Lexical Marking and the Recovery of Discourse Structure

Kathleen Dahlgren
InQuizit Technologies, Inc.
725 Arizona Avenue, Suite 204
Santa Monica, California 90401
email – kd@inquizit.com

Introduction

In the theory presented here, discourse relations are equated with coherence relations. The relata are taken to be sets of events or entities introduced into the discourse, as in SDRT (Asher, 1993). Our empirical studies of commentary, narrative and news texts have shown that coherence relations are frequently signaled syntactically or semantically rather than lexically. In a full natural language understanding design, this much discourse structure can be recognized computationally. However, there remain discourses in which the coherence relations are unmarked even by syntax, tense or aspect. Some of these relations cannot be recognized computationally because they require extensive world knowledge. Ultimately any relation between events or objects which is common knowledge among discourse participants can form the basis for a felicitous request for an unmarked coherence inference. In order to recognize all coherence relations, a computational system needs full world knowledge.

Theory of Discourse Structure

Other theories have defined the relata in discourse structure as clauses (Trabasso and Sperry, 1985), pieces of text (Hobbs, 1985, Mann and Thompson, 1987), pieces of text plus connectives (Cohen, 1984, Reichman, 1985), propositions (Van Dijk and Kintsch, 1983, Polanyi, 1988), plans (Lochbaum, Grosz and Sidner, 1990) and segmented discourse representation structures, the SDRT theory (Asher, 1993).

The theory that we adopt here proposes that the relata in discourse structure are sets of events, states or entities introduced into the discourse, along the lines of SDRT. The reader builds a "cognitive model" of the text content. In the cognitive model or "situation model", the events in a discourse are connected by inferences concerning the surrounding events (causes, goals, parts of events, enabling conditions, and so on), as shown in many studies by Graesser and colleagues, such as Graesser and Zwaan, 1995. In narrative, the cognitive model forms a causal chain of events (Trabasso and Sperry, 1985). A discourse is coherent to the extent that a cognitive model of the discourse content can be built by a qualified reader, that is one with the requisite background knowledge.

In our theory, coherence inferences are added to the discourse representation as predications added to the DRS (Dahlgren, 1996). A discourse segment is a set of discourse events and entities that cohere among themselves, and share a single coherence relation to another discourse segment (which could consist of just one event or entity, such as the discourse topic). In this theory the same set of coherence relations relate the events introduced by individual sentences and also by sets of sentences (segments), because it was found that the same naive theories of the relatedness of things and events explained both local and global coherence. Surface rhetorical relations such as "turning to" were not considered in this theory.

The theory of coherence informing the empirical studies summarized below claims that the basis of coherence relations and discourse structure is the naive view (or naive theory) of events and the relatedness of objects in the real world. It is supposed that coherence inferences during discourse understanding are made according to the same naive theories people use to understand real events. A coherent discourse is then one for which a cognitive model of events can be built in which events and things relate in ways people naively expect them to relate. By "naive" we mean non-scientific and non-truth-conditional (Hayes, 1985). The cognitive model is built by the reader is again a naive theory—a belief structure about the way the world would be if the writers' story were true.

The set of coherence relations in Table I are justified from the above philosophical point of view and also by noticing that each of them can be grounded in a known psychological process, and that each of them can be marked by an overt lexical cue phrase, as summarized in Table I.

Studies of Coherence, Discourse Structure and Anaphora Resolution

Preliminary studies were conducted in order to facilitate the design of a computation text understanding system. We examined texts in three genres: 1) commentary text (13000 and 20000 words), narrative (the novel *Wheels* by Alex Hailey), and wire service reports (MUC-3 terrorism texts). The commentary corpus was drawn from Wall Street Journal articles which might be called "news", but in every case there was a key section that had commentary or evaluation of events, and the discourse structure differed from that of the terrorism news reports. The discourse segment boundaries, the coherence (discourse) relations and the anaphoric expressions throughout the commentary and narrative corpora were labeled and analyzed by two individuals (the author and another linguist).

