Connective-based Local Coherence Analysis: A Lexicon for Recognizing Causal Relationships

Manfred Stede

University of Potsdam (Germany)

email: stede@ling.uni-potsdam.de

Abstract

Local coherence analysis is the task of deriving the (most likely) coherence relation holding between two elementary discourse units or, recursively, larger spans of text. The primary source of information for this step is the connectives provided by a language for, more or less explicitly, signaling the relations. Focusing here on causal coherence relations, we propose a lexical resource that holds both lexicographic and corpus-statistic information on German connectives. It can serve as the central repository of information needed for identifying and disambiguating connectives in text, including determining the coherence relations being signaled. We sketch a procedure performing this task, and describe a manually-annotated corpus of causal relations (also in German), which serves as reference data.

1 Introduction

"Text parsing" aims at deriving a structural description of a text, often a tree in the spirit of Rhetorical Structure Theory (Mann and Thompson, 1988). For automating this task (see, e.g., Sumita et al. (1992); Corston-Oliver (1998); Marcu (2000)), the central source of information are the *connectives* that the author employed to more or less specifically signal the type of coherence relation between adjacent spans. For illustration, consider this short text:[1].

> Because well-formed XML does not permit raw less-than signs and ampersands, if you use a character reference such as `&#60;` or the entity reference `<` to insert the < character, the formatter will output `<` or perhaps `&#60;`.

Supposing that we are able to identify the connectives and punctuation symbols correctly (here in particular: note that *to* is not a spatial preposition; distinguish between commas in enumerations and those finishing clauses), we can identify the "scaffold" of this short text as the following:

`Because A, if B or C to D, E or F`

with `A` to `F` representing the minimal units of analysis. Next, fairly simple rules will be sufficient to guess the most likely overall bracketing of this string:

`(Because A, (if ((B or C) to D)), (E or F))`

And finally, it happens that the connectives *because, if, to* and *or* are quite reliable signals of the coherence relations *Reason, Condition, Purpose* and *Disjunction*, respectively. Combining this information with the bracketing, we can obtain a tree structure in spirit of RST.

Texts of this level of complexity could be handled by early text parsers (see Section 2). But, obviously, not too many texts behave as nicely as our example does. In general, constructing a discourse tree is highly complicated even without trying to find semantic/pragmatic labels for the relationships; the discussion by Polanyi et al. (2004) demonstrates that just the structural decisions are often very difficult to make. Taking a different viewpoint, this author argues in Stede (2008) that constructing "the" tree structure for a text should not be regarded as such an important goal and that coherence should rather be explained as the interplay of different levels of (possibly partial) description, such as referential and thematic structure, intentional structure, and a level of *local coherence analysis* that records the clearly recognizable relationships between adjacent text spans but does not aim at constructing a complete and well-formed tree. In the present paper, this viewpoint is taken to the task of automatic analysis, which aims at identifying individual coherence relations and the spans related. We restrict ourselves here to *causal* relationships and moreover to those that are explicitly signaled by a connective. The central resource used in our approach is a lexicon that collects the information associated with individual connectives and makes it available to applications such as a coherence analysis or text generation.

The paper is organized as follows. After reviewing some earlier research on text parsing in Section 2, we turn to connectives in Section 3 and point out a number of problems that sophisticated coherence analyzers have to reckon with. Then, Section 4 explains the connective lexicon we developed, and Section 5 describes a corpus we collected and annotated manually for causal connectives and the relations they signal.

[1] Source: `http://www.cafeconleche.org/books/bible2/chapters/ch17.html`

It serves as a reference for designing the analysis procedure, which is finally sketched in Section 6. Our analysis and implementation target German text, but most of the phenomena apply equally to English.

2 Related Work

In the late 1990s. the best-known work on "text parsing" was that of Marcu, which is collected in Marcu (2000). He had used surface-based and statistical methods to identify elementary discourse units, hypothesize coherence relations between adjacent segments, and finally compute the most likely overall "rhetorical tree" for the text. Surface-based methods were highly popular at the time, but with the recent advances in robust and wide-coverage sentence parsing, it seems sensible to cast local coherence analysis as a problem of linguistic analysis, drawing on the results of syntactic parsing (or even, on top of that, semantic analysis).

An early approach in this spirit was implemented in the RASTA analyzer (Corston-Oliver, 1998). It perused the output of the 'Microsoft English Grammar' to guess the presence of coherence relations on the basis of accumulated evidence from a variety of more or less deep linguistic features. For instance, a hypotactic clause would always figure as the satellite of some nucleus-satellite relation in RST terms. For some relations (e.g., *Elaboration*), the type of referring expressions, especially in subject position, was considered a predictive feature. In general, RASTA employed a set of necessary criteria for each relation to hold in a particular context, and for those relations passing the filter, a voting scheme accumulated evidence to decide on the most likely relation. The system worked on *Encarta* articles, hence on expository text; 13 relations were being used.

