Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying

Marilisa Amoia, Kerstin Kunz, Ekaterina Lapshinova-Koltunski
Department of Applied Linguistics, Saarland University
{m.amoia,k.kunz,e.lapshinova}@mx.uni-saarland.de

Abstract

In this paper, we discuss some linguistic phenomena that pose potential problems for multilevel linguistic annotation of parallel corpora in general and specifically for data encoding with state-of-art multilevel corpus querying tools such as CQP. We describe the strategy we use for integrating the standard hierarchical XML representation used to annotate such phenomena in our aligned bilingual corpus GECCo into a timeline-based format as used in CQP. Thus, our framework supports efficient multilevel representation as well as corpus exploitation and querying of linguistic data of arbitrary complexity.

1 Introduction

Gathering and providing a natural language corpus of good quality requires the definition of data models that mirror the complexity of natural language data from written as well as spoken discourse. In recent years, much work has been done to develop standards for annotations, annotation schemes and coding practice guidelines (c.f. (McKelvie et al., 2001), (Blache et al., 2010)) with the aim of allowing data exchange between different annotations tools and portability of corpora to other platforms as well as the integration of corpora. Yet, relative little attention has been devoted to interfacing annotation schemes with the encoding formats required by corpus query engines.

Although several efficient automatic systems for parallel corpus exploitation have been developed, these systems are generally specialized for the storage and retrieval of a very limited number of annotation levels. For instance, UNITEX (Paumier, 2000) only allows alignment on sentence level, and although EMDROS (Petersen, 2004) is a system for storing and retrieving annotated texts that is very generic and applicable to almost any kind of linguistic annotation, it does not allow alignment.

In fact, very few corpus query tools such as CQP (Christ, 1994), ANNIS2 (Zeldes et al., 2009) or MATE (McKelvie et al., 2001) exist that can handle multilevel annotated corpora. To our knowledge, ANNIS2 is still in the development phase, at the moment of writing, and MATE (McKelvie et al., 2001) does not easily support alignment of parallel corpora.

In this paper, we present our experience with the multilevel query engine CQP developed within the CWB Open Corpus Workbench (Christ, 1994), a collection of open-source tools for managing and querying large text corpora (ranging from 10 million to 2 billion words).

Our focus will be on some problematic issues that have been raised by our attempt to automatically encode our multilevel-annotated bilingual parallel corpus into CQP. The GECCo corpus, which was developed in our research group for the contrastive study of cohesion in English and German combines automatic and manual annotation on different layers of linguistic knowledge ranging from pos-tagging, syntax chunking to semantic information such as linguistic chains and coreference. We noticed that certain annotations were difficult to encode employing state-of-art query tools, namely those representing discontinuous segments.

The paper is structured as follows: Section 2 gives an overview of linguistic phenomena that might lead to discontinuous segments. Section 3 describes the XML-based data format on which multi-layer annotations in the corpus are based. Section 4 deals with the strategy we adopted to encode corpus annotations into CQP and in particular describes the strategy for encoding problematic constituents such as discontinuous segments into a timeline-based data format, as the one used

Proceedings of Recent Advances in Natural Language Processing, pages 2–10,
Hissar, Bulgaria, 12-14 September 2011.

by CQP, so as to allow corpus querying and exploitation. Section 5 concludes with pointers for future research.

2 Differences in Information Distribution between English and German

This section is concerned with differences in information distribution between English and German as these complicate annotation and exploitation of parallel corpora. Here, structural shifts between originals and translations have turned out to be particularly problematic in view of semi-automatic annotation and querying of translational equivalents or extraction for further processing. They cause discontinuity in cases where the translational equivalents are aligned on the basis of semantic criteria[1].

