ited attention constraint. On the other hand, centered segmentation may replace the cache model entirely, since both are competing models of the attentional state. Centered segmentation has also the additional advantage of restricting the search space of anaphoric antecedents to those discourse entities actually referred to in the discourse, while the cache model allows unrestricted retrieval in the main or long-term memory.

Many studies on discourse segmentation highlight the role of cue words for signaling segment boundaries (cf., e.g., Passonneau & Litman (1993)) or the use of overspecified referential expressions to indicate a thematic shift (Vonk et al., 1992; Walker, 1996b). However useful these strategies might be, we see the danger that such a surface-level description may actually hide structural regularities at deeper levels of investigation illustrated by access mechanisms for centering data at different levels of discourse segmentation.

We consider our proposal a first attempt to unmask the mystery behind the *oracle* Webber (1991) introduced in order to decide with which existing segment the current utterance is actually connected. This revelation is achieved by the formalization of the *linguistic structure* of discourses, while the *intentional structure* as proposed by Grosz & Sidner (1986) seems as hard to grasp as before. The methodology we propose might also constitute a starting point for the psycholinguistic community to conduct discourse processing experiments which go beyond extremely brief (and hence artificial) *pseudo texts* (cf., e.g., Greene et al. (1992)) and to apply their empirical methods to texts of considerable length instead.

Conclusions

We have developed a methodology for extending the centering model to incorporate the global referential structure of discourse for reference resolution. The hierarchy of discourse segments we compute realizes certain constraints on the reachability of antecedents. Moreover, the claim is made that the hierarchy of discourse segments implements an intuitive notion of the limited attention constraint, as we avoid a simplistic, cognitively implausible linear backward search for potentional discourse referents.

The model, nevertheless, still has several restrictions. First, it has been developed on the basis of a small corpus of written texts. Though these cover diverse text sorts (*viz.* technical product reviews, newspaper articles and literary narratives), we currently do not account for spoken monologues as modelled, e.g., by Passonneau & Litman (1993) or even the intricacies of dyadic conversations Rosé et al. (1995) deal with. Second, a thorough integration of the referential and intentional description fo discourse segments still has to be worked out.

Acknowledgments. We like to thank our colleagues in the CLIF group for fruitful discussions and instant support, Joe Bush who polished the text as a native speaker, and, in particular, Bonnie Webber for supplying invaluable comments to an earlier draft of this paper. Michael Strube is supported by a post-doctoral grant from DFG (Str 545/1-1).

References

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

Brennan, S. E., M. W. Friedman & C. J. Pollard (1987). A centering approach to pronouns. In *Proc. of the 25th Annual Meeting of the Association for Computational Linguistics; Stanford, Cal., 6–9 July 1987*, pp. 155–162.

Daneš, F. (1974). Functional sentence perspective and the organization of the text. In F. Daneš (Ed.), *Papers on Functional Sentence Perspective*, pp. 106–128. Prague: Academia.

Gordon, P. C., B. J. Grosz & L. A. Gilliom (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

Greene, S. B., G. McKoon & R. Ratcliff (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):266–283.

Grosz, B. J., A. K. Joshi & S. Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz, B. J. & C. L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hahn, U. & M. Strube (1996). Incremental centering and center ambiguity. In *Proc. of the 18th Annual Conference of the Cognitive Science Society; La Jolla, Cal., 12–15 August 1996*, pp. 568–573.

Hahn, U. & M. Strube (1997). Centering in-the-large: Computing referential discourse segments. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics; Madrid, Spain, 7–12 July 1997*.

Markert, K., M. Strube & U. Hahn (1996). Inferential realization constraints on functional anaphora in the centering model. In *Proc. of the 18th Annual Conference of the Cognitive Science Society; La Jolla, Cal., 12–15 August 1996*, pp. 609–614.

Passonneau, R. J. (1996). Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering in Discourse*. Preprint.

Passonneau, R. J. & D. J. Litman (1993). Intention based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics; Columbus, Ohio, 22–26 June 1993*, pp. 148–155.

Rosé, C. P., B. Di Eugenio, L. S. Levin & C. Van Ess-Dykema (1995). Discourse processing of dialogues with multiple threads. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics; Cambridge, Mass., 26–30 June 1995*, pp. 31–38.