While RASTA employed a relation-centric approach, the recent work by Lüngen et al. (2006) places the connectives at the center of the analysis, recording information about them in a specific lexicon (similar to our own earlier work (Stede, 2002)). In the lexicon used by Lüngen et al., an entry consists of three zones: the *identification* zone gives the textual representation of the connective, its lemma and part-of-speech tag; the *filter* zone encodes necessary conditions for particular discourse relations, in the form of context descriptions; the *allocation* zone then specifies a default relation to be assumed if no other relation can be derived on the basis of further (soft) conditions. It also encodes constraints on the size of units to be related, the nuclearity assignment, and the information whether the segment including the connective attaches to the left or to the right in the text. Each entry gives rise to a rule used by a shift-reduce parser that tries to build a complete rhetorical tree. This parser works in close cooperation with a module identifying logical document structure, and the context conditions specified in lexicon entries often refer to this level of structure, or to a syntactic dependency analysis provided by the *Connexor* parser[2].

We share with these approaches (and with that of Polanyi et al. (2004)) the desire to derive as much information about discourse relations as possible *without* resorting to non-linguistic knowledge, so that the role of local coherence analysis in effect can be seen as extending the realm of robust sentence parsing. Our approach is to represent as much of the necessary information as possible in a declarative resource: a lexicon

[2] `http://www.connexor.com`

of connectives.

3 Complications with Connectives

Connectives are closed-class lexical items that can belong to four different syntactic categories: coordinating and subordinating conjunction, adverbial, and preposition (such as *despite* or *due to*). They have in common that semantically they denote two-place relations, and the text spans they relate can at least potentially be expressed as full clauses (Pasch et al., 2003). As mentioned in the beginning, they are not always as easy to interpret as in our "well-formed XML" example. In this section, we suggest an inventory of the complications that a thorough local coherence analysis procedure needs to deal with. We group them into four categories.

Ambiguity. Here we need to distinguish two kinds: (i) ambiguity as to whether a word is used as a connective or not, and (ii) ambiguity as to the semantic reading of a connective. Certain cases of (i) correspond to the distinction between 'sentential use' and 'discourse use' that Hirschberg and Litman (1994) had proposed not for connectives but more generally for 'cue phrases' in spoken language. For example, German *denn* can be a coordinating conjunction (sentential use) or a particle often used in questions without a recognizable semantic effect (discourse use). Other cases of (i) reflect ambiguity between different 'sentential' uses. Sometimes this coincides with a syntactic difference (e.g., English *as* is a connective only when used as subordinator), but with many adverbials it does not (e.g., German *daher* can be a locative adverbial 'from there' or a causal adverbial 'therefore'). Also, sometimes the distinction coincides with semantic scope, as with the focus particle / connective *nur* ('only'):

(1) Es war ein schöner Sommertag. Nur die Vögel sangen nicht.
('It was a nice summer day. Only the birds weren't singing.')

In a narrow-scope reading of 'only', the message is that everybody was singing except for the birds; in a wide-scope reading, 'only' connects the two sentences and signals a restrictive elaboration. Ambiguity of type (i) is more widespread than one might think; in Dipper and Stede (2006), we report that 42 out of 135 frequent German connectives also have a non-connective reading, and we point out that many of the problems cannot be handled with off-the-shelf part-of-speech taggers.

Concerning ambiguity (ii), some connectives can have more than one semantic reading, which we regard as a difference in the coherence relation being signaled. Sometimes, the relation can be established on different levels of linguistic description (see, e.g., Sweetser (1990)). For example, *finally* can be used to report the last one in a sequence of events, or it can be used by the author as a device for structuring the discourse ("and my last point is..."). Interestingly, the very similar German word *schließlich* in addition has a third reading: It can also be an argumentative marker conveying that a presented reason is definitive or self-evident, which in English may be signaled with 'after all': *Vertraue ihr. Sie ist schließlich die Chefin.* ('Trust her. She is the boss, after all.')

Pragmatic features. In addition to the relational differences, connectives can sometimes be distinguished by more fine-grained pragmatic features, which are usually not modeled as a difference in coherence relation. A well-known case in point is the difference between *because* and *since* (corresponding to German *weil / da*), where only

the latter has a tendency to mark the following information as hearer-old (not necessarily discourse-old). The same pair of connectives serves to illustrate the feature of non-/occurrence within the scope of focus particles:

(2) Nur weil/?da es regnet, nehme ich das Auto
('Only because/?since it's raining, I take the car.')

While in German, the *da* variant is hardly acceptable at all, in English there is a tendency for *since* to be interpreted in its temporal reading when used within the scope of *only*.

Also, connectives can convey largely the same information yet differ in terms of stylistic nuances, for instance in degree of formality. Thus a concessive relation in English may be signaled in a standard way with *although*, or with a rather formal, and in that sense "marked" *notwithstanding* construction.