2.1 Contrasts between English and German

General differences between English and German such as case marking and word order (see e.g. (Hawkins, 1986), (Koenig and Gast, 2007), (Steiner, 2001) and (Teich, 2003)) are believed to have implications with respect to the positional options for the integration of information into sentences. For instance, (Doherty, 2004) but also (Teich, 2003) and (Steiner and Teich, 2004) note that the order of information is more flexible in English at the beginning of declarative main clauses where more than one constituent may occur before the verb complex. In contrast, German offers more structuring options after the finite verb (in the Mittelfeld) and an additional option behind the non-finite verb (Nachfeld). This is due to the topological peculiarity of the German verbal bracket. Fabricius-Hansen (1999) highlights the tendency of German to structure experiential meaning more vertically and metaphorically in contrast to a more horizontal and congruent distribution of information in English. She indicates "recursive compounding, repeated nominalization, heavy prenuclear and postnuclear noun phrase modification, and accumulation of adverbial adjuncts" (Fabricius-Hansen, 1999) as grammatical features that enhance hierarchical information packaging. In summary, the differences between English and German described above may provoke the following relevant shifts between originals and translation: Meaning that is expressed inside phrases in German may be expressed by a subordinate or main clause or may appear in a separate sentence in English. Meaning that is expressed in a medium sentence position in German may be shifted to the beginning or end of a sentence or be incorporated in a separate sentence in English. As a consequence of these shifts we assume that meaning may be conveyed in English by a contiguous element at one particular position, while corresponding meaning may be realized in German by separate elements in different syntactic positions. We thus expect a higher number of discontinuous segments in German, both at phrase and at clause level. In the following section of this paper some examples of discontinuous segments will be discussed.

[1] We opted for a semantic alignment as we assume that only this kind of annotation provides the information necessary for studying English-German contrasts in information distribution and further phenomena of cohesion.

2.2 Some examples for discontinuous segments

We now go on to examine some stretches of text from the GECCo corpus in which discontinuous segments are encoded in case of semantic alignment. In German, discontinous segments at sentence level may be caused by a tendency to encode relevant information in the form of complex appositions in a middle position of the sentence:

(1) a. Dieser Lösung - und das ist für mich das Wunder - haben zum Schluss alle zugestimmt:

b. The miraculous thing for me was that in the end everyone agreed to this solution:

In the German original (1a), relevant and focused information is inserted as a clausal apposition into another sentence, without being related to one specific constituent. The same meaning is expressed in the English translation (1b) by the subject and predicate of the main clause turning the predicate plus arguments of the German main clause in a subordinate clause. The alignment of translational equivalents therefore requires the annotation of a discontinuous segment. Below is an example of a discontinuous segment in German that is not only annotated for alignment of translational equivalents but also for the annotation of coreference.

(2) a. Sehr erfolgreich ist - und das bestätigen mir vor Ort nicht nur sozialdemokratische Kommunalpolitiker - das Förderprogramm InnoRegio.

b. The InnoRegio funding programme has been very successful – something local politicians, and not just Social Democrats, have confirmed.

In the German original (2a), a clausal apposition is again inserted into another main clause. However, the anaphoric pronoun *das* in the apposition refers to the whole main clause. Thus, the latter has to be annotated as discontinuous segment in order to mark it as the antecedent of the pronoun. In the English translation (2b), the apposition is retained but appears after the main clause, without splitting it into two linear parts. Thus only one continuous element needs to be segmented in the translation for the annotation and alignment of the antecedent. Another cause of discontinuous segments on sentence level are prepositional phrases which, are again distributed more freely in German than in English.

(3) a. Dieser Konsens ist trotz aller möglichen Vorbehalte ein hohes politisches Gut.

b. Despite all possible reservations, this consensus is a key political asset.

A prepositional phrase functioning as an adverbial occurs after the predicate (in the Mittelfeld) in German (3a) but is moved to the beginning of the sentence in the English translation (3b). Consequently, alignment of the main clause according to semantic criteria would result in a discontinuous segment in German but not in English.

(4) a. Dieser Konsens ist, trotz aller möglichen Vorbehalte, ein hohes politisches Gut, das die Stiftung "Erinnerung, Verantwortung und Zukunft" im Kuratorium unter Leitung von Botschafter Kastrup und durch den Vorstand aus Dr. Jansen, Dr. Bräutigam und Botschafter Primor erhalten muß.

b. Despite all possible reservations, this consensus is a key political asset which the Foundation "Remembrance, Responsibility and the Future" must preserve on its Board of Trustees. The latter is chaired by Ambassador Dieter Kastrup. Board of Trustees members Michael Jansen, Hans-Otto Bräutigam, and Ambassador Avi Primor were elected the foundation's executive officers.

In the German excerpt, a prepositional phrase functioning as an adverbial occurs in the Mittelfeld (4a) before the right verbal bracket. The meaning of this PP is realized in a separate sentence in the English translation (4b). A meaning-based alignment of the first English sentence therefore includes a discontinuous element in the German original.

Discontinuous segments on phrase level in German may be due to distinct NP pre-modification conventions (see (Koenig and Gast, 2007), (Doherty, 2004), (Fabricius-Hansen, 1999) and (Teich, 2003)). In contrast to English, merely prepositional phrases and finite relative clauses follow the head noun in German. Constructions of medium complexity are usually placed before the head noun. These contrasts may complicate coreference chaining, on the one hand, and alignment of elements of these chains in the parallel corpora, on the other hand.

