Chapter 1

DiMLex: A lexical approach to discourse markers

1.1 Introduction

1.1.1 Discourse markers

The notion of *discourse relation* is uncontroversially taken as an important descriptive tool for characterizing coherent text: Relations provide the "glue" between adjacent text spans. Finding a precise *inventory* of relations, on the other hand, is by no means uncontroversial; proposals range from very few general relations (e.g., ADDITIVE, ADVERSARY, CAUSAL, CONDITIONAL) to sets of more than twenty quite specific relations such as VOLITIONALCAUSE, BACKGROUND, ANTITHESIS (e.g., [Mann, Thompson 1988]).

Building upon the idea of discourse relations, we here use the term *discourse marker* for those lexical items that can signal the presence of a relation at the linguistic surface. As such, discourse markers are semantically 2-place predicates whose arguments can be entire propositions.[1]

A genre that is particularly rich in discourse markers is *instructional* text. Figure 1.1 gives an excerpt from an automobile manual, where the discourse markers have been underlined. Notice that removing the markers — and the punctuation[2] — from the text leaves essentially a sequence of individual proposi-

[1] The term 'discourse marker' is sometimes used to characterize items occuring in spoken language to signal hesitations, topic shifts, etc.: *well, I mean, mmh, yeah* and more. Such items, which we prefer to call 'discourse particles', are not considered in this paper.

[2] This qualification points to the interesting relationship between discourse markers and punctuation marks, which can serve to signal various kinds of relations as well. In this paper,

Wait until the engine is cool, then turn the radiator cap clockwise until it stops. DO NOT PRESS DOWN WHILE TURNING THE CAP. After any remaining pressure has been relieved, remove the cap by pressing down and again turning it counterclockwise. Add enough coolant to fill the radiator, and reinstall the cap. Be sure to tighten it securely. Fill the reserve tank up to the max mark with the engine cold.

Figure 1.1: Instructional text from automobile manual

tions, expressed in various syntactic forms; exactly which forms are used is then dependent on the syntactic features of the marker. In the text, there are adverbs, conjunctions, and prepositions identified as discourse markers; accordingly, the words of interest to us here do not constitute a syntactically homogeneous class. A term possibly covering them quite well is *connective*, as defined by Crystal [1985] as lexical items that "link linguistic units at any level."

What makes the issue of discourse markers interesting is the fact that an individual discourse relation is typically associated with a wide range of such markers; consider, for instance, the following variety of CONCESSIONS, which all express the same underlying propositional content. Curly brackets indicate a choice between alternatives, and again, the markers are underlined.

- *We were in SoHo;* {*nevertheless* | *nonetheless* | *however* | *still* | *yet*}, *we found a cheap bar.*
- *We were in SoHo, but we found a cheap bar anyway.*
- {*Despite* | *Notwithstanding*} *the fact that we were in SoHo, we found a cheap bar.*
- {*Although* | *Even though*} *we were in SoHo, we found a cheap bar.*

If one accepts these sentences as paraphrases, then the ten alternative discourse markers used (counting *but...anyway* as a single marker) all need to be associated with the information that they signal a concessive relationship between the two propositions involved. This example, too, demonstrates the syntactic variety of markers (of the very same relation). On the other hand, even though the sentences denote the same facts and thus can be labelled as paraphrases, they are clearly not equally felicitous in different contexts. Therefore, the next step is to discern the fine-grained differences between these sentences and thus between the markers; this is a central part of our work and will be discussed in Section 1.3.

though, we leave punctuation aside.

1.1.2 Motivation for a lexicon of discourse markers

In computational linguistics, discourse markers are usually treated as closed-class items and thus not much attention is given to them from the lexical point of view. Resources such as WordNet [Fellbaum 1998] deal only with 'content words' and thus yield no information on discourse markers. Probably, this lack of interest is due to the fact that there is little theoretical research to draw upon: While a number of studies of individual relations and their markers exist, as well as a wealth of linguistic analyses of individual markers (see Section 1.2), there is no broader account that would

- provide general descriptions of discourse markers and their role in language,
- account for the precise differences between similar markers,
- explain the relationships between markers and their linguistic environment (phrase, clause, sentence, text).