Form. While the majority of connectives consist of a single word, some of them have two parts. Well-known instances are *either .. or* and *if .. then*. For the German version of the latter (*wenn .. dann*), a coherence analyzer must account for the possibility of its occurring in reverse order: *Dann nehme ich eben das Auto, wenn Du so bettelst.* ('Then I'll take the car, if you're begging so much'.) Further, looking at highly frequent collocations such as *even though* or *even if*, it is difficult to decide whether we are dealing with a single-word connective and a focus particle, or with a complex connective; one solution is to check in such cases whether the meaning is in fact derived compositionally and then to prefer the focus particle analysis. From "regular" two-word connectives it is only a small step to the shady area of *phrasal* connectives, which can allow for almost open-ended variation and modification: *for this reason / for these reasons / for all these very good reasons / ...*.

For German, we have dealt with the issue of differentiating between types of multi-token connectives in a separate paper Stede and Irsig (2008).

Discourse structure. As is well-known, the structural description of a text can also be more complicated than in our "well-formed XML" example shown at the beginning. For one thing, discourse units can be embedded into one another, using parenthetical material or appositions. Further, connectives can occasionally link text segments that are non-adjacent — a phenomenon that has been studied intensively by Webber et al. (2003) and also by Wolf and Gibson (2005). An example from Webber et al.: *John loves Barolo. So he ordered three cases of the '97. But he had to cancel the order because then he discovered he was broke.* Here, the *then* is to be understood as linking the discovery event back to the ordering event rather than to the (adjacent) canceling. In German, many adverbial connectives have an overt anaphoric affix (e.g., *des*wegen, *da*her, trotz*dem*), and the ability to link non-adjacent segments appears to be restricted to these. Non-adjacency also leads to the issue of crossing dependencies, which is also discussed by the two teams of authors mentioned above. It correlates with the problem of two connectives occurring in the same clause, as it happens in the *Barolo* example *(because then)*, which renders the parsing task significantly more complex than in the "well-formed XML" example.

A different problem is to be found in situations where a single coherence relation is signaled twice, by two different connectives, where one typically is to be read cataphorically:

(3) Ich nehme deshalb$_i$ das Auto, weil$_i$ Du so bettelst.
('I take the car (for that reason)$_i$ because$_i$ you're begging so much.')

This phenomenon is difficult to reproduce in English; again, in German it is also limited to a certain class of connectives that can serve as cataphoric 'correlates'. Obviously, in such examples, a coherence analyzer will have to be very careful not to hypothesize two separate causal relationships. The same danger applies when multiple causes are enumerated for the same consequence, or multiple consequences arising from the same cause. The mere insertion of the focus particle *auch* ('also') in example 3 can fundamentally change the discourse structure to stating two reasons for taking the car:

(4) Es regnet sehr stark. Ich nehme deshalb das Auto, auch weil Du so bettelst.
('It's raining heavily. I therefore take the car, also because you're begging so much.')

Finally, it is to be noted that certain connectives convey information about the discourse structure *beyond* the local relation between two segments. A case in point is the first word of this paragraph, which not only makes a 'List' or 'Enumeration' relation explicit, but also provides the information that this very list is now coming to an end. A smart coherence analyzer could thus reduce the search space for linking the subsequent text segment — it will definitely *not* be part of the same 'List' configuration.

4 A Rich Lexical Resource for Connectives

For building programs to perform local coherence analysis on texts that display the complexities discussed above, our approach is to clearly divide the labor between a declarative connective lexicon on the one hand, and a flexible analysis procedure on the other. In this section, we describe our *Discourse Marker Lexicon* (DIMLEX), whose first version was described in Stede (2002). At the time, it was used for relatively simple text parsing as outlined at the beginning of the paper, and also for a language generation application. The multi-functionality results from using a rather abstract XML encoding for the "master" lexicon, which is transformed by XSLT scripts to the format needed by a specific application — both in terms of technical format (e.g., programming language) and the amount and granularity of information needed for the application. With our current focus on causal relations, we extended the DIMLEX entries of the causal connectives to a richer scheme, which will gradually be transferred to the remaining connectives as well.

It is not trivial to define an inventory of causal connectives, due to the grey area of words marking a semantic relationship that readers *can* also interpret causally — after all, causality is very often not explicitly signaled but being left for the reader to reconstruct. For example, in *The wind shook the shed for a few seconds, and then it collapsed* there certainly is causality involved in the relationship between the sentences, but we would not want to treat *and* or *then* as causal connectives. With the help of the 'Handbook of German Connectives' (Pasch et al., 2003), we determined a set of 66 German connectives that *primarily* convey causality.

The DIMLEX entries for these connectives consist of the following zones of information: (1) orthography, syntax, and structural features; (2) non-/connective disambiguation rules; (3) semantic and pragmatic features, including information on disambiguating different readings, and on role linking. As for the type of information, entries contain both binary features and probabilities derived from corpus analyses.

