


Centering in-the-Large: Computing Referential Discourse Segments

Udo Hahn & Michael Strube

 Computational Linguistics Research Group

Freiburg University, Werthmannplatz 1

D-79085 Freiburg, Germany

<http://www.coling.uni-freiburg.de/>

Abstract

We specify an algorithm that builds up a hierarchy of referential discourse segments from local centering data. The spatial extension and nesting of these discourse segments constrain the reachability of potential antecedents of an anaphoric expression beyond the local level of adjacent center pairs. Thus, the centering model is scaled up to the level of the global referential structure of discourse. An empirical evaluation of the algorithm is supplied.

1 Introduction

The centering model (Grosz et al., 1995) has evolved as a major methodology for computational discourse analysis. It provides simple, yet powerful data structures, constraints and rules for the *local* coherence of discourse. As far as anaphora resolution is concerned, e.g., the model requires to consider those discourse entities as potential antecedents for anaphoric expressions in the current utterance U_i , which are available in the forward-looking centers of the *immediately preceding* utterance U_{i-1} . No constraints or rules are formulated, however, that account for anaphoric relationships which spread out over non-adjacent utterances. Hence, it is unclear how discourse elements which appear in utterances preceding utterance U_{i-1} are taken into consideration as potential antecedents for anaphoric expressions in U_i .

The extension of the search space for antecedents is by no means a trivial enterprise. A simple linear backward search of all preceding centering structures, e.g., may not only turn out to establish illegal references but also contradicts the cognitive principles underlying the limited attention constraint (Walker, 1996b). The solution we propose starts from the observation that additional constraints on valid antecedents are placed by the *global* discourse structure previous utterances are embedded in. We want to emphasize from the beginning that our proposal considers only the *referential* properties underlying

the global discourse structure. Accordingly, we define the *extension* of referential discourse segments (over several utterances) and a *hierarchy* of referential discourse segments (structuring the entire discourse).¹ The algorithmic procedure we propose for creating and managing such segments receives local centering data as input and generates a sort of superimposed index structure by which the reachability of potential antecedents, in particular those prior to the immediately preceding utterance, is made explicit. The adequacy of this definition is judged by the effects centered discourse segmentation has on the validity of anaphora resolution (cf. Section 5 for a discussion of evaluation results).

2 Global Discourse Structure

There have been only few attempts at dealing with the recognition and incorporation of discourse structure beyond the level of immediately adjacent utterances within the centering framework. Two recent studies deal with this topic in order to relate attentional and intentional structures on a larger scale of global discourse coherence. Passonneau (1996) proposes an algorithm for the generation of referring expressions and Walker (1996a) integrates centering into a cache model of attentional state. Both studies, among other things, deal with the supposition whether a correlation exists between particular centering transitions (which were first introduced by Brennan et al. (1987); cf. Table 1) and intention-based discourse segments. In particular, the role of SHIFT-type transitions is examined from the perspective of whether they not only indicate a shift of the topic between two immediately successive utterances but also signal (intention-based) segment boundaries. The data in both studies reveal that only a weak correlation between the SHIFT transitions and segment boundaries can be observed. This finding precludes a reliable prediction of segment boundaries based on the occurrence of

¹Our notion of *referential* discourse segment should not be confounded with the *intentional* one originating from Grosz & Sidner (1986), for reasons discussed in Section 2.

SHIFTS and *vice versa*. In order to accommodate to these empirical results divergent solutions are proposed. 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 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. Walker, finally, claims that the content of the cache, rather than the intentional discourse segment structure, determines the accessibility of discourse entities for anaphora resolution.

	$C_b(U_n) = C_b(U_{n-1})$ OR $C_b(U_{n-1})$ undef.	$C_b(U_n) \neq$ $C_b(U_{n-1})$
$C_b(U_n) =$ $C_p(U_n)$	CONTINUE (C)	SMOOTH-SHIFT (SS)
$C_b(U_n) \neq$ $C_p(U_n)$	RETAIN (R)	ROUGH-SHIFT (RS)

Table 1: Transition Types

As a working hypothesis, for the purposes of anaphora resolution we subscribe to Walker's model, in particular to that part which casts doubt on the hypothesized dependency of the attentional from the intentional structure of discourse (Grosz & Sidner, 1986, p.180). We diverge from Walker (1996a), however, in that we propose an alternative to the caching mechanism, which we consider to be methodologically more parsimonious and, at least, to be equally effective (for an elaboration of this claim, cf. Section 6).