The purpose of investigation was to discover:

1. human mechanisms for coherence relation assignment (including between segments of discourse) during interpretation
2. human mechanisms for discourse segmentation during interpretation
3. constraints on resolution of anaphoric expressions during interpretation

The goal was to clearly define the mechanisms so that they could be imitated by computational algorithms. These studies and results are fully described in (Dahlgren, 1996, Lord and Dahlgren, 1997, and Iwanska, et al, 1991).

Lexical Marking of Coherence Relations

In our theory, as in Knott and Dale (1994), it is not possible to have a coherence relation which is NEVER signaled lexically. Membership in the set of relations is justified first by having some overt marker in English. If coherence relations which never have an overt marker are allowed into the theory, a hopeless multiplication of invisible relations could ensue. Thus the theory trivially answers the symposium question, "Are there any that are never lexically signaled?" As for the other part of that question, "Are there discourse relations that are always lexically signaled?" our study replies in the negative. Any of the coherence relations in Table I can occur without overt lexical cues, as illustrated in the examples in Figure I.

The two-sentence examples in Figure I are replicated with discourse segments in the study corpora.

Non-lexical Marking of Discourse Structure

Our corpus studies directly answer one of the symposium questions, "What non-lexical (i.e., syntactic or prosodic) means are used

Table I: Evidence for Coherence Relations

Coherence Relation	Cognitive Capacity	Connective
cause	naive theories of causation	because, as a result, consequently
background	perception of figure/ground and salience	when, while
goal	naive theories of intentionality	in order to, so that
enablement	naive theories of causation	because
constituency	recognition of part/whole	in summary, for example, first, second
contrast	recognition of similarity	similarly, likewise, in contrast
elaboration	perception of spatio-temporal contiguity	then, next, that is to say
evaluation	preference (goodness ratings)	evidently, that means, modality, negation

Figure I: Coherence Relations without Lexical Cues

Cause	Fred died. Harry stabbed him.
Background	It was raining outside. Fred rushed in the front door.
Goal	Fred bought at book. He read it.
Enablement	Fred was well-heeled. He invested heavily.
Constituency	Fred was tried for murder. The prosecution opened the case with a diatribe.
Contrast	Fred loves rain. Mary hates it.
Elaboration	Fred rushed in the front door. He threw off his raincoat.
Evaluation	The networks are losing money. They should cut back on the ads.

to signal a relation?" In our study of segments, coherence relations, and pronouns in *Wheels*, there was a cue phrase at only 41% of the segment boundaries. Other indicators were non-lexical (see Table II). In our study of 13000 words of commentary genre text, cue phrases signaled segment boundary at only 16% of the segment boundaries.

Table II: Lexical and Non-Lexical Discourse Markers in *Wheels*

change of coherence relation	88%
change of sentence subject	72%
segmenting cue phrase	41%
change in tense or aspect	58%

By non-lexical discourse marker we mean an indicator of a structural boundary in the discourse, one that requires the reader to begin a new segment, and also find a plausible attachment point for the new segment in the discourse structure tree built so far (Asher, 1993). Examples of lexical discourse markers would be "while" (indicating the beginning of a background segment), "then" (indicating the continuation of the same segment with a constituent of the larger event described in the segment), and so on.

The non-lexical discourse markers found in all three genres are shown in Table III.

In the study of *Wheels* it was noted that change of sentence subject marked 72% of the new segments. This relates as well to the pattern of pronoun use. There is a complex relation between personal pronouns and discourse structure in *Wheels*. Antecedents of personal pronouns were found either in the same segment, in a dominating segment, or in an unclosed prior segment for which the use of the pronoun signals that the prior segment should be reopened (popped). Infrequently, the antecedent could be found in an

Table III: Non-lexical Discourse Markers

Change of sentence subject (change of local topic)
and consequently, absence of personal pronoun reference to prior sentence
Event anaphora
Change of coherence relation
Change of time
Change of place
Tense
Aspect

immediately adjacent closed sister segment in the discourse tree.