Sidner, C. L. (1983). Focusing in the comprehension of definite anaphora. In M. Brady & R. Berwick (Eds.), *Computational Models of Discourse*, pp. 267–330. Cambridge, Mass.: MIT Press.

Strube, M. & U. Hahn (1996). Functional centering. In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics; Santa Cruz, Cal., 23–28 June 1996*, pp. 270–277.

Suri, L. Z. & K. F. McCoy (1994). RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics*, 20(2):301–317.

Vonk, W., L. G. Hustinx & W. H. Simons (1992). The use of referential expressions in structuring discourse. *Language and Cognitive Processes*, 7(3/4):301–333.

Walker, M. A. (1996a). Centering, anaphora resolution, and discourse structure. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering in Discourse*. Preprint.

Walker, M. A. (1996b). Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264.

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

	IT	Spiegel	Lit	Σ
U_{i-1}	94	15	15	124
U_{i-2}	42	6	8	56
U_{i-3}	16	0	0	16
U_{i-4}	14	0	0	14
U_{i-5}	8	0	0	8
U_{i-6} to U_{i-10}	14	1	0	15
U_{i-11} to U_{i-15}	7	0	0	7

Table 5: Functional Anaphoric Antecedent in Utterance U_x

Table 6 and Table 7 give the success rate of the centered segmentation algorithm for (pro)nominal anaphors and functional anaphors, respectively. The numbers in these tables indicate at which segment level anaphoric expressions were correctly resolved. The category of *errors* covers erroneous analyses the algorithm produces, while the one for *false positives* concerns those resolution results where a referential expression was resolved with the hierarchically most recent antecedent but not with the linearly most recent (obviously, the targeted) one (both of them denote the same discourse entity). The categories $C_f(s, U_{i-1})$ in Tables 6 and 7 contain more elements than the categories U_{i-1} in Tables 4 and 5, respectively, due to the mediating property of functional anaphora in functional centering (Strube & Hahn, 1996).

	IT	Spiegel	Lit	Σ
U_i	10	7	32	49
$C_f(s, U_{i-1})$	161	78	125	364
$C_p(s-1, U_{DS[s-1.end]})$	14	9	24	47
$C_f(s-1, U_{DS[s-1.end]})$	7	5	9	21
$C_p(s-2, U_{DS[s-2.end]})$	1	0	1	2
$C_p(s-3, U_{DS[s-3.end]})$	1	0	1	2
$C_p(s-4, U_{DS[s-4.end]})$	0	0	1	1
$C_p(s-5, U_{DS[s-5.end]})$	0	1	0	1
errors	3	1	5	9
false positives	(1)	(3)	(7)	(11)

Table 6: Anaphoric Antecedent in Center$_x$

	IT	Spiegel	Lit	Σ
$C_f(s, U_{i-1})$	156	18	17	191
$C_p(s-1, U_{DS[s-1.end]})$	18	0	4	22
$C_f(s-1, U_{DS[s-1.end]})$	10	1	2	13
$C_p(s-2, U_{DS[s-2.end]})$	7	1	0	8
$C_p(s-3, U_{DS[s-3.end]})$	3	0	0	3
errors	1	2	0	3
false positives	(2)	(0)	(3)	(5)

Table 7: Functional Anaphoric Antecedent in Center$_x$

The centered segmentation algorithm reveals a pretty good performance. This is to some extent implied by the structural patterns we find in expository texts, *viz.* their single-theme property (e.g., *"1260"* in the sample text). In contrast, the literary texts in the test exhibited a much more difficult internal structure which resembled the multiple thread structure of dialogues discussed by Rosé et al. (1995). The good news is that the segmentation procedure we propose is capable of dealing even with these more complicated structures. While only one antecedent of a pronoun was not reachable given the superimposed text structure, the remaining eight errors are characterized by full definite noun phrases or proper names. The vast majority of these phenomena can be considered *informationally redundant utterances* in the terminology of Walker (1996b) for which we currently have no solution at all. It seems to us that these kinds of phrases may override text-grammatical structures as evidenced by referential discourse segments and, rather, trigger other kinds of search strategies.