Ultimately, the challenge is to explain the link between sentence grammar and discourse phenomena, of which discourse markers are but one aspect.

Even though discourse markers certainly are closed-class items, they nonetheless display interesting lexical properties. For instance, the most important lexical relations, which are used in lexicology to structure the realm of the vocabulary, can be observed not only for the 'content words' but for discourse markers as well:

- Synonymy: The German words *obzwar* and *obschon* (both more formal variants of *obwohl* = *although*) can be regarded as synonymous.
- Plesionymy (near-synonymy): *although* and *though*, according to Martin [1992], differ in formality; *although* and *even though* differ in terms of emphasis.
- Antonomy: *if/unless*, according to Barker [1994], have opposite polarity, as in *He will not attend unless he finishes his paper* vs. *He will attend if he finishes his paper.*
- Hyponomy: Some markers are more specific than others; *but* can signal a general CONTRAST or a more specific CONCESSION.
- Polysemy: Other than being more or less specific, some markers can signal quite different relations; e.g., *while* can be used for TEMPORAL CO-OCCURRENCE, and also for CONTRAST.

Exploiting such relationships, our MARKER project at TU Berlin is developing a DIscourse Marker LEXicon (DiMLex), which aims at producing lexical

entries for German and English discourse markers on a uniform level of representation, so that it can be employed as a declarative resource for various NLP applications (at present, our focus is on text generation). Our approach is to represent the commonalities and differences between discourse markers by sets of features. Markers expressing the same relation (such as the ten English CONCESSION markers listed above) can then be distinguished by assigning different values to features characterizing their syntactic, semantic, and pragmatic behaviour. In effect, we transfer the methodology of lexical-field theory ('Wortfeldtheorie', [Trier 1931]), as it has been employed successfully for content words, to the realm of function words, in particular to discourse markers.

From the perspective of *text generation*, the task is to choose the most appropriate verbalization of a discourse relation in a given context, for which a generator needs knowledge about the fine-grained differences between similar markers of the same relation. It furthermore needs to account for the interactions between marker choice and other generation decisions and hence needs knowledge about the syntagmatic constraints associated with different markers. We have implemented a first version of DiMLex and incorporated it into a text generator; this approach will be presented in Section 1.4.

From the perspective of *text understanding*, a sophisticated system should be able to derive the discourse relations holding between adjacent text spans, and also to notice the additional semantic and pragmatic implications stemming from the usage of a particular discourse marker. We will briefly characterize such applications in Section 1.5.

1.2 Related research

While traditional research in natural language understanding has paid relatively little attention to discourse markers, they have been a topic of active research in language generation. Elhadad and McKeown [1990], for instance, investigated the influence of *argumentative intent* on marker choice and from this angle looked at *but* and at the difference between *since* and *because*. Vander Linden and Martin [1995] dealt with the discourse relation PURPOSE and proposed a framework for choosing among candidate English verbalizations, along with the grammatical consequences. Dorr and Gaasterland [1995] worked on English temporal markers and their interaction with tense and aspect of the clause; Grote [1998] similarly analyzed German temporal markers. Stede [1995] looked at English and German SUBSTITUTION markers and the translation problems they pose; Grote *et al.* [1997] proposed an analysis of the CONCESSION relation and its possible realizations in English and German.

Such analyses of individual relations are an important resource to draw from, but these studies address the issues from different perspectives and assume different grammatical frameworks; the same holds for the analyses of individual markers in the Linguistics literature (which are far too numerous to list here),

which tend to concentrate on particular features rather than on broader description.