Event anaphora refers to the use of adverbs like "so" and demonstratives like "this" to close a segment and then refer to the summation of all of the events in the segment pronominally. An example is from a Wall Street Journal article about Brazil (shortened), where "this" refers to the sum of the events in the prior segment :

> Brazil suspended debt payments. Mexico followed suit. Chile threatened to cancel all of its foreign debt. This caught international bankers unprepared.

Computational Recovery of Discourse Structure

The symposium organizers ask, "In analysis, is it possible to reliably infer discourse relations from surface cues?" Superficially, the answer is "no", because as described above, frequently there are no surface cue phrases at segment boundaries. Even paragraph boundaries in highly edited text are not reliable cues of segment boundaries.

But at a deeper level, the answer is conditional. If by "surface cues" we mean all of the syntactic and semantic information available in the sentences of the discourse, it is conceivable to somewhat reliably (at least as reliably as humans can) infer discourse relations from that information. This task requires a full linguistic interpretive module that parses and disambiguates the discourse, producing a logical form. The logical form is input to a formal semantic module (such as an SDRT module). In the resulting SDRS for the discourse, along with the meaning representations of the word senses in the discourse (naive semantics), all of the information required to recover the discourse relations is available.

In this design, world knowledge is encoded in a "naive semantic" lexicon, which reflects a shallow layer of knowledge, just enough to interpret the text. "Just enough" means enough to disambiguate the word meanings and the syntactic structure, and enough to recover the antecedents of anaphoric expressions (Dahlgren, 1991).

Naive semantic representations capture some of the naive theories of the world which people associate with word sense meanings in a given culture (Dahlgren, 1988). Naive semantics is a lexical theory which equates the meaning of a word sense with a concept, so that concept representations and word sense meaning representations have the same form. In the contrasting classical tradition, word meanings are conjunctions of primitives which form conditions for being in the extension (or class of objects) named by a word sense (Katz and Fodor, 1963). The meaning of "water" is a formula

water(X)
$\Longleftrightarrow$ $clear(X)\&colorless(X)\&liquid(X)$

The classical theory doesn't work because: 1) true scientific theories of the nature of categories are not necessarily known by speakers of a language; 2) the categories concepts name are gradient, with some members better examples than others; and 3) typical

properties of objects aren't necessary properties. For example, muddy water is still water.

Naive semantics posits that word meanings are shallow, limited naive theories of objects and events. The meaning of "water" has naive propositions equivalent to the following:

> Water is a clear, tasteless liquid; you find it in rivers, you find it at the tap; you drink it; you wash with it.

The features in the representation are psycholinguistically justified. These are the types of propositions subjects list when asked to give the "characteristics" of nouns. (In our computational lexicon, features are represented in a first-order logic form with temporal markings.)

In naive semantics, the content of verb concepts is based upon psycholinguistic studies of story comprehension (Graesser and Clark, 1985). A verb is understood and recalled in terms of other events and states which typical surround the event type it names (rather than being understood as a metaphor for motion, as in other theories). For example, the verb "stab" is associated with the goal of harming someone, the goal of killing someone, the constituent event of piercing someone with a sharp instrument, the consequence of killing someone, the consequence of someone bleeding, the enabling state of having a knife and so on. These surrounding events are elicited by the wh-questions such as "What caused X?" and "What was the goal of X?". The corresponding features are those employed in our computational naive semantic lexicon, namely "cause", "goal", "what_next", "consequence", "time", "location", and "how", along with selectional restrictions.

