Deciding on units of analysis within Centering Theory

MAITE TABOADA and LORELEY HADIC ZABALA

Abstract

Many efforts in corpora annotation start with segmenting discourse into units of analysis. In this paper, we present a method for deciding on segmentation units within Centering Theory (Grosz et al. 1995). We survey the different existing methods to break down discourse into utterances and discuss the results of a comparison study among them. The contribution of our study is that it was carried out with spoken data and in two different languages (English and Spanish). Our comparison suggests that the best unit of analysis for Centering-based annotation is the finite clause. The final result is a set of guidelines for how to segment discourse for Centering analysis, which is also potentially applicable to other analyses.

Keywords: corpus annotation; discourse units; anaphora; Centering Theory; Spanish.

1. Introduction

Any project that codes and analyzes discourse data needs to solve one major problem: how to segment discourse. Decisions in this area are sometimes based on theoretical grounds: following a particular theory of discourse implies adopting that theory's units of discourse. In other situations, decisions are practical: the method needs to be consistent across coders, and maybe apply to different languages. Identifying units of analysis is often desirable, as Chafe points out: "Researchers are always pleased when the phenomena they are studying allow them to identify units. Units can be counted and their distributions analyzed, and they can provide handles on things that would otherwise be obscure." (Chafe 1994: 58).

In this paper, we use the term 'utterance' or 'segment' to refer to the units of analysis in Centering Theory. In other applications, 'segment'

refers to the broad parts into which a discourse can be divided (e. g., introduction, thesis statement), or to discourse segments, recognizable because each has a purpose, an intention on the part of the speaker or writer (Grosz and Sidner 1986). We are not concerned with those high-level discourse segments here, but only with minimal units of analysis, typically interpreted to be entire sentences or finite clauses. There are, in fact, a number of possible minimal discourse units: the complete sentence (most often found in carefully constructed written text); the matrix clause plus dependent adjuncts, similar to the T-unit used in developmental studies and text analysis (Hunt 1977); the finite clause; the clause (finite or non-finite); the clause, with any adjuncts (including prepositional adjuncts); the speech act; or the turn-constructional unit (Sacks et al. 1974). In Section 2, we discuss which of those units were considered in our analysis.

We describe an effort to find the best segmentation method for applying Centering Theory (Grosz et al. 1995), a theory of focus of attention and anaphora, to spoken language. This is the first step in a large-scale corpus annotation project, with a view towards automatic anaphora resolution. We believe that deciding on units of analysis should be a first step in any large-scale analysis of Centering. One such effort (Poesio et al. 2004) considered mostly written texts in English. Our main focus is Spanish, but we also considered the applicability of different methods to English, to ensure a minimum of cross-linguistic validity. This paper is concerned with both English and Spanish data equally, although our future annotation efforts will involve mostly Spanish. Our contribution is a tested, validated method for utterance segmentation that can serve as the starting point for any Centering analysis. Although the focus of our analysis is quite narrow (an application of Centering Theory to spoken language analysis), it has wider implications. We show how one can devise methods and tests to decide on the minimal unit of corpus coding. The tests may vary from one project to the next, but overall the types of questions that we ask and the type of tests that we use should be applicable to a variety of projects in corpus coding.

The corpus used for this paper consists of two sets of English and Spanish telephone conversations, distributed by the Linguistic Data Consortium.[1] The CallHome corpus was an effort by the Linguistic Data Consortium to collect spontaneous telephone conversations. Participants were given thirty minutes of long-distance calling time, to call relatives or friends, provided they agreed to being recorded. There are CallHome-style recordings for a variety of languages. For this particular study, we chose five conversations in (American) English and five in Spanish. There is no detailed information on place of origin for the Spanish speakers, but we were able to identify a variety of dialects. In this sense,

the English corpus is more homogeneous, since most callers were speakers of American English. The conversations are about 30 minutes long, but only five minutes of each conversation were transcribed (Kingsbury et al. 1997; Wheatley 1996). We performed our analyses on the data that was already transcribed for each of the 10 conversations. This resulted in about 2,000 utterances for English and 1,500 for Spanish.[2]

The paper is organized as follows: In Section 2 we present some of the candidates for minimal unit of analysis. In Section 3, we provide a brief introduction to Centering Theory. In Section 4, we discuss the different segmentation methods proposed within Centering. Section 5 describes the measures used to test the best method, and Section 6 presents the results of different methods, when applied to a corpus of Spanish and English conversations. We conclude that a method following Kameyama's (1998) proposals (with the finite clause as basic unit) would be the most adequate for coding our corpus.

2. Defining units of analysis

In Centering, the issue of segmenting discourse into units of analysis is of foremost importance, since all other constructs within the theory are calculated utterance-by-utterance ('utterance' being the word used to refer to the minimal unit of analysis). Before we explain Centering Theory, and the issues around segmenting discourse into utterances, we would like to point out that determining units of analysis is not a problem exclusive to Centering; most studies of discourse and most efforts in corpus annotation need to grapple with segmentation.[3] In this section, we review issues around segmenting discourse into minimal units.

Mosegaard Hanse (1998) lists a number of potential units, grouped into: form-based units (sentence, clause, turn, tone unit, utterance); content-based unit (proposition); and action-based units (speech act and communicative act). We discuss some of those in this section.

In written discourse, units are often associated with sentences. A sentence is considered to be the smallest independent unit. In a reinterpretation of the sentence, many functional and developmental studies (Fox 1987; Fries 1994; Klecan-Aker and Lopez 1985), including studies of second language learners (Larsen-Freeman 1978), have taken the T-unit, or "minimum terminable unit" (Hunt 1977) as the basic unit of analysis, defined as a main or matrix clause with all its modifiers, including subordinate clauses. Sentences with two independent (coordinated) clauses contain two T-units. T-units have been adopted in other fields, such as studies of brain damage (Coelho 2002), language impairment (Restrepo and Gutierrez-Clellen 2001), or language disabilities in general (Scott and Windsor 2000).

The choice of T-units in analysis is not without its critics, who point out that traditional sentences may be a better choice (Bardovi-Harlig 1992), that the T-unit is not clearly defined (Barnwell 1988; Gaies 1980), or that it is not stable in length across genres (Witte 1980). The T-unit is typically used to show how first language learners increase T-unit length and complexity in the course of language development (O'Donnell 1976). Barnwell (1988) suggests that T-units are not useful measures in second language learning, because, on the one hand, they fail to capture other problems with L2 learners, such as morphological errors, while, on the other hand, they do not reward increased sophistication in vocabulary. Reed and colleagues (Reed et al. 2001), in a comparison of different methods, have shown that the choice of unit of analysis, or 'utterance', had an effect on the results of a syntactic complexity study.