Orthography and syntax. Orthographic variants that we store in the lexicon result from the recent official German spelling reform and from frequent mistakes made by speakers/authors (as found in corpora). Also, we list both upper and lower case spellings because this difference plays a role in many disambiguation rules (see below). Each variant has a unique identifier that is being used in those rules. Also, one of the variants is marked as 'canonical' for co-reference purposes. Here is a sample excerpt from the entry for *aufgrund*, corresponding to the English *due to*:

```
<orth type="simple" canon="1" onr="k2v1">
   <part type="cont">aufgrund</part> </orth>
<orth type="complex" canon="0" onr="k2v2">
   <part type="cont">auf Grund</part> </orth>
<orth type="simple" canon="0" onr="k2v3">
  <part type="cont">Aufgrund</part> </orth>
<orth type="complex" canon="0" onr="k2v4">
  <part type="cont">Auf Grund</part> </orth>
```

Each `orth` is of type 'simple' or 'complex', depending on the number of tokens involved. For simple connectives (single tokens), the `part type` is always 'cont' (continuous), whereas for complex connectives it may also be 'discontinuous' if linguistic material can intervene between the parts (which is not the case for the two complex variants above).

Syntactically, connectives can be subordinating conjunctions; *Postponierer*; pre-, post- and circumpositions; and adverbials, some of which can occur only in specific positions (characterized in accordance with the *Feldermodell* that is often used to describe German sentence structure in terms of *Vorfeld, Mittelfeld, Nachfeld*). We encode this information following the classification by Pasch et al. (2003)), whose primary criterion is whether the connective can be *integrated* into the clause, and if so, at what positions it can occur. Here is the information for the prepositional adverb ('padv') *dadurch* ('by means of this'):

```
<padv>
  <vorfeld>1</vorfeld>
  <mittelfeld>1</mittelfeld>
  <nacherst>0</nacherst>
  <nachfeld>1</nachfeld>
  <nullstelle>0</nullstelle>
  <nachnachfeld>0</nachnachfeld>
  <satzklammer>0</satzklammer>
</padv>
```

The binary features say that the connective can be in the Vorfeld (preceding the finite verb or auxiliary: *Dadurch ist es geschehen*), Mittelfeld (between auxiliary and

verb: *Es ist dadurch geschehen*), and Nachfeld (following the verb phrase: *Es ist geschehen dadurch*).

As a representation more directly usable for computational purposes, we also specify patterns of the connective being situated in a syntax tree in TIGER format (Brants et al., 2004). This format is used both in large hand-annotated German corpora as well as in an automatic parser[3]. The idea of the patterns in the lexical entry thus is to find instances of the word in a TIGER-tree, whether coming from a treebank or from a parser. For illustration, here is the pattern for the complex connective *so .. dass* ('so .. that'):

```
(#avp:[cat="AVP"] > [lemma="so"])
&
((#avp > #s:[cat="S"])
 |
 ((#avp > #cs:[cat="CS"]) &
  (#cs > #s:[cat="S"]))
)
&
(#s > [lemma=("dass")])
```

This expression looks for an adverbial phrase (AVP) that dominates both *so* and a sentence (S), or a coordination of sentences (CS) that in turn dominate *dass*. Between the *so* and *dass*, any material can intervene. An examples matched by this expression in the TIGER corpus is: *Der Kanzler hat China so gern , daß er ihm sogar die höchsten Berge der Welt zu schenken vermöchte.* ('The chancellor likes China so much, that he even wants to give the world's highest mountains as a present to the country.')

Besides the syntactic structure of individual conjuncts, we also need to represent the possibilities on linear order of the conjuncts. This is also based on the terminology of Pasch et al. (2003), who distinguish between the *internal* conjunct (the clause or phrase that the connective syntactically belongs to) and the *external* one. Sometimes, this a hard constraint: With the conjunction *denn* (causal 'for'), the internal conjunct can only follow the external one. With other connectives, e.g., *weil* ('because'), both orderings are possible, i.e., the *because*-clause giving a reason can precede or follow the clause giving the effect. In these cases we include probabilities derived from a corpus analysis, which the coherence analysis module can use for disambiguating scope when it has no other information available.

The syntactic representations become somewhat more complicated in case of *complex* connectives. For instance, there is a variant of *dadurch* that co-occurs with a subsequent (but not necessarily adjacent!) complement clause headed by *dass* ('that'). Similarly, as shown in the previous section, certain causal conjunctions and adverbials can co-occur and redundantly mark the same relation. Our lexicon entries contain features representing those possible pairings. For a more general discussion on German complex connectives, see Stede and Irsig (2008).

Finally, we include a feature stating whether the connective can be in the scope of a focus particle. This information can sometimes support non-/connective disambiguation.

[3] http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/BitPar.html

Non-/connective disambiguation. In Dipper and Stede (2006), we reported on an approach to disambiguating non-/connective use for nine connectives by incrementally training a Brill tagger, which lead to F-measures of 81% (+connective) and 95% (–connective) in the best of four training scenarios. During this work it became clear that the part-of-speech context of the word often indeed provides enough information for making the decision. The main reason why off-the-shelf taggers, however, do not perform very well is that tagsets do not reflect the distinction — recall the syntactic heterogeneity of the "class" of connectives. From our findings we thus constructed for each connective a set of patterns over part-of-speech and lemma information, leading to regular expressions associated with probabilites (again gathered from corpus studies). These expressions become part of the *DiMLex* entries and can be used by the coherence analyzer. Starting from the Dipper/Stede results, we manually created classes of connectives with apparently-equivalent behavior, rather than studying each of the 66 connectives in detail. For illustration, here is the pattern set for *daher*, which can be a causal connective ('therefore') or a locative adverb ('from there'):

```
<conn-disambi>
   <pros>
      <pro value="90" ref="k5v2"> $. $$/PROAV </pro>
      <pro value="90" ref="k5v1"> VVFIN $$/PROAV </pro>
   </pros>
   <cons>
      <con value="99" ref="k5v1 k5v2">
         $$/PROAV $, {'dass'}/KOUS
       </con>
      <con value="95" ref="k5v2">
        $. $$/PROAV .* {'kommen' 'ruehren'} .+ $, {'dass'}/KOUS
      </con>
      <con value="99" ref="k5v1"> $$/PROAV $. </con>
   </cons>
</conn-disambi>
```

