Inferable Centers, Centering Transitions, and the Notion of Coherence

Laurel Fais*
NTT Communication Science Laboratories

A centering analysis of the corpus of Japanese e-mail that is examined in this article relies heavily on the inclusion of inferable centers. However, utilizing this type of center results in a high level of indeterminacy in labeling transitions and thus in characterizing the coherence of the corpus. The difficulty lies in the requirement of identity of discourse entities in the definitions of transition states. Lexical cohesion is proposed as a well-defined notion to replace the intuitions captured by the use of inferable centers. Two new transitions, based on lexical relatedness instead of identity, supplement the standard definitions and more adequately characterize coherence in this corpus. Implications and extensions of the proposal are discussed.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

1. Introduction

Centering has been proposed as a model of the local attentional states of speakers and hearers involved in the mutual construction of conversation (Brennan, Friedman, and Pollard 1987; Grosz and Sidner 1986, 1998; Walker 1998). Centering mechanisms are designed to model the coherence of discourse by characterizing transitions between utterances in terms of their inferential load and hence their naturalness. These characterizations are intended to capture intuitions about the "flow" (Chafe 1979) or the "ongoing process of meaning" (Halliday 1994) in discourse.

In this work, we examine a corpus of Japanese e-mail to investigate the mechanisms by which coherence is achieved. Because this corpus contains a high number of discourse elements that are inferable from the discourse context, we have an opportunity to examine the interplay between standard centering transition definitions and the presence of inferable discourse entities. We claim on the basis of intuitions of native speakers that the actual level of coherence in the corpus is much higher than the centering account implies, primarily by virtue of the fact that transitions involving inferable entities are often difficult to specify. We conclude that the standard centering account cannot accurately model the coherence in this corpus. Detailed analysis reveals that one major problem lies in the requirement of identity of discourse elements in adjacent utterances in order for those elements to contribute to coherence. We describe this problem and propose two additions to the usual repertoire of transitions that enable a more authentic account of coherence in this corpus, while remaining within a centering framework.

The article is organized as follows. In Section 2, we briefly describe centering mechanisms and their role in modeling coherence. We go on to outline the features of the corpus in Section 3 and illustrate how standard centering mechanisms characterize transitions and coherence in this corpus, suggesting that these mechanisms are not adequate for the task. In Section 4, we describe more general problems with the

* Infant Studies Centre, University of British Columbia, Room 1401, 2136 West Mall, Vancouver, British Columbia, V6T 1Z4 Canada. E-mail: jwlab@psych.ubc.ca.

© 2004 Association for Computational Linguistics

inclusion of inferable discourse entities in centering theory and propose a revision to the standard set of transitions that more accurately describes the corpus. In this section as well, we explore the implications of this proposal for other areas of discourse analysis. In Section 5, we outline some possibilities for improvement and extension of the proposal, and we conclude in the final section.

2. Centering Mechanisms

The central intuitions of centering concern the relationships among the discourse entities appearing or represented in adjacent utterances in a discourse (Walker, Joshi, and Prince 1998). Each utterance U_i in a discourse is considered to contain a set of discourse entities called forward-looking centers, or Cfs. These entities are ranked in the Cf list for each utterance according to language-specific ranking principles. We follow, in general, the ordering principles for Japanese given in Walker, Iida, and Cote (1994) (with some adjustments for possessive phrases as noted in example (1)):

> Cf ranking for Japanese:
> (Grammatical OR $\emptyset$) topic $>$ empathy $>$ subject $>$ object2 $>$ object $>$ others

A special member of the Cf list, the backward-looking center, or Cb, represents the "topic"[1] of U_i and is the highest-ranked Cf on the Cf list of U_{i-1} which is realized in U_i. In addition, the preferred center of U_i, or Cp, is the highest-ranked Cf in U_i.

Given U_i and U_{i-1}, then, there are four different ways in which their Cbs and Cps may be related; each of these is defined as a type of transition state (Table 1).

There are two rules in a centering approach:

Rule 1: If some entity in the Cf list for U_{i-1} is realized as a pronoun in U_i, then so is the Cb for U_i.

Rule 2: Transition states are ordered such that CONTINUE is most preferred, followed in order by RETAIN, SMOOTH SHIFT, and ROUGH SHIFT (Walker, Joshi, and Prince 1998).

Rule 2 captures the centering intuitions concerning coherence: Utterances that CONTINUE the topic of a previous utterance in a prominent position impose a lower inferential load, and are thus more coherent, than utterances which relegate the topic to less prominent positions or which change the topic.

The vast majority of the sentences in our corpus are complex sentences. Thus, the question of how to interpret centering principles in complex sentences cannot be ignored. We will consider the basic utterance unit of centering to be the tensed clause

Table 1
Transition definitions.

	Cb(U_i) = Cb(U_{i-1}) OR Cb(U_i) = ?	Cb(U_i) $\neq$ Cb(U_{i-1})
Cb(U_i) = Cp(U_i)	CONTINUE	SMOOTH SHIFT
Cb(U_i) $\neq$ Cp(U_i)	RETAIN	ROUGH SHIFT

[1] Whenever the word *topic* is used in this article, we mean the general notion of topic as the subject of a discourse unit, not the specialized meaning of topic/comment.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

and will assume that these clauses form a flat, linear sequence of discourse units, such that the centering output of the first clause in the sentence is the input to the next, and so on, in the spirit of Kameyama (1998) and Suri and McCoy (1994).[2]

Because the notion of *backward-looking center* will be critical to the discussion that follows, we look at this notion in greater detail here. According to Grosz, Joshi, and Weinstein (1986, quoted in Walker, Joshi, and Prince [1998]), a center is realized in an utterance *U* if it "is an element of the situation described by *U* or the semantic interpretation of some subpart of *U*" (page 4). As Walker, Joshi, and Prince (1998) point out, this covers "pronouns, zero pronouns, explicitly realized discourse entities, and... entities inferable from the discourse situation" (page 4). The definition proposed by Grosz, Joshi, and Weinstein allows inferable entities, that is, entities that are not expressed at the surface level of the utterance or immediately recoverable from the subcategorization properties of the verb (as, for example, zero pronouns are) to constitute centers of an utterance. However, the theory does not make explicit the parameters within which to characterize the class of permissible inferable elements or the constraints on doing so. We will return to this difficulty later.

3. The Corpus

3.1 The Nature of the Corpus

The data examined consist of a collection of 32 e-mail messages exchanged among five employees of a Japanese company from June 5 to June 16, 1995. The messages constitute a collective attempt to schedule a sports-watching outing convenient for and interesting to all five in the group. Thus, the tone is usually casual. The authors combine standard aspects of written text with various strategies for encoding speech-like information in the messages: nonstandard uses of punctuation, katakana (the syllabary for writing foreign words), and English; nonstandard spelling; emoticons; discourse markers, sentence-final particles, tense, and formality typical of speech; and fillers (Fais 2001; Fais and Yamura-Takei 2003). Quantitative information for the corpus is given in Table 2.

Table 2
Number of messages, paragraphs, sentences, clauses, and characters per author.

Authors	Messages	Paragraphs	Sentences	Clauses	Characters
H	5	9	20	60	880
I	10	46	50	122	1,904
M	5	27	39	84	1,268
R	5	29	56	125	1,805
U	7	35	60	127	2,015
Total	32	146	225	518	7,872

[2] Kameyama (1998) also mentions two cases in which a hierarchical interpretation is warranted: reported speech and nonreport complements. Our corpus does not contain examples of these kinds of utterances. On the other hand, it does contain tensed clauses acting as "relative clauses" and as verbal complements. We have not separated these from their heads or matrix clauses. This has no effect on the analysis presented in this article, except for the fact that had we separated these clauses, it would have made the problematic situation we describe later even more marked. The status of these clauses vis-à-vis centering is a topic in need of extensive investigation.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

3.2 Inferables and Transition States in the Corpus

Example (1) illustrates how centering mechanisms apply to the beginning of one of the messages and how the CONTINUE and RETAIN transitions can model coherence:

1a. 私	は[3]	18日	の	ヤクルトー横浜戦	は	Cb = (beginning)
watashi	wa	18nichi	no	yakuluto-yokohamasen	wa	Cf = Swallows
me	TOP	18th	on	Swallows vs. Baystars game	TOP	vs. Baystars game,[4]
						18th, enthusiasm

もともと	乗り気	だった	ので	
motomoto	noriki	datta	node	CONTINUE
actually	enthusiasm	was	since	

1b. ＯＫです。	Cb = SvsB game ∅
∅ OK desu	Cf = SvsB game ∅
∅ OK is	
"As for me, since I'm really up for the Swallows vs.	
Baystars game on 18th, that is OK with me."	RETAIN

1c. でも、	テコンドー	に	興味津々	だった	U-さん	や	Cb = it = Swallows
demo	tekondoh	ni	kyoumishinshin	datta	U-san	ya	vs. Baystars game
but	taekwondo	in	keen interest	was	U-san	and	Cf = U-san, M-san, it,
							interest, taekwondo

M-さんは	これでいいんでしょうか？
M-san wa	koredeiindeshouka
M-san TOP	I'm wondering if it is ok

"But I'm wondering if it is ok for U-san and M-san, who have shown a keen interest in taekwondo."

Clause (1a) has no Cb, since it occurs at the beginning of the message. The Cp of (1a) is *Swallows vs. Baystars game*, by virtue of the fact that it is topic-marked. The subject of the copular *desu* in (1b) is omitted; rule 1 implies that the referent for this zero argument is *Swallows vs. Baystars game*, which is a correct assignment. A similar process resolves the referent for *kore* 'this' in (1c). Note that the CONTINUE and RETAIN transitions do, in fact, capture the intuition that this segment of the message is coherent and is "about" the Swallows vs. Baystars game.

Example (2) is much more typical of the messages in the corpus and is not as well behaved as (1). Notice that all the Cbs in this example are inferable "from the discourse situation" of U_{i-1}. The preponderance of inferable Cbs is typical; out of 330 Cbs in the corpus, 250 (more than 75%) are entities *other than* pronouns, zero

[3] We will ignore first-person arguments in this analysis, since centering is intended to handle only third-person arguments. How topic-marked first-person arguments affect attentional states in discourse is an interesting question, though beyond the scope of this article.

We use the following abbreviations for Japanese case markers in this article: TOP, topic; SUBJ, subject; OBJ, object.