T-units are also used in the analysis of spoken data, such as Bäcklund's study of telephone conversations (Bäcklund 1992). However, in spoken language other units are commonly adopted. Researchers within Conversation Analysis (Sacks et al. 1974) segment spoken discourse into Turn-Constructional Units (TCUs), defined as "[…] unit-types with which a speaker may set out to construct a turn. Unit-types for English include sentential, clausal, phrasal, and lexical constructions." (Sacks et al. 1974: 702). According to Schegloff (1996), both syntax and prosody contribute to the identification of TCUs, with syntax taking precedence. Other researchers have drawn attention to problems in the definition. Selting (2000) says the "possible completion point" and syntactic completeness criteria are sometimes in conflict: Some constructions are syntactically continued, but prosodically independent (Selting 2000: 481). She characterizes TCUs as "the result of the interplay of syntax and prosody in a given semantic, pragmatic, and sequential context" (Selting 2000: 511). Ford and Thompson (1996) and Ford, Fox and Thompson (1996) have also pointed out that a TCU is defined through the interaction of syntactic, prosodic and pragmatic completion. The characterization by Selting and Ford and colleagues is probably accurate, but it makes consistent and reliable segmentation difficult, since it involves a holistic approach to identifying each possible unit. Having different criteria converge would probably mean more chances for annotators to disagree when criteria are in conflict. One annotator may give more importance to, for instance, phonological criteria, whereas another one may consider syntactic criteria to be overriding.

In a recent paper, Thompson and Couper-Kuhlen (2005) argue for the clause as the basic unit of interaction. Clauses are "understood as [predicate + phrases that accompany it], while 'sentence' is a term reserved for a unit that can consist of either a clause or a combination of clauses" (Thompson and Couper-Kuhlen 2005: 499). Their argument hinges on

the notion of predicate (i. e., verb): hearers can anticipate where a predicate occurs, and often what the nature of the predicate will be. Hearers can then anticipate when an utterance (= a clause) will end, because they know what typically accompanies that predicate. This projectability characteristic of the clause will vary from language to language (Thompson and Couper-Kuhlen analyze English and Japanese); however, regardless of how the language constructs clauses, the clause will be the locus of the interaction.

Philosophers of language have contributed speech acts, units of talk defined by the speaker's intention behind uttering the act (Austin 1962; Searle 1969). Speech acts were adopted by researchers building spoken dialogue systems (e. g., Stolke et al. 2000; Traum and Hinkelman 1992), who found the usual difficulties in matching stretches of talk to speech or dialogue acts. The common practice is to have a very precise definition, and an annotation manual with examples and potential problem cases. For instance, the Linguistic Data Consortium provides a lengthy annotation manual for their Metadata Extraction (MDE) project.[4] Their definition of units is as follows: "One of the goals of metadata annotation is the identification of all units within the discourse that function to express a complete thought or idea on the part of the speaker. Often times this unit corresponds to a sentence, but sometimes a sentence will contain multiple units; other times a unit is semantically complete but smaller than a sentence.[5]" Eckert and Strube (2000: 69) define a dialogue act as "each main clause plus any subordinated clauses, or a smaller utterance". The definition is mostly syntactic, but it allows for "smaller utterances", that is, backchannels and elliptical utterances that are common in spoken language.

There is also an extensive literature, especially in computational linguistics, on the issue of segmenting discourse into 'chunks', 'segments', or larger units. These larger units are typically chunks that revolve around a single topic (Hearst 1994; Passonneau and Litman 1997). The chunks are decomposable into smaller units, what we here call utterances. In some models, the difference between 'chunks' and 'utterances' is one of length. For instance, in Grosz and Sidner's model (Grosz and Sidner 1986), discourse can be broken down into discourse segments, with each having a purpose (the intention that the speaker or writer has in producing that segment). Segments can be of varying length, and longer segments may be further broken down into smaller segments, with more specific purposes behind them.

In other approaches to discourse, discourse segments are the stages into which a particular discourse unfolds. Such is the case with Houtkoop and Mazeland's (1985) Discourse Units, larger 'projects' such as stories, jokes, extended descriptions and pieces of advice. In studies of

genre within Systemic Functional Linguistics (Eggins and Martin 1997; Eggins and Slade 1997), a text or a piece of spoken discourse that belongs to a particular genre has different stages, some of them optional and some obligatory. The stages are the chunks or discourse segments of each discourse, each one of them fulfilling specific goals, and showing characteristic linguistic properties in terms of syntax, vocabulary or discourse structures. Taboada, for instance, has studied the different stages of task-oriented conversation (Taboada 2004a), and of bulletin board posts (Taboada 2004b). We are not concerned with those larger chunks of discourse here, but only with the minimal units into which discourse can be broken down for the purpose of analysis.

It is obvious that there are multiple proposals for the problem. It seems that each group of researchers, each field that studies discourse, or even each new research project, devises a definition to suit their purposes. We illustrate the problem of segmenting discourse into units with a particular application: the annotation of conversations according to Centering Theory. A systematic approach to segmentation, even when it is borne out of a particular study, may have uses for other approaches.

The units we considered for our study are towards the smaller, more self-contained end of the spectrum. Since Centering is concerned with anaphora across units, our aim is to find a sentence or clause-like unit. Turns may contain more than one sentence, and therefore multiple instances of anaphoric references. As we discuss in Section 4, finite clauses and sentences (a main clause and any embedded or subordinate clauses attached to it) are the main contenders in our study.

3. Centering Theory[6]

Centering (Grosz et al. 1995; Walker et al. 1998a) was developed within a theory of discourse structure (Grosz and Sidner 1986) that considers the interaction between (i) the intentions or purposes of the discourse and the discourse participants, (ii) the attention of the participants and (iii) the structure of the discourse. Centering is concerned with the participants' attention and how the global and local structures of the discourse affect referring expressions and the overall coherence of the discourse. It models the structure of local foci in discourse, i. e., foci within a discourse segment. It has been applied to pronoun resolution (Brennan et al. 1987; Eckert and Strube 1999; Kameyama 1986a; Kehler 1997; Kim et al. 1999; Tetreault 2001; Walker 1998; Wolf et al. 2004; Yamura 1996), generation of referring expressions (GNOME 2000; Henschel et al. 2000; Hitzeman and Poesio 1998; Yuksel and Bozsahin 2002), text planning and natural language generation (Chiarcos and Stede 2004; Karamanis 2003; Kibble 1999), and modeling of coherence in discourse

(Barzilay and Lapata 2005; Hurewitz 1998), including measuring coherence for automatic scoring of essays (Miltsakaki and Kukich 2004).