Lexical naive theories arise in a culture or subculture, and are limited to those properties and propositions shared among the members of the subculture. In addition to the shared naive theory of an object or event, speakers of a dialect may hold individual beliefs which are at odds with the shared naive theory, but they have to use the shared theory in order to communicate. In other words, a scientist may know that an object which appears to be falling is not (such as the Sun), but must still understand such statements as "The Sun is setting" in terms of the incorrect naive theory underlying the use of "set" in the context. While naive semantic representations contain far more information than meaning representations in the classical theory, they are limited as well to that knowledge which is very closely associated with a word sense, and used to recover the interpretation of sentence structure and meaning while listening or reading. Included are the most typical propositions describing an object or event, those which inform word sense disambiguation, structural disambiguation and anaphora resolution processing, but not the elements which are used in deep inferencing or recollection of personal episodes.

The shallowness of naive semantic representations is particularly important in explaining the use of lexical markers in discourse. Writers tend to employ markers when they cannot assume that the reader will easily and readily draw coherence inferences without them. Readers will be able to do so if the shared naive theory of events includes enough information. If the naive theory says that an event E1 typically causes an event E2, then it is felicitous to write two sentences describing just the two events, with no discourse markers relating the events, i.e., E1 E2 or E2 E1. But if the naive semantics of the events does not provide the connection, writers tend to make it explicit at some point, in order to aid the reader in building the intended cognitive model.

The surrounding events in naive semantic verb representations are precisely the information required to trigger unmarked coherence inferences. A causal relation can be inferred by inspection of lexical information alone when no other cue is available as in the discourse below which has no cue phrase, no change in tense or aspect, and the reverse of temporal order.

Fred died. Harry stabbed him.

Humans know to make the required coherence inference (required because all felicitous literal discourses must cohere), and they infer that the cause of Fred's death was the stabbing. The naive semantic information associated with senses of "stab" and "die" enable a computational inference of the same kind to be made. This is reflected by adding to the DRS below a coherence predicate cause(e2,e1).

u1,e1,u2,e2 r1,r2,he1
r1 < now fred(u1) e1 die(u1) e1 included in r1 harry(u2) e2 stab(u2,him1) r1 < r2 e2 included in r2

After assignment of the coherence relation the segmented DRS (or cognitive DRS) has an added coherence predicate "cause(e2,e1)", which indicates that the cause of dying was stabbing. Also, the anaphoric expression "him1" is resolved to the same entity as Fred, namely u1 in the equation u1 = u2 in the cognitive DRS.

u1,e1,u2,e2 r1,r2
r1 < now fred(u1) e1 die(u1) e1 included in r1 harry(u2) e2 stab(u2,u1) r1 < r2 e2 included in r2 u2=u1 cause(e2,e1)

In another example, the precursor of our current implementation was able to build a shallow, topic-related discourse structure tree for MUC-3 message number 99 by noticing change of time, change of place, or segmenting cue phrase (Iwanska et al, 1991).

However, events and individuals in the world relate in indefinitely many ways. No matter how large the naive semantic lexicon would get, no matter how detailed the knowledge would become, a natural language understanding system would encounter discourses which required additional knowledge. The gap would prevent the system from drawing a coherence inference which would be easy for humans to draw. When they do have difficulty building the cognitive model, humans have a huge store of knowledge, and they dig deeper (while taking more time). Even in simple secular texts which require no knowledge of jargon, it is possible to find many segments related by coherence inferences which could not be drawn using a shallow naive semantic lexicon.

The problem lies in the fact that coherence inferences are based upon naive theories of the relatedness of events and objects in the world. Until a computer system can be taught the complete system of naive theories of the world, it can't form the full cognitive model of all discourses. It can only guarantee the derivation of the structure in those cases where lexical marking, change in sentence subject, event anaphora, change in time or place, tense or aspect are present as indicators. Nevertheless, a capability to derive that much of the structure is useful for many computational goals, including improved anaphora resolution, temporal reasoning and locative reasoning.

Conclusion

Discourse relations are often not marked lexically. However, other indicators, including syntax, semantics and world knowledge, are available in commentary, narrative and news genre texts. These can be used by a computational system that has a full syntax, formal semantics and a naive semantic lexicon, to recover much of the discourse structure. Complete recovery of discourse structure computationally awaits machine learning systems which can teach computers extensive knowledge about objects, events and their relations in the world.