As for broader surveys, Knott and Mellish [1996] provided an overview of English markers and proposed a feature-based classification of coherence relations on the basis of substitution tests; this instrument is also central to our work (see Section 1.3), even though the methodology requires some modifications when dealing with German, where clauses and discourse markers cannot be separated as nicely as in English. Halliday and Hasan [1976] undertook steps towards a general typology of clause linkage. Elaborating their approach, Martin [1992] provided extensive systemic networks that differentiate between many English markers. For German, an extensive study of connectives is currently underway [Pasch et al., in prep.].

Recently, discourse markers have attracted attention from the perspective of large-scale text understanding and summarization. Marcu [1997] proposed an algorithm for rhetorical parsing of unrestricted text on the basis of markers and of knowledge on typical co-occurrences of discourse relations. Corston-Oliver [1998] followed a similar goal but also used some shallow linguistic analysis to infer discourse structure from the presence of markers in conjunction with certain surface-linguistic features. Clearly, this line of research is very relevant to DiMLex, even though our initial efforts concentrate on generation.

1.3 Developing the lexicon

When pursuing the goal of building a discourse marker lexicon, methodological considerations pertain first to the task of assembling the set of markers, and subsequently to that of finding a set of features to characterize and differentiate markers. In this section, we will consider these steps in turn and then describe an initial set of features that we have obtained so far.

1.3.1 Assembling sets of markers

As indicated in Section 1, the class of discourse markers is a semantic one and cannot be defined by clear syntactic criteria. Accordingly, it is not a straightforward task to assemble the set of markers that should be included in DiMLex. The standard linguistic resources always employ syntactic criteria for their classifications; for instance, Quirk et al. [1972] (for English) and Helbig and Buscha [1991] (for German) offer a classification of *conjunctions* into semantic groups, which for our purposes is only a first step.

A very helpful resource for German is the 'Handbook of German Connectives' [Pasch et al., in prep.]. It establishes quite clear — and mostly syntactic — criteria for the decision whether a lexical item x is a connective:

- x is not inflectable,

- x does not assign case to its syntactic environment,
- x expresses some specific two-place semantic relation,
- the arguments of the relational meaning of x are propositional structures,
- the verbalizations of the arguments of the relational meaning of x can be clauses.

These criteria define a class that is clearly broader than conjunctions — it includes adverbs and conjunctives, but deliberately excludes prepositions (as these assign case to their NP). This decision allows for drawing a relatively clear boundary around the class of connectives, but again leaves out items that for our purposes should be included.

The difficulty of demarcating the class of discourse markers is also demonstrated by the relatively complicated 'test for relational phrases' proposed by Knott [1996]. Instead of seeking syntactic features, Knott suggests a corpus-based identification procedure. Somewhat condensed, it works as follows:

1. Isolate candidate phrase and its 'host clause'.
2. In the host clause, substitute anaphors with their antecedents.
3. If the candidate phrase is a relational phrase, the resulting text is incomplete in that it needs one or more extra clauses to form a coherent message.
4. Exclude phrases that refer directly to the text itself, such as *in the next section*.
5. Exclude phrases that include comparatives.
6. If more than one cue phrase is contained in the clause, they should each pass the test individually.

This test also rules out prepositions, as only clauses are considered as candidates for text spans related by markers. Thus, in *Despite the heavy rain, we went for a walk* the test would not detect a marker, and the motivation is, again, to keep the test reasonably simple. Knott's test furthermore is geared towards the English language and would cause problems when applied to some German subordinators.

From the DiMLex perspective, prepositions such as the German *trotz, wegen* or the English *despite, due to* are important means for expressing relationships and thus are to be admitted into the lexicon. At this point, we therefore take the criteria given by Pasch *et al.* as a starting point and add to the resulting set those prepositions that paraphrase tests suggest as equivalent means for expressing a relation. With this assumption we are at present focusing our

attention on adversative markers (for the relations CONTRAST and CONCESSION), for which we are developing lexicon entries; on the basis of these results we will later broaden the view to other relations and markers and re-examine the identification criteria if necessary.