Centers are semantic entities that are part of the discourse model of each utterance in the segment. Centers are always nominals; event and propositional reference have not been widely studied within the theory. For each utterance, Centering establishes a ranked list of entities mentioned or evoked, the *forward-looking center list* (Cf). The list is ranked according to salience, defined most often in terms of grammatical relations. The first member in the Cf list is the *preferred center* (Cp). Additionally, one of the members of the Cf list is a *backward-looking center* (Cb), the highest-ranked entity from the previous utterance that is realized in the current utterance.

Example (1) illustrates these concepts.[7] Let us assume that the utterances in the example constitute a discourse segment.[8] In the first utterance, (1a), there are two centers: *Harry* and *snort*. (1a) does not have a backward-looking center (the center is empty), because this is the first utterance in the discourse segment. In (1b), two new centers appear: *the Dursleys* and *their son, Dudley*. The lists include centers ranked according to two main criteria: grammatical function and linear order. (Ranking will be further discussed in Section 5.) The Cf list for (1b) is: Dursleys, Dudley.[9] The preferred center in that utterance is the highest-ranked member of the Cf list, i.e., Dursleys. The Cb of (1b) is empty, since there are no common entities between (1a) and (1b). In (1c), a few more entities are presented, and they could be ranked in a number of ways. To shorten the discussion at this point, we will rank them in linear order, left-to-right (i.e., the order in which they are mentioned). In any event, the most important entities seem to be the subject, which is the same as in (1b), Dursleys; and Dudley, realized in the possessive adjective *his* (twice). The Cp is Dursleys, since it is the highest-ranked member of the Cf list, and the Cb is also Dursleys, because it is the highest-ranked member of (1b) repeated in (1c). The next utterance, (1d), reintroduces Harry to the discourse, and links to (1c) through Dudley, which is the Cb in (1d).

(1)   a.   Harry suppressed a snort with difficulty.
      b.   The Dursleys really were astonishingly stupid about their son, Dudley.
      c.   They had swallowed all his dim-witted lies about having tea with a different member of his gang every night of the summer holidays.
      d.   Harry knew perfectly well that Dudley had not been to tea anywhere;
      e.   he and his gang spent every evening vandalising the play park, [...]

In (2) we see the Cf, Cp and Cb for each of the utterances in the segment, plus the transitions between utterances (discussed below):

(2) a. Harry suppressed a snort with difficulty.
         Cf: HARRY, SNORT
         Cp: HARRY − Cb: Ø
         Transition: ZERO
     b. The Dursleys really were astonishingly stupid about their son, Dudley.
         Cf: DURSLEYS, DUDLEY
         Cp: DURSLEYS − Cb: Ø
         Transition: ZERO
     c. They had swallowed all his dim-witted lies about having tea with a different member of his gang every night of the summer holidays.
         Cf: DURSLEYS, DUDLEY, LIES, TEA, MEMBER, GANG, NIGHT, HOLIDAYS
         Cp: DURSLEYS − Cb: DURSLEYS
         Transition: CONTINUE
     d. Harry knew perfectly well that Dudley had not been to tea anywhere;
         Cf: HARRY, DUDLEY, TEA
         Cp: HARRY − Cb: DUDLEY
         Transition: ROUGH SHIFT
     e. he and his gang spent every evening vandalising the play park, [...]
         Cf: DUDLEY, GANG, EVENING, PARK
         Cp: DUDLEY − Cb: DUDLEY
         Transition: CONTINUE

In addition to the different types of centers, Centering proposes transition types, based on the relationship between the backward-looking centers of any given pair of utterances, and the relationship of the Cb and Cp of each utterance in the pair. Transitions, shown in Table 1, capture the introduction and continuation of new topics. Cb_i and Cp_i refer to the centers in the current utterance. Cb_{i-1} refers to the backward-looking center of the previous utterance. Thus, a CONTINUE occurs when the Cb and Cp of the current utterance are the same and, in addition, the Cb

Table 1. *Transition types*

	$Cb_i = Cb_{i-1}$ or $Cb_{i-1} = Ø$	$Cb_i \neq Cb_{i-1}$
$Cb_i = Cp_i$	CONTINUE	SMOOTH SHIFT
$Cb_i \neq Cp_i$	RETAIN	ROUGH SHIFT

of the current utterance is the same as the Cb of the previous utterance. Transitions capture the different ways in which a discourse can progress: from how an utterance refers to a previous topic, the Cb_{i-1}, and it is still concerned with that topic, the Cp_i, in a CONTINUE, to how it can have a weaker link to the previous topic, in a ROUGH SHIFT. Transitions are one explanation[10] for how coherence is achieved: a text that maintains the same centers is perceived as more coherent.

We provide the transitions for each utterance in (2) above. The first utterance has no Cb, because it is segment-initial, and it therefore has no transition (or a zero-Cb transition). The transition between (2a) and (2b) is also zero. Between (2b) and (2c) there is a CONTINUE transition, because the Cb of (2b) is empty, and the Cp and Cb of (2c) are the same, DURSLEYS. Utterance (2d) has a different Cb from (2c), and it also shows different Cb and Cp, producing then a ROUGH SHIFT in the transition between (2c) and (2d). Finally, (2e) and (2d) are linked by a CONTINUE transition.

Because transitions capture topic shifts in the conversation, they are ranked according to the processing demands they pose on the reader. The ranking is: CONTINUE > RETAIN > SMOOTH SHIFT > ROUGH SHIFT. This transition ranking is often referred to as Rule 2 in the Centering paradigm. Centering predicts that CONTINUE will be preferred to RETAIN, and RETAIN to SHIFTS, all other things being equal. The preference applies both to single transitions and to sequences of transitions.

Other transition types have been proposed: Kameyama (1986b) added a center-establishment transition; Poesio et al. (2004) make a distinction between NULL and ZERO transitions. The difference hinges on whether to include $Cb_{i-1} = \emptyset$ in the CONTINUE and RETAIN transitions. In current work (Taboada 2008) we are exploring more refined transition classifications; for this paper we coded the corpora using the transitions in Table 1.