Weights range from 0 to 100, so 99 represents basically a strict rule. Notice the `ref` attribute, which restricts the rules to orthographic variants (in this case to upper and lower case ones). The first two rules support a +connective reading: *daher* tagged as pronominal adverb (PROAV) following a full stop or a finite verb, respectively. The following three rules support a –connective reading: *daher* followed by the subordinating conjunction (KOUS) *dass*; occurring in a collocation like *kommt daher, dass* ('stems from'); occurring before a full stop, i.e., sentence-final.

Semantics and Pragmatics. As stated earlier, we identify a difference in readings with a difference in *coherence relation* signaled by the connective. As for the inventory of relations, we take inspiration from Mann and Thompson (1988), Asher and Lascarides (2003), and especially for the causal relations, from the taxonomic approach of Sanders et al. (1992). Not every distinction made in the literature can be traced to connectives; so we do for instance not follow RST's distinction between 'Volitional Cause' and 'Non-volitional Cause' in DIMLEX. But we find differences in connective use for semantic versus pragmatic causal relations (Sanders et al., 1992).

For instance, the *denn* used in (4) below is quite typical for pragmatic relations (see, e.g. Pasch, 1989).

(5) Er wird bestimmt pünktlich kommen, denn er ist doch immer so gewissenhaft.
('Surely he will arrive on time, for he is always so assiduous.')

Thus, in the realm of causality we use coherence relations labeled 'Argument-Claim' (pragmatic) and 'Reason-Consequence' (semantic). Further, if the consequence is a yet-unrealized intended effect, we assign the relation 'Purpose' as it has been suggested by Mann and Thompson (1988). The connectives associated with Purpose are mostly quite specific (e.g., English *in order to*; German *um .. zu*), but there can also be ambiguity between Purpose and "other" causality (e.g., English *so that*; German *damit*).

Disambiguation between the semantic and the pragmatic relation is usually very difficult and thus a matter of heuristically weighing the evidence. Similar to our handling non-/connective disambiguation (see above), we use a scheme of weight accumulation for features indicating the presence of a relation. For example, for the connective *schließlich* we found that with the main verb of the clause elided, the pragmatic reading is very unlikely; on the other hand, if the verb is in present tense and the Aktionsart is 'state', it very likely signals the pragmatic 'Claim-Argument' relation. Other evidence for this relation includes modal particles signaling the epistemic status of the proposition(s), often in conjunction with present or future tense. This is illustrated in example 4 above, where the speaker expresses her confidence that the event will materialize with *bestimmt* ('surely'), while *doch* ăin the second clause marks the information has hearer-old, so that the difference between claim and argument in this case is quite transparent. Other features we modeled are inspired by the empirical work of Frohning (2007). They include position, tense and aspect of the clause, mood and modality, and lexical collocations; Frohning derived their weights from corpus analyses.

Often, however, no compelling evidence for either of the three relations can be found, and for these cases we use a neutral relation called 'Cause-Caused', which is thus meant to subsume the two others.

In addition to relation(s), a lexicon entry specifies the *role linking* for connectives: the mapping from the syntactically internal or external conjunct (see above) to its function in the relation. We label these functions in accordance with the relations: 'Argument', 'Claim', 'Reason', and so forth. Since causal relations are directed, and the mapping cannot be predicted from syntactic features, it is crucial to represent this information explictly.

Besides, we use a number of more idosyncratic features to represent information that is relevant only for certain connectives, in particular to distinguish very similar ones. An example mentioned in the previous section is the information-structural difference between *weil* ('because') and *da* ('since'). For other families of connectives, this "miscelleneous features" section is more important; with temporal connectives, for instance, we specify in addition to the coarse-grained coherence relation more fine-grained distinctions such as whether the time spans of the related events meet or not, etc.

Having discussed our treatment of syntax and semantics separately, we now have to attend to the relationship between the two, i.e., to the issues of ambiguity and polysemy. The majority of connectives has one syntactic description and can convey one or two similar coherence relations (the typical ambiguity between semantic and pragmatic reading). We do, however, also find other configurations:

- Two syntactic descriptions: *weil* used to be a subordinating conjunction, but in spoken German is now widely accepted as a coordinating conjunction as well. Since the meaning is the same, it suffices to simply list both syntactic variants in DIMLEX.

