and German: Unifying the Contrasts.* London: Croom Helm.

Heidolf, K.E., Flämig, W. and W. Motsch, eds. 1981. *Grundzüge einer deutschen Grammatik.* Berlin: Akademie Verlag.

Helbig, G. and J. Buscha. 2001. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* Berlin, etc.: Langenscheidt.

Huddleston, R. D. and G.K. Pullum. 2002. *The Cambridge Grammar of the English Language.* Cambridge University Press.

König, E. and V. Gast. 2007. *Understanding English–German Contrasts. Grundlagen der Anglistik und Amerikanistik.* Berlin: Erich Schmidt (revised second edition: 2009).

Königs, K. 2000. *Übersetzen Englisch - Deutsch. Systemischer Ansatz.* München: Oldenbourg.

Linke, A., Nussbaumer, M. and P. Portmann. 2001. *Studienbuch Linguistik.* 4th edition. Tübingen: Niemeyer.

Mair, C. 2006a. *Twentieth-Century English. History, variation and standardization.* Cambridge: Cambridge University Press.

---. 2006b. Grammatical Change in 20th. Century English. In: Anglistik. Mitteilungen des Deutschen Anglistenverbandes. 17.2. 2006: 11- 34.

Nevalainen, T., Taavitsainen, I., Pahta, P. and M. Korhonen, eds. 2008. *The Dynamics of Linguistic Variation. Corpus evidence on English past and present.* Amsterdam: John Benjamins.

Neumann, S. 2008. Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German. Habilitationsschrift, Universität des Saarlandes.

Quirk, R., Greenbaum, S., Leech, G., and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Harlow: Longman.

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

Steiner, E. and E. Teich. 2004. Metafunctional profile of the grammar of German. In: Caffarel, A., Martin, J.R. and C.M.I.M. Matthiessen, eds. 2004. Language Typology. A Functional Perspective. Amsterdam: Benjamins.

Steiner, E. and W. Ramm. 1995. On Theme as a grammatical notion for German. In: Functions of Language Vol. 2 (1). 57-93.

Steiner, E. 2006. Construing Contextualization through Meaning: Some Thoughts on a Semantics for Theme. In: S.-Y. Cho and E. Steiner, eds. 2006. Information Distribution in English Grammar and Discourse and other Topics in Linguistics. Festschrift for Peter Erdmann on the Occasion of his 65th. Birthday. Frankfurt/M. etc.: Peter Lang. 267 – 288.

Teich, E. 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts.* Berlin, New York: de Gruyter.

Vater, H. 2001. *Einführung in die Textlinguistik.* 3rd edition. München: Fink.

Vela, M., Hansen-Schirra, S. and S. Neumann. 2007. Querying multi-layer annotation and alignment in translation corpora. In: Proceedings of the Corpus Linguistics Conference CL 2007, University of Birmingham, UK, 27-30 July 2007.

Vinay, J.-P. and J. Darbelnet. 1958. *Stylistique Comparée du Francais et de l'Anglais. Méthode de Traduction.* Paris: Didier.

Weinrich, H. 1993. *Textgrammatik der deutschen Sprache.* Mannheim: Dudenverlag.

Zifonun, G., Hoffmann, L., and B. Strecker. 1997. *Grammatik der deutschen Sprache*. Berlin/New York: de Gruyter. (http://home.edo.uni-dortmund.de/~hoffmann/GDS.html)

Dictionaries:

Collins COBUILD English dictionary. 1995. London: HarperCollins.

Duden – Deutsches Universalwörterbuch. 2006. 6th edition. Mannheim: Dudenverlag.

Duden-Oxford-Großwörterbuch Englisch. 2005. 3rd edition. Mannheim: Dudenverlag.

Langenscheidt Collins Großwörterbuch Englisch. 2004. 5th edition. Berlin: Langenscheidt.

Merriam-Webster's collegiate dictionary. 2003. 11th edition. Springfield, Mass.: Merriam-Webster.

Oxford advanced learner's dictionary of current English. 2006. 7th edition. Oxford: University Press.

Pons - Großwörterbuch Englisch-Deutsch, Deutsch-Englisch. 2005. Stuttgart: Klett.

Wahrig Deutsches Wörterbuch. 2006. 8th edition. Gütersloh: Wissen-Media-Verlag.

Websites:

http://fr46.uni-saarland.de/croco/index_en.html
(Croco project)

http://www.cas.uio.no/research/coming.php
(research group “Meaning and understanding across languages” at the Centre of Advanced Study, Oslo)

http://www.uni-hamburg.de/sfb538/
(SFB multilinguism at Hamburg university)

http://www.uni-stuttgart.de/linguistik/sfb732/
(SFB on “Incremental Specification in Context”)

http://wordnet.princeton.edu/
(WordNet)

http://www.sfs.uni-tuebingen.de/GermaNet/
(GermaNet)

http://www.illc.uva.nl/EuroWordNet/
(EuroWordNet)

http://www.dwds.de/
(digital dictionary of the German language)