In addition to transitions, Centering proposes rules and constraints. Among them, the most relevant for us is Rule 1. Rule 1 captures the preference for pronouns when the same topic of discourse is continued. Its formulation is as follows:

Rule 1   For each U_i in a discourse segment D consisting of utterances $U_1, ..., U_m$, if some element of $Cf(U_{i-1}, D)$ is realized as a pronoun in U_i, then so is $Cb(U_i, D)$.

Rule 1 is sometimes referred to as the Pronoun Rule. It captures the fact that a topic that is continued from a previous utterance does not need to be signalled by more explicit means than a pronoun (or a zero pronoun, in languages that allow those). Other pronouns are of course al-

lowed in the same utterance, but the most salient entity must be realized by the least marked referring expression. In (2c) above, the backward-looking center, DURSLEYS, is realized as a pronoun, following Rule 1, since other pronouns are also present in the utterance (*his* to refer to DUDLEY).

The Pronoun Rule has different interpretations. The definition cited above comes from Grosz, Joshi and Weinstein (1995).[11] In the original formulation by Grosz, Joshi and Weinstein (1983), the rule stated that if the Cb of the present utterance was the same as that of the previous utterance, a pronoun must be used. Gordon et al. (1993) carried out experiments that indicated that the processing of utterances with non-pronominalized Cbs was slower. This they dubbed the Repeated Name Penalty, and established that Rule 1 should indicate a preference for pronominalizing the Cb, in all cases. Poesio et al. (2004), in a comparison of different versions of Rule 1, found that the version from Grosz et al. (1995) was the one most often satisfied in their corpus, that is, that the Cb tends to be pronominalized if anything else is. A generalization of Rule 1 is that the most salient entity (the Cb) tends to be pronominalized. This is useful in both anaphora resolution and in the generation of referring expressions.

All of these formulations of Rule 1 do not take into account zero pronouns in pro-drop languages. A reformulation of Grosz et al. (1995) for pro-drop languages is:

Rule 1 (reform.)   For each U_i in a discourse segment D consisting of utterances U_1, ..., U_m, if some element of $Cf(U_{i-1}, D)$ is realized as a zero pronoun in U_i, then so is $Cb(U_i, D)$. If there are no zero pronouns, and some element of $Cf(U_{i-1}, D)$ is realized as a pronoun in U_i, then so is $Cb(U_i, D)$.

One of the applications of Centering is to pronominal anaphora resolution. Brennan et al. (1987) proposed a method for ranking possible antecedents, based on transition rankings (e. g., that a pronoun interpretation that yields a CONTINUE transition is preferred to one that yields a RETAIN). In Example (2) above, most instances can be resolved through syntactic constraints: *they* and *his* in (2c) can be resolved as references to *Dursleys* and *Dudley* in (2b) through gender and number. In (2e), the pronoun *he* is ambiguous, since it could refer to either *Harry* or *Dudley*. Centering would predict that *he* refers to *Dudley*, since this results in a CONTINUE transition, and the previous and current Cbs refer to *Dudley*.

The introduction to Centering Theory outlined in this section is very brief, and it only contains the material necessary to follow the rest of the paper. More detailed descriptions can be found in Grosz et al. (1995) and Walker et al. (1998b).

4. Variations in Centering Theory

There are variations and disagreements in the interpretation of a number of the constructs in Centering, from the exact definition of backward-looking center (Cb) to how to populate the list of forward-looking centers, the Cf list (Poesio et al. 2004). The starting point of any application of Centering is the definition of 'utterance'. The notions of discourse segment and utterance are very important: Centering predicts the behaviour of entities within a discourse segment; centers are established with respect to the utterance. As it turns out, precisely what an utterance is has been a matter of debate and interpretation. As Centering came to be applied in different areas and to different languages, it underwent some transformations, most significantly in the method for utterance segmentation and in the ranking of the Cf list. We concentrate here on the segmentation methods for utterances, named after the author(s) who proposed them. We evaluated four of the proposals, which we discuss in the next sections, excluding the very general proposal by Grosz, Joshi and Weinstein (1995) that we describe in Section 4.1.

4.1. Grosz, Joshi and Weinstein (1995)

The original formulation of Centering Theory proposed that discourse is divided into utterances, which are the units of analysis in Centering. Segmenting is based on the discourse structure theory of Grosz and Sidner (1986), which divides discourse, first of all, into discourse segments. A discourse segment is recognizable because it always has an underlying intention associated with it. Discourse segments can also be embedded. They exhibit local coherence (among the utterances in the segment), and global coherence (with other segments in the discourse).

Each discourse segment can, in turn, be composed of utterances. No particulars are given as to how to segment discourses into utterances. In the 1995 paper, a few examples suggest that the sentence is the basic unit of analysis. Examples (3) and (13) in that paper, reproduced below as (3) and (4), show compound sentences with two coordinated clauses as single utterances.

(3) a. Terry really goofs sometimes.
   b. Yesterday was a beautiful day and he was excited about trying out his new sailboat.
   c. He wanted Tony to join him on a sailing expedition.
   d. He called him at 6 AM.
   e. He was sick and furious at being woken up so early.

(4) a. Have you seen the new toys the kids got this weekend?
     b. Stuffed animals must really be out of fashion.
     c. Susie prefers the green plastic tugboat to the teddy bear.
     d. Tommy likes it better than the bear too, but only because the silly thing is bigger.

This segmentation method constitutes the basis for all the other methods described below. Because other methods have been more explicit, we did not consider it in our comparison.

4.2. Kameyama (1998)

Kameyama tackles the problem of solving intra-sentential anaphora, and proposes to break down complex sentences into their constituent clauses which then become center-updating units, equivalent to utterances in the original formulation. She proposed a number of hypotheses that can be summarized as follows, where 'segment' means 'separate into a new utterance':

- Segment all coordinated clauses (finite or non-finite).
- Segment all finite subordinated clauses, in the order in which they appear.
- Do not segment non-finite subordinated clauses.
- Do not segment clausal complements (noun clauses) and relative clauses.

In summary, Kameyama's method is based on the finite clause. The only place where non-finite clauses become their own units is when they are in coordinate structures: Two coordinated non-finite clauses become two separate units. Finite clauses are those that contain a finite verb as their main predicate.