- One syntactic description, many coherence relations: When used as an adverb, the connective *damit* can signal Purpose ('so that') or Reason-Consequence ('thus'). This situation is similar to the previous one: We provide a disjunction of semantic readings (including the disambiguation information) and a single syntactic description.[4]

- Two syntactic descriptions, several coherence relations: These cases are the only serious complications, as a difference in syntax can correlate with one in semantics, so that we cannot simply specify disjunctions for the syntactic and semantic descriptions. Instead, we use multiple lexical entries, in accordance with the intuition that we are dealing with fairly unrelated items (polysemy). An example is *dann* ('then'), which on the one hand is a temporal adverbial, and on the other hand can express a Condition relation (optionally with a corresponding *wenn* ('if') in the other clause). In the latter case, it does not behave as an adverb, though, but it governs a verb-second clause. So, distributing the information across two separate lexicon entries seems to be appropriate.

Finally, to enhance the maintainability of DIMLEX, we inlcude with the entries a range of linguistic examples that illustrate the relevant distinctions, and we also citee information that is provided by standard dictionaries — especially in those cases where our formalization is not yet complete. One of the XSLT scripts for converting DIMLEX maps the base lexicon to an HTML format that allows for inspecting the entries, including the information just mentioned, which is intended for the human eye rather than for automatic parsers or generators.

5 A Corpus Annotated with Causal Relations

As a preparatory step for implementing a local coherence analyzer that aims specifically at identifying causal relations, we built a corpus with causal connectives annotated manually. We selected 200 short texts from a product review web site[5], where travelers comment on various tourist destinations. Since they often give reasons for the opinions they express, this genre offers more instances of causal connectives than, say, newspaper text. On the other hand, there is the undeniable drawback of frequent

[4] As a matter of fact, the situation is more difficult: *Damit* is one of the most complicated words in our lexicon, as it also has a reading as subordinator where it signals Purpose, as well as a non-connective adverbial reading ('with it/that').

[5] `http://www.dooyoo.de`

mistakes in grammar and orthography, which makes any automatic analyis quite hard, and also sometimes poses challenges to the human annotator.

Creating the corpus involved several steps. First, potential causal connectives were searched automatically (using the list from DIMLEX) and manually filtered. Subsequently, *identifying* causal connectives was not an issue for the annotation process, as they were already presented to annotators as "anchors" for their task. We then designed annotation guidelines with instructions for identifying causes and effects. As for the length of spans, annotators were encouraged to prefer a shorter span in cases where the boundary of a cause or effect is not quite clear. At the same time, they were asked to mark two discontinuous spans in cases where a cause/effect was interrupted by extraneous material such as authors' remarks on their own text production. Thus, in the following example, the C1 and C2 indices mark the intended cause, and E the intended effect.

(6) [The beach was not very pleasant]$_E$, as [it was,]$_{C1}$ I just have to say this here, [utterly littered with remains of picnics.]$_{C2}$

When multiple reasons are given for the same effect (or vice versa), annotators had to mark them separately, so that each cause-effect pair can be derived individually from the annotated data. Sometimes this multiplicity can involve separate connectives, as in the following example. In such cases, annotators had to choose a central connective (the one linking the adjacent cause and effect) and then add additional ones as secondary connectives, possibly forming a chain. This ensures easy retrievability of all pairs from the data.

(7) [We reached the hotel late]$_E$ [due to]$_{Co1}$ [the flight's delay]$_{Ca}$ and also [because]$_{Co2}$ [it took so long to find a cab.]$_{Cb}$

Further, annotators had to identify possible redundant markings of the *same* cause-effect pair (as with the cataphoric correlates discussed above) as well as focus particles that modify connectives. Thus, in example (7), they would mark *also* and link it to the modified *because*.

Our first version of the annotation guidelines was subject to an informal evaluation with annotators who had not been involved in the project. On the basis of the results we clarified several aspects in the guidelines and thus wrote the final version. Furthermore, we prepared two instructional videos: one for using the annotation tool MMAX2[6], and one for our specific annotation scenario, illustrating the handling of a fairly complicated text passage. In the formal evaluation with two annotators, they received no training other than by the guidelines and the two videos. Of 78 connectives, 34 were analyzed identically. The vast majority of the mismatches (36 of 44) resulted from different span length: There was overlap between the spans chosen by the annotators, but the boundaries were not exactly identical. Other mismatches, which occurred only a few times, included different decisions on secondary connectives and the resulting chains of causes/effects.

Finally, with the guidelines having become stable, experienced annotators created the "official" annotation of the entire corpus of 200 texts (containing some 1,200

[6] `http://mmax2.sourceforge.net`

causal connectives). It is now is available as a resource for training and evaluation of automatic procedures. We also developed a web-based viewer (essentially translating the MMAX2 format to HTML and Javascript) that allows for manually browsing the corpus comfortably.[7]

6 Towards recognizing causal relations automatically

Having described DIMLEX as the central resource for local coherence analysis, and the corpus as reference and evaluation tool, we now briefly sketch a procedure for recognizing causal relations, whose implementation is currently under way in our text analysis workbench (Chiarcos et al., 2008), a standoff XML architecture for fusing linguistic annotations coming from different manual or automatic annotation tools. In this highly modular approach, the output of each individual analysis module is stored in a separate layer, using our standoff XML format PAULA (Dipper, 2005). Analysis tools can use previously computed layers for their own task, which usually involves creating one or more new layers.