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's trichotomy between *given information*, *theme* and *rheme* in terms of the centering model. The $C_b(U_n)$, the most highly ranked element of $C_f(U_{n-1})$ realized in U_n , corresponds to the element which represents the *given information*. The *theme* of U_n is represented by the preferred center $C_p(U_n)$, the most highly ranked element of $C_f(U_n)$. The *theme/rheme hierarchy* of U_n corresponds to the ranking in the C_f s. As a consequence, utterances without any anaphoric expression do not have any *given* elements and, therefore, no C_b . But independent of the use of anaphoric expressions, each utterance must have a theme and a C_f as well.

The identification of the *preferred center* with the *theme* implies that it is of major relevance for determining the thematic progression of a text. This is reflected in

our reformulation of the two types of thematic progression (TP) which can be directly derived from centering data (the third one requires to refer to conceptual generalization hierarchies and is therefore beyond the scope of this paper, cf. Daneš (1974) for the original statement):

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.

$C_f(U_{i-1}) :$	$[c_1, \dots, c_j, \dots, c_s]$
	↓
$C_f(U_i) :$	$[c_1, \dots, c_k, \dots, c_t]$
$C_f(U_{i-1}) :$	$[c_1, \dots, c_j, \dots, c_s] \quad 1 < j \leq s$
	↙
$C_f(U_i) :$	$[c_1, \dots, c_k, \dots, c_t]$

Table 2: Thematic Progression Patterns

Table 2 visualizes the abstract schemata of *TP patterns*. In our example (cf. Table 8 in Section 4), U_1 to U_3 illustrate the *constant theme*, while U_7 to U_{10} illustrate the *linear thematization of rhemes*. In the latter case, the theme changes in each utterance, from "*Handbuch*" (*manual*) via "*Inhaltsverzeichnis*" (*table of contents*) to "*Kapitel*" (*chapter*) etc. Each of the new themes are introduced in the immediately preceding utterance so that local coherence between these utterances is established.

Daneš (1974) also allows for the combination and recursion of these basic patterns; this way the global thematic coherence of a text can be described by recurrence to these structural patterns. These principles allow for a major extension of the original centering algorithm. Given a reformulation of the TP constraints in centering terms, it is possible to determine referential segment boundaries and to arrange these 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, one which 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. Anaphora resolution can then be performed (a) with the forward-looking centers of the linearly immediately preceding utterance, (b) with the forward-looking centers of the end point of the hierarchically immediately reachable discourse segment, and (c) with the preferred center of the end point of any hierarchically reachable discourse segment (for a formalization of this constraint, cf. Table 4).

3 Computing Global Discourse Structure

Prior to a discussion of the algorithmic procedure for hypothesizing discourse segments based on evidence from local centering data, we will introduce its basic building blocks. Let x denote the anaphoric expression under consideration, which occurs in utterance U_i associated with segment level s . The function $Resolved(x, s, U_i)$ (cf. Table 3) is evaluated in order to determine the proper antecedent *ante* for x . It consists of the evaluation of a reachability predicate for the antecedent on which we will concentrate here, and of the evaluation of the predicate $IsAnaphorFor$ which contains the linguistic and conceptual constraints imposed on a (pro)nominal anaphor (viz. agreement, binding, and sortal constraints) or a textual ellipsis (Hahn et al., 1996), not an issue in this paper. The predicate $IsReachable$ (cf. Table 4) requires *ante* to be reachable from the utterance U_i associated with the segment level s .² Reachability is thus made dependent on the segment structure DS of the discourse as built up by the segmentation algorithm which is specified in Table 6. In Table 4, the symbol “ $=_{str}$ ” denotes string equality, \mathbb{N} the natural numbers. We also introduce as a notational convention that a discourse segment is identified by its index s and its opening and closing utterance, viz. $DS[s.beg]$ and $DS[s.end]$, respectively. Hence, we may either identify an utterance U_i by its linear text index, i , or, if it is accessible, with respect to its hierarchical discourse segment index, s (e.g., cf. Table 8 where $U_3 = U_{DS[1.end]}$ or $U_{13} = U_{DS[3.end]}$). The discourse segment *index* is always identical to the currently valid segment *level*, since the algorithm in Table 6 implements a stack behavior. Note also that we attach the discourse segment index s to center expressions, e.g., $C_b(s, U_i)$.