Though we fed the centered segmentation algorithm with rather long texts (up to 84 utterances), the antecedents of only two anaphoric expressions had to bridge a hierarchical distance of more than 3 levels. This coincides with our supposition that the overall structure computed by the algorithm should be rather flat. We could not find an embedding of more than seven levels.

Related Work

There has always been an implicit relationship between the local perspective of centering and the global view of focusing on discourse structure (cf. the discussion in Grosz et al. (1995)). However, work establishing an explicit account of how both can be joined in a computational model has not been done so far. The efforts of Sidner (1983), e.g., have provided a variety of different focus data structures to be used for reference resolution. This multiplicity and the on-going growth of the number of different entities (cf. Suri & McCoy (1994)) mirrors an increase in explanatory constructs that we consider a methodological drawback to this approach because they can hardly be kept control of. Our model, due to its hierarchical nature implements a stack behavior that is also inherent to the above mentioned proposals. We refrain, however, from establishing a new data type (even worse, different types of stacks) that has to be managed on its own. There is no need for extra computations to determine the "segment focus", since that is implicitly given in the local centering data already available in our model.

A recent attempt at introducing global discourse notions into the centering framework considers the use of a cache model (Walker, 1996b). This introduces an additional data type with its own management principles for data storage, retrieval and update. While our proposal for centered discourse segmentation also requires a data structure of its own, it is better integrated into centering than the caching model, since the cells of segment structures simply contain "pointers" that implement a direct link to the original centering data. Hence, we avoid extra operations related to feeding and updating the cache. The relation between our centered segmentation algorithm and Walker's (1996a) integration of centering into the cache model can be viewed from two different angles. On the one hand, centered segmentation may be a part of the cache model, since it provides an elaborate, non-linear ordering of the elements within the cache. Note, however, that our model does not require any *pre*fixed size corresponding to the lim-

U_i	Sample Text	Centering-Data	Segment Levels 1 2 3	Block
(1)	Brother HL-1260	**Cb:** – **Cf:** [**1260**]		
(2)	Ein Detail fällt schon beim ersten Umgang mit dem großen *Brother* auf: One particular – is already noticed – in the first approach to – the big *Brother*:	**Cb:** 1260 **Cf:** [**1260**, Umgang, Detail]		**1**
(3)	Im Betrieb macht *er* durch ein kräftiges Arbeitsgeräusch auf sich aufmerksam, das auch im Stand-by-Modus noch gut vernehmbar ist. In operation – draws – *it* – with a heavy noise level – attention to itself – which – also – in the stand-by mode – is still well audible.	**Cb:** 1260 **Cf:** [**1260**, Betrieb, Arbeitsgeräusch, Stand-by-Modus]		**1**
(4)	Für Standard-Installationen kommt man gut ohne Handbuch aus. As far as standard installations are concerned – gets – one – well – by – without any manual.	**Cb:** – **Cf:** [**Standard-Installation**, Handbuch]		**2c**
(5)	Zwar erläutert das dünne *Handbüchlein* die Bedienung der *Hardware* anschaulich und gut illustriert. Admittedly, gives – the thin *leaflet* – the operation of the *hardware* – a clear description of – and – well illustrated.	**Cb:** Handbuch **Cf:** [**Handbuch**, 1260, Hardware, Bedienung]		**3**
(6)	Die *Software-Seite* wurde im *Handbuch* dagegen stiefmütterlich behandelt: The *software part* – was – in the *manual* – however – like a stepmother – treated:	**Cb:** Handbuch **Cf:** [**Handbuch**, 1260, Software]		**1, Lift**
(7)	bis auf eine karge *Seite* mit einem Inhaltsverzeichnis zum *HP-Modus* sucht man vergebens weitere Informationen. except for one meagre *page* – containing the table of contents for the *HP mode* – seeks – one – in vain – for further information.	**Cb:** Handbuch **Cf:** [**Handbuch**, Seite, 1260, HP-Modus, Inhaltsverzeichnis, Informationen]		**1**

Table 2: Sample Text, Centering-Data and Discourse Structure

U_6 at level 2. As a consequence, the centering data of U_5 are excluded from further consideration as far as subsequent matching processes are concerned. U_7 simply continues the same segment, since the functional anaphor *"Seite" (page)* refers to *"Handbuch" (manual)*, the $C_p(2, U_6)$.