To illustrate each of the methods that we considered, we provide examples from our corpus. Following Kameyama's method, segmentation would proceed as shown in the Examples (5)−(8). Example (5) shows two coordinated sentences, each separated into a new utterance (a new line indicates a new utterance). In (6), the second conjunct does not have an explicit subject, but we still consider it as a new unit. This is similar to having a zero subject in Spanish (because Spanish allows pro-drop in general). Examples (7) and (8) illustrate two subordinate structures, the first one with the subordinate clause preceding the matrix, and the second one in the opposite order.

(5) a. but um we didn't have any trees
   b. so we don't didn't lose anything

(6) a. and then Friday we saw a thing or two
   b. and came back

(7) a. I mean unless you're going to be really trying to make money off them
   b. there's no point

(8) a. Isn't that a pain in the ass
   b. when they do that?

In (9), we see a longer stretch of text, with a subordinate clause in (9b). The *yeah* of speaker A is disregarded in the Centering analysis, since it is a backchannel (Yngve 1970) or an agreement with what the other speaker is saying. In our analysis, we do not count backchannels as units for the analysis, unless they introduce new entities in the discourse or refer to entities already present (Hadic Zabala and Taboada 2006; Taboada 2008). The utterance is relevant for the overall corpus analysis and it has a clear function, but it is not part of the progression or introduction of discourse entities. The 'previous utterance' for (9d) is (9b). We will see that the analysis for this example is different under the segmentation system of Suri and McCoy (1994), since for them (9a) and (9b) form one unit.

(9) B: a. You know I've got to wait for that to calm down
       b. before I do anything with it
   A: c. yeah
   B: d. And uh they're selling it like for uh six million uh uh …

On the other hand, non-finite subordinate clauses belong to the same utterance unit as their matrix clause. In Example (10), from the Spanish corpus,[12] the non-finite subordinate clause *para enganchar todo* ('to hook up everything') does not constitute a separate utterance and belongs in the same unit with the matrix clause.

(10) *porque   tenemos   unos   por   ahí   para*
     because have.1PL.PRES one.MASC.PL around there to
     enganchar   todo,
     hook.up.INF everything,
     'because (we) have some (modems) around to hook up everything,'

Clausal complements, whether finite or not, are part of the clause in which they are arguments. In Example (11) the matrix verb *prefiero* ('(I)

prefer') introduces a non-finite complement, *quedarme*. Matrix and embedded verb form a single utterance.

(11)   *prefiero   quedar-me   con lo   que tengo*
       prefer.1SG.PRES stay.INF-CL.1SG with CL.3SG that have.1SG.PRES
       '(I) prefer to keep what (I) have ...'

Complements create embedded discourse segments. They constitute Centering units (i. e., they have their own Cf lists), but do not update the centers for the following utterance. In (12), (12b) is a finite complement, with its own Centering structure. However, to create Centering structures for (12c), the previous utterance for (12c) is (12a), not (12b). The same analysis applies to relative clauses. We will see that Suri and McCoy (1994) use a similar updating strategy for subordinate clauses.

(12)   a. *sí, yo lo   quiero   convencer*
          yes, I   CL.3SG.MASC want.1SG.PRES convince.INF
      b. *[que me   haga   un préstamo a mí]*
          that CL.1SG do.3SG.PRES.SUBJ a loan   to me
      c. *ahí sí que se me   acaba   la ...*
          there yes that SE CL.1SG finish.3SG.PRES the.SG.FEM
      'Yes, I want to convince him to give me a loan, then (it) is really over for me...'

In (13), we show two contiguous instances of embedded reported speech (quoted as it was originally said), included together with the reporting part into a single unit.

(13)   a. And then they said can you come out
      b. I said no I can't leave work

We should also point out that, although Kameyama discusses complements in general, we believe that some 'complements' do not qualify as such. As Thompson (2002) has argued, predicates such as *think, guess, realize, believe, hope, mean,* and *be interesting* cannot be considered complement-taking predicates, but rather markers of epistemic stance, evidentiality or evaluation. In the sentence *I thought she might pull it out of the garbage* (Thompson 2002: 126), *thought* is not the main predicate; it is instead expressing the speaker's attitude towards the main predicate *pull*. We follow this view, and consider the above example a single utterance, with *pull* as its main predicate. The distinction between these attitudinal expressions and true predicates that take a complement is not

always clear-cut, however. We currently annotate as attitudinal expressions predicates such as the ones mentioned by Thompson, and only when they have a first person singular subject.

Other researchers also propose the finite clause as the unit of analysis in Centering. Hurewitz (1998) annotated excerpts from the Brown corpus (written language) and the Switchboard corpus (spoken) and defined utterances as finite clauses, with the exception of restricted and medial position relative clauses, which were joined in an utterance with their matrix clauses. Passonneau (1998) explains that, in her coding of the Pear stories corpus (Chafe 1980), identification of utterance units involved syntactic, prosodic, and performance factors. The syntactic clause unit was "roughly any tensed clause that was not a verb argument, not a restrictive relative clause, and not one of a small set of formulaic clauses that I refer to as interjection clauses" (Passonneau 1998: 334). The formulaic clause example provided is *you know.*

4.3. Suri and McCoy (1994)

Suri and McCoy (1994) present an alternative approach to Centering for solving pronominal anaphora, RAFT/RAPR (Revised Algorithms for Focus Tracking and Revised Algorithms for Pronoun Resolution). Although there are differences between their approach and Centering, they propose segmentation rules that can be applied to Centering. Theirs is not a full-scale approach to segmentation. Rather, Suri and McCoy raise one issue with regard to subordinate clauses in adjunct position. Specifically, they address complex sentences with the conjunction because (S1 because S2). Their proposal consists of processing linearly up to S2, but using S1 as the previous utterance for the sentence that follows the because sentence (S3).[13] They argue that pronoun antecedents are most often found in matrix clauses. Their rules are as follows:

− Segment complex clauses linearly.
− Consider a because clause as an utterance when computing Centering structures.
− Continue segmentation and processing, but ignore the because clause for Centering purposes: when processing S3, the previous utterance is S1, not the because S2 clause.

For example, in (14), there is a matrix clause (… *we'll try to let you know what the date is*), and a subordinate *because* clause. The *because* clause would receive its own Centering structure, separate from the matrix. But what concerns us here is that the antecedent for *otherwise he'll videotape you* is (14a). In this example, the antecedent for the subject of (14c), *he,*

is found in the embedded clause, (14b), which would be ignored under this segmentation method.