$Resolved(x, s, U_i) :=$	
$\left\{ \begin{array}{l} ante \text{ if } \\ undef \text{ else} \end{array} \right.$	$\left\{ \begin{array}{l} IsReachable(ante, s, U_i) \\ \wedge IsAnaphorFor(x, ante) \end{array} \right.$

Table 3: Resolution of Anaphora

$IsReachable(ante, s, U_i)$	
$\left\{ \begin{array}{l} \text{if } \\ \text{else if } \\ \text{else if } \end{array} \right.$	$\left\{ \begin{array}{l} ante \in C_f(s, U_{i-1}) \\ ante \in C_f(s-1, U_{DS[s-1.end]}) \\ (\exists v \in \mathbb{N} : ante =_{str} C_p(v, U_{DS[v.end]}) \\ \wedge v < (s-1)) \\ \wedge (\neg \exists v' \in \mathbb{N} : ante =_{str} C_p(v', U_{DS[v'.end]}) \\ \wedge v < v') \end{array} \right.$

Table 4: Reachability of the Anaphoric Antecedent

Finally, the function $Lift(s, i)$ (cf. Table 5) determines the appropriate discourse segment level, s , of an utter-

ance U_i (selected by its linear text index, i). $Lift$ only applies to structural configurations in the centering lists in which themes continuously shift at three different consecutive segment levels and associated preferred centers at least (cf. Table 2, lower box, for the basic pattern).

$Lift(s, i) :=$	
$\left\{ \begin{array}{l} Lift(s-1, i-1) \text{ if } \\ s \text{ else} \end{array} \right.$	$\left\{ \begin{array}{l} s > 2 \wedge i > 3 \\ \wedge C_p(s, U_{i-1}) \neq C_p(s-1, U_{i-2}) \\ \wedge C_p(s-1, U_{i-2}) \neq C_p(s-2, U_{i-3}) \\ \wedge C_p(s, U_{i-1}) \in C_f(s-1, U_{i-2}) \end{array} \right.$

Table 5: Lifting to the Appropriate Discourse Segment

Whenever a discourse segment is created, its starting and closing utterances are initialized to the current position in the discourse. Its end point gets continuously incremented as the analysis proceeds until this discourse segment DS is *ultimately closed*, i.e., whenever another 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. In Table 8, e.g., the first two discourse segments at level 3 (ranging from U_5 to U_5 and U_8 to U_{11}) are closed, while those at level 1 (ranging from U_1 to U_3), level 2 (ranging from U_4 to U_7) and level 3 (ranging from U_{12} to U_{13}) are open.

The main algorithm (see Table 6) consists of three major logical blocks (s and U_i denote the current discourse segment level and utterance, respectively).

1. **Continue Current Segment.** The $C_p(s, U_{i-1})$ is taken over for U_i . If 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 by the function $Lift$ to a higher level s' . As a result of lifting, the entire sequence (including the final two utterances) forms a single segment. This is trivially true for cases of a constant theme.
2. **Close Embedded Segment(s).**
 - (a) *Close the embedded segment(s) and continue another, already existing segment:* If U_i does not include any anaphoric expression which is an element of the $C_f(s, U_{i-1})$, then match the antecedent in the hierarchically reachable segments. Only the C_p of the utterance at the end point of any of these segments is considered a potential antecedent. Note that, as a side effect, hierarchically lower segments are ultimately closed when a match at higher segment levels succeeds.
 - (b) *Close the embedded segment and open a new, parallel one:* If none of the anaphoric expressions under consideration co-specify the

²The C_f lists in the functional centering model are *totally ordered* (Strube & Hahn, 1996, p.272) and we here implicitly assume that they are accessed in the total order given.

$C_p(s - 1, U_{[s-1.end]})$, then the entire C_f at this segment level is checked for the given utterance. If an antecedent matches, 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 of the preceding parallel segments from the search for a valid antecedent and just visit the currently open one.