(5) a. über zwei Zeilen Lagerhäuser blicken wir auf [das strömungslose Grau des Hafenbeckens und auf die Landzunge, die sich zwischen ihm und dem Fluss erstreckt]_$A1$. Seit Menschengedenken gehört [dieses auf drei Seiten von Wasser umgebene Gelände]_$B1$ der chemischen Industrie.

b. We look across two rows of warehouses at [the motionless grey surface of the harbor basin and the tongue of land that extends between it and the river]_$A2$. Enclosed on three sides by water, [this area]_$B2$ has been a preserve of the chemical industry for as long as anyone can remember.

The German antecedent (A1) and its English translational equivalent (A2) exhibit similar NP structures. At the same time, there are some positional differences between the German anaphor (B1) and the corresponding anaphor in the English translation (B2): While the German noun phrase contains several premodifying elements, the English anaphor only consists of the demonstrative determiner and the head noun. The reason for this is that the non-finite predicate argument construction "auf drei Seiten von Wasser umgebene" inserted between the demonstrative determiner and the nominal head in German could not be realized as premodifier in English. The translator chose to separate it from the rest of the noun phrase

and transformed it into an adverbial clause functioning as a clausal adverbial at sentence level. Hence, semantic alignment of the English subject "this area" results in the annotation of a discontinuous segment in German, consisting of "diese" and "Gelände". Flexible positioning of complex NP postmodifiers in German may also yield discontinuous segments:

(6) a. This occurred just after I took a turning and found myself on a road curving around the edge of a hill.

b. Dies geschah kurz nach einer Abzweigung, als ich mich plötzlich auf einer Straße befand, die in Kurven an einem Hang entlangführte.

The relative clause occurs after the nominal head in both the English original and its German translation. However, the heavy NP shift enables the German relative clause to be postponed after the predicate. The alignment of the corresponding relative clauses entails annotating a discontinuous element in German.

(7) a. Aber wenn die Notwendigkeit von Reformen besser verstanden wird, als die Bereitschaft verbreitet ist, diese zu unterstützen (...)

b. However, if the awareness of necessary reforms is greater than the willingness to support these reforms (...)

In the example above, the infinitive plus argument postmodifying the NP head "Bereitschaft" occurs after the predicate, while the corresponding infinitive construction appears directly after the NP head "willingness" in English. The alignment of both noun phrases requires the creation of a discontinuous segment in German.

Although we assume that the number of discontinuous segments may be higher in the German than in the English corpus, for the reasons highlighted above, note should be made of the fact that English-German contrasts may also trigger discontinuous elements in the English corpus as illustrated by the following example:

(8) a. What is now clear from the historical evidence of the last century is that in every case where a poor nation has significantly overcome its poverty, this has been achieved while engaging in production for export markets and opening itself to the influx of foreign goods, investment and technology; that is, by participating in globalization."

b. Anhand der historischen Beweise des letzten Jahrhunderts ist jetzt klar, da in jedem Fall, in dem eine arme Nation ihre Armut in beträchtlichem Maße überwunden hat, dies durch die Produktion für Exportmärkte und die eigene Öffnung für ausländische Waren, Investitionen und Technologie geschah - das heißt, durch die Beteiligung an der Globalisierung."

Pseudo-cleft constructions as employed in the example above are a rather frequent strategy in English for realizing clauses as subjects in Theme position (see (Teich, 2003)). Equivalent constructions are relatively rare in German, and indeed, the meaning of the English pseudo-cleft clause is realized as a main clause in the German translation. As a consequence, the complex prepositional phrase of the English pseudo-cleft is moved to the beginning of the sentence in German. An alignment of these two PPs therefore entails the creation of other discontinuous segments in the English original.

Other differences between English and German causing discontinuous segments especially in English may result from the greater availability of non-finite verb constructions or a more verbal realization of meaning in general.

3 Annotation of Parallel Corpora

3.1 GECCo: A Multilingual Parallel Corpus

Our multilingual parallel corpus GECCo, which is an extended version of the CroCo corpus (cf. (Neumann, 2005)), was specifically designed to support contrastive studies of English and German texts as described in the above examples. To our knowledge, it represents one of the few existing resources containing annotation of cohesive devices in parallel multilingual corpora. This type of information plays a crucial role not only in contrastive linguistics and translation studies but also in numerous NLP research areas. Most of the information encoded in the corpus was annotated manually. Further, the corpus includes manual clause alignment.