1.3.2 Determining lexical features

Methodology In order to determine a set of lexical features for discourse markers, we follow two different methods. On the one hand we exploit standard dictionaries and grammars, and work with the research literature (as sketched in Section 1.2; on the other hand, we derive additional information from *substitution tests* on corpus data and on specifically constructed examples.

As for the literature, grammars such as Quirk et al. [1972] or Helbig and Buscha [1991] give some syntactic descriptions of markers but offer only very little information on their precise meaning and usage conditions. Standard dictionaries provide a start, but do not help in differentiating similar markers. For the English language, a useful resource is the Longman Activator [1993], a dictionary specifically designed from the language production perspective. As it tries to help with choosing the 'right' word in a particular situation (similar to thesauri), its goal is to shed light on the subtle differences between similar words. For example, under the heading 'but' we find 6 groups of different kinds of adversative situations, associated with different (if overlapping) markers, which are characterized by brief descriptions and typical examples. Even though these cannot be immediately translated into a set of features, they offer good insights and orientation.

The research literature on individual markers or groups of markers, as indicated in Section 1.2, offers a wealth of individual studies, from which we are trying to synthesize a uniform level of description. The overall goal can be characterized as the aim to synthesize two strands of research that so far are rather disconnected:

- "Top-down": Text linguistics considers markers as a means to signal coherence, and provides us with insights on the semantic and pragmatic properties of marker classes.
- "Bottom-up": Grammars as well as the linguistic research literature provide syntactic, semantic and stylistic properties of individual markers, comparative studies of related markers, etc.

Aside from literature studies, we try to elicit the nature of semantic and pragmatic differences between similar markers by means of substitution tests. Taking the example of adversative markers, we gather from online newspaper corpora occurrences of the markers we are interested in and try to replace them with the others. To accommodate for differences in the markers' syntactic classes, this might involve some reordering of words; subjects are then asked

Die Diskussion brachte die Selbstkritik der Mitgliedschaft zum Ausdruck und gab eine ganze Reihe von Anregungen fr die kommende Arbeit. Es muß ALLERDINGS gesagt werden, daß einige Genossen nicht konkret genug diskutierten.
[+*aber*, +*ALLERDINGS*, +*dennoch*, +*doch*, +*hingegen*, +*jedoch*]
('The discussion ...resulted in many suggestions for the upcoming work. It must be said, though, that several comrades did not discuss concretely enough.')

Mindestens 40000 Kriegsgefangene, wahrscheinlich ABER bedeutend mehr, seien für die französische Fremdenlegion verpflichtet worden.
[+*ABER*, +*allerdings*, −*dennoch*, +*doch*, −*hingegen*, +*jedoch*]
('At least 40000 prisoners of war, but probably many more, were recruited for the French foreign legion.')

Figure 1.2: Example from substitution test for German contrastive markers

to judge whether sentences assume a different meaning when the marker is replaced. (See also [Knott 1996] for a description of a substitution test, which he used to infer a set of discourse relations.) For illustration, Figure 1.2 gives an example from our test set for differentiating the German markers *aber, allerdings, dennoch, doch, hingegen, jedoch*, using data from the 1946 edition of the newspaper *Neues Deutschland*. The original marker appears in upper case letters; in the first case, any of the other markers can be used without a noticable shift in meaning, whereas in the second case, neither *dennoch* nor *hingegen* can be substituted. *Dennoch* has a concessive flavour, which is not compatible with the context, and *hingegen* is similar to *on the other hand*, which is also inappropriate here. Extensive tests of this kind lead to classes of contexts that allow for various subsets of adversative markers to be used, and from these sets of contexts, we can then deduce features present in them that license the usage of some markers but not of others. In this way, differentiating features can be gathered.