- (c) *Open new, embedded segment*: If there is no matching antecedent in hierarchically reachable segments, then for utterance U_i a new, embedded segment is opened.
3. **Open New, Embedded Segment**. If none of the above cases applies, then for utterance U_i a new, embedded segment is opened. In the course of processing the following utterances, this decision may be retracted by the function *Lift*. It serves as a kind of “garbage collector” for globally insignificant discourse segments which, nevertheless, were reasonable from a local perspective for reference resolution purposes. Hence, the centered discourse segmentation procedure works in an incremental way and revises only locally relevant, yet globally irrelevant segmentation decisions on the fly.

```

s := 1
i := 1
DS[s.begin] := i
DS[s.end] := i
while ¬ end of text
  i := i + 1
  R := {Resolved(x, s, U_i) | x ∈ U_i}
  if ∃ r ∈ R : r = str C_p(s, U_{i-1})
    then s' := s
        i' := i
        DS[Lift(s', i').end] := i
    else if ¬∃ r ∈ R : r ∈ C_f(s, U_{i-1})
      then found := FALSE
          k := s
          while ¬found ∧ (k > 1)
            k := k - 1
            if ∃ r ∈ R : r = str C_p(k, U_{[k.end]})
              then s := k
                  DS[s.end] := i
                  found := TRUE
            else if k = s - 1
              then if ∃ r ∈ R : r ∈
                    C_f(k, U_{[k.end]})
                  then DS[s.begin] := i
                      DS[s.end] := i
                      found := TRUE
          if ¬found
            then s := s + 1
                DS[s.begin] := i
                DS[s.end] := i
          else s := s + 1
              DS[s.begin] := i
              DS[s.end] := i

```

Table 6: Algorithm for Centered Segmentation

4 A Sample Text Segmentation

The text with respect to which we demonstrate the working of the algorithm (see Table 7) is taken from a German computer magazine (*c't*, 1995, No.4, p.209). For ease of presentation the text is somewhat shortened. Since the method for computing levels of discourse segments depends heavily on different kinds of anaphoric expressions, (pro)nominal anaphors and textual ellipses 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 entire text.

The centered segmentation analysis of the sample text is given in Table 8. The first column shows the linear text index of each utterance. The second column contains the centering data as computed by functional centering (Strube & Hahn, 1996). The first element of the C_f , the *preferred center*, C_p , is marked by bold font. The third column lists the centering transitions which are derived from the C_b/C_f data of immediately successive utterances (cf. Table 1 for the definitions). The fourth column depicts the levels of discourse segments which are computed by the algorithm in Table 6. Horizontal lines indicate the beginning of a segment (in the algorithm, this corresponds to a value assignment to $DS[s.begin]$). Vertical lines show the extension of a segment (its end is fixed by an assignment to $DS[s.end]$). The fifth column indicates which block of the algorithm applies to the current utterance (cf. the right margin in Table 6).

The computation starts at U_1 , the headline. The $C_f(U_1)$ 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), so *Lift* does not apply. The C_p is set to “1260” in all utterances of this segment. Since U_4 does neither contain any anaphoric expression which co-specifies the $C_p(1, U_3)$ (block (1)) nor any other element of the $C_f(1, U_3)$ (block (2a)), and as there is no hierarchically preceding segment, 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 phrase “*das dünne Handbüchlein*” (*the thin leaflet*) in U_5 does not co-specify the $C_p(2, U_4)$ but co-specifies 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 anaphor “*Handbuch*” (*manual*) in U_6 co-specifies 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 function *Lift* lifts the embedded segment one level, so the