[4] There is no consensus as to the ordering of the arguments in a Japanese *A no B* construction (roughly equivalent to possessives; Tetreault 2001; Matsui 1999; but see Fais 2002). We have listed the arguments in the order suggested in Fais (2002); this has little effect on the present discussion.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

arguments, or explicitly realized entities.[5] This raises a question undiscussed in the centering literature: How do we interpret inferable Cbs in the context of assigning transition types? Example (2) illustrates the problems involved:

2a.	ところで	宴席 なん	ですが,
	tokorode	enseki nan	desuga
	by the way	restaurant	is

Cb = ?
Cf = restaurant

2b.	千駄ヶ谷	近 辺	に	酒	を	飲 める
	sendagaya	kinpen	ni	sake	wo	nomeru
	sendagaya	neighborhood	in	alcohol	OBJ	can drink

Cb = ?
Cf = soba shop, Sendagaya neighborhood, alcohol

旨い	蕎麦屋	が	あると	耳にしました.
umai	sobaya	ga	aruto	miminishimashita
good	soba shop	SUBJ	is	have heard

"By the way, about the restaurant, I've heard there is a good soba shop, which also serves alcohol, around Sendagaya."

CONTINUE?
SMOOTH SHIFT?

2c.	閉店	時刻	が	早いそうなので、
	heiten	jikoku	ga	hayaisounanode
	shop closing time		SUBJ	early seems because

Cb = shop closing time (inferable)
Cf = shop closing time

SMOOTH SHIFT

2d.	ちょっと	まずい	かもしれません	が
∅	chotto	mazui	kamoshiremasen	ga
∅	a little	bad	might be	but

Cb = choosing this restaurant ∅ (inferable)
Cf = choosing this restaurant ∅

2e.	I-さんあたり	ご存じでは	ないでしょうか？
	I-sanatari	gozonjidewa	naideshouka
	I-san	information	have

"Choosing this restaurant may not be good because it closes early. I-san, do you have information about this restaurant?"

ROUGH SHIFT

Cb = information (about restaurant) (inferable)
Cf = I-san, information (about restaurant)

The Cb of (2a) is null, since that clause is the first in the discourse segment. Clause (2b) has three centers, listed in the Cf list. All three of these entities are inferable from the discourse context of (2a). We mentioned previously that there is no principled way to determine the list of inferables of an utterance; it is even more difficult, then, to

[5] In the subsequent discussion, for the sake of simplicity, we will refer to this group as "explicit" centers, realizing that zero arguments are not precisely "explicit," but setting them off in this way from inferable centers.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

Table 3
Transitions occurring in the corpus, by type.

Transition type	Number	Percentage
CONTINUE	54	16.4
RETAIN	22	6.7
SMOOTH SHIFT	2	0.6
ROUGH SHIFT	2	0.6
Transition to inferable Cb	138	41.8
NULL	112	33.9
Total	330	

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

determine the order in which inferables should be listed on a Cf list. Therefore, we cannot say which of the Cfs in (2b) is the "highest ranked" and thus the Cb for (2b). Since we cannot determine which is the Cb, we likewise cannot determine whether the Cb of (2b) is the same as its Cp, and so we cannot label the transition at all.

The fact that there is only one Cf for (2c) simplifies the problem somewhat. That Cf, which is inferable from (2b), must also be the Cb of (2c), but since we could not ascertain the Cb of (2b), we do not know if the transition from (2b) to (2c) is a CONTINUE or a SMOOTH SHIFT. Again, the presence of only one Cf in (2d) makes matters easier; we are able to label the transition from (2c) to (2d) a SMOOTH SHIFT.

We surmise that *I-san* may not be inferable from the discourse context,[6] though *information (about the restaurant)* is, and thus the latter becomes the Cb of (2e). Because this Cb is neither the same as the Cb of (2d) nor the same as its own Cp, the transition from (2d) to (2e) is a ROUGH SHIFT.

Using the standard centering definitions, supplemented with the notion of NULL transitions to label transitions to utterances that contain no Cb, we hand-tabulated the number of transitions of each type occurring in the corpus (Table 3). The figures for CONTINUE, RETAIN, SMOOTH SHIFT, and ROUGH SHIFT are those for transitions to utterances containing either explicit, pronominal, or zero-argument Cbs. Given the difficulties in accurately labeling transitions to utterances containing inferable Cbs, we grouped these latter together separately from transitions to utterances with explicit, pronominal, or zero-argument Cbs.

There are two points of special interest in Table 3. First, of course, is the particularly high number of transitions to utterances containing inferable Cbs and the number of NULL transitions. At over 40%, the utterances involved in transitions to utterances containing inferable Cbs make up a substantial portion of the corpus. Second are the relative proportions of CONTINUE, RETAIN, SMOOTH SHIFT, and ROUGH SHIFT transitions. Note that, considering just these transitions to utterances containing explicitly realized centers, these proportions are roughly what we expect of coherent text: Most of the shifts are CONTINUE, followed by a respectable number of RETAINs, and a very few SHIFTs.

[6] Of course, in the absence of any principled way to determine this, this is merely conjecture. In fact, because I-san is a part of the ongoing e-mail exchange, he could possibly be considered part of the discourse context. The difficulty is, of course, the lack of any rigorous way to determine what the elements in the discourse context are.

4. Lexical Cohesion and Discourse Coherence: A Revision of Standard Transition Types

4.1 Inferable Centers in Centering Theory

We saw in Section 3.2 that the introduction of inferable centers into a centering analysis leads to indeterminate transition identification. We also saw that transitions to utterances with inferable centers make up a large proportion (over 40%) of the transitions in this corpus. It is important, then, if a centering analysis is to represent the nature of coherence in this corpus accurately, that we make principled provisions for the notion captured by inferable centers in the theory.

Only rarely are inferable entities actually listed in analyses in the literature, and then usually when their presence is supported by previous explicit mention and by a clear, tight semantic or syntactic relationship between the entities involved, as in (3), taken from (17) in Kameyama (1998):

3a. It is the apparent intention of the Republican Party to campaign on the carcass of what they call Eisenhower Republicanism.	Cb = Republican Party
3b. but the heart stopped beating	Cb = Republican Party (inferable Possessor of *the heart*)
3c. and the lifeblood congealed	Cb = Republican Party (inferable Possessor of the *lifeblood*)

Note that (3) avoids at least one of the difficulties we encountered in (2); when there is only one center to deal with, as there is in (2c) and (2d) and (3b) and (3c), the choice of Cb is trivial. However, in cases such as (2a) and (2e), or in (3b) if it had been something like *but the heart and lifeblood stopped pumping*, we cannot determine what the Cb should be. Further, in these cases, it is impossible to identify the appropriate transition to the following utterance; when the Cb of U_{i-1} is undetermined (as it is for (2b)), the transition to U_i (in this case, (2c)) is undeterminable.

4.2 Logical Difficulties with Inferable Centers

There are further problems resulting from taking the use of inferable centers to its logical extreme. In our analysis so far, we have been concerned with explicit entities in U_i that realize centers in U_{i-1} that are inferable from the discourse context. To be accurate, those inferable centers need to be listed in the Cf list for U_{i-1}. However, if we process discourse incrementally, this leads to the conclusion that since we do not know which inferable entity from U_{i-1} will be evoked in U_i, we need to list every inferable entity in the Cf list of U_{i-1}. This is both computationally untenable and, in view of the lack of any parameters for determining what constitutes an allowable inferable Cb, impossible. Even if it were possible and desirable, how could we define for inferable entities the type of grammatical-role information essential to determining the placement of these entities on a Cf list?

Again taking the definition of inferable centers to its extreme, we note another problem. Every utterance, not just U_{i-1}, evokes inferable centers. There is nothing in the theory to preclude a situation such as that shown in (4) (a version of (2) simplified for illustrative purposes) (implicit inferable centers are given in italics):

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

4a. 千駄ヶ谷 近 辺 に	Cb = ?
sendagaya kinpen ni	Cf = soba shop,
sendagaya neighborhood in	Sendagaya
	neighborhood,
旨い 蕎麦屋 が あると 耳にしました.	*udon shop, sake*
umai sobaya ga aruto miminishimashita	*shop, shop clerk,*
good soba shop SUBJ is have heard	*menu, food,*
"I've heard there is a good soba shop around Sendagaya."	*customers. . .*
4b. でも 閉店 時刻 が 早いそうなんです。	Cb = *shop clerk*
demo heiten jikoku ga hayaisounandesu	Cf = shop closing time,
but shop closing time SUBJ is early	*shop opening time,*
"But the shop closes early."	*shop, shop clerk,*
	door. . .

In (4), an inferable center from the Cf list of (4b) matches an inferable center from the Cf list of (4a) and is chosen as the Cb. Of course, there could be numerous identical inferable centers on the Cf lists of U_i and U_{i-1}, all "vying" for Cb of U_i. Although this is obviously an absurd extension of the inclusion of inferable centers in centering theory, there is nothing, unfortunately, in the theory itself to rule it out.

4.3 Possible Solutions to the Problems of Inferable Centers

4.3.1 Inferable Centers as Bridging References. Inferable centers are similar to bridging references (Clark 1977); they have a conceptual relationship to entities in a previous utterance. There is a sense that bridging references should participate in the creation of coherence in a discourse (Hahn, Markert, and Strube 1996). But the work on bridging references characterizes this relationship as referential or anaphoric; this can be seen in the various terms under which this phenomenon is discussed: bridging references, indirect anaphora, functional anaphora, and partial anaphora. Bridging references, however, unlike the usual case of anaphora, may be mediated not only by a strict identity condition, but also by any number of other semantic relationships (is-a, has-a, made-of, at-time, etc.).

Unfortunately, the establishment of the semantic relationship between an anchor and its bridging reference is notoriously difficult. Poesio et al. (2000), even after severely restricting the types of relationships to be labeled, had extremely poor interlabeler reliability on a first pass. Every account in which bridging references are addressed restricts allowable relationships to a small, relatively well-defined set (Vieira and Poesio 2001; Poesio et al. 2000; Murata, Isahara, and Nagao 1999; Strube and Hahn 1999). Cote (1998) proposes the use of lexical-conceptual primitives instead of grammatical relations in Cf templates and suggests that the conceptual information that this approach provides might be rich enough to supply part-whole information necessary to the resolution of bridging references. She points out as well, however, that a number of other types of semantic relationships manifested in bridging references would *not* be identifiable from lexical-conceptual information. Thus, although work on bridging references has attempted to provide a characterization of the possible semantic relationships involved, what success has been achieved is limited to a small subset of cases.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

4.3.2 Restrictions on the Notion of Inferable Center. One possible way around the logical problems introduced by inferable centers is to take into account only inferable centers that are explicitly realized in an utterance when determining the Cb of that utterance. This would mean, for example, considering *soba shop*, *Sendagaya neighborhood*, and *alcohol* from (2b) in choosing the Cb for (2c). But how do we make the choice among these possibilities? Any of these three different, explicit, inferable centers could be chosen to be the Cb. But this is exactly the difficulty: We have no principled way to make such a choice. We have no way of knowing which of these inferable centers is ranked highest in the Cf list for (2a) so that we can select that Cf to be the Cb of (2b).