Aligned Clauses

English:

[when they put it back in] $_{cl:53\_EN}$

German:

[wenn sie es wieder einsetzten] $_{cl:40\_GE}$

Word Layer

English:

```
<token id="t310" string="when"/>
<token id="t311" string="they"/>
<token id="t312" string="put"/>
<token id="t313" string="it"/>
<token id="t314" string="back"/>
<token id="t315" string="in"/>
```

German:

```
<token id="t326" string="wenn"/>
<token id="t327" string="sie"/>
<token id="t328" string="es"/>
<token id="t329" string="wieder"/>
<token id="t330" string="einsetzten"/>
```

Chunk Layer

English:

```
<chunk id="ch132" type="conj" gf="conj">
  <tok xlink:href="t310"/>
</chunk>

<chunk id="ch133" type="np" gf="subj">
   <tok xlink:href="t311"/>
</chunk>

<chunk id="ch134" type="vp_fin"
gf="fin">
  <tok xlink:href="t312"/>
  <tok xlink:href="t315"/>
</chunk>

<chunk id="ch135" type="np" gf="dobj">
   <tok xlink:href="t313"/>
</chunk>

<chunk id="ch136" type="advp"
gf="adv_loc">
  <tok xlink:href="t314"/>
</chunk>
```

German:

```
<chunk id="ch123" type="conj" gf="conj">
  <tok xlink:href="t326"/>
</chunk>

<chunk id="ch124" type="np" gf="subj">
  <tok xlink:href="t327"/>
</chunk>

<chunk id="ch125" type="np" gf="dobj">
  <tok xlink:href="t328"/>
</chunk>

<chunk id="ch126" type="advp"
gf="adv_temp">
   <tok xlink:href="t329"/>
</chunk>

<chunk id="ch127" type="vp_fin"
gf="fin">
   <tok xlink:href="t330"/>
</chunk>
```

Figure 1: Example of Corpus Annotation Layers in GECCo.

For the time being, GECCo contains 10 different registers, i.e. the eight registers of written language of the CroCo corpus and two new registers (interviews and academic discourse) of spoken language (see (Kunz and Koltunski, 2011) for a more detailed description of the GECCo corpus architecture). We are currently trying to enhance the automatic annotation of the new registers by means of manual annotation. Encoding the different layers of manual annotation into CQP, we are faced with the difficulty of encoding discontinuous constituents as illustrated in section 2.

In conclusion we can say that the complexity of linguistic annotations required for studying contrasts in English-German cohesive devices necessitates both

(i) an annotation scheme capable of coping with multilevel annotations, i.e. graph structures and

(ii) a multilevel corpus query engine that can cope with the complexity of our annotation layers and data model.

3.2 Annotation data model

XML is generally considered to be a useful tool for encoding complex structured language data. Indeed, XML is a widely used standard for encoding annotations of natural language corpora. Although the base formalism cannot describe overlapping structures since it was originally designed to represent tree structures only, its extension (Isard and Thompson, 1998) with hyperlinks (*href*) enables the representation of crossing and overlapping structures.

In our corpus annotation framework we have adopted a modular strategy. Each annotation layer is represented as a different XML file generated by MMAX2 (Müller and Strube, 2006) that supports the manual annotation. The mapping of different representation layers (the graph structure) is guaranteed by the (href) hyperlinks between the different XML files. Figure 1 shows some example annotations from the corpus.

In order to allow further corpus query and exploitation, the linguistic information contained in the XML files needs to be merged into a format readable by a corpus query engine. As this operation is not straightforward in the case of discontinuous segments, an overview of the potential difficulties will be provided in the following section.
.

4 Interfacing XML Annotations of Discontinuous Segments in CQP

4.1 CQP data model

CQP is based on an XML-like corpus encoding language that is compatible with the data model we use for corpus annotation.

The primary data used in CQP are tokens. The CQP language is a rigid positional system on the token positions, i.e. the tokens are totally ordered, providing a timeline for the incremental encoding of structural attributes. CQP provides annotations of two types of attributes:

- positional attributes: describe features related to the tokens or token position such as part-of-speech, morphological features, etc.
- structural attributes: describe features related to ordered sets of tokens, such as syntactic chunks, clauses, sentences, etc.

CQP allows for incremental information merging, i.e. structural attributes can be sequentially integrated with the positional attributes so as to refine the linguistic information present in the corpus. Figure 2 displays an example of incremental annotation encoding in CQP.

