DiMLex: A lexicon of discourse markers for text generation and understanding

Manfred Stede and Carla Umbach*
Technische Universität Berlin
Projektgruppe KIT
Sekr. FR 6-10
Franklinstr. 28/29
D-10587 Berlin, Germany
email: {stede|umbach}@cs.tu-berlin.de

Abstract

Discourse markers ('cue words') are lexical items that signal the kind of coherence relation holding between adjacent text spans; for example, *because*, *since*, and *for this reason* are different markers for causal relations. Discourse markers are a syntactically quite heterogeneous group of words, many of which are traditionally treated as function words belonging to the realm of grammar rather than to the lexicon. But for a single discourse relation there is often a set of similar markers, allowing for a range of paraphrases for expressing the relation. To capture the similarities and differences between these, and to represent them adequately, we are developing DiMLex, a lexicon of discourse markers. After describing our methodology and the kind of information to be represented in DiMLex, we briefly discuss its potential applications in both text generation and understanding.

1 Introduction

Assuming that text can be formally described (and represented) by means of *discourse relations* holding between adjacent portions of text (e.g., [Mann, Thompson 1988]), we use the term *discourse marker* for those lexical items that (in addition to non-lexical means such as punctuation, aspectual and focus shifts, etc.) can signal the presence of a relation at the linguistic surface. Typically, a discourse relation is associated with a wide range of such markers; consider, for instance, the following variety of CONCESSIONS, which all express the same underlying propositional content. The words treated here as discourse 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 the fact that we were in SoHo, we found a cheap bar.

Notwithstanding the fact that we were in SoHo, we found a cheap bar.

Although we were in SoHo, we found a cheap bar.

If one accepts these sentences as paraphrases, then the various discourse markers all need to be associated with the information that they signal a concessive relationship between the two propositions involved. Next, the fine-grained differences between similar markers need to be represented; one such difference is the degree of specificity: for example, *but* can mark a general CONTRAST or a more specific CONCESSION. We believe that a dedicated discourse marker lexicon holding this kind of information can serve as a valuable resource for natural language processing. Our efforts in constructing that lexicon are described in Section 2.

From the perspective of text generation, not all paraphrases listed above are equally felicitous in specific contexts. In order to choose the most appropriate variant, a generator needs knowledge about the fine-grained differences between similar markers for the same relation. Furthermore, it 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 will discuss this perspective in Section 3.

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

* This paper appears in the Proceedings of the COLING-ACL '98 Conference, Montréal, August 1998

from the usage of a particular discourse marker. We will briefly characterize such applications in Section 4.

2 Building a Discourse Marker Lexicon

2.1 The idea

The traditional distinction between content words and function words (or open-class and closed-class items) relies on the stipulation that the former have their "own" meaning independent of the context in which they are used, whereas the latter assume meaning only in context. Then, content words are assigned to the realm of the lexicon, whereas function words are treated as a part of grammar.

For dealing with discourse markers, we do not regard this distinction as particularly helpful, though. As we have illustrated above and will elaborate below, these words can carry a wide variety of semantic and pragmatic overtones, which render the choice of a marker meaning-driven, as opposed to a mere consequence of structural decisions. Furthermore, a number of lexical relations that are customary used to assign structure to the universe of "open class" lexical items, most prominently synonymy, plesionymy ("near-synonymy"), antonymy, hyponymy and polysemy, can be applied to discourse markers as well:

- Synonymy: It can be argued that true synonyms do not exist at all. However, the German words *obzwar* and *obschon* (both more formal variants of *obwohl* = *although*) certainly come very close to being synonymous.
- Plesionymy: *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; recall the example of *but* given above. Knott and Mellish [1996] deal with the issue of "taxonomizing" discourse markers.
- 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.

Accordingly, we propose that the proper place for describing discourse markers is a dedicated lexicon that provides a classification of their syntactic, semantic and pragmatic features and characterizes the relationships between similar markers. To this end, our group is developing a DIscourse Marker LEXicon (DiMLex), which aims at assembling the various information associated with markers and describing it on a uniform level of representation. Our initial focus is on German, but English will also be a target language.

2.2 Methodology

Methodological considerations pertain to the two tasks of determining the set of words we regard as discourse markers and thus are to be included in the lexicon, and determining the lexical entries for these words.

Finding the "right" set of discourse markers is not an easy task, since the common lexicographic practice of taking part of speech as the primary criterion for inclusion or exclusion does not apply. Knott and Mellish [1996] provide an apt summary of the situation. Their 'test for relational phrases' is a good start, but geared towards the English language (we are investigating German as well), and furthermore it catches only items relating clauses; in *Despite the heavy rain, we went for a walk* it would not detect a cue phrase.

To arrive at a more comprehensive set, we began by consulting standard grammars such as Quirk et al. [1972] and Helbig and Buscha [1991], which provide descriptions of function words grouped according to semantic class — but these are far from "complete". A very good source for German is [Brausse et al. in prep.], which investigates a huge set of connectives from a grammatical viewpoint.

As for determining lexical descriptions, the research literature offers a large number of helpful, even though quite heterogeneous, sources. There are several detailed studies of individual groups of markers, such as [Vander Linden, Martin 1995] for PURPOSE markers. Besides, the Linguistics literature offers fine-grained analyses of individual markers, which are far too

numerous to list. We are drawing upon all these sources, trying to place them in a single unified framework. 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.