A second possibility is to allow only explicitly realized (i.e., noninferable) centers. This seems to be the approach taken by Passonneau (1998); her definition of a null Cb seems to imply that a Cb must be (noninferable and) explicitly realized, and her null Cbs constitute the cases in which there is no explicit Cb. In her examination of the Pear Stories (recordings of people describing to another person a movie they had seen; Chafe 1980), NULL transitions (transitions to an utterance with a null Cb) represent the majority of transitions. Although her concern is discourse segmentation, Passonneau does note that the patterning of transition types does not accurately reflect the coherence of the stories.

Allowing only explicit centers would mean, for the corpus studied in this article, that inferable Cbs become null Cbs and the proportion of NULL transitions becomes 75.7%. Under this assumption, this corpus would be characterized as extremely incoherent, a claim belied both by native-speaker intuitions (an acceptable level of coherency for these texts was confirmed by three native speakers) and also by the fact that the task that was the central concern of these messages was successfully completed; the group exchanged a number of opinions and pieces of information and came to a consensus regarding their sports outing, with no message showing confusion about information contained in previous messages. Thus, the solution of allowing only explicit centers does not yield an accurate characterization of the coherence of this corpus.

Hurewitz (1998) chose to define allowable inferable Cbs fairly narrowly in her English data, requiring functional dependency or a poset relationship to hold in order for a Cb to be recognized. Even with this definition, which is more constrained than we have taken "inferable" to be in our previous discussion, she finds that 21% of the Brown corpus (a variety of written texts) and 28% of the switchboard corpus (taped telephone conversations) consist of what she calls a no-Cb condition. Poesio et al. (2000) report a similarly high proportion of nonexplicit Cbs in their English text corpora. They test a number of configurations of parameters of centering theory to attempt to minimize the number of null Cbs (reasoning that the best configuration of parameters would result in the fewest violations of the constraints of the theory, in this case, the constraint that all utterances in the discourse except for the first have at least one Cb). One way in which they are able to improve their results significantly is by allowing a restricted set of three types of nonidentity relationships between centers, that is, by recognizing three types of well-defined inferables. However, simply limiting the type of inferables allowed still does not address the issue of the indeterminacy of transitions to utterances containing inferable Cbs. And the central question raised by the high number of nonexplicit Cbs found in naturally occurring texts remains unaddressed: How can we characterize coherence in a text in which Cbs are so often inferable and thus in which transition types are often indeterminate?

The crux of the problem lies in the application of standard centering processes to inferable centers. In a nonproblematic case, a Cb in U_i is recognized by virtue of its

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

identity to a Cf in U_{i-1}. This is (relatively) straightforward in the case of explicit centers. But now apply this process to inferable centers. In a standard centering approach, a Cb in U_i can also be recognized by virtue of its identity to an inferable center in U_{i-1}. This implies that we must somehow make available all the possible inferable centers in U_{i-1} in order to recognize (possibly) one of them as the Cb for U_i. We have already noted that this position is untenable. Even if we recognize the inferable centers of U_{i-1} a posteriori by considering only those that appear in U_i, it is still impossible to select which of these centers is highest ranked in U_i. What we need, instead, is a way to recognize a Cb in U_i not by virtue of its identity with a preestablished list of (explicit and inferable) centers, but by virtue of a relationship, *other than identity*, with the *explicit* centers of U_{i-1}. We propose the relationship of lexical cohesion to fill this function. The recognition of a lexically cohesive relationship, then, admits inferable centers without allowing the virtually uncontrollable proliferation of hard-to-define inferable centers in the Cf lists for utterances. We propose a principled way to define this relationship that not only avoids the problems discussed above but also more accurately characterizes the coherence of this corpus.

4.4 Coherence and Cohesion

Halliday (1994) characterizes cohesion in text as the establishment of "relations within the text that are not subject to [grammatical] limitations; relations that may involve elements of any extent, both smaller and larger than clauses, from single words to lengthy passages of text; and that may hold across gaps of any extent. . . without regard to the nature of whatever intervenes" (page 309).[7] Cohesion is that aspect "whereby the flow of meaning is channelled into a traceable current of discourse instead of spilling out formlessly in every possible direction" (page 311). It is this "traceable current of discourse" that centering is meant to model.

Lexical cohesion contributes to textual coherence. In other words, strong semantic and structural relationships among words in a text help to make that piece of text "make sense." Coherence is a property of discourse; cohesion is a property of discourse elements. Centering models coherence by characterizing relationships between elements of discourse. We claim that it is not only the continuation of identical explicit discourse elements that creates coherence, but also strong cohesion among discourse elements.[8]

Halliday identifies four features of text that create cohesion among discourse elements: conjunction, reference, ellipsis, and lexical cohesion. Insights concerning conjunction types and their interactions with the processes of referent resolution have been elaborated in a number of works (Nariyama 2000; Nakaiwa and Shirai 1996; Kuno 1973) but have not been well integrated into the centering approach. Reference and ellipsis are, of course, some of the mainstays of centering research. Lexical cohesion, however, is an aspect of text coherence that has had only a trivial application in a centering approach, although it has been incorporated into other aspects of natural language processing (see subsequent discussion). "Lexical cohesion," according to Halliday, "comes about through the selection of items that are related in some way to those that have gone before" (page 330). In centering, that relationship has been

[7] We follow Halliday in assuming that "[f]or a text to be coherent, it must be cohesive; but it must be more besides." He characterizes the "more" as being socially, semantically, and structurally appropriate. We will not deal with these elements here but rather will limit ourselves to the contribution of lexical cohesion to coherence.

[8] Of course it is possible to have cohesion without coherence and vice versa; Morris and Hirst (1991) give some nice examples. However, as they assert, "most sentences that relate coherently do exhibit cohesion as well" (page 26).

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

assumed to be one of identity of centers; an element in U_i is required to be the same as an element in U_{i-1} (or in its discourse context) in order for it to be considered a Cb. However, earlier we saw the difficulties of admitting inferables and the inability of a centering approach to characterize coherence for inferable centers. Applying the notion of lexical relatedness to cases involving inferables allows us to capture what seems intuitively to constitute the relationship between clauses containing inferable centers.

A number of other accounts provide relevant information concerning lexical cohesion in text. These accounts are based upon the characterization of semantic relations among discourse elements by reference to semantic information contained in WordNet (Harabagiu 1998, 1999), thesauruses (Harabagiu 1998; Morris and Hirst 1991; Okumura and Honda 1994), or dictionaries (Kasahara et al. 1996; Kozima 1993; Kozima and Furugori 1993). In all of these approaches, the semantic distance or similarity between (or among) words is computed, and in most of these accounts, the results are applied to the segmentation of discourse.

The intuition behind the importance of lexical relatedness has been applied to a number of other tasks in natural language processing and analysis as well. Lotfipour-Saedi (1997) uses lexical cohesion to develop a rigorous notion of "translation equivalence"; Boguraev and Neff (2000) to improve document summarization techniques; Sack (1999) to create "diagrams of social cohesion" for newsgroup postings; and Okumura and Honda (1994) to disambiguate word senses.

Halliday asserts that "this interaction between lexical cohesion and reference. . . is the principal means for tracking a participant through the discourse" (page 332), that is, for modeling focus. Centering has provided us with a principled way of characterizing the tracking of reference; the addition of the notion of lexical cohesion allows centering to function in an even more empirically comprehensive way, by making possible the principled inclusion of what have been called inferable centers. In the next section, we outline how lexical cohesion can be incorporated into a centering theory.

4.5 COHESIVE Transition and COMPLETE SHIFT

With the use of the sorts of techniques to establish semantic distance described in the works cited earlier, it is possible to be precise about the notion of lexical cohesion. In this section, we discuss how semantic distance is established using one of these techniques, a semantic similarity measure derived from the Gainen Base ('Concept Database') (Kasahara et al. 1996). We then indicate how semantic distance can be used to define the notion of lexical cohesion as a crucial element in the creation of two new types of transition: COHESIVE and COMPLETE SHIFT, which allow us to adequately characterize coherence in a corpus containing a high proportion of nonexplicit Cbs.

The Gainen Base is a knowledge base built from machine-readable dictionaries of Japanese. Each word in the knowledge base is defined by a list of weighted keywords extracted from the dictionary definition of the word. The number of times a keyword appears in the word's definitions determines the weight for the keyword. Keywords are standardized to take into account the presence of semantically similar words in the definitions, and their weights are normalized to take into account the differing lengths of definitions in the dictionaries. The semantic distance between two words is calculated as a function of the nearness of the two words in a vector space.[9] For

[9] Each Word$_i$ is defined by a list of standardized, weighted keywords from which is generated a vectorized-Word$_i$. More specifically, then, semantic similarity is measured by a function R that satisfies

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

Table 4
Transition definitions.

	$\text{Cb}(U_i) = \text{Cb}(U_{i-1})$ OR $\text{Cb}(U_{i-1}) = ?$ and $\text{Cb}(U_i) \neq ?$	$\text{Cb}(U_i) \neq \text{Cb}(U_{i-1})$	$\text{Cb}\ (U_i) = ?$	
$\text{Cb}(U_i) = \text{Cp}(U_i)$	CONTINUE	SMOOTH SHIFT	COHESIVE	$\exists\text{Cf}(U_i) \approx \text{Cf}(U_{i-1})$
$\text{Cb}(U_i) \neq \text{Cp}(U_i)$	RETAIN	ROUGH SHIFT	COMPLETE SHIFT	$\sim(\exists\text{Cf}(U_i) \approx \text{Cf}(U_{i-1}))$

our present purposes, then, we say that there is lexical cohesion between U_i and U_{i-1} when $\text{Cf}(U_i)$ is semantically close to $\text{Cf}(U_{i-1})$ as determined using a well-defined semantic similarity measure over the Gainen Base.[10] We will examine in more detail in Section 4.6 whether semantic similarity is best viewed as holding between the sets of Cfs of utterances or between individual discourse entities in those utterances. For now, we will talk equally of lexical cohesion between utterances and lexical cohesion among the Cfs that participate in defining cohesion for those utterances. We define the relation $\approx$ as indicating strong lexical cohesion, with a lexical cohesion factor of one indicating identity.

We supplement the standard table of transition states, shown in the left side of Table 4, with transitions defined in the right side. The sense of this table is as follows. We assume all centers to be explicit, that is, pronouns, zero arguments, or explicitly realized entities. The left portion of Table 4 allows us to model transition states in well-behaved explicit contexts, tracking the focus of the discourse in a specific, local way. It includes the cases in which the Cb of U_{i-1} might be "?," that is, that U_{i-1} might be the discourse-initial utterance or might simply have no Cb, while U_i does have (an explicit) Cb. However, where there is not strict identity between any (explicit) element in U_i and any (explicit) element in U_{i-1}, we have the situation in which $\text{Cb}(U_i) = ?$. We propose a new interpretation for these cases, described in the right portion of Table 4.