(1) Brother HL-1260	(8) Kein Wunder: unter dem <i>Inhaltsverzeichnis</i> steht der lapidare Hinweis, man möge sich die Seiten dieses Kapitels doch bitte von Diskette ausdrucken – Frechheit. No wonder: beneath the <i>table of contents</i> – one finds the terse instruction, one should – oneself – the pages of this section – please – from disk – print out – – impertinence.
(2) Ein Detail fällt schon beim ersten Umgang mit dem großen <i>Brother</i> auf: One particular – is already noticed – in the first approach to – the big <i>Brother</i> .	(9) Ohne diesen <i>Ausdruck</i> sucht man vergebens nach einem Hinweis darauf, warum die <i>Auto-Continue-Funktion</i> in der <i>PostScript-Emulation</i> nicht wirkt. Without this <i>print-out</i> , looks – one – in vain – for a hint – why – the <i>auto-continue-function</i> – in the <i>PostScript emulation</i> – does not work.
(3) Im Betrieb macht <i>er</i> durch ein kräftiges Arbeitsgeräusch auf sich aufmerksam, das auch im Stand-by-Modus noch gut vernehmbar ist. In operation – draws – <i>it</i> – with a heavy noise level – attention to itself – which – also – in the stand-by mode – is still well audible.	(10) Nach dem Einschalten zeigt das <i>LC-Display</i> an, daß diese praktische <i>Hilfsfunktion</i> nicht aktiv ist; After switching on – depicts – the <i>LC display</i> – that – this practical <i>help function</i> – not active – is;
(4) Für Standard-Installationen kommt man gut ohne Handbuch aus. As far as standard installations are concerned – gets – one – well – by – without any manual.	(11) <i>sie</i> überwacht den Dateientransfer vom Computer. <i>it</i> monitors the file transfer from the computer.
(5) Zwar erläutert das dünne <i>Handbüchlein</i> die Bedienung der <i>Hardware</i> anschaulich und gut illustriert. Admittedly, gives – the thin <i>leaflet</i> – the operation of the <i>hardware</i> – a clear description of – and – well illustrated.	(12) Viele der kleinen Macken verzeiht man dem <i>HL-1260</i> wenn man erste Ausdrücke in Händen hält. Many of the minor defects – pardons – one – the <i>HL-1260</i> , when – one – the first print outs – holds in [one's] hands.
(6) Die <i>Software-Seite</i> wurde im <i>Handbuch</i> dagegen stiefmütterlich behandelt: The <i>software part</i> – was – in the <i>manual</i> – however – like a stepmother – treated:	(13) Gerasterte Graufächen erzeugt der <i>Brother</i> sehr homogen ... Raster-mode grey-scale areas – generates – the <i>Brother</i> – very homogeneously ...
(7) bis auf eine karge <i>Seite</i> mit einem Inhaltsverzeichnis zum <i>HP-Modus</i> sucht man vergebens weitere Informationen. except for one meagre <i>page</i> – containing the table of contents for the <i>HP mode</i> – seeks – one – in vain – for further information.	

Table 7: Sample Text

segment which ended with U_4 is now continued up to U_6 at level 2. As a consequence, the centering data of U_5 are excluded from further consideration as far as the co-specification by any subsequent anaphoric expression is concerned. U_7 simply continues the same segment, since the textual ellipsis “Seite” (*page*) refers to “Handbuch” (*manual*). The utterances U_8 to U_{10} exhibit a typical thematization-of-the-rhemes pattern which is quite common for the detailed description of objects. (Note the sequence of SHIFT transitions.) Hence, block (3) of the algorithm applies to each of the utterances and, correspondingly, new segments at the levels 3 to 5 are created. This behavior breaks down at the occurrence of the anaphoric expression “sie” (*it*) in U_{11} which co-specifies the $C_p(5, U_{10})$, viz. “auto-continue function”, denoted by another anaphoric expression, namely “Hilfsfunktion” (*help function*) in U_{10} . Hence, block (1) applies. The evaluation of *Lift* succeeds with respect to two levels of embedding. As a result, the whole sequence is lifted up to level 3 and continues this segment which started at the discourse element “Inhaltsverzeichnis” (*list of contents*). As a result of applying *Lift*, the whole sequence is captured in one segment. U_{12} does not contain any anaphoric expression which co-specifies

an element of the $C_f(3, U_{11})$, hence block (2) of the algorithm applies. The anaphor “HL-1260” does not co-specify the C_p of the utterance which represents the end of the hierarchically preceding discourse segment (U_7), but it co-specifies an element of the $C_f(2, U_7)$. The immediately preceding segment is ultimately closed and a parallel segment is opened at U_{12} (cf. block (2b)). Note also that the algorithm does not check the $C_f(3, U_{10})$ despite the fact that it contains the antecedent of “1260”. However, the occurrences of “1260” in the C_f s of U_9 and U_{10} are mediated by textual ellipses. If these utterances contained the expression “1260” itself, the algorithm would have built a different discourse structure and, therefore, “1260” in U_{10} were reachable for the anaphor in U_{12} . Segment 3, finally, is continued by U_{13} .