Further, CQP enables the representation of overlapping structures, which is not allowed in standard XML. However, as CQP uses the positions of tokens for storage and retrieval, discontinuous segments cannot be directly represented.

In conclusion we can say that, in order to encode the GECCo corpus annotation data into CQP, the hierarchical XML representation used for encoding multi-layer annotations needs to be translated into the CQP timeline-based corpus representation on the basis of the position of tokens. The next section describes the strategy we employ for encoding discontinuous constituents into CQP.

4.2 Representing discontinuous segments in CQP

As we have seen previously, structural attributes are encoded in CQP as ordered sets of token positions. Thus, a structural attribute TAG describing an XML tag (e.g. *token* or *chunk*) can be defined as the following sequence of token positions:

$$TAG = [t_1, t_2, ..., t_n],$$

with $[1, 2, ..., n]$ being a continuous sequence. Therefore, in a TAG attribute no gaps are allowed.

Step1: tokens
311: they 312: put 313: it 314: back 315: in
Step 2: morphology
311: they_Pro_plural 312: put _Verb 313: it _Pro_singular 314: back_Adv 315: in_Adv
Step 3: syntax
<np> 311: they_Pro_plural </ /np> <vp> 312: put _Verb 313: it _Pro_singular 314: back_Adv 315: in_Adv </ /vp>

Figure 2: Merging multi-layer XML annotations into CQP.

In order to describe the strategy used to encode discontinuous segments into CQP, we first give the formal definition of a discontinuous structural attribute.

Let TAG be a sequence of tokens describing the structural attribute represented by an XML tag

$$TAG = [t_1, ..., t_j, ..., t_{j+n}, ..., t_k]$$

and $GAPS$ a set of integer pairs such that

$(x_i, y_i) \in GAPS$ iff
$[x_i, x_i+1, ..., y_i]$ is a sequence of integer numbers without gaps

Then, the definition of a discontinuous sequence is as follows:

TAG is discontinuous iff
$\mid GAPS \mid > 1$

As CQP does not support the representation of such discontinuous segments we adopt the following strategy: First, we split a TAG containing gaps into the set of its continuous subsets ($\cup TAG_i$), i.e. sequences of tokens without gaps

```
<chunk id="ch133" gf=subj>
   <token id="t311" string="they"/ >
< /chunk>
<chunk id=”ch134” gap_id=”ch134-gap” gf=fiv>
   <token id=”t312” string= ”put”/ >
< /chunk>
<chunk id="ch135" gf=dobj>
   <token id="t313" string="it"/ >
< /chunk>
<chunk id="ch136" gf=adv_loc>
    <token id="t314" string="back"/ >
< /chunk>
<chunk id=”ch134” gap_id=”ch134-gap” gf=fiv>
    <token id=”t315” string=”in”/ >
< /chunk>
```

Figure 3: CQP XML-like representation of discontinuous segments.

$$TAG = \cup TAG_i, \text{ e.i.}$$
$$= \cup [x_i, ..., y_i], \forall (x_i, y_i) \in GAPS$$

Then, after having assigned an identical coindex *gap_id* to all the subsets of a discontinuous TAG, we represent each of them as a standard CQP structural attribute. At the query stage, the segments that have been split are linked together into a unique segment by a query macro that selects structural attributes with the same *gap_id*.

Summing up, the strategy we adopt consists of three steps:

- partitioning the discontinuous segment into a set of continuous subsets,
- representation of the continuous partitions of the original set as standard CQP structural attributes,
- reconstruction of the original discontinuous segment at the query stage.

An example of a structure that cannot be directly encoded in CQP was given in Figure 1. The English aligned clause contains a discontinuos TAG segment representing a finite verb vp_fin (*put in*).

vp_fin = [t312,t315],
Gap_{vp_fin}=[(312,312),(315,315)]

Figure 3 shows the CQP encoding of the continuous subsets of vp_fin defined by Gap_{vp_fin} for this example.