Empirical Findings

In this section, we present some empirical data concerning the centered segmentation algorithm. Our study was based on the analysis of twelve texts from the information technology domain (IT), of one text from a German news magazine (Spiegel)[4], and of two literary texts[5] (Lit). Table 3 summarizes the total numbers of (pro)nominal and functional anaphors, utterances, and words in the test set.

Table 4 and Table 5 consider the number of (pro)nominal anaphors and functional anaphors, respectively, and the linear distance they have to their corresponding antecedents. Note that common centering algorithms (e.g., the one by Brennan et al. (1987)) are specified only for the resolution of anaphors in U_{i-1}. They are neither specified for anaphoric antecedents in U_i, not an issue here, nor for anaphoric antecedents beyond U_{i-1}. In the test set, 139 (pro)nominal anaphors (28%) and 116 functional anaphors (48,3%) fall out of the (intersentential) scope of those common algorithms. So, the problem we consider is not a marginal one.

	IT	Spiegel	Lit	Σ
(pro)nominal anaphors	197	101	198	496
functional anaphors	195	22	23	240
utterances	336	84	127	547
words	5241	1468	1610	8319

Table 3: Test Set

	IT	Spiegel	Lit	Σ
U_i	10	7	32	49
U_{i-1}	117	70	121	308
U_{i-2}	28	14	24	66
U_{i-3}	18	5	10	33
U_{i-4}	6	1	5	12
U_{i-5}	6	0	1	7
U_{i-6} to U_{i-10}	8	1	3	12
U_{i-11} to U_{i-15}	3	1	1	5
U_{i-15} to U_{i-20}	1	2	1	4

Table 4: Anaphoric Antecedent in Utterance U_x

[4] Japan – Der Neue der alten Garde. In *Der Spiegel*, Nr. 3, 1996.

[5] The first two chapters of a short story by the German writer Heiner Müller (Liebesgeschichte. In Heiner Müller. *Geschichten aus der Produktion 2*. Berlin: Rotbuch Verlag, 1974, pp.57-63) and the first chapter of a novel by Uwe Johnson (*Zwei Ansichten*. Frankfurt/Main: Suhrkamp Verlag, 1965.)

analysis proceeds until this discourse segment DS is *ultimately closed*, i.e., whenever another discourse segment DS' exists at the *same* or a *hierarchically higher* level of embedding such that the end point of DS' exceeds that of the end point of DS. Closed segments are inaccessible for the antecedent search. Considering the sample text analysis in Table 2, the discourse segment at level 3 (ranging from utterance 5 to 5) is ultimately closed, while those at level 1 (ranging from utterance 1 to 3) or level 2 (ranging from utterance 4 to 7) are considered open.

In the following, s and U_i denote the currently considered segment level and utterance, respectively. The main algorithm consists of three major logical blocks.

1. **Continue Current Segment.** If one of the referents from U_i contained in $\mathcal{R}$ matches the $C_p(s, U_{i-1})$, then the current segment is extended by U_i at the same segment level s. If, in addition, U_{i-1} and U_i indicate the end of a sequence in which a series of thematizations of rhemes have occurred, all embedded segments are *lifted* to the higher level s'. As a result of such a lifting, the entire sequence (including the final two utterances) forms a single segment. This is, of course, trivially true for cases of a constant theme. Lifting is the most severe operation for restructuring already established segment structures.

2. **Close Embedded Segment(s).** If the referents from U_i contained in $\mathcal{R}$ do not match an element of the $C_f(s, U_{i-1})$, then three possibilities emerge:

   (a) *Close the embedded segment:* If one of the referents from U_i contained in $\mathcal{R}$ match the $C_p(s-1, U_{[s-1.end]})$, the segment at level s is closed while the segment at level $s-1$ is extended by U_i. If none of the referents from U_i contained in $\mathcal{R}$ match the $C_p(s-1, U_{[s-1.end]})$, then the remaining elements of the C_f at this segmentation level are checked for the given utterance. If a match succeeds, the segment which contains U_{i-1} is ultimately closed, since U_i opens a parallel segment at the *same* level of embedding. Subsequent anaphora checks exclude any preceding closed or parallel segments from consideration for the matching process and just consider the currently open one.