Table 4 defines two new types of shift to utterances that do not contain an explicit Cb. If there is at least one Cf in U_i that has a high lexical cohesion value with some Cf(s) in U_{i-1}, then the transition from U_{i-1} to U_i is a COHESIVE transition. This situation is illustrated in clauses (2b), (2c), (2d), and (2e) of example (2). The transitions to these clauses, under the present proposal, are reanalyzed as COHESIVE, as shown in a reanalyzed version of example (2) (entities claimed to bear close semantic relationships to one another are shown in boldface here and in subsequent examples):

the following conditions:

$$0 \leq R(\text{Word}_a, \text{Word}_b) \leq 1$$

$$R(\text{Word}_a, \text{Word}_b) = R(\text{Word}_b, \text{Word}_a)$$

$$R(\text{Word}_a, \text{Word}_a) = 1$$

Word_b is more similar to Word_c than to Word_a if $R(\text{Word}_a, \text{Word}_b) \leq R(\text{Word}_c, \text{Word}_b)$

The function R = cosine A, where A is the angle defined by the vectors for Word_a and Word_b, satisfies these conditions and is therefore chosen as the function to define the similarity between Word_a and Word_b.

We refer the reader to Kasahara et al. (1996) for a full discussion of the algorithms used to weight and normalize keywords and to calculate semantic distance.

[10] What constitutes "semantically close" is an issue we will discuss briefly in Section 5.4; exactly how semantic distance is measured is elaborated in Section 4.6.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

2a.	ところで	宴席 なん	ですが,				Cb = ?
	tokorode	enseki nan	desuga				Cf = **restaurant**
	by the way	restaurant	is				
							COHESIVE

2b.	千駄ヶ谷	近辺	に	酒	を	飲める	Cb = ?
	sendagaya	kinpen	ni	sake	wo	nomeru	Cf = **soba shop**,
	sendagaya	neighborhood	in	alcohol	OBJ	can drink	**Sendagaya neighborhood, alcohol**

旨い	蕎麦屋	が	あると	耳にしました.	
umai	sobaya	ga	aruto	miminishimashita	
good	soba shop	SUBJ	is	have heard	

"By the way, about the restaurant, I've heard there is a good soba shop, which also serves alcohol, around Sendagaya."

COHESIVE

2c.	閉店	時刻	が	早いそうなので、	Cb = ?
	heiten	jikoku	ga	hayaisounanode	Cf = **shop**
	shop closing time		SUBJ	early seems because	**closing time**
					COHESIVE

2d.	ちょっと	まずい	かもしれません	が	Cb = ?
∅	chotto	mazui	kamoshiremasen	ga	Cf = choosing **this**
∅	a little	bad	might be	but	**restaurant** ∅
					COHESIVE

2e.	I-さんあたり	ご存じでは	ないでしょうか？	Cb = ?
	I-sanatari	gozonjidewa	naideshouka	Cf = I-san,
	I-san	information	have	information (about **restaurant**)

"Choosing this restaurant may not be good because it closes early. I-san, do you know about this restaurant?"

Each COHESIVE transition in the modified version of (2) is justified by the presence of a strong semantic relation between at least one Cf in U_i and at least one Cf in the previous utterance. For example, *soba shop* in (2b) is semantically related to *restaurant* in (2a), as is *alcohol*, probably to a lesser extent, and *Sendagaya* (if our database includes the information that this is the name of a restaurant). Note that the transition from (2d) to (2e) in this interpretation is a COHESIVE one, rather than a ROUGH SHIFT. Certainly the Cp changes from *restaurant* to *I-san*, but since *restaurant* is still present in the Cf list for (2e), the transition is by no means as abrupt as the designation ROUGH SHIFT implies.

The presence of a null Cb in and of itself, then, is not necessarily indicative of incoherence. The level of coherence is captured instead by the proportions of the various transition states present in a corpus, including COHESIVE transitions.

Not all utterances containing null Cbs are felt to be cohesive with previous utterances, of course. If there is no explicit Cb and no Cf in U_i such that it has strong lexical cohesion with a Cf in U_{i-1}, then the shift is considered COMPLETE. This is illustrated in example (5), which shows the continuation of example (1):

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

5a.	でも、	テコンドー	に	興味津々	だった	U-さん	や	Cb = it = game
	demo	tekondoh	ni	kyoumishinshin	datta	U-san	ya	Cf = U-san, M-san, it, taekwondo, interest
	but	taekwondo	in	keen interest	was	U-san	and	

M-さん	は	これでいいんでしょうか？	
M-san	wa	koredeiindeshouka	COMPLETE SHIFT
M-san	TOP	I'm wondering if it is OK	

"But I'm wondering if it is OK for U-san and M-san, who have shown a keen interest in taekwondo."

5b.	どうやら	梅雨 入り	したらしいという	宣 言 も出た	Cb = ?
	douyara	tsuyuiri	shitarashii toiu	sengenmodeta	Cf = rain, today, declaration, beginning of rainy season
	somehow	rainy season	has come	declaration	

ようで、	今日	も	しっかり	雨	が	降っています。	
youde	kyou	mo	shikkari	ame	ga	futteimasu	
with	today also		heavily	rain	SUBJ	is raining	RETAIN

"With the declaration that the rainy season has come, it is raining heavily today."

5c.	18日	が	雨	という	確率	も	高	そうですよね。	Cb = rain
	18nichi	ga	ame	toiu	kakuritsu	mo	taka	soudesuyone	Cf = 18th, probability, rain
	18th	SUBJ	rain		probability		high	seems	

"There seems to be a high probability that it will rain on the 18th."

There is very low cohesion between the Cfs of (5b), that is, *today, rain, declaration*, and *beginning of rainy season*, and those of (5a), namely, *U-san, M-san, game, taekwondo*, and *interest*. Thus, we designate the transition from (5a) to (5b) as a COMPLETE SHIFT, which matches our intuition that, in fact, the topic of the message has changed. This is further corroborated by the fact that (5c) RETAINs the Cp of (5b) as its Cb; in other words, it goes on to develop the new topic begun in (5b).

Since the two sides of Table 4 are complementary, it is perfectly possible for coherence to be characterized by various combinations of the transitions in the table. Example (6) illustrates the interaction between the two sides of Table 4:

6a.	このうち	神宮	の	方	は、	１７日	と	１８日	CONTINUE, from above
	konouchi	jinguu	no	hou	wa	17 nichi	to	18 nichi	Cb = these[11]
	of these	Jingu			TOP	17th	and	18th	Cf = these, **Jingu Stadium, Yokohama**

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

の	横浜戦	が	いいかと	思っていますが、	**game**, 17th, 18th
no	yokohamasen	ga	iikato	omotteimasu ga	
game versus	Yokohama	SUBJ	is good	suppose but	COHESIVE

6b.	Ｊリーグ	の	方	は	カード	など	Cb = ?
	J-rihgu	no	hou	wa	kahdo	nado	Cf = **J-league,**
	J-league			TOP	match-ups	and others	**match-up**
							information

詳しいこと	が	不明	なので	RETAIN
kuwashiikoto	ga	fumei	nanode	
further information	SUBJ	unknown	since	

6c.	わかる	人	は	教えて	ください。	Cb = match-up
∅	wakaru	hito	wa ∅	oshiete	kudasai	information
have information		someone	TOP	let me know	please	Cf = person,
						match-up
						information ∅

"Of these, for Jingu Stadium I suppose the game versus Yokohama on the 17th or the 18th would be good. As for J-league, since I don't have further information such as the match-ups, please let me know if you have any information."

COHESIVE

6d.	ボクシング	は	今　月	は	いい	カード	が	な い	ので	Cb = ?
	bokushingu	wa	kongetsu	wa	ii	kahdo	ga	nai	node	Cf = **boxing**,
	boxing	TOP	this month	TOP	good	matches	SUBJ	no	since	this month,
										match-ups
										CONTINUE

6e.	私	は	パス	したい	Cb = boxing
	watashi	wa ∅	pasu	shitai	Cf = boxing
	I	TOP	skip	would like	

"Since there are no good boxing match-ups in this month, I would like to skip it this time."

The transition from (6a) to (6b) is COHESIVE, since (6b) has no Cb and yet the Cfs of the two utterances are semantically related. (6c) RETAINs *match-up information* from (6b), which is then COHESIVE with *match-ups* and *boxing* in (6d). Since (6e) does in fact have an explicit Cb identical both to the Cb of (6d) and to its own Cp, the transition to (6e) is CONTINUE. This latter transition is an example of a non-discourse-initial utterance

[11] It is a departure from the standard Cf template for Japanese to designate *these* as the Cb instead of the topic-marked *Jingu Stadium*; however, this pronominal form comes immediately after the mention of two options in the previous clause and so seems to be the Cb regardless of its lack of marking. This choice of Cb has no bearing on the point of this example.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

without a Cb that licenses a CONTINUE transition by virtue of the second part of the OR disjunction in the second column of Table 4, namely, that the $\mathrm{Cb}(U_{i-1})$ = ? and $\mathrm{Cb}(U_i) \neq$?. (Note that (5) also demonstrates the interaction between the two sides of Table 4, with a RETAIN transition following a COMPLETE SHIFT.)

The proposal to include COHESIVE and COMPLETE SHIFT in centering theory is motivated by concerns about having to specify two identical entities in adjacent utterances in order for those utterances to exhibit coherence. The requirement of identity leads to the need to allow inferable entities to play a part. However, it is often impossible to characterize transition states to utterances containing inferable Cbs. By including the notion of COHESIVE transitions, we capture the relatedness of two entities without the need to invoke inferable centers, and we can far better characterize the apparent coherence in this corpus.

4.5.1 Transition States in the Corpus, Revisited. Given the inclusion of COHESIVE and COMPLETE SHIFTs, we reinterpreted the labeling of transitions in the corpus (as represented in Table 3) and hand-tabulated a revised distribution of transition types (Table 5). The decision to designate transitions to inferable Cbs and NULL transitions as COHESIVE or COMPLETE SHIFT was made on the basis of an intuitive assessment of the possible semantic relations between entities in adjacent utterances. However, a possible method for automatic determination is described in the next section.