In this setting, the task of local coherence analysis involves the following layers. The first four are to be built in the pre-processing phase, and the last two are the result of the coherence analyzer:

1. Token layer (including sentence boundaries)
2. Part-of-Speech
3. Logical document structure (headlines, paragraph breaks, etc.)
4. Dependency syntax analysis
5. Elementary discourse units
6. Connectives and (sets of) EDUs they relate

The procedure of coherence analysis consists of the following three sequential steps, which at various points make use of information from DIMLEX:

Connective identification. All the words listed in DIMLEX as some orthographic variant of a causal connective are identified in the text. This includes a check for complex connectives as listed in the lexicon, i.e., two corresponding words in adjacent clauses (amongst others, the *if .. then* type). It also includes a check for correlates, i.e., a connective that according to DIMLEX can be a correlate occurring in a clause immediately preceding a subordinate clause governed by a connective that according to DIMLEX can have a correlate (the *deshalb .. weil* type. For these checks, the syntax layer (4) is used to identify adjacent clauses.

Next, the single-word connective candidates are run through the disambiguation filters, i.e., the PoS/token regular expressions specified in their lexical entries are matched against the text's PoS representation on the corresponding layer (2). Those items that appear to be words in non-connective use are removed from the connective list. Finally, a new layer (6) is created, for now holding only the words that were recognized as connectives.

[7] All material can be found at `http://www.ling.uni-potsdam.de/~stede/kausalkorpus.html`.

Segmentation. The basic idea of our approach follows that of the module implemented by Lüngen et al. (2006) for German. We first overgenerate, guessing segment boundaries at every possible position, according to the dependency parse result; then, contextual rules remove those boundaries that appear to be wrong (e.g., commas in enumerations). We are, however, using somewhat different definitions of segments, namely a variant of Jasinskaja et al. (2007), and the corpus annotated according to those segmentation guidelines will be used to evaluate our module. One issue where we diverge from both Lüngen *et al.* and from Jasinskaja *et al.* is in our handling of prepositions: We do admit certain prepositional phrases as elementary discourse units, but only those that are headed by a preposition listed in DIMLEX, e.g., the causal markers *wegen* ('due to') or *durch* ('through'). The resulting sequence of segments is represented on a new layer of analysis (5).

Relation and scope identification. Next, the connective layer (6) is extended with information on relations and scopes: Every connective is associated with one or more attribute-value structures listing possible coherence relations along with probabilities. To this end, all relations stored with the connective in DIMLEX are recorded as hypotheses, and weights are accumulated as the result of evaluating the associated disambiguation rules, which largely operate on the syntax layer, as explained in Section 4.

Finally, for each relation we also hypothesize its scope: the thematic roles are associated with sequences of minimal units from layer (5). Given a reliable syntactic analysis, scope determination is usually straightforward for coordinating and subordinating conjunctions. For adverbials, we hypothesize different solutions and rank them according to size: The most narrow interpretation is taken as most likely. In this step, we also consider the layer of logical document structure in order to avoid segments that would stretch across paragraphs or other kinds of boundaries. Similarly, a layer with the results of "text tiling" (breakdown of the text in terms of thematic units, in the tradition of Hearst (1994)) could be used for this purpose, as well as as an 'attribution' layer that identifies those modal contexts that attribute a span of text to a particular source (as in indirect speech).

In this way, the module will generate hypotheses of coherence relations and related spans, for the time being solely on the basis of connectives occurring in the text. As explained, this information is represented in two additional analysis layers. Modules following in the processing chain may combine the various hypotheses into the most likely overall relational tree structure for the paragraph (or a set of such tree structures, see Reitter and Stede (2003)), or they may use the hypotheses directly for some application that does not rely on a spanning tree.

7 Discussion

The central idea behind the separation of the declarative DIMLEX resource and the (ongoing) implementation of an analysis procedure is to facilitate a smooth extensibility of the overall approach towards further kinds of connectives and coherence relations. When the lexicon is extended — while the underlying scheme remains unchanged — coverage of the analyzer grows without adaptations to the analysis procedure. An important benefit of the XML-based organization of the lexicon is its suitability for a variety of applications (parsing, generation, lexicography), which can each select from

the master lexicon exactly those types of information that are relevant for them. On the other hand, an obvious drawback of the present "flat" XML format is a relatively high degree of redundancy. The good reasons for introducing inheritance-based representation formalisms in "standard" computational lexicons of content words largely apply to the realm of connectives (and possibly to other function words) as well. For the time being, however, the more mundane task of lexical description still offers a great many open questions for individual connectives and families thereof; the issue of more intelligent storage should become prominent later, when the groundwork has stabilized.