(14)   a. We'll I'll we'll try to let you know what the date is
        b. [because he he he really would prefer to either have you there]
        c. otherwise he'll videotape you

Suri and McCoy discuss *because* clauses exclusively. We have decided to generalize their approach to all complex clauses, regardless of the subordinate conjunction used. All complex sentences of the form S1 Sub-Conj S2 are separated into two different units, but the SubConj S2 clause is ignored for computing the following utterance (S3). Example (15), the same data as in (9) above, now shows slightly different segmentation. This is a subordinate clause with the conjunction *before*. The difference with (9) is that now the *before* clause is ignored when it comes to calculating the backward-looking center in (15d), and thus the previous clause for (15d) is (15a). Recall that (15c) is not part of the Centering analysis, since it is a backchannel.

(15)   B: a. You know I've got to wait for that to calm down
              b. [before I do anything with it]
        A: c. yeah
        B: d. And uh they're selling it like for uh six million uh uh …

As Poesio et al. (2004) point out, Suri and McCoy also do not address complex sentences of the form SubConj S1 S2. We decided to divide them into two separate units, which behave as regular utterances (i. e., SubConj S1 is used as the previous utterance for computing structures in S2; S2 is used as the previous utterance for computing S3). This is equivalent to a Kameyama-like analysis, as we have seen in Example (7) above.

   Kameyama (1998) provides evidence to support the segmentation of SubConj S1 S2 sentences into two clauses with arguments from anaphora resolution: a pronoun in the subordinate adjunct clause usually refers to an entity already in the discourse, and not to an entity introduced by the main clause (S2). By considering S1 as a separate unit, we can use Centering to resolve the anaphora in S1. This applies even in cases of backward anaphora, such as When he woke up, Bill was very tired. The pronoun he typically realizes an entity already in the discourse context, according to Kameyama.

For all other aspects of segmentation in this method (non-finite clauses, coordinated clauses), we followed Kameyama. This results in an equal number of utterances for both methods, but different Centering structures for some utterances.

4.4. Miltsakaki (2002)

This model proposes the sentence, with any adjuncts, as the basic discourse unit, supported with examples from Greek, English and Japanese. The main argument is that this method results in discourse structures and transitions that "reflect our intuitions about perceived discourse coherence" (Miltsakaki 2002: 329). Miltsakaki argues that anaphora resolution is not the only goal of Centering, and that accounting for discourse coherence should be more important. She presents examples to illustrate that treating subordinate clauses as utterances results in counterintuitive transitions.

The rules for utterance segmentation are then quite simple: an utterance corresponds to a sentence. The Cf list is constructed independently of surface order: the entities in the main clause are always higher up in the list than the entities in subordinate clauses. Entities in subordinate clauses are available as potential antecedents, but the most likely antecedents are in main clauses, since they are higher up in the Cf list.

Miltsakaki explains that the phenomena observed by Suri and McCoy (1994) are easily captured in a segmentation method where the sentence is the basic unit. When a pronoun in S3 refers to the main clause in a S1 because S2 structure, the pronoun can be resolved by considering the complex sentence as a Centering utterance. In cases of backward anaphora, Miltsakaki proposes to solve the antecedent internally within the Cf list, thus treating it as intrasentential anaphora, not discourse anaphora.

We repeat below some of the examples that we have already shown, but this time with Miltsakaki-style segmentation. In (16), the two coordinated clauses are part of the same unit.[14] Examples (17) and (18) show subordinate + matrix and matrix + subordinate structures, all part of a single unit.

(16)   but um we didn't have any trees so we don't didn't lose anything

(17)   I mean unless you're going to be really trying to make money off them there's no point

(18)   Isn't that a pain in the ass when they do that?

The example that we have already seen in the previous two segmentation methods, reproduced as (19) below, displays now a single unit for the matrix + subordinate clause.

(19) B: a. You know I've got to wait for that to calm down before
                 I do anything with it
      A: b. yeah
      B: c. And uh they're selling it like for uh six million uh uh …

One potential problem with this method is that, just like 'utterance' is problematic, 'sentence' is also difficult to instantiate, especially in spoken data. In the example that we have just seen, Example (19), it is not clear whether (19c) should also be part of the structure in (19a), since it seems to be coordinated to the previous unit. One reason for breaking (19c) off into its own unit is that speaker A may have felt that (19a) finished a unit, since that is where he inserted a backchannel. We find more of a problem in (20), a long string of coordinated and subordinated sentences.[15] The example could be a single sentence from a syntactic point of view, but it could also be broken down into smaller units. It is, however, difficult to say where, unless one performs a clause-based analysis.

(20) Y, ahora recién tuvo un accidente, la semana pasada, que rompió
      todo el auto, pero fue la culpa del otro así es que ahí están viendo
      si le van a pagar, para que le, porque le rompieron todo su au-
      tito, pues.'

'So, now he just had an accident, last week, that wrecked the whole car, but it was the other (person)'s fault so now they're seeing if they're going to pay him, so that, because they wrecked the whole car, you know.'

4.5. Poesio (2000)

Poesio (2000) presents a segmentation approach within the GNOME project. The goal of the GNOME project was to develop a system for generation of nominal expressions: proper names, definite descriptions, and pronouns (GNOME 2000). Reports on segmentation methods are varied. The annotation manual (Poesio 2000) describes a very detailed segmentation method, with units that could be used in an RST analysis (Mann and Thompson 1988). The units are clauses, whether finite or non-finite, including clausal subjects and clausal NP modifiers. Also units are PP modifiers of NPs, comparative phrases, and parentheticals. Other reports (Henschel et al. 2000) suggest that the basic unit is the finite clause. Clauses that contain complement clauses or relative clauses are single utterances. The final project report (GNOME 2000) states that the best results in terms of Centering were obtained when the unit was the sentence.

Since the methods reported in Henschel et al. (2000) and in GNOME are similar to others already considered, we decided to test the first

method proposed in the annotation manual (Poesio 2000), and see whether it fared better in spoken Spanish and/or spoken English.[16] This is a very detailed method in which segmentation is as follows:

− Segment all conjoined and adjunct clauses (whether finite or non-finite).
− Segment complement, subject, and relative clauses (whether finite or not).
− Segment coordinated VPs, and treat the second VP as having a zero subject.

In summary, the method takes the clause as the minimal unit of analysis, whether finite or not. This method would, then, differ from a Kameyama-style segmentation in that there would be more units. We should point out that this is called the "Poesio" method after the author of the publication manual (Poesio 2000), but Poesio has proposed different segmentation methods elsewhere, including an extensive comparison of several methods (Poesio et al. 2004).