   (b) *Close the embedded segment(s) and continue another, already existing segment:* If one of the referents from U_i contained in $\mathcal{R}$ matches a discourse referent in one of the remaining hierarchically reachable segments, that particular segment which contains the matching discourse referent is extended by U_i. Only the C_p of the utterance at the end point of any of these segments is considered for a match. Note that, as a side effect, hierarchically lower segments are ultimately closed when a match at higher levels succeeds.

   (c) *Open new, embedded segment:* If there is no matching referent in hierarchically reachable segments, then for utterance U_i a new, embedded segment is opened.

3. **Open a New, Embedded Segment.** If none of the above cases applies, then for utterance U_i a new embedded segment is opened. While processing further utterances, this decision may be retracted by *lifting* the embedded segments. This realizes a kind of "garbage collection" for globally insignificant discourse segments which, nevertheless, were reasonable from a local perspective for reference resolution purposes. Hence, the discourse segmentation procedure we propose works in a strictly incremental way and revises previously made segmentation decisions on the fly. We consider this kind of dynamic adjustment of global discourse structure a particularly interesting cognitive issue.

A Sample Text Segmentation

For illustrative purposes, we here consider a text fragment (see Table 2) taken from a German computer magazine (*c't*, 1995, No. 4, p.209). Since the method for computing levels of discourse segments heavily depends on different kinds of anaphoric expressions, (pro)nominal and functional anaphors are marked by italics, and the (pro)nominal anaphors are underlined, in addition. In order to convey the influence of the German word order, we provide a rough phrase-to-phrase translation of the sample text.

Table 2 depicts a complete segmentation of the sample text. The first column contains the number of each utterance, the second column gives the sample text, while the third column shows the centering data which are computed by the functional centering approach (Strube & Hahn, 1996). The first element of each C_f, the *preferred center*, C_p, is marked by bold font. The fourth column depicts the levels of discourse segments which are computed by our algorithm. The horizontal lines indicate the beginning of a segment, the vertical ones mark the extension of a segment. The fifth column indicates which block of the algorithm applies to the current utterance.

The computation starts at U_1, the headline. The $C_p(1, U_1)$ is set to *"1260"* which is meant as an abbreviation of *"Brother HL-1260"*. Upon initialization, the beginning as well as the ending of the initial discourse segment are both set to "1". U_2 and U_3 simply continue this segment (block (1) of the algorithm). The C_p is set to *"1260"* in all utterances of this segment. Since the referents for U_4 do neither match the $C_p(1, U_3)$ (block (1)) nor is any hierarchically preceding segment available, block (2c) applies. The segment counter s is incremented and a new segment at level 2 is opened, setting the beginning and the ending to "4". The referent of the phrase *"das dünne Handbüchlein" (the thin leaflet)* in U_5 does not match the $C_p(2, U_4)$ but it matches an element of the $C_f(2, U_4)$ instead (*viz. "Handbuch" (manual)*). Hence, block (3) of the algorithm applies, leading to the creation of a new segment at level 3. The referent of the anaphor *"Handbuch" (manual)* in U_6 matches the $C_p(3, U_5)$. Hence block (1) applies (the occurrence of *"1260"* in $C_f(U_5)$ is due to the assumptions specified by Strube & Hahn (1996)). Given this configuration, the embedded segment is *lifted* one level, so the segment which ended with U_4 is now continued up to further be complemented by a cache mechanism. She thus intends to extend the scope of centering in accordance with cognitively plausible limits of the attentional span.

As a working hypothesis, for the purpose of anaphora resolution we subscribe to Walker's model, especially to that part which casts doubt on the hypothesized dependency of the (local) attentional from the (global) intentional structure of discourse (Grosz & Sidner, 1986, p.180). We diverge from Walker (1996a), however, by proposing an alternative to the cache mechanism, which we consider to be methodologically more parsimonious and, at least to be equally effective.

The proposed extension of the centering model builds on the methodological framework of *functional centering* (Strube & Hahn, 1996). This is an approach to centering in which issues such as thematicity or topicality are already inherent. Its linguistic foundations relate the ranking of the forward-looking centers and the functional information structure of the utterances, a notion originally developed by Daneš (1974). Strube & Hahn (1996) use the centering data structures to redefine Daneš' trichotomy between given information, theme and rheme in terms of the centering model. In this framework, the $C_b(U_i)$ corresponds to the element which represents the *given* information. The *theme* of U_i is represented by the preferred center $C_p(U_i)$, the most highly ranked element of $C_f(U_i)$. The *theme/rheme hierarchy* of U_i corresponds to the ranking in the C_fs.