As with the vast majority of coherence relations, causal ones often need not be explicitly signaled at the linguistic surface by a connective. Thus the approach proposed in this paper will of course only partially solve the problem of local coherence analysis. An important challenge for future work is to identify linguistic features of discourse units *other than* connectives that can also serve to at least constrain the range of admissible coherence relations (see, e.g. Asher and Lascarides, 2003). Investigating these with empirical methods is an important next step in the overall program of partially deriving coherence relations in authentic text *without* resorting to non-linguistic knowledge.

Acknowledgments

The following people from the Potsdam Applied Computational Linguistics Group contributed to the work described in this paper (in alphabetical order): André Herzog (causality corpus); Kristin Irsig (DiMLex entries for causal connectives); Andreas Peldszus (causality corpus and annotation guidelines); Uwe Küssner (implementation of LCA module).

References

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Brants, S., S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit (2004). Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation 2*(4), 597–620.

Chiarcos, C., S. Dipper, M. Götze, J. Ritz, and M. Stede (2008). A flexible framework for integrating annotations from different tools and tagsets. In *Proc. of the First International Conference on Global Interoperability for Language Resources*, Hongkong.

Corston-Oliver, S. (1998). *Computing of Representations of the Structure of Written Discourse*. Ph. D. thesis, University of California at Santa Barbara.

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In R. Eckstein and R. Tolksdorf (Eds.), *Proceedings of Berliner XML Tage*, pp. 39–50.

Dipper, S. and M. Stede (2006). Disambiguating potential connectives. In M. Butt (Ed.), *Proceedings of KONVENS '06*, Konstanz, pp. 167–173.

Frohning, D. (2007). *Kausalmarker zwischen Pragmatik und Kognition. Korpusbasierte Analysen zur Variation im Deutschen.* Tübingen: Niemeyer. (Im Erscheinen).

Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, Las Cruces/NM.

Hirschberg, J. and D. J. Litman (1994). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics 19*(3), 501–530.

Jasinskaja, K., J. Mayer, J. Boethke, A. Neumann, A. Peldszus, and K. J. Rodríguez (2007). Discourse tagging guidelines for German radio news and newspaper commentaries. Ms., Universität Potsdam.

Lüngen, H., H. Lobin, M. Bärenfänger, M. Hilbert, and C. Puskas (2006). Text parsing of a complex genre. In B. Martens and M. Dobreva (Eds.), *Proc. of the Conference on Electronic Publishing (ELPUB 2006)*, Bansko, Bulgaria.

Lüngen, H., C. Puskas, M. Bärenfänger, M. Hilbert, and H. Lobin (2006). Discourse segmentation of German written text. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Phikkala (Eds.), *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*, Berlin/Heidelberg/New York. Springer.

Mann, W. and S. Thompson (1988). Rhetorical structure theory: Towards a functional theory of text organization. *TEXT 8*, 243–281.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge/MA: MIT Press.

Pasch, R. (1989). Adverbialsätze – kommentarsätze – adjungierte sätze. eine hypothese zu den typen der bedeutungen von 'weil', 'da' und 'denn'. In W. Motsch (Ed.), *Wortstruktur und Satzstruktur*, Linguistische Studien des ZISW: Reihe A – Arbeitsberichte 194, pp. 141–158. Berlin: Akademie der Wissenschaften der DDR.

Pasch, R., U. Brauße, E. Breindl, and U. H. Waßner (2003). *Handbuch der deutschen Konnektoren*. Berlin/New York: Walter de Gruyter.

Polanyi, L., C. Culy, M. van den Berg, G. L. Thione, and D. Ahn (2004). A rule based approach to discourse parsing. In *Proceedings of the SIGDIAL '04 Workshop*, Cambridge/MA. Assoc. for Computational Linguistics.

Reitter, D. and M. Stede (2003). Step by step: underspecified markup in incremental rhetorical analysis. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC)*, Budapest.

Sanders, T., W. Spooren, and L. Noordman (1992). Toward a taxonomy of coherence relations. *Discourse Processes 15*, 1–35.

Stede, M. (2002). DiMLex: A lexical approach to discourse markers. In *Exploring the Lexicon - Theory and Computation*. Alessandria: Edizioni dell'Orso.

Stede, M. (2008). RST revisited: Disentangling nuclearity. In C. Fabricius-Hansen and W. Ramm (Eds.), *'Subordination' versus 'coordination' in sentence and text*. Amsterdam: John Benjamins.

Stede, M. and K. Irsig (2008). Identifying complex connectives: Complications for local coherence analysis. In A. Benz, P. Kühnlein, and M. Stede (Eds.), *Proceedings of the Workshop on Constraints in Discourse*, Potsdam, pp. 77–84.

Sumita, K., K. Ono, T. Chino, T. Ukita, and S. Amano (1992). A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pp. 1133–1140.

Sweetser, E. (1990). *From etymology to pragmatics*. Cambridge: Cambridge University Press.

Webber, B., M. Stone, A. Joshi, and A. Knott (2003). Anaphora and discourse structure. *Computational Linguistics 29*(4), 545–587.

Wolf, F. and E. Gibson (2005). Representing discourse coherence: a corpus-based study. *Computational Linguistics 31*(2), 249–287.