For instance, non-finite clauses would be separated from their matrix clauses. Example (11), reproduced here as (21), consists now of two separate utterances. There is also a headless relative clause, *con lo que tengo,* which also has its own Centering structure.

(21)   a. *prefiero*
          prefer.1SG.PRES

       b. *quedar-me   [con lo   que tengo]*
          stay.INF-CL.1SG [with CL.3SG that have.1SG.PRES]
          '(I) prefer to keep what (I) have ...'

There have been other attempts at comparing different segmentation methods. Poesio et al. (2004) describe an evaluation of all of Centering's parameters, including segmentation. They found that it is difficult to decide on the best parameter setting, since there is a trade-off between Constraint 1 (that every utterance should have a Cb) and Rule 1 (that the Cb is a pronoun, if other pronouns are present). Even the definitions of Constraint 1 and Rule 1 depend on further specifications of what types of pronouns are affected by Rule 1 (e. g., relative pronouns, traces).

The results of their evaluation are, in general, that the best definition of utterance is the sentence. However, they point out that some of the results lead to counterintuitive or narrow-focused assumptions. Identifying utterances with sentences may produce better Centering structures, but it is in disagreement with other theories of discourse, such as Rhetorical Structure Theory (Mann and Thompson 1988). They conclude

that further studies in other genres are needed. We add that studies in other languages are also necessary. This is precisely what we provide here.

5. Measuring the performance of segmentation methods

There is some agreement on a few measures for what a good Centering method is. When there is a choice between Centering methods, the most appropriate has some or all of the following characteristics:

1. It produces few empty backward-looking centers (Cb).
2. It results in Cbs that coincide with the topic of the sentence/utterance, for the most part.
3. It produces more cheap than expensive transitions.
4. It provides Centering structures that can be used by anaphora resolution methods.

Most of these measures are taken from the comparison carried out by Byron and Stent (1998). They considered 1, 2, and 3, as methods for deciding between different variants of Centering.[17] Poesio et al. (2004) also performed comparisons of Centering parameters. Whereas Poesio et al. tested variants by computing violations of Centering constraints (e. g., Rule 1, Constraint 1), our evaluation was based on practical criteria: the structures produced should be intuitively plausible (e. g., Cb = topic) and they should contribute to a method for anaphora resolution.

Characteristic 1 is Constraint 1, that all utterances should have a Cb (Brennan et al. 1987). This does not concern utterances at the beginning of a discourse segment, which naturally do not have a Cb, or have one that is underspecified (to be established within the discourse segment). There are two versions of Constraint 1 (Poesio et al. 2004). The strong version is that all utterances in a discourse segment have exactly one Cb. The weak version is that all utterances should have at most one Cb. Grosz, Joshi and Weinstein (1995: 209−210) suggest that each utterance should have only one Cb, and disambiguation methods should be used when more than one candidate exists. This has been our approach, i. e., we have used the weak version of Constraint 1.

Characteristic 2 captures the fact that the Cb is the most salient entity from the previous utterance continued in the current utterance. The term 'backward-looking center' was used to avoid confusion with other definitions of topic and with the notion of focus.[18] Typically, the topic is also a salient entity. Thus, we can expect that a 'traditional' definition of topic will coincide with Cb. Although there may be new topics being introduced, and a sentence may have no Cb or no topic, we expect topic and Cb to coincide often.

Characteristic 3 refers to a distinction proposed by Strube and Hahn (1996; 1999), where transitions are ranked based on inference load. They introduced two new transitions, in addition to those in Table 1. Since the Cp of an utterance is the expected Cb of the next utterance, a transition is cheap if the prediction holds, and expensive otherwise. In other words, in a cheap transition, $Cb(U_i) = Cp(U_{i-1})$; in an expensive transition, $Cb(U_i) \neq Cp(U_{i-1})$.

Characteristic 4 is obviously necessary, if our final goal is to provide a good method for anaphora resolution. Eventually, we want to use Centering to solve anaphora, maybe with an existing method (Brennan et al. 1987; Strube 1998; Strube and Hahn 1999), or with a new algorithm. This is not the goal of all Centering-based analyses. Poesio et al. (2004) are concerned with generation issues, but we are more interested in parsing and anaphora resolution. Nevertheless, we would like to explore the possibility of a unified Centering method for both parsing and generation. There are different possible interpretations of this constraint, but ours was focused on the starting point of anaphora resolution, finding the antecedent for an anaphor. Centering deals exclusively with anaphoric relations across utterances: The antecedent for a pronoun in the current utterance can be found by examining the possible antecedents in the previous utterance. Intra-sentential anaphora has to be addressed through other methods than Centering (e. g., Binding Theory, agreement and syntactic constraints in general). Therefore, the first constraint to find the antecedent for a pronoun in the current utterance is that the antecedent cannot be in the current utterance. When Centering utterances are entire sentences, it is more likely that both pronoun and antecedent are in the same utterance. Miltsakaki (2002) understands that this is a problem with her segmentation proposal, but proposes to address it by using methods other than Centering for solving those anaphors.

We have chosen these characteristics as the basis for our comparison. The characteristics are not necessarily restricted to constraints within Centering, and can be widely interpreted to be characteristics of coherent discourse. Characteristic 1 refers to the preference to maintain a topic: Sequences of utterances that concern themselves with the same entity form more coherent discourse segments. Usually, an empty backward-looking center means that there was no entity in common between the current and the previous utterance. Since that is supposed to be rare within discourse segments,[19] a method that minimizes the number of empty Cbs is more likely to capture coherent discourse.

Characteristic 2 is motivated by a desire to bring Centering in line with other approaches to discourse. The Cb has been characterized as a

salient entity, which is elsewhere considered a topic. Thus, a Cb which coincides with a topic is probably a well-motivated Cb.

Characteristic 3 refers to the establishment of expectations about where the discourse is going. The Cp, the most salient entity of the Cf list, is usually expected to be the Cb, the most salient entity of the next utterance. When $Cp_{i-1} = Cb_i$, the transition is considered cheap. A higher number of cheap transitions in our different configurations could indicate that the Cf list have been ranked correctly. This is, at least, what Strube and Hahn expected would be the case (Strube and Hahn 1996, 1999). A caveat with this measure is that corpus studies have not borne this out, even with different instantiations of the theory: Poesio et al.'s (2004) study found many more instances of expensive than cheap transitions.

Finally, Characteristic 4 is, as we already mentioned, motivated by the desire to apply this annotation effort to anaphora resolution, and the constraint that Centering deals only with intersentential anaphora.