After segment reconstruction, CQP will extract the expected aligned finite verb chunks from the clause-aligned German/English corpus:

199090: <clause id="GO _SPEECH_009-cl67" align="G2E_SPEECH_009-cl67-cl93">
Dieser Lösung
</clause>
→etrans:<clause id="ETRANS_SPEECH_009-cl93" align="G2E_SPEECH_009-cl67-cl93">
that in the end everyone agreed to this solution
</clause>

199093: <clause id="GO_SPEECH_009-cl68" align="G2E_SPEECH_009-cl68-cl92">
und das ist für mich das Wunder
</clause>
→etrans:<clause id="ETRANS_SPEECH_009-cl92" align="G2E_SPEECH_009-cl68-cl92>
The miraculous thing for me was
</clause>

199101: <clause id="GO_SPEECH_009-cl67" align="G2E_SPEECH_009-cl67-cl93">
haben zum Schluss alle zugestimmt
</clause>
→etrans:<clause id="ETRANS_SPEECH_009-cl93" align="G2E_SPEECH_009-cl67-cl93">
that in the end everyone agreed to this solution
</clause>

Figure 4: CQP representation of alignment in GECCo.

$vp_fin_$EN = *[put in]*
$vp_fin_$GE = *[einsetzen]*

Figure 4 represents the output obtained by querying the GECCo corpus with CQP. In particular, it shows how the framework described in this paper permits both an efficient encoding and querying of linguistic annotations (e.g. the alignment of linguistic discontinuous constituents such as (1) with (2)) in CQP.

5 Conclusion

In this paper, we have discussed problematic issues that may arise in connection with the automatic encoding of a manually annotated corpus into the multilevel corpus query engine CQP. Manual corpus annotation often produces complexly structured representations of the linguistic information displayed in the corpus that are difficult to encode using general state-of-art corpus query tools.

While much research has addressed the issue of providing annotation standards for linguistic corpora, only a few resources (e.g. ANNIS2 and MATE) exist that provide efficient interfacing of those multi-layer annotations standards with corpus query engines. However, MATE (McKelvie et al., 2001) does not support parallel corpora encoding. ANNIS 2 (Zeldes et al., 2009) for instance provides translation utilities from arbitrary XML data structures to the ANNIS format. The Annis2 representation format allows the representation and graphs and discontinuous constituents of arbitrary complexity. However, the corpus query language provided by this system is highly complex and requires a high level of expertise on the part of the user.

In this paper we proposed a CQP-based alternative to ANNIS2. We described the strategy we implemented that allows the encoding and querying in CQP of multi-layer parallel corpora that include linguistic phenomena of arbitrary complexity.

Our framework compares well with frameworks such as the one implemented into ANNIS 2 in that it combines all the advantages of the corpus query engine CQP, e.g. efficient querying of very large text corpora, efficient querying of parallel corpora and an intuitive and user-friendly corpus query language, with a framework for encoding arbitrary complex data structures into CQP.

Acknowledgments

The authors thank the DFG (Deutsche Forschungsgemeinschaft) for supporting this project.

References

Philippe Blache, Brigitte Bigi, Laurent Prévot, Stéphane Rauzy, and Julien Seinturier. 2010. A general scheme for broad-coverage multimodal annotation. In *Proceedings of ICGL-10*.

Oli Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*.

M. Doherty. 2004. Strategy of incremental parsimony. *SPRIKreports*, No 25.

C. Fabricius-Hansen. 1999. Information packaging and translation: Aspects of translational sentence splitting (german/english/norwegian). *Studia Grammatica*, 47:175–214.

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

McKelvie D. Isard, A. and H.S. Thompson. 1998. Dialogue transcripts: A new sgml architecture for the hcrc map task corpus. In *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP98*, Sydney.

E. Koenig and V. Gast. 2007. *Understanding English-German Contrasts. Grundlagen der Anglistik und Amerikanistik*. Schmidt (revised 2nd edition: 2009), Berlin.

Kerstin Kunz and Ekaterina Lapshinova Koltunski. 2011. Tools to analyse german-english contrasts in cohesion. In *Hamburg Working Papers in Multilingualism*.

David McKelvie, Amy Isard, Andreas Mengel, Morten Baun Mller, Michael Grosse, and Marion Klein. 2001. The mate workbench - an annotation tool for xml coded speech corpora. *Speech Communication*, pages 97–112.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

S. Neumann. 2005. Corpus design. *Deliverable of the CroCo Project*, No 1.

Sébastien Paumier. 2000. Nouvelles méthodes pour la recherche d'expressions dans de grands corpus. In A. Dister, editor, *Actes des 3èmes Journées INTEX. Revue Informatique et Statistique dans les Sciences Humaines, 36éme annèe, n 1 à 4*.

Ulrik Petersen. 2004. Emdros: a text database engine for analyzed or annotated text. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics.

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

Erich Steiner. 2001. Translations englishgerman: investigating the relative importance of systemic contrasts and of the text type translation. *SPRIKreports*, No 7:1–49.

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

Amir Zeldes, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. Annis: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009.*