Identifying the preferred center with the theme implies that the *preferred center* is of major relevance for determining the thematic progression of a text (a term also introduced by Daneš (1974)), from which discourse segments should be easily derivable. Daneš (1974) distinguishes three types of thematic progression (TP), two of which can directly be derived from centering data, while the third requires to refer to conceptual generalization hierarchies (not an issue in this paper). In particular, a distinction is made between

1. *TP with a constant theme:* Successive utterances continuously share the same C_p.

2. *TP with linear thematization of rhemes:* An element of the $C_f(U_{i-1})$ which is not the $C_p(U_{i-1})$ appears in U_i and becomes the $C_p(U_i)$ after the processing of this utterance.

$C_f(U_{i-1})$:	[c_1, ..., c_j, ..., c_s]	
	$\downarrow$	
$C_f(U_i)$:	[c_1, ..., c_k, ..., c_t]	
$C_f(U_{i-1})$:	[c_1, ..., c_j, ..., c_s]	$1 < j \leq s$
	$\swarrow$	
$C_f(U_i)$:	[c_1, ..., c_k, ..., c_t]	

Table 1: Thematic Progression Patterns

Table 1 gives the abstract schemata of *TP patterns*, while in Table 2 U_1 to U_3 illustrate the *constant theme* and U_4 to U_5 reflect the *linear thematization of rhemes*. In the latter case, the theme changes from *"Standard-Installation" (standard installation)* to *"Handbuch" (manual)*.

Furthermore, Daneš (1974) allows for the combination and recursion of TP patterns; this way, the global thematic coherence of a text can be described by recurrence to these basic structural patterns. Their use allows for a major extension of the original centering algorithm, since given a reformulation of the corresponding constraints in centering terms, it is possible to determine segment boundaries and to arrange the segments in a nested, i.e., hierarchical manner on the basis of which reachability constraints for antecedents can be formulated. According to the segmentation strategy of our approach, the C_p of the end point (i.e., the last utterance) of a discourse segment provides the major theme of the whole segment. This theme is particularly salient for anaphoric reference relations. Whenever a relevant new theme is established, however, it should reside in its own discourse segment, either embedded or in parallel to another one.

Computing Referential Discourse Segments

The computation of referential discourse segments can be split into two major processes. The first process determines the set of referents $\mathcal{R}$, i.e., the valid antecedents for the anaphoric expressions in the current utterance U_i, which are *reachable* from the segment level s associated with U_i by considering:[3]

- the forward-looking centers of the *linearly immediately* preceding utterance (as in the standard centering model),
- the forward-looking centers at the end point of the *hierarchically immediately* reachable discourse segment,
- the preferred center at the end point of *any hierarchically* reachable, though minimally distant discourse segment.

Reachability is thus made dependent on the current segment structure DS of the discourse as determined by the segmentation algorithm. The second process builds up the segment structure DS depending on the *reachability patterns* that characterize a referent from $\mathcal{R}$ and its possible occurrence in one of the above-mentioned centers according to the specification in the algorithm given below (for a more technical treatment, cf. Hahn & Strube (1997)).

Note that, as a result, the current discourse segment structure may not only be "monotonically" refined but can also be fundamentally revised, thus making our approach a truly *incremental* model of anaphora resolution. In the following discussion, the discourse segment index s is usually attached to any center expression, e.g., $C_f(s, U_i)$.

We will now concentrate on the construction part of the segmentation algorithm. Whenever a discourse segment is created at level s, its starting and closing utterance, $U_{[s.beg]}$ and $U_{[s.end]}$, are initialized to the current position in the discourse. Its end point gets continuously incremented as the

[3] We abstract here from the linguistic constraints imposed on (pro)nominal or functional anaphors (cf. Strube & Hahn (1996), Hahn & Strube (1996) and Markert et al. (1996)).