The corpus substitution test has the advantage of leading us to contexts that we otherwise had not thought of, and that display telling differences in substitutability. On the other hand, often it is useful to compare markers with respect to some specific, pre-conceived features. Thus we use substitution tests on hand-crafted sample utterances to test markers on their behaviour within intensional contexts, with specific tense/aspect combinations, etc.

With these methods, we are lead to determine two classes of features: those that can be generally used in the descriptions of all markers; and those that characterize fine-grained semantic/pragmatic differences between similar mark-

ers, and which are applicable only to one particular lexical field. Although our classification of lexical features is still under development, we give here a tentative list of such features in order to illustrate the range of phenomena under consideration (cf. [Stede, Umbach 1998]).

Syntax The *part of speech* of a marker (conjunctive, subordinating conjunction, coordinating conjunction, preposition) determines the possibilities of *positioning* the marker within the constituent: conjunctives (especially the German 'Konjunktionaladverbien') can float to various positions, whereas the positions of others are fixed. The *linear order* of the conjuncts is fixed for some markers and flexible for others; this is independent of the aforementioned two features. Some markers show a specific behavior towards *negation*, e.g., the German *sondern* (which corresponds to certain uses of *but*) requires an explicit negation in the antecedent clause. Some markers impose constraints on *tense and aspect* of the clauses, either by requiring specific temporal/aspectual attributes in one clause, or by constraining the relationship between the two conjuncts (e.g., *after*).

Semantics and Pragmatics Several grammars suggest classifications of markers according to the *semantic relation* they express: adversative, alternative, substitution, causal, conditional, etc. Within these groups, some markers exhibit opposite *polarity*, i.e., have an incorporated negation or not (e.g., *if* versus *unless*). *Commentability* is a feature that sometimes distinguishes a single marker within a semantic class in that it can be negated or focused on by scalar particles (e.g., in German, the causal *weil* is commentable, whereas *denn* is not).

Moving towards pragmatics, the *intention* behind using a marker can vary. A well-known example is the contrast between German *aber* and *sondern* (in English, they both correspond to *but*), where the former merely states a contrast, whereas the latter corrects an assumption on the hearer's side (e.g., [Helbig, Buscha 1991]). Another dimension concerns the *presuppositions* associated with markers; a well-known case is the contrast between *because* and *since*, where only the latter marks the subsequent proposition as *given*. The German CAUSE markers *weil* and *denn* differ in terms of the *illocutions* they connect: The former applies to propositions, the latter to epistemic judgements [Pasch et al., in prep.]. Certain very similar markers differ only *stylistically*. The German example of *obwohl / obzwar / obschon* was given above, and another one is the English *notwithstanding*, which is more formal than *despite* and moreover is more flexible in positioning, as it can be postponed.

The final but crucial feature to be mentioned here is the *discourse relation* expressed by a marker. Rhetorical Structure Theory (RST) [Mann, Thompson 1988] offers an inspiring theory of such relations, but we do not fully subscribe to this account. Rather, we think that the relationship between semantic relations (see above) and pragmatic ones needs to be clarified (e.g., [Asher 1993]),

which can be done by teasing apart the various dimensions incoporated in RST's definitions, for example in the spirit of Sanders et al. [1992].

Sample analyses of English PRECONDITION markers along these lines can be found in [Grote, Stede 1998].

1.4 DiMLex in text generation

In the research field of text generation, a consensus has emerged that the overall task is best split into three distinct phases:

1. Text planning
2. Sentence planning
3. Surface realization

The output of the text planning phase is commonly a tree structure with propositions at the leafs and discourse relations at the internal nodes. The subsequent generation phase linearizes this tree into a sequence of sentence plans, which are then given to the realization module. Sentence planning is thus in charge of determining clause boundaries and clause structure, and of lexicalization; these tasks obviously interact with choosing discourse markers. Accordingly, DiMLex is to be used as one resource in sentence planning, where markers are chosen to signal the discourse relations present in the tree.