4.6 A Preliminary Implementation

4.6.1 Preparation of the Corpus. In order to explore the computational feasibility of COHESIVE and COMPLETE SHIFT, we subjected the corpus to an analysis of semantic distance using the Gainen Base. Before implementing the analysis, we processed the corpus so that it contained only the canonical forms of the words in each utterance, the forms accepted by the Gainen Base algorithms. The first step in this procedure was to run the corpus through a morphological analyzer, ALT-JAWS (Nippon Telegraph and Telephone Corporation 1996), which rendered all word forms in their canonical (usually kanji [Chinese character]) form. This is a necessary and standard preprocessing step for most computational analyses of Japanese text, which contains no spaces between word forms to indicate their morphological structure. In addition, writers of e-mail may use hiragana or katakana (the two syllabaries used for writing primarily function words and foreign words, respectively) even for words that have standard kanji forms (Fais and Yamura-Takei 2003). These words need to be rendered in kanji as well. The results of this analysis were checked and corrected by native speakers of Japanese in order to eliminate cases in which the analyzer chose the incorrect kanji form for a homophonous hiragana representation.

Table 5
Reanalysis of transitions occurring in the corpus, by type.

Transition type	Number	Percentage
CONTINUE	54	16.4
RETAIN	22	6.7
SMOOTH SHIFT	2	0.6
ROUGH SHIFT	2	0.6
COHESIVE	138	41.8
COMPLETE SHIFT	112	33.9
Total	330	

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

The next step was to delete all but noun forms from the corpus. This step substantiates the notion that the central factors in the centering approach are the discourse entities of utterances (we will return to this in Section 5.3). Further, the antecedents of all pronominal and zero-argument references were made explicit in the text.[12] This was done by inserting the kanji forms for these antecedents, as they appeared in the text, into the utterance in which the pronominal or zero-argument form appeared. Antecedents were determined by a native speaker of Japanese. The same native speaker assisted in dividing the text into clauses. Both of these steps, identifying antecedents and parsing sentences into clauses, can be completed, in theory, by automatic, computational methods (Huls, Bos, and Claassen 1995; Nakaiwa and Shirai 1996; Paul and Sumita 2001; Yamura-Takei et al. 2002), but the success rate of these approaches is not high enough to rely on them for completely accurate analyses of this type of corpus at this time. Because our intent is to examine the effectiveness of the use of the Gainen Base in determining lexical cohesion, and not to implement a fully automatic process, we did not attempt to use entirely automatic methods in the preprocessing of the text.

4.6.2 Determining Semantic Similarity with the Gainen Base: The Problem of Coverage. Of the 670 types of nouns present in the e-mail corpus, only 235 are found in the machine-readable dictionaries with which the Gainen Base was constructed. What makes this ratio even more problematic is that a number of these missing words (e.g., *supo-tsu kansen*, 'sports-watching event,' *rakurosu*, 'lacrosse,' and the names of sports teams) are high-frequency words in this corpus.

The problem of coverage in automatic language-processing systems is a common one (Hutchins 1995; Sag et al. 2002; Fujita and Bond 2002). At the level of coverage provided by the Gainen Base, however, we cannot usefully assess how well semantic similarity characterizes coherence in this corpus as a whole. However, we can make this assessment for those clauses in which every noun can be found in the Gainen Base. In order to get an accurate assessment of lexical cohesion between adjacent clauses that fall into this subset of the corpus, we included those cases in which all the nouns in the clause itself as well as those in the clause before it were found the Gainen Base. Out of an original 443 clauses, there are 66 clauses that meet these two criteria.

We measured semantic similarity between adjacent clauses in two ways. In the first, we measured the semantic distance between the group of nouns in U_{i-1} and the group of nouns in U_i. In the second, we measured the semantic distances between each individual noun in U_i and the group of nouns in U_{i-1}.[13] The first method is far "lighter" computationally, but the second method gives us useful information about the contribution of each noun in U_i to the lexical cohesion between utterances. We will compare the information derived from these two approaches hereafter.

4.6.3 Evaluation of the Use of Semantic Similarity. We assessed in three ways how well semantic similarity can define COHESIVE transitions and thus contribute to the characterization of coherence in the text. First, in the 66 clauses with full coverage by the Gainen Base, we examined the 18 instances of what we had, on the basis of intuitive human judgment, designated COHESIVE transitions. In particular, we focused

[12] We did not, however, include the entities involved in event deixis; the identity of these entities is much harder to determine, both for human judges and for automatic language-processing systems.

[13] We actually measured semantic similarity in a third way as well, that is, between every possible combination of the individual nouns in U_{i-1} and those in U_i. This yielded the same results as the individual-to-group method reported on here and, of course, is much "heavier" computationally, so we have restricted our discussion to the first two methods.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

on the noun(s) we had singled out in our revised hand-tabulation of the corpus (Table 5) as providing the lexical cohesion in these transitions.[14] Using results from the individual measures of semantic similarity (the second method described earlier), we determined the noun in U_i with the highest level of semantic similarity to the nouns in U_{i-1}. We then compared the nouns picked out by human judgment as providing lexical cohesion with those determined computationally to see if they matched. We eliminated three cases in which there was only one noun in U_i and thus only one possible choice for the lexically cohering entity, which would, by default, have had the highest semantic similarity to the preceding utterance. Out of the 15 remaining examples of COHESIVE transitions, in 13 cases, or 87% of the time, the human judgments and the computationally determined choices matched.

The other two assessments of the results are based on centering claims for the relative ease of processing of the different transitions, claims captured in rule 2: CONTINUE transitions impose the lowest inferential load on processors, ROUGH SHIFTs the highest. Since COHESIVE transitions act like CONTINUE transitions but replace the identity condition on Cbs with a similarity condition, we conjecture that they place only a slightly higher load on processing than CONTINUE transitions. Likewise, since COMPLETE SHIFTs represent an even greater discontinuity than ROUGH SHIFTs (there being not only no Cb in the utterance, but also no entity even similar to entities in the previous utterance), we conjecture that COMPLETE SHIFTs impose a higher processing load than ROUGH SHIFTs.

We reason that greater semantic similarity corresponds to a lower processing load.[15] Granted this assumption, then, utterances that are connected by CONTINUE transitions have the highest semantic similarity, since CONTINUE represents the lowest processing load; those connected by COMPLETE SHIFTs, the lowest semantic similarity, since COMPLETE SHIFT has the highest processing load. Although the particular ranking of RETAIN, SMOOTH, ROUGH, and COHESIVE transitions is problematic, we are saved from having to make an exact determination by the fact that, among the 66 clauses with full coverage, only CONTINUE, COHESIVE, and COMPLETE SHIFT transitions are well-represented (we simply note the results for the two RETAIN transitions, since this is hardly a large enough sample to be meaningful). Our reasoning then predicts that CONTINUE transitions have the highest semantic similarity measures, followed by COHESIVE SHIFTs, followed by COMPLETE SHIFTs.[16]

In our first assessment of this prediction, we averaged semantic measures for each transition type over the 66 clauses with full coverage by the Gainen Base. Table 6 gives the averages obtained for both the groupwise and individual analyses. These results support the ranking of transitions for processing load predicted on the basis of similarity measures. However, averaging over all messages provides only a gross approximation of the values for each transition type. To get a more detailed look at this claim, we examined the semantic distances between entities involved in each type of transition within messages.

The 66 clauses with full coverage include six messages with only one clause and four messages with only one transition type represented; this resulted in the elimina-

[14] Where there was more than one noun, as, for example in (2b), we chose the one that seemed, intuitively, to have the strongest semantic connection to a center in the previous utterance.

[15] This claim must, of course, be tested empirically and independently supported. However, it seems a reasonable assumption and allows us to further evaluate the characterization of coherence by semantic similarity.

[16] We do not have nearly enough data to determine absolute values for each transition type (or even to determine whether this would be a desirable course of action), and so we base our evaluation on relative values for transition types (see Section 5.4 for further discussion).

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

Table 6
Average similarity measures for each transition type, for both groupwise and individual analyses.

Transition type	Number	Groupwise analysis Average similarity	Individual analysis Average similarity
CONTINUE	10	0.605	0.639
RETAIN	2	0.530	0.496
COHESIVE	18	0.204	0.231
COMPLETE SHIFT	36	0.068	0.069
Total	66		

tion of 19 clauses (since it is impossible to determine *relative* values of transitions in either of these cases), leaving us with a total of 47 clauses grouped into 12 messages. We sorted the types of transitions to these clauses within each message by similarity measure. For each transition, we ascertained whether it fulfilled the prediction for ranking transitions by inferential load that was made on the basis of our earlier assumptions concerning semantic distance. We assigned two scores for each transition: whether it was appropriately positioned with respect to the preceding transition and with respect to the following transition in the sort. Those transitions with the highest and the lowest semantic measures were scored only with respect to the following and the preceding transitions, respectively, and thus received just one score. Two consecutive identical transitions were scored "correct."

Table 7 gives an example of how this scoring was performed for the clauses from one representative message that have full coverage in the Gainen Base. It lists the type of transition to each clause, the semantic distance to the previous clause, and the scores designating whether that transition is appropriately positioned with respect to the previous and following transitions. (Recall that the predicted order is CONTINUE, COHESIVE, COMPLETE.) So, for example, the CONTINUE transition to clause 235 fulfills the ranking prediction with respect to both the previous and the following clauses; its semantic measure is lower than that of only one other clause, which is also a CONTINUE, and is higher than that of a COHESIVE transition. The COMPLETE SHIFT to clause 227, on the other hand, fulfills the prediction vis-à-vis the previous transition on the list

Table 7
Ranking transitions in a representative message by similarity measure.

Clause	Transition to clause	Semantic measure (semantic distance to previous clause)	Score
220	CONTINUE	0.44113	Correct v
235	CONTINUE w/?	0.13974	Correct ↑ Correct v
228	COHESIVE	0.06226	Correct ↑ Correct v
227	COMPLETE	0.03922	Correct ↑ Incorrect v
216	COHESIVE	0.01370	Incorrect ↑ Correct v
234	COMPLETE	0.00000	Correct ↑

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

Table 8
Evaluation of relative similarity measures for transition types.

Transition type	Groupwise analysis		Individual analysis	
	Number	Percentage	Number	Percentage
CONTINUE	12/12	100	11/12	92
RETAIN	Ranked before cohesive shift		Ranked before cohesive shift	
COHESIVE	26/29	90	25/29	86
COMPLETE SHIFT	25/28	89	24/28	89
Total	63/69[a]	91	60/69[a]	87

[a] Since we make no claim as to whether the placement of RETAIN is correct or not, we do not count it in these totals.

(i.e., its semantic measure is lower than that of a COHESIVE transition) but violates the prediction with respect to the following transition (i.e., its semantic measure is higher than that of another COHESIVE transition). For 47 clauses over 12 messages, then, the total number of such scores was 70: 24 scores for the transitions having the highest and lowest measures in each message, and two each for the remaining 23 transitions, one for their positions relative to the transitions above, and the other for their positions relative to the transitions below them.