5 Empirical Evaluation

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 Ger-

U_i	Centering Data	Trans.	Levels of Discourse Segments					Block
			1	2	3	4	5	
(1)	Cb: – Cf: [1260]	—						
(2)	Cb: 1260 Cf: [1260, Umgang, Detail]	C						1
(3)	Cb: 1260 Cf: [1260, Betrieb, Arbeitsgeräusch, Stand-by-Modus]	C						1
(4)	Cb: – Cf: [Standard-Installation, Handbuch]	—						2c
(5)	Cb: Handbuch Cf: [Handbuch, 1260, Hardware, Bedienung]	C						3
(6)	Cb: Handbuch Cf: [Handbuch, 1260, Software]	C						1, Lift
(7)	Cb: Handbuch Cf: [Handbuch, Seite, 1260, HP-Modus, Inhaltsverzeichnis, Informationen]	C						1
(8)	Cb: Inhaltsverzeichnis Cf: [Inhaltsverzeichnis, Hinweis, Seiten, Kapitel, Diskette, Frechheit]	SS						3
(9)	Cb: Kapitel Cf: [Kapitel, Ausdruck, Hinweis, 1260, Auto-Continue-Funktion, PostScript-Emulation]	SS						3
(10)	Cb: 1260 Cf: [Auto-Continue-Funktion, 1260, LC-Display]	RS						3
(11)	Cb: Auto-Continue-Funktion Cf: [Auto-Continue-Funktion, Dateien-Transfer, Computer]	SS						1, Lift
(12)	Cb: – Cf: [1260, Macken, Ausdruck]	—						2b
(13)	Cb: 1260 Cf: [1260, Graufächen]	C						1

Table 8: Sample of a Centered Text Segmentation Analysis

man news magazine (Spiegel)³, and of two literary texts⁴ (Lit). Table 9 summarizes the total numbers of anaphors, textual ellipses, utterances, and words in the test set.

	IT	Spiegel	Lit	Σ
anaphors	197	101	198	496
ellipses	195	22	23	240
utterances	336	84	127	547
words	5241	1468	1610	8319

Table 9: Test Set

Table 10 and Table 11 consider the number of anaphoric and text-elliptical expressions, 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

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

⁴The 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.)

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 anaphors (28%) and 116 textual ellipses (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
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 10: Anaphoric Antecedent in Utterance U_x

Table 12 and Table 13 give the success rate of the centered segmentation algorithm for anaphors and textual ellipses, respectively. The numbers in these tables indicate at which segment level anaphors and textual ellipses were correctly resolved. The category of errors

	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 11: Elliptical Antecedent in Utterance U_x

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 12 and 13 contain more elements than the categories U_{i-1} in Tables 10 and 11, respectively, due to the mediating property of textual ellipses 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 12: 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 13: Elliptical 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.

6 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 *prefixed size* corresponding to the limited 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.

Text segmentation procedures (more with an information retrieval motivation, rather than being related to reference resolution tasks) have also been proposed for a coarse-grained partitioning of texts into contiguous, non-overlapping blocks and assigning content labels to these blocks (Hearst, 1994). The methodological basis of these studies are lexical cohesion indicators (Morris & Hirst, 1991) combined with word-level co-occurrence statistics. Since the labelling is one-dimensional, this approximates our use of preferred centers of discourse segments. These studies, however, lack the fine-grained information of the contents of C_f lists also needed for proper reference resolution.

Finally, many studies on discourse segmentation highlight the role of cue words for signaling segment boundaries (cf., e.g., the discussion in Passonneau & Litman (1993)). However useful this strategy 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.

7 Conclusions

We have developed a proposal 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 potential discourse referents. Since we operate within a functional framework, this study also presents one of the rare formal accounts of thematic progression patterns for full-fledged texts which were informally introduced by Daneš (1974).

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 of 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, the three anonymous reviewers for their critical comments, 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., 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.
- 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., K. Markert & M. Strube (1996). A conceptual reasoning approach to textual ellipsis. In *Proc. of the 12th European Conference on Artificial Intelligence (ECAI '96); Budapest, Hungary, 12-16 August 1996*, pp. 572-576. Chichester: John Wiley.
- Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics; Las Cruces, N.M., 27-30 June 1994*, pp. 9-16.
- Morris, J. & G. Hirst (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21-48.
- 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.
- 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.