Obviously, marker selection also includes the decision whether to use any marker at all or leave the relation implicit (e.g., [Di Eugenio et al. 1997]). When these decisions can be systematically controlled, the text can be tailored much better to the specific goals of the generation process. Our approach is to perform marker choice in tandem with other sentence planning decisions, and to employ DiMLex as a declarative resource for this task.

To demonstrate the utility of this idea, we have implemented a first version of DiMLex and integrated it into a text generation environment. The nature of the generation task imposes a particular *view* of the information coded in DiMLex: The entry point to the lexicon is the discourse relation to be realized, and the lookup yields the range of alternatives. But many markers have more semantic and pragmatic constraints associated with them, which have to be verified in the generator's input representation for the marker to be a candidate. Then, discourse markers place (predominantly syntactic) constraints on their immediate context, which affects the interactions between marker choice and other realization decisions. And finally, markers that are still equivalent after evaluating these constraints are subject to a choice process that can utilize preferential (e.g. stylistic) criteria; for example, when a goal is to produce concise text, a nominalization with *despite* will be preferred over a more lengthy *although* construction. Therefore, under the generation view, the information in DiMLex is grouped into the following three classes (cf. [Grote, Stede 1998]):

- *Applicability conditions:* The necessary conditions for using a discourse marker, i.e., the features or structural configurations that need to be present in the input specification. Chiefly, this is the semantic/discourse relation to be expressed, and also (if applicable) features pertaining to presuppositions and intentions.

- *Combinability constraints:* The constraints regarding the combination of a marker and the neighbouring constituents; most of them are syntactic (part of speech, linear order, etc.).

- *Preferential features:* Features that label the differences between similar markers sharing the same applicability conditions, such as stylistic features or degrees of emphasis.

Given a tree structure produced by text planning, the first task of the sentence planner is to find *verbalization options*: lexemes for expressing the propositions and discourse markers for realizing the discourse relations. To what extent fine-grained lexical-semantic features are used here depends obviously on the granularity of the input representation and hence on the underlying application. Going back to our CONCESSION example in Section 1.1.1, suppose that one part of the input tree were

```
(CONCESSION (FIND(WE,CHEAP-BAR))
            (LOCATED(WE,SOHO)))
```

The first argument of the CONCESSION is the nucleus[3] of the relation and the second its satellite; the propositions are written here in a simplified notation. In the matching step, the 10 markers given in Section 1.1.1 are found as candidates for expressing this relation node. The second step in sentence planning evaluates the syntagmatic constraints associated with the verbalization options across the tree and tries to find one option per node such that the complete tree is covered by the collective options, and all syntagmatic constraints are respected. In cases where choices among options remain, the paradigmatic features are considered. If this results in a preferred option, it is chosen; otherwise, a random selection is made. In our example, the decision would for instance be constrained by considerations of thematic development. If the part expressing our sub-tree were at the very beginning of the text, the information about being in SoHo would preferably be expressed as a main clause opening the text; otherwise, if SoHo were already *given*, a subordinate clause would be appropriate. Once the structural decisions are made, preferential parameters can give rise to choosing between *although* and *even though* on the grounds of desired emphasis, or between *despite* and *notwithstanding* on the grounds of formality.

[3] In the terminology of RST, most discourse relations link a nucleus and a satellite element; the former is the more central one, whereas the latter can often be removed without loosing the main line of argumentation.

In this way, the prototype of our generator can produce a range of English and German paraphrases for input involving a discourse relation (for an example with detailled representations, see [Grote, Stede 1998]); the additional step of embedding this functionality in a more comprehensive model of linearization, in order to produce paragraph-size text, is under way.

1.5 DiMLex in text understanding

On the side of natural language understanding, the notion of *rhetorical parsing* has recently gained popularity. Traditional text understanding systems, when faced with essentially unrestricted text, will inevitably have to cope with gaps in both its dictionary and grammar. In response, one way of achieving robustness is to only perform a shallow analysis that largely relies on closed-class lexemes and on punctuation; discourse markers can in this way be exploited to infer the most likely rhetorical structure of a text (which is typically conceived in the way of RST [Mann, Thompson 1988]).