References

Asher, N. 1993. *Reference to Abstract Objects in English*. Boston, MA: Kluwer Academic Publishers.

Britton, B. and J. Black (Eds.) 1985. *Understanding expository text*. Hillsdale, NK: Erlbaum.

Cohen, R. 1984. A computational theory of the function of clue words in argument understanding. *Proceedings of COLING-84*, 251–258.

Dahlgren, K. 1988. *Naive Semantics for Natural Language Understanding*. Boston, MA: Kluwer Academic Publishers.

Dahlgren, K. 1991. The autonomy of shallow lexical knowledge. In J. Pustejovsky and S. Bergler (Eds.), *Lexical Semantics and Knowledge Representation*. New York: Springer Verlag.

Dahlgren, K. 1996. Discourse coherence and segmentation. In E. Hovy and D. Scott (Eds.), *Burning Issues in Discourse* Hillsdale, NJ: Erlbaum.

Graesser, A., and L. Clark. 1985. *Structures and Procedures of Implicit Knowledge*. Norwood, NJ: Ablex.

Graesser, A., and G.H. Bower 1990. *Inferences and Text Comprehension*. San Diego, CA: Academic Press.

Graesser, A., and R.A. Zwaan. 1995. Inference Generation and the Construction of Situation Models. In Weaver, C.A., S. Mannes and C.R. Fletcher *Discourse Comprehension* Hillsdale, NJ: Erlbaum.

Grosz, B. and C. Sidner. 1986. Attention, Intensions and the Structure of Discourse: A Review. *Computational Linguistics* 7:85–98; 12:175–204.

Hayes, P.J. 1985. The Second Naive Physics Manifesto. In J.R. Hobbs and R.C. Moore (Eds.) *Formal Theories of the Commonsense World*. Norwood, NJ: Ablex.

Hobbs, J.R. 1985. *On the Coherence and Structure of Discourse*. CSLI Report CSLI-85-37.

Iwanska, L., D. Appelt, D. Ayuso, K. Dahlgren, B. Stalls, R. Grishman, G. Krupka, C. Montgomery, and E. Riloff. 1991. Computational aspects of discourse in the context of MUC-3. *Proc. of the Third Message Understanding Conference (MUC-3)*, 256–282.

Katz, J.J. and J.A. Fodor. 1963. The Structure of Semantic Theory. *Language* 39:170-210.

Knott, A. and R. Dale. 1994. Using Linguistic Phenomena to Motivate a Set of Rhetorical Relations. *Discourse Processes* 18(1):35-62.

Lochbaum, K.E., B.J. Grosz and C.L. Sidner. 1990. Models of plans to support communication: an initial report. Proc. AAAI: 485–490.

Lord, C. and K. Dahlgren. 1997. Participant and event anaphora in newspaper articles. In J. Bybee et al. (Eds.) *Essays on Language Function and Language Type Dedicated to T. Givon*. Amsterdam: John Benjamins.

Mann, W. and S. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. ISI Reprint Series: ISI-RS-87-190.

Morrow, D.G., S.L. Greenspan, and G.H. Bower. 1987. Accessibility and situation models in narrative comprehension. *Journal of Memory and Language* 26:165-87.

Polanyi, L. 1988. A formal model of the structure of discourse. *Journal of Pragmatics* 12:601–638.

Reichman, R. 1985. *Getting Computers to Talk Like You and Me*. Cambridge, MA: MIT Press.

Trabasso, T. and L.L. Sperry. 1985. Causal relatedness and the importance of story events. *Journal of Memory and Language* 24:595–611.

van den Broek, P., P.J. Bauer, and T. Bourg (Eds.) 1997. *Developmental Spans in Event Comprehension and Representation: Bridging Fictional and Actual Events*. Mahwah, NJ: Lawrence Erlbaum.

Van Dijk, T. and W. Kintsch. 1983. *Strategies of Discourse Comprehension*. New York: Academic Press.