Centered Segmentation: Scaling up the Centering Model to Global Referential Discourse Structure

Michael Strube & Udo Hahn
Computational Linguistics Research Group
Freiburg University, Werthmannplatz 1
D-79085 Freiburg, Germany
http://www.coling.uni-freiburg.de/

Abstract

We introduce a methodology for determining referents in full-length texts in a computationally parsimonious way. Based on the centering model, whose focus is on the local coherence of discourse, we build up a hierarchy of referential discourse segments from the local centering data. The spatial extension and nesting of these discourse segments constrain the reachability of potential antecedents of an anaphoric expression above the level of adjacent center pairs. Thus, the centering model is scaled up to the level of global discourse structure.

Introduction

Referent identification in discourse is a cognitively challenging task, since the more a discourse unfolds the greater the number of possible antecedents which are available as candidates for reference resolution. Any naive account, e.g., a simple linear backward search of all previously established discourse entities coupled with linguistic and conceptual compatibility tests at each check point, is error-prone and far too ressource-consumptive. Cognitively plausible models of reference identification have, therefore, always tried to reduce the search efforts and inference load imposed on the hearer (Greene et al., 1992). The centering model (Grosz et al., 1995) explicitly addresses these issues, as it supplies simple, yet powerful data structures, constraints and rules for the *local* coherence of discourse. In essence, candidate antecedents in an utterance are preferentially ranked, and a discourse increases in coherence the least expensive a valid antecedent is found based on this ranking. These assumptions of the centering model have already been experimentally validated (Gordon et al., 1993; Brennan, 1995).

There is, however, a structural limitation inherent in these models, the centering model in particular. Their focus is on anaphoric relations that hold *within* a discourse segment composed of only few utterances (usually, two- or three-sentence discourses are examined). However, well-formed anaphoric relationships may well *extend over* segment boundaries in full-length texts. In these cases, cognitive parsimony becomes even more important, as the number of possible antecedents tends to explode. The procedure we propose for creating and managing large referential discourse segments[1] provides a sort of superimposed index structure by which the search space for potential antecedents can be effectively restricted to a *minimal* number of discourse elements. This model fully subscribes to the inference-reducing efforts of the basic centering model but extends it to account for the referential coherence covering the *entire* discourse.

[1] Our notion of *referential* discourse segments should not be confounded with the *intentional* one originating from Grosz et al. (1995), for reasons discussed in Hahn & Strube (1997).

Centering and Global Discourse Structure

In the centering model, each utterance U_i is assigned a set of *forward-looking centers*, $C_f(U_i)$, and a unique *backward-looking center*, $C_b(U_i)$. A ranking is imposed on the elements of the C_f which reflects the assumption that the most highly ranked element of $C_f(U_i)$ is the most preferred antecedent of an anaphoric expression in U_{i+1}, while the remaining elements are ordered according to decreasing preference for establishing referential links. The most highly ranked element of $C_f(U_i)$ that is *realized* in U_{i+1} (i.e., has a valid interpretation in the underlying semantic representation) is the $C_b(U_{i+1})$. Given these widely acknowledged conditions, anaphora resolution in U_i depends on accessing the *immediately preceding* forward-looking centers, $C_f(U_{i-1})$, only.

There have been only few attempts at explicitly dealing with the recognition and incorporation of discourse structure beyond the level of immediately adjacent utterances within the centering framework. Two recent studies (Passonneau, 1996; Walker, 1996a) deal with this topic by considering, among other things, whether a correlation exists between particular configurations of local centering data and the global structure of discourse segments. In particular, the supposition is investigated whether two different backward-looking centers, $C_b(U_{i-1}) \neq C_b(U_i)$, for two immediately adjacent utterances, U_{i-1} and U_i, [2] indicate the presence of a discourse segment boundary. Empirical data, however, yield only a weak support for this claim, if at all. Hence, this finding precludes a reliable prediction of segment boundaries based on the occurrence of local centering data (i.e., shift transitions) and *vice versa*. In order to accommodate to these empirical results Passonneau suggests that the centering data structures need to be modified appropriately, while Walker concludes that the local centering data should be left as they are and

[2] This pattern is usually considered a particular form of centering transitions, so-called "shifts"; cf. Brennan et al. (1987) for a detailed discussion of transition types.

CogSci '97: Proc. of the 19th Annual Conf. of the Cognitive Science Society. Palo Alto, Cal., Aug 7-10 1997, to appear.