In general, though, the depth of analysis need not stop at the "shallow" surface level: Inferring discourse structure trees can be regarded as one aspect of "deep" text understanding, and it can for instance provide a valuable basis for summarization algorithms that perform more analysis work than merely extracting sentences from text based on statistical measures (as almost all contemporary programs do). The decision as to what portions of a text are more relevant than others is greatly facilitated if the rhetorical structure, and in particular the assignments of nuclearity (in RST terms) are known.

Approaches differ in what kind of information is used for rhetorical parsing. Marcu [1997] largely relies on the presence of lexical discourse markers and on knowledge about typical nuclearity configurations in RST trees. Corston-Oliver [1998] in addition considers some surface-linguistic features extracted by a robust parser. Rehm [in prep.] treats discourse markers as one source of information, besides paragraph structure and layout, for automatically analyzing texts into SGML format, where discourse relations are identified as far as possible.

For tasks of this kind, DiMLex can supply the set of cue words to be looked for and support the initial disambiguation of cues in the text. Depending on the depth of the syntactic and semantic analysis carried out by the text understanding system, different features provided by DiMLex can be taken into account. Certain structural or surface-syntactic properties can be tested without any deep understanding; for instance, the German marker *während* is generally ambiguous between a CONTRAST and a TEMPORALCOOCCURRENCE reading, but when followed by a noun phrase, only the latter reading is available (*während* corresponds not only to the English *while* but also to *during*). In practice, a rhetorical parser would first identify discourse markers, and then consult DiMLex to determine which relation(s) can be expressed by the marker. In case

more than one relation is possible, it would check whether the lexicon entries offer some discriminating features for disambiguation. Then the system would analyze the relevant text span further and try to verify the presence or absence of the feature(s), so that the right relation can be associated with the marker in question.

1.6 Summary

Discourse markers are words that signal the kind of relationship holding between adjacent spans of text and thereby are an important instrument for achieving cohesion. Yet they are not very well understood, and in natural language processing they are just beginning to attract greater attention. An important difficulty with discourse markers is that they do not constitute a syntactically homogeneous group; rather, we find conjunctions, different kinds of adverbs, and prepositions. Conceiving markers (like those for CONCESSION given in Section 1.1.1) as a uniform group requires stepping beyond purely structural classifications, taking a more functional perspective, investigating more deeply the structure of text and the underlying notions of coherence and cohesion, and explaining the link between textual phenomena and the grammar of sentences.

While it is possible to establish sets of discourse relations, and to group discourse markers into broad semantic categories (such as CONTRAST or CONDITION), the more fine-grained differences between markers of the same group invite closer scrutiny. Furthermore, many markers are ambiguous (e.g. *while*), and there appear to be hyponymy relations between some markers (or relations), as between CONTRAST and CONCESSION.

We see this situation as calling for a dedicated lexicon of discourse markers: one that cuts across syntactic categories and assembles information that allows for distinguishing broad semantic-pragmatic classes as well as pinning down more fine-grained distinctions between similar markers. We therefore assemble information on markers from a variety of sources and conflate it into DiMLex, the DIscourse Marker LEXicon. Our target languages are German and English, and in the first phase of the project we are focusing our attention on CONTRAST and CONCESSION markers. In this paper, we have described our methodology; listed a preliminary inventory of features for characterizing markers; outlined our initial implementation of a partial DiMLex used by a text generator; and sketched how the lexicon can likewise support rhetorical parsing, for instance for the purpose of summarization. In the upcoming second phase of the project, our efforts will focus on extending the coverage of the lexicon to other discourse relations, and on devising suitable (inheritance-based) lexical representations that support the usage of DiMLex in automatic text generation and understanding.

References

Language Activator. The World's First Production Dictionary. 1993. Burnt Mill: Longman.