Recall that we determined similarity in two different ways: first, for the group of nouns in U_{i-1} and the group of nouns in U_i, and second, for the group of nouns in U_{i-1} and each individual noun in U_i.[17] Table 8 reports the total number and percentage of correct and incorrect scores for each transition type in both the groupwise and the individual analyses (the measure taken for U_i in the latter case is the maximum similarity measure out of the measures for all the nouns in U_i). The results in this table suggest that similarity scores can accurately represent relative coherence as characterized by the transitions in this small sample. That is, the similarity scores we have examined here reflect the relative load on processing imposed by each type of transition with between 86% and 100% accuracy, with groupwise scores being slightly more accurate than those based on the highest individual score. In addition, similarity measures never predict a CONTINUE transition with a higher processing load (i.e., lower similarity score) than a COMPLETE SHIFT. That is, the relative positions of CONTINUE and COMPLETE SHIFT are always correct.

We examined the cases in which similarity scores make an incorrect prediction about relative placement in the scale of processing load. The one incorrect prediction concerning a CONTINUE (in the individual-analysis method) involves an example in which U_{i-1} contains eight nouns; the similarity of each individual noun in U_i to the group of nouns in U_{i-1} is "diluted" by the high number of nouns in U_i in this case (the average number of nouns per clause in the corpus is 2.3).

The remaining incorrect predictions involve the assignment of lower similarity scores to COHESIVE transitions than to COMPLETE SHIFTs. Some of these lower scores for COHESIVE transitions are the result of the fact that world knowledge is necessary to infer a connection between the two clauses. This is the case, for example, in (7):

[17] Table 7 reported groupwise scores.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

7a.	遠出	でも	東京	に	戻って	来たところで
	toode	demo	Toukyo	ni	modotte	kitatokorode
	at a distance	though	Tokyo	to	come back	

Cb = ? (from previous clause)
Cf = **distance**, Tokyo

COHESIVE

7b.	宴会	タイム	には	丁度	いいですよね。
	enkai	taimu	niwa	choudo	iidesuyone
	drinking party	time	for TOP	exact	right

"Even if it's far we can come back to Tokyo just the right time for a drinking party."

Cb = ?
Cf = party, **time**

It requires world knowledge to understand that there is a connection between how far away something is and the time that it will take to travel there. Although humans can make this inference intuitively, that understanding is not represented in the Gainen Base.[18]

Other incorrect judgments are the result of how the database handles determining the similarity scores, "quirks" that don't seem to match our intuitive judgments, as in (8):

8a.	時間的には	ちょうど	タイミング	は	いいだろう	し、
	jikantekiniwa	choudo	taimingu	wa	iidarou	shi
	timewise TOP	exact	timing	TOP	seems good	and

Cb = ? (from previous clause)
Cf = **time, timing**

COHESIVE

8b.	地理的に	も	∅	遠くは	ないんで。
	chiritekini	mo		tookuwa	nainde
	geographically	too		far TOP	not

"Timewise, the timing seems good, and the night game is not far away either."

Cb = ?
Cf = **night** game ∅, geography

The Gainen Base yielded a relatively low score for the similarity between *jiken*, 'time,' and *yoru*, 'night,' despite our strong sense that these two words should be closely related.[19]

Overall, however, similarity scores seem to provide a fairly accurate measure of the relative coherence of this subset of the corpus. This result, coupled with the high

[18] Recall that in 13 out of 15 cases, the Cf judged by humans to license a COHESIVE transition and the Cf picked out by the similarity measure matched. (7a) is one of the two clauses in which the human-chosen Cf did *not* match that chosen by the Gainen Base (the second is given in (8)). The previous clause is *If it is a day game*. The discourse entity *tode*, 'at a distance,' was the Cf chosen by human judgment to license the COHESIVE transition, since a human can make the connection that it is the *day game* that is *at a distance* (as evidenced in the translation). However, the use of the Gainen Base determined *Tokyo* to be more semantically similar to *day game*, with a score of 0.105 as compared to 0.005 for *tode*.

[19] This is the second of the two cases in which human judgment and computational choice for lexically cohering entity did not match.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

level of correlation discussed earlier between lexically cohesive entities designated by human judgment and those determined by semantic similarity, supports our proposal that lexical cohesion, as measured by semantic distance, can feasibly be included as a well-defined notion to capture crucial aspects of text coherence.

4.7 Exploring the Implications of the COHESIVE Transition and COMPLETE SHIFT

4.7.1 The Need to Identify Relation Type. The usefulness of relaxing the notion of identity for Cbs has been recognized by Poesio et al. (2000) and others (Hahn, Markert, and Strube 1996; Murata, Isahara, and Nagao 1999). Poesio et al. supplemented the identity relation with three different possible semantic relations between Cfs in the utterances: set membership, subset, and “generalized possession.” Murata, Ishara, and Nagao induced a number of possible relations between bridging reference and anchor using a verb case frame dictionary and a corpus of Japanese *A no B* expressions (where *A* and *B* are nominal arguments and the *A no B* construction encodes a wide variety of semantic relations between the two nominal arguments [Shimazu, Naito, and Nomura 1987]).

As noted in Section 4.3.1, however, in all of these approaches, only a small subset of examples of bridging relations can be handled, because they all attempt to identify some particular relation existing between two elements. This is a necessary move for resolving bridging references and building text understanding systems, but neither of those is our aim here. We merely need to identify the level of semantic closeness or similarity between Cfs in U_i and Cfs in U_{i-1}. Utilizing instead the more general semantic-distance measure proposed here, then, has several advantages over the explicit choice of particular relation labels. First, it avoids the need to make choices about which relations can or cannot, should or should not be included, as well as the difficulties with interlabeler reliability that Poesio et al. note, since there is no labeling in our approach. This is actually closely tied to another advantage: There is no need to limit the types of semantic relationships into which the Cfs can enter. Thus, there is no need to restrict our analysis to a subset of the phenomenon; our account will handle inferable centers having any kind of semantic relationship to the centers in the previous utterance.

4.7.2 Overestimation of COMPLETE SHIFTs. In examining Table 5, we see a high number of COMPLETE SHIFTs. Is this number an accurate estimation of the (in)coherence in this corpus? Of course, as we saw in (5), when a message makes a shift in topic, we expect a COMPLETE SHIFT to occur. The writers of the messages in this corpus made 146 paragraphs (see Table 2); although we know that there is no guarantee that writers’ paragraphing will coincide with shifts in cohesion, this number at least gives us a general estimation of a possible maximal number of COMPLETE SHIFTs (we would expect writers to err on the side of more paragraphs than topics rather than on the side of more topics than paragraphs). The number of COMPLETE SHIFTs is comfortably within that range.

But there are two confounding factors that make this number higher than is actually appropriate for the nature of the corpus. The first is the inability of a centering approach to handle event deixis (Fais and Yamura-Takei 2003). Consider (9):

9a.	急きょ	U-さんの	仕事の 都 合	上	１７日	の	Cb = ?
	kyukyo	U-sanno	shigotonotsugou	jou	17nichi	no	Cf = U-san, business,
	urgent	U-san	business	because of	17th	on	17th, match-up

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

同	カード	に	変更したい	とのことです。	
dou	kahdo	ni	henkoushitai	tonokotodesu	
same	match-up	to	want to change	thing is	
“Because of U-san’s urgent business, she would like to change to the same match-up on the 17th.”					COMPLETE SHIFT

9b.	わたし	は	ＯＫなん	です	が	Cb = ?
	watashi	wa ∅	OKnan	desu	ga	Cf = that ∅
	me	TOP ∅	OK	is	but	
“That’s OK with me.”						
						RETAIN W/?

9c.	みんな	は	どうでしょう？	Cb = that ∅
	minna	wa	doudeshou	Cf = everyone, that ∅
	everyone	TOP	how is	
“How is it for everyone else?”				

The zero argument translated as *that* in (9b) is an example of event deixis. When we examine the Cf list for (9a) to ascertain the Cb of (9b), we do not encounter *that* or its referent. In fact, we cannot encounter *that*. It is not possible for the discourse element represented by *that* to have appeared in the Cf list in (9a); clauses do not contain self-referential discourse elements. Thus, it is impossible to recognize, within the theory, that the discourse element *that* in (9b) is functioning as a strong cohesive element in the discourse. This is a problem to be resolved within centering theory regardless of whether COHESIVE transitions and COMPLETE SHIFTs are countenanced. However, sentential deixis does contribute to a slight skewing of the proportion of COMPLETE SHIFTs found in this corpus, for example, the COMPLETE SHIFT resulting from this problem in (9a)–(9b) and similar examples in the corpus.

The second confounding factor is actually simply a byproduct of the nature of the corpus. Example (10), which is an entire message, illustrates:

10a.	I-さんの	提案	で	すべて	ＯＫです。	Cb = ?
	I-sanno	teian	de	subete	OK desu	Cf = all, I-san
	I-san’s	proposal		all	OK are	suggestion
“All of I-san’s proposals are OK with me.”						
						COMPLETE SHIFT

10b.	１塁側でも	３塁側でも	かまいません。	Cb = ?
	1ruigawademo	ruigawademo	kamaimasen	Cf = first-base side,
	first-base side	third-base side	don’t mind	third-base side
“I don’t mind the first-base side or the third-base side.”				
				COMPLETE SHIFT

10c. 千駄ヶ谷なら、	多少	遅くなっても	平気ですね。	Cb = ?
sendagayanara	∅ tashou	osokunattemo	heikidesune	Cf = Sendagaya,
Sendagaya if	∅ quite	late even if	don’t care	party
				∅, lack of caring
“If it’s Sendagaya, we don’t have to care even if the party goes late.”				

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

The COMPLETE SHIFTs in this passage are perfectly appropriate; there is no cohesion among any of the discourse entities in this message. Why would anyone send a message that was completely, by this account, incoherent? In fact, the statements in (10) refer to the outcomes of various discussions held in the course of exchanging the messages in this corpus. This message occurs toward the end of the exchange of messages, as resolution of the questions of what sports event to go to, where to sit, and what restaurant to go to afterward is in sight. This writer is simply adding his opinions on each of these apparently unrelated topics. The relationship among them all holds only in the understanding of the coparticipants in the message exchange. Although we might be able to imagine the sorts of mechanisms required to model this level of understanding, we are a very long way from realizing them.

4.7.3 Lexical Cohesion and Text Understanding. The incorporation of a COHESIVE transition into our centering account gives us a flexibility that is important for full-text understanding of the discourse. Consider example (11):

11a.	わたし	も	テレビ	で	しか見たことはない	ので、	Cb = ?
	watashi	mo ∅	terebi	de	shika mitakotowanai	node	Cf = lacrosse ∅,[20] TV
	I	too ∅	TV	on	only have seen	since	

11b.	生で	観戦	したい気もする	のです が、	Cb = ?
	namade	kansen	shitaikimosuru	nodesu ga	Cf = live game
	live game	watching	feel like	because but	watching

11c.	２５日	だと	Cb = live game
∅	25 nichi	dato	watching ∅
∅	25th	is if	Cf = live game
			watching ∅, 25th

11d.	M-さん、	H-さん	が	参加	できない	ですね	Cb = ?
	M-san	H-san	ga ∅	sanka	dekinai	desune	M-san, H-san,
	M-san	H-san	SUBJ ∅	join	cannot	isn't it	game ∅

"Since I have seen it only on TV, I feel like watching a live game, but if it is on the 25th, M-san and H-san cannot join us for the game, can they?"