2.3 The lexicon

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. The list is loosely ordered from syntactic to semantic and pragmatic features; for now, we do not explicitly assign such categories.

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*).

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 often 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 [Brausse et al., in prep.]. Certain very similar markers differ only *stylistically*. One German example was given above, and another one is the English *notwithstanding*, which is more formal than *despite* but 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. 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].

Once the range of dimensions has been described, we will deal with questions of representation; we envisage using some inheritance-based formalism that allows for a compact representation of individual descriptions, hyponymic relations between them, and polysemous entries.

3 Using DiMLex in text generation

Present text generation systems are typically not very good at choosing discourse markers. Even though a few systems have incorporated some more sophisticated mappings for specific relations (e.g., in DRAFTER [Paris et al. 1995]), there is still a general tendency to

treat discourse marker selection as a task to be performed as a "side effect" by the grammar, much like for other function words such as prepositions.

To improve this situation, we propose to view discourse marker selection as one subtask of the general lexical choice process, so that — to continue the example given above — one or another form of CONCESSION can be produced in the light of the specific utterance parameters and the context. 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.

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. Therefore, under the generation view, the information in DiMLex is grouped into the following three classes:

— *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.

— *Syntagmatic constraints:* The constraints regarding the combination of a marker and the neighbouring constituents; most of them are syntactic and appear at the beginning of the list given above (part of speech, linear order, etc.).

— *Paradigmatic features:* Features that label the differences between similar markers sharing the same applicability conditions, such as stylistic features and degrees of emphasis.

Very briefly, we see discourse marker choice as one aspect of the *sentence planning* task (e.g., [Wanner, Hovy 1996]). In order to account for the intricate interactions between marker choice and other generation decisions, the idea is to employ DiMLex as a declarative resource supporting the sentence planning process, which comprises determining sentence boundaries and sentence structure, linear ordering of constituents (e.g., thematizations), and lexical choice. All these decisions are heavily interdependent, and in order to produce truly adequate text, the various realization options need to be weighted against each other (in contrast to a simple, fixed sequence of making the types of decisions), which presupposes a flexible computational mechanism based on resources as declarative as possible. This generation approach is described in more detail in a separate paper [Grote, Stede 1998].

4 Using DiMLex in text understanding

In text understanding, discourse markers serve as cues for inferring the rhetorical or semantic structure of the text. In the approach proposed by Marcu [1997], for example, the presence of discourse markers is used to hypothesize individual textual units and relations holding between them. Then, the overall discourse structure tree is built using constraint satisfaction techniques. 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 configurations 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*).

Similarly, we envisage applications of DiMLex for dialogue processing. For example, within the VERBMOBIL project, Stede and Schmitz [1997] have analysed the various pragmatic functions that German discourse particles fulfill in dialogue; many of these particles are discourse markers, and DiMLex can provide

valuable information for their disambiguation, which in turn facilitates the recognition of underlying speech acts.

5 Summary and Outlook

Discourse markers, words that signal the presence of a coherence relation between adjacent text spans, play important roles in human text understanding and production. Due to their being classified as "non-content words" or "function words", however, they have not received sufficient attention in natural language processing yet. In response to this situation, we are assembling pieces of information on German and English discourse markers from grammars, dictionaries, and the linguistics research literature. This information is classified and organized into a discourse marker lexicon, DiMLex.

The first phase of our project runs until mid-1999. At present, we are on the theoretical side focusing our attention on German CONTRAST and CONCESSION markers; on the implementational side, we have assembled a generation testbed that allows for exploring the role of DiMLex in producing paragraph-size text. By the end of the first phase, we plan to have completed a system that produces German and English text, with a prototypical DiMLex specified for contrastive markers. For a potential follow-up phase of the project, we envisage enlarging DiMLex to other groups of markers; working out systematic lexical representations within a suitable formalism; and giving more attention to the requirements for text understanding in addition to those of generation.

References

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

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

U. Brausse, E. Breindl-Hiller, R. Pasch. "Handbuch der deutschen Konnektoren." Institut für deutsche Sprache, Mannheim. In preparation.

B. Di Eugenio, J. Moore, M. Paolucci. "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. Grote, M. Stede. "Discourse marker choice in sentence planning." To appear in: Proceedings of the 9th International Workshop on Natural Language Generation, Niagara-on-the-lake/Canada, 1998.

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

A. Knott, C. Mellish. "A feature-based account of the relations signalled by sentence and clause connectives." In: *Language and Speech* 39 (2-3), 1996.

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

D. Marcu. "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. *English Text – System and Structure.* Philadelphia/Amsterdam: John Benjamins, 1992.

C. Paris, K. Vander Linden, M. Fischer, A. Hartley, L. Pemberton, R. Power, D. Scott. "A support tool for writing multilingual instructions." In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, 1995.

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

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

M. Stede, B. Schmitz. "Discourse particles and routine formulas in spoken language translation." In: Proceedings of the ACL/ELSNET Workshop on Spoken Language Translation, Madrid, 1997.

K. Vander Linden, J. Martin. "Expressing rhetorical relations in instructional text" In: *Computational Linguistics* 21(1):29–58, 1995.

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