N. Asher. 1993. *Reference to abstract objects in Discourse.* Dordrecht: Kluwer.

K. Barker. 1994. Clause-level relationship analysis in the TANKA system. Technical report, Dept. of Computer Science, University of Ottawa, TR-94-07.

S. Corston-Oliver. 1998. Identifying the linguistic correlates of rhetorical relations. In: Proceedings of the Coling-ACL '98 Workshop on Discourse Relations and Discourse Markers, Montréal, August 1998.

D. Crystal. 1985. *A dictionary of linguistics and phonetics.* Oxford: Blackwell.

B. Di Eugenio, J. Moore, M. Paolucci. 1997. Learning features that predict cue usage. In: Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the ACL, Madrid, July 1997.

B. Dorr, T. Gaasterland. 1995. Selecting Tense, Aspect and Connecting Words in Language Generation. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1299–1305.

M. Elhadad, K. McKeown. 1990. Generating Connectives. In Proceedings of the 13th Conference on Computational Linguistics, Helsinki, 97–101.

C. Fellbaum (ed.). 1998. *WordNet – An electronic lexical database.* Cambridge/MA: MIT Press.

B. Grote. 1998. Representing temporal discourse markers for generation purposes. In: Proceedings of the COLING-ACL 98 Workshop on Discourse Relations and Discourse Markers, Montréal.

B. Grote, M. Stede. 1998. Discourse marker choice in sentence planning. In: Proceedings of the Ninth International Workshop on Natural Language Generation, Niagara-on-the-lake, Canada, August 1998.

B. Grote, N. Lenke, M. Stede. 1997. Ma(r)king concessions in English and German. *Discourse Processes* 24(1).

M. Halliday, R. Hasan. 1976. *Cohesion in English.* London/New York: Longman.

G. Helbig, J. Buscha. 1990. *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht.* Berlin, Leipzig: Langenscheidt, Verlag Enzyklopädie.

A. Knott. 1996. A data-driven methodology for motivating a set of coherence relations. Doctoral dissertation, University of Edinburgh.

A. Knott, C. Mellish. 1996. A feature-based account of the relations signalled by sentence and clause connectives. *Language and Speech* 39:143-183.

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

D. Marcu. 1997. The rhetorical parsing of natural language text. In: Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the ACL, Madrid, July 1997.

J. Martin. 1992. *English Text: System and Structure.* Philadelphia/Amsterdam: John Benjamins.

R. Pasch, U. Brausse, E. Breindl. In preparation. Handbuch der deutschen Konnektoren. (Preliminary version.) Mannheim.

R. Quirk, S. Greenbaum, G. Leech, J. Svartvik. 1992. *A Grammar of Contemporary English.* Harlow: Longman (20th ed.).

G. Rehm. In preparation. Vorüberlegungen zur automatischen Zusammenfassung deutschsprachiger Texte mittels einer SGML-und DSSSL-basierten Repräsentation von RST-Relationen. Master's thesis, University of Osnabrück.

T. Sanders, W. Spooren, L. Nordman. 1992. Towards a taxonomy of coherence relations. In: *Discourse Processes* 15.

M. Stede. 1995. Kontrastive Untersuchung einiger kontrastiver Diskurspartikel. *Kognitionswissenschaft* 5(3):127–140

M. Stede, C. Umbach 1998. DiMLex: A lexicon of discourse markers for text generation and understanding. In: Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (Coling–ACL 98), Montréal.

J. Trier. 1931. *Der deutsche Wortschatz im Sinnbezirk des Verstandes.* Heidelberg.

K. Vander Linden, J. Martin. 1995. Expressing rhetorical relations in instructional text: a case study of the purpose relation. *Computational Linguistics* 21(1):29–58.

L. Wanner, E. Hovy. 1996. The HealthDoc sentence planner. In: Proceedings of the Eighth International Workshop on Natural Language Generation, Herstmonceux Castle/UK.