Lacrosse in (11a), *(some) live game watching* in (11b) and (11c), and *(a) game (on the 25th)* in (11d) are actually semantically distinct elements, and recognizing their distinctiveness is important for full-text understanding or summarizing. However, maintaining these distinctions in a standard account means characterizing the transitions between the utterances containing them as NULL SHIFTs. (It is not clear that we would even want to say that *(a) game (on the 25th)* was an inferable center from *(some) live game watching*.) Being able to designate these transitions as COHESIVE allows us both to maintain the

[20] The message that this is taken from constitutes a sort of lesson on lacrosse from one of the authors to all the others. The ∅ in (11a) refers to this global topic; however, neither *lacrosse* nor *TV* appears in the preceding utterance, and so (11a) has no Cb.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

semantic distinctiveness of the discourse elements and to capture the coherence in this portion of the discourse.

4.7.4 Lexical Cohesion and Discourse Segmentation. As we saw earlier, the use of some notion of lexical cohesion to delineate discourse structure is fairly well researched. In the lexical chain approach (Morris and Hirst 1991), a new discourse segment is hypothesized where the chain "breaks," that is, where subsequent entities do not bear a semantic relationship to previous entities that would allow them to be added to the chain. Kozima (1993) provides an algorithm for determining where, on the graph of semantic cohesion values of words in a text, likely topic breaks occur and validates that determination against human judgment. We would say, then, that once a chain breaks or a significant dip in the semantic cohesion value graph occurs, the utterance following such a break is considered the first utterance of a new discourse segment.

In terms of semantic distance as determined by the Gainen Base, we suggest that a sufficiently low similarity measure might characterize both COMPLETE SHIFTs and the beginning of a new discourse segment. Once again, "sufficiently" must be defined; in light of the preliminary results we saw previously, we conjecture that the definition of "low" will be relative to the measures for similarity in the message under scrutiny and not an absolute value (see Section 5.4).

In addition, examination of the particular entities contributing to high levels of semantic similarity might also allow us to characterize the notion of "global topic," albeit in a differentiated way. That is, if we determine not just one semantic distance measure for the sets of entities in two adjacent utterances, but the individual distances for each combination of those entities (as briefly described in note 13), we can determine those entities that are contributing the greatest amount of semantic similarity to the measure and identify a cluster of entities that can be taken to represent a global topic.

This concept is worth examining more closely, since it bears on some of the very foundations of centering theory. The Cb of an utterance "represents the discourse entity that the utterance U_i most centrally concerns, similar to what is elsewhere called the 'topic'" (Walker, Joshi, and Prince 1998, page 3). The presence of a Cb is taken to be both a necessary and a sufficient condition for topic coherence. However, in our account, utterances that have a coherent relationship to the immediate context of discourse may nonetheless have no Cb; that is, the presence of a Cb is, in our approach, only a sufficient condition. Our claim, then, is that this more accurately reflects the nature of how coherence is maintained in discourse: not only through the explicit repetition of a central entity, but also through the successive use of entities that are closely related semantically. In our account, Cbs are recognized and function just as in standard centering theory, but their absence, a common situation in at least some kinds of discourse, does not signal a breakdown in coherence. Coherence may be maintained as well by semantically similar entities, which can become Cbs in their own right, as *match-up information* and *boxing* do in (6). Using an individuated approach to determining semantic similarity, we can identify these particular entities.

5. Future Work

5.1 Refinement of the COHESIVE Transition

There are a number of aspects of the proposal that need further scrutiny. Note that we have defined the COHESIVE transition by reference to any Cf in U_i. It might be the case, in fact, that there is motivation to define two different COHESIVE transitions: a

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

"CONTINUE COHESIVE," in which the Cp(U_i) has the highest similarity measure with respect to Cf(U_{i-1}), and a "RETAIN COHESIVE," in which some other Cf has the highest similarity to Cf(U_{i-1}). In our preliminary implementation with this corpus, the entity in U_i with the maximum semantic similarity to U_{i-1} was the Cp of U_i 55% of the time. Whether such a distinction is a necessary or meaningful one is a completely open question.

In a similar vein, we might suggest that the Cf in U_i with the highest similarity to the Cfs in U_{i-1} be designated as a type of Cb (say, Cb′). So, for example, if *soba shop* in (2b) had the highest level of similarity to *restaurant* in (2a), *soba shop* would be the Cb′ of that utterance.[21] What would be the ramifications of this move for the theory?

Certainly there would be far fewer utterances with Cb = ?. In terms of tracking focus in the discourse, one of the major aims of the centering approach, this is a positive result. However, if the Cb′ is chosen simply on the basis of semantic similarity, we lose another major insight of centering theory, namely, that focus is not dependent upon semantics (Hudson-D'Zmura and Tanenhaus 1998) or upon word order (Gordon et al. 1999; but see also Gordon, Grosz, and Gilliom 1993), but upon the grammatical roles played by the Cfs. It is possible that for any given language, speakers tend to place Cfs that have strong semantic ties to the previous utterance in grammatical roles high on the Cf template, but this is an empirical question that we do not have the data to answer here.

5.2 Scalability

The incorporation of scalable considerations into a centering account provides enormous flexibility. It remains to be seen how best to use this added capability. For example, some COHESIVE transitions seem more cohesive than others. Compare the COHESIVE transition in (12) with the COHESIVE transitions in (13):

12a.	どうやら	梅 雨入り	したらしい	という	Cb = ?
	douyara	tsuyuiri	shitarashii	toiu	Cf = **beginning of rainy season,** declaration
	rainy season beginning has happened				

宣 言 も出た	ようで、	
sengenmodeta	youde	
declaration	with	COHESIVE

12b.	今日	も	しっかり	雨	が	降っています。	Cb = ?
	kyou	mo	shikkari	ame	ga	futteimasu	Cf = **rain**, today
	today		heavily	rain	SUBJ	is raining	

"With the declaration that the rainy season has come, it is raining heavily today."

[21] We can't corroborate this, because *sobaya* is not in the Gainen Base.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

13a.	私	は、	何より必要なもの	は	Cb = ?
	watashi	wa	naniyorihitsuyounamono	wa	Cf = **feeling,**
	me	TOP	more important than anything else	TOP	most important
					thing, **strength**

体力 なり!!	ということ	を	見せつけられた	ような	
tairyoku nari	toiukoto	wo	misetsukerareta	youna	
physical strength		OBJ	was shown		

気	が	したのです	が、	
ki	ga	shitanodesu	ga	
feeling	SUBJ	got	but	COHESIVE

13b.	∅ 見た	方	が	いたら	Cb = ?
	∅ mita	kata	ga	itara	Cf = **person,**
	watched	person	SUBJ	if exists	game ∅
					COHESIVE

13c.	感想	を	聞かせてください	Cb = ?
	kansou	wo	kikasetekudasai	Cf =
	impression	OBJ	let me know	**impressions**

"That gave me a strong impression that PHYSICAL STRENGTH IS MORE IMPORTANT THAN ANYTHING ELSE. If someone else watched (that game), please let me know your impressions."

We have the intuition that the COHESIVE transition in (12) can be characterized as "more" COHESIVE than the ones in (13). Both of these examples come from the same message, and the semantic similarity measures for each of these examples confirms our intuitions: The measure for the COHESIVE transition between (12a) and (12b) is 0.583; for that between (13a) and (13b), 0.166; and for that between (13b) and (13c), 0.011.[22] It is not clear to what use we might put this more detailed information about the varying levels of strength of connection among utterances; however, it parallels the observation in Fais (2001) that some e-mail authors make hierarchical distinctions in marking paragraphs in their messages, using line breaks to separate utterance clusters having some semantic connection and full spaces to separate utterance clusters having weak or no semantic connection.

Recall, too, that lexical cohesion, as measured by semantic distance, is only one aspect of coherence. Merging this measure with information provided by conjunctions,

[22] We chose examples from the same message because comparison of raw measures across messages may not be informative (see Section 5.4). However, we expect that measures within messages can be compared with one another usefully.

The very low measure for the transition between (13b) and (13c) is consistent with the observation that the cohesion between these two clauses actually may not reside in the similarity between *person* and *impressions*, but in our intuitive understanding of a relationship between *game* and *impressions (of the game).*

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

referential form, and other aspects of discourse would allow a more comprehensive account of coherence.

5.3 Beyond Simplex Nominals

The ramifications of exploiting lexical cohesion within a centering theory are exciting. Up to this point, we have considered only the cohesion exhibited by discourse entities. But not only (simplex) nominal arguments enter into cohesive relationships. Consider (14), in which the recognition of the contribution of the verb to the cohesion of the excerpt could allow us to account more accurately for the coherence in the passage:

14a.	わたし	は	明日の	土曜日	は	出社しない	ので、	Cb = ?
	watashi	wa	asuno	doyoubi	wa	shusshashinai	node	Cf = **tomorrow, Saturday**
	I	TOP	tomorrow	Saturday	TOP	won't go to work	since	
								COHESIVE

14b.	チケット	や	集合場所	など	は	Cb = ?
	chiketto	ya	shuugoubasho	nado	wa	Cf = ticket, meeting place, discussion, **Monday**
	tickets	and	meeting place	and so on	TOP	

月曜日	に	相談	ということに	したいんですが	
getsuyoubi	ni	soudan	toiukotoni	shitaindesuga	
Monday	on	discuss		would like to	COMPLETE SHIFT

14c.	（働く	人	ゴメンナサイ）。	(Cb = ?
	hataraku	hito	gomennasai	Cf = working
	working	people	sorry	people)

"Since I won't go to the office tomorrow (Saturday), I'd like to discuss tickets and a meeting place on Monday (sorry to those who are working)."

The first transition shown is COHESIVE by virtue of the Cfs *tomorrow* and *Saturday* in (14a) and *Monday* in (14b). But considering only the discourse entities appearing in the Cf lists, *working people* in (14c) is not cohesive with any of the other elements in this passage, although the inclusion of (14c) in this portion of the message seems quite coherent and natural. However, if we allow verbal elements to contribute to cohesion, *working people* then shows cohesion with *shusshashinai*, 'won't go to work,' and we can explain why the passage seems coherent.

In a similar vein, the incorporation of an adequate analysis for complex nominals would allow us to refine our measurement of cohesion as well. Recall (2), in which lexical cohesion exists among the entities *enseki* 'restaurant,' *sake* 'alcohol,' and *sobaya*, 'soba shop.' While the recognition of this cohesion allows us to characterize the coherence in the message, in fact, we probably underestimate the cohesion present if we base our measures on the individual lexical items listed. If we could calculate lexical cohesion not just for simplex nominals, but for complex nominals as well, we would calculate cohesion for *enseki* and for the entire phrase *sake wo nomeru umai sobaya*, 'a

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

good soba shop, which also serves alcohol,' possibly a much stronger cohesive connection than that among the individual items.[23]

5.4 Improving Implementation

We demonstrated in Section 4.6 that semantic distance as measured by the Gainen Base provides a feasible basis for a rigorous definition of lexical cohesion. In order for this or any similar implementation of semantic distance measure to be fully effective, however, two major areas need to be addressed. The first is coverage; while the lack of complete coverage of the corpus by the Gainen Base does not prevent us from making an assessment of the Gainen Base's effectiveness over a subset of the data, it does make it impossible to characterize cohesion over the corpus as a whole. The second area that needs to be addressed has to do with definitions of distance for each transition type. Throughout our discussion we have examined semantic distances as relative strengths within messages; it would be useful to determine empirically whether it is possible to set definitive, independent levels for each transition type. These levels might be absolute (e.g., SMOOTH SHIFTs are those transitions having a semantic measure of 0.15–0.3) or, what seems more likely, relative, such that a portion of the range of semantic distance in a particular message is defined for each transition type (e.g., in a message in which semantic distances measure from 0.005 to 0.875, SMOOTH SHIFTs correspond to distances falling between 15% and 25% of the total range of 0.87, that is, from 0.13 to 0.22, for that message).

6. Conclusion

Upon subjecting a corpus of Japanese e-mail data to a centering analysis, it became clear that a centering description of these messages had to rely heavily on inferable centers and was not adequate to capture the coherence this corpus displays. We couched the notion of connectedness that inferable centers were intended to capture in terms of the more principled and explicit relation of lexical relatedness. We used this relation, then, to supplement the standard inventory of transitions with well-defined transition types that more accurately characterize the nature of coherence in this corpus and demonstrated the computational feasibility of this approach in a preliminary implementation. The proposed transition types provide a characterization of a previously unaccounted-for situation in centering theory, namely, the coherence of sequences of utterances containing inferable Cbs. This is a crucial improvement over the standard model because of the high number of nonexplicit Cbs in this corpus (and others; see Passonneau 1998). The inclusion of a COHESIVE transition and COMPLETE SHIFT allows us to characterize the 76% of the corpus in which nonexplicit Cbs play a part, a portion of the corpus undescribed in the standard approach, while maintaining the standard operation of the usual transition states of centering theory in cases in which explicit Cbs are present.

It may be the case that lexical cohesion and the notion of a COHESIVE transition can make other contributions to discourse analysis as well, such as allowing us to characterize coherence in a discourse while still recognizing referentially distinct discourse elements. Further, these notions can augment the definition of the topic of a discourse segment by virtue of the semantic information contained in a cohesion analysis. Finally, such an analysis can provide clues to discourse segment boundaries.

[23] I would like to thank an anonymous reviewer for pointing this out.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

The incorporation of the notion of lexical cohesion has been shown to be crucial to the characterization of the coherence in this corpus. In addition, it can make a variety of further contributions to discourse analysis. And finally, it opens the door to a number of useful related strategies that can allow us to come closer to understanding and comprehensively modeling coherence in discourse.

Acknowledgments
I would like to thank Mitsuko Yamura-Takei for critical observations concerning the major points in this work, Francis Bond for his insightful and gentle comments, and an anonymous translator for the translations. I would also like to acknowledge the unflagging aid of Shigeaki Amano, Kaname Kasahara, Takashi Hamo, Tomoko Kawaguchi, and Yuuko Kanasugi in providing the Gainen Base analysis.

References

Boguraev, Branimir K. and Mary S. Neff. 2000. Lexical cohesion, discourse segmentation, and document summarization. Paper presented at RIAO-2000, Paris.

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, Stanford, CA.

Chafe, Wallace. 1979. The flow of thought and the flow of language. In Talmy Givon, editor, *Syntax and Semantics: Discourse and Syntax*, volume 12. Academic Press, New York, pages 159–182.

Chafe, Wallace. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Productions*. Ablex, Norwood, NJ.

Clark, Herbert H. 1977. Inferences in comprehension. In David LaBerge and S. Jay Samuels, editors, *Basic Processes in Reading: Perception and Comprehension*. Erlbaum, Mahwah, NJ, pages 83–112.

Cote, Sharon. 1998. Ranking forward-looking centers. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 55–69.

Fais, Laurel. 2001. Discourse issues in the translation of Japanese email. In *Proceedings of PACLING 2001*, pages 93–102, Kitakyushu, Fukuoka, Japan.

Fais, Laurel. 2002. The Japanese *A no B* construction and centering. In *Proceedings, Eighth Annual Meeting of the Association for Natural Language Processing (NLP 2002)*, pages 603–606, Keihanna, Japan.

Fais, Laurel and Mitsuko Yamura-Takei. 2003. The nature of referent resolution in Japanese email. *Discourse Processes*, 36(3):167–204.

Fujita, Sanae and Francis Bond. 2002. A method of adding new entries to a valency dictionary by exploiting existing lexical resources. In *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, pages 45–52, Keihanna, Japan.

Gordon, Peter C., Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

Gordon, Peter C., Randall Hendrick, Kerry Ledoux, and Chin Lung Yang. 1999. Processing of reference and the structure of grammar: An analysis of complex noun phrases. *Language and Cognitive Processes*, 14(4):353–379.

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

Grosz, Barbara J. and Candace L. Sidner. 1998. Lost intuitions and forgotten intentions. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 39–51.

Hahn, Udo, Katja Markert, and Michael Strube. 1996. A conceptual reasoning approach to textual ellipsis. In *Proceedings of the 12th European Conference on Artificial Intelligence*, pages 572–576, Budapest.

Halliday, Michael A. K. 1994. *An Introduction to Functional Grammar*. Arnold, London.

Harabagiu, Sanda M. 1998. WordNet–based inference of textual cohesion and coherence. In *Proceedings of FLAIRS-98*, Sanibel Island, FL, pages 265–269.

Harabagiu, Sanda M. 1999. From lexical cohesion to textual coherence: A data driven perspective. *Journal of Pattern Recognition and Artificial Intelligence*, 13(2):241–265.

Hudson-D'Zmura, Susan and Michael K. Tanenhaus. 1998. Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince,

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 199–226.

Huls, Carla, Edwin Bos, and Wim Claassen. 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79.

Hurewitz, Felicia. 1998. A quantitative look at discourse coherence. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 273–291.

Hutchins, John. 1995. Reflections on the history and present state of machine translation. In *The MT Summit V Proceedings*, pages 89–96, Luxembourg.

Kameyama, Megumi. 1998. Intrasentential centering: A case study. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 89–112.

Kasahara, Kaname, Kazumitsu Matsuzawa, Tsutomu Ishikawa, and Tsukasa Kawaoka. 1996. Viewpoint-based measurement of semantic similarity between words. In Douglas H. Fisher and Hans Lenz, editors, *Learning from Data: AI and Statistics V*. Springer-Verlag, New York, pages 433–442.

Kozima, Hideki. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 286–288, Columbus, OH.

Kozima, Hideki and Teiji Furugori. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the Sixth Conference of the European Chapter, Association for Computational Linguistics*, pages 232–239, Utrecht, the Netherlands.

Kuno, Susumu. 1973. *The Structure of the Japanese Language*. MIT Press, Cambridge, MA.

Lotfipour-Saedi, Kazem. 1997. Lexical cohesion and translation equivalence. *Meta*, 42(1):185–192.

Matsui, Tomoko. 1999. On the role of context in relevance-based accessibility ranking of candidate referents. In Paolo Bouquet, Luciano Serafini, Patrick Brezillon, Massimo Benerecetti, and Francesca Castellani, editors, *CONTEXT '99* (Springer Lecture Notes in Artificial Intelligence, no. 1688), pages 228–241.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Murata, Masaki, Hitoshi Isahara, and Makoto Nagao. 1999. Resolution of indirect anaphora in Japanese sentences using examples "X *no* Y (Y of X)." In *Proceedings of the ACL99 Workshop on Coreference and Its Applications*, College Park, MD.

Nakaiwa, Hiromi, and Satoshi Shirai. 1996. Anaphora resolution of Japanese zero pronouns with deictic reference. In *Proceedings of COLING-96*, pages 812–817.

Nariyama, Shigeko. 2000. *Referent Identification for Ellipted Arguments in Japanese*. Ph.D. thesis, University of Melbourne, Melbourne, Australia.

Nippon Telegraph and Telephone Corporation, Communication Science Laboratories. 1996. *Reference Manual for Morphological Analysis Program ALT-JAWS*. Keihanna, Japan: Nippon Telegraph and Telephone.

Okumura, Manabu, and Takeo Honda. 1994. Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 755–761, Kyoto, Japan.

Passonneau, Rebecca J. 1998. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 327–358.

Paul, Michael and Eiichiro Sumita 2001. A trainable method for pronominal anaphora resolution using shallow information. *Shizen Gengo Shori* [Journal of Natural Language Processing], 8(3):59–85.

Poesio, Massimo, Hua Cheng, Renate Henschel, Janet Hitzeman, Rodger Kibble, and Rosemary Stevenson. 2000. Specifying the parameters of centering theory: A corpus-based evaluation using text from application-oriented domains. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, pages 400–407.

Sack, Warren. 1999. Diagrams of social cohesion. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference (CICLing-2002)*, Springer-Verlag, Heidelberg/Berlin, pages 1–15.

Shimazu, Akira, Shozo Naito, and Hirosato

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022

Nomura. 1987. Semantic structure analysis of Japanese noun phrases with adnominal particles. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 123–130, Stanford, CA.

Strube, Michael, and Udo Hahn. 1999. Functional centering—Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

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

Tetreault, Joel R. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Vieira, Renata, and Massimo Poesio. 2001. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Walker, Marilyn A. 1998. Centering anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 401–435.

Walker, Marilyn A., Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–231.

Walker, Marilyn A., Aravind Joshi, and Ellen F. Prince. 1998. Centering in naturally occurring discourse: An overview. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*. Clarendon, Oxford, pages 1–28.

Yamura-Takei, Mitsuko, Miho Fujiwara, Makoto Yoshie, and Teruaki Aizawa. 2002. Automatic linguistic analysis for language teachers: The case of zeros. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING '2002)*, pages 1114–1120, Taipei, Taiwan.

Downloaded from http://direct.mit.edu/coli/article-pdf/30/2/119/1798022/089120104323093267.pdf by guest on 05 October 2022