We (the two authors) annotated the transcripts for five Spanish and five English conversations from the CallHome corpus following each of the methods outlined above: Kameyama, Suri and McCoy, Miltsakaki, and Poesio. We segmented them, and then built Centering structures for each method. Once that work was done, we tested the four characteristics, as measures of goodness for each segmentation method. We examined Characteristic 1, percentage of empty Cbs; Characteristic 2, percentage of Cb and topic matches; Characteristic 3, cheap and expensive transitions; and Characteristic 4, the percentage of pronouns that had an antecedent inside the sentence.

The most difficult part was deciding on the topic for an utterance. Topic, in general, is 'what the sentence is about', an entity given in the relational sense, as compared to a new entity, the focus (Gundel and Fretheim 2004).[20] The topic considered is the sentence topic, not the discourse topic. In order to assign topic to an utterance we used the following four criteria:[21] (i) we tried to determine what question the utterance was answering (Gundel 1977); (ii) we used the *as for* test (Gundel 1977); (iii) the *say about X that S* test (Reinhart 1981); and (iv) pseudo-cleft tests (Cohan 2000).[22]

To illustrate the tests, we show how the topic for (1) was determined. The speaker is talking about the weather, and complaining that it is very muggy. She then says that the house breathes (it lets a lot of air in). The topic for the example was considered to be *house*. It answers the question *What happens to the house?* or *What about the house?* It can be placed in an *As for* expression: *As for the house, it breathes*. It is also the topic picked in the *Say about X that S* test: *Speaker B said about the house*

Table 2. *Utterances for which no topic could be identified*

	English		Spanish	
	Total utterances	Utts. without a topic	Total utterances	Utts. without a topic
Kameyama	1268	399 (31.47%)	1024	315 (30.76%)
Suri and McCoy	1268	404 (31.86%)	1024	318 (31.05%)
Miltsakaki	1187	364 (30.66%)	954	278 (29.14%)
Poesio	1625	778 (47.88%)	1276	531 (41.61%)

that it breathes. There are several possible pseudo-clefts, among them: *What the house does is breathe*.

(1)   but uh the house here uh breathes

The topics were assigned in consultation. We assigned topics separately, and we decided together on difficult cases, where we had had no agreement. We found that many utterances did not have a clear topic, and those were left outside the comparison. Topic assignment was, in a sense, easier for the Poesio/GNOME segmentation, because there was often no topic, or only one clear topic, without competitors in the same utterance. In the Kameyama-style segmentation, assignment was also straightforward, because utterances usually contained at least one topic candidate. The most difficult was Miltsakaki's segmentation, since the tests we used are not well suited for complex clauses. We determined the topic by looking at the main clause. For example, in (23), which is shown here segmented according to Miltsakaki, we could not determine a topic for the utterance. There were too many possible entities that could be the topic (*you, the Pleasantville job, resume*). Since we could not agree, we labelled this example as having no topic (or no clear topic).[23] Table 2 lists the total number of utterances for each segmentation method, and how many of them did not have a clear topic in our analysis, and were thus left out of the comparison.

(23)   but a- also it seems to me if you can survive the Pleasantville job you'd certainly a great uh uh on your resume.

We did not experiment with different methods for ranking the list of forward-looking centers (Cf). We used a standard method for English, based on grammatical function (Subject > Indirect Object > Direct Object > Other), and linear order for disambiguation (e. g., when there is

more than one Other). The Spanish Cf template, shown in (24), is based on our ongoing work in Spanish (Hadic Zabala and Taboada 2006; Taboada 2002, 2005, 2008). It relies on grammatical function, except for psychological verbs (the equivalent of *It pleases me* in English), where the Experiencer is higher than the Subject.

(24)   Experiencer > Subj > Animate IObj > DObj > Other > Impersonal/Arbitrary pronouns

To populate the Cf list, we allowed indirect realization of entities: null subjects; member:set relations (Mom:Mom and Dad) and part:whole relations (branches:trees). We found that a strict direct realization (where the entities have to be mentioned explicitly in the utterance) resulted in a large number of empty Cbs. What exactly an indirectly realized entity is may, of course, not be obvious. We used the relations identified by Halliday and Hasan (1976) as lexical cohesion (synonymy, hyponymy, and superordinate, but not collocation, which does not necessarily link entities with the same referent). As described below, we tested our agreement on this task. We also include first and second person pronouns in the Cf list.

Coding was done by the two authors, first separately, and then comparing results and reaching consensus. To make sure that our coding was reliable, we also compared this agreed-upon coding with the coding of a third analyst, who had received a few sessions of training with the coding.[24] We compared one English and one Spanish conversation, for all the potentially subjective measures: segmentation, Cf ranking and topic assignment. We compared the conversation using Kameyama's segmentation method (because it is the one we eventually decided on). It is not clear to us that standard measures of agreement, such as kappa (Carletta 1996), are appropriate for this comparison, since there are no pre-defined categories to assign, such as a specific number of speech acts, or a yes/no distinction (e. g., is the sentence subjective or not?). Rather, the categories are defined by each sentence: each will have a different list of entities, and a different topic, depending on the entities present. For this reason, we calculated percentages of agreement.

Agreement on segmentation was 91.89 % in the English conversation, and 92.89 % in Spanish. The full details are as follows: In the English conversation, the number of segments for the composite of the first two coders was 407. Of those, Coder 3 had exactly the same segmentation in 374 of the utterances (91.89 %). In Spanish, the first two coders divided the conversation in 422 utterances. Coder 3 performed the same segmentation for 392 of those (92.89 %). Disagreements were mostly related to spoken language issues: whether to separate false starts and some

backchannels into a new utterance, or whether to make them part of the previous utterance.

The two sets of coders (i. e., the two authors vs. Coder 3) agreed 77.34 % of the time when it came to populating the Cf list in English. An error analysis of the disagreements (92 out of 406 utterances) shows interesting results: only 6 were due to a disagreement on selecting completely different entities (e. g., ambiguous pronouns). In Example (2), the last utterance had a different set of entities for the two codings, having to do with the referent for *her*. The first set of coders thought *her* referred to *Susan*, whose phone speaker B was using. The third coder thought it referred to *Gal*, who was the person that the speaker had called.

(2)   B: Gal got pierced ears
            she she told me
      A: Oh my God you're kidding me
      B: yeah I I called from Susan's
            because um because it's cheap from Friday night right
            so I was at Susan's house Friday night
            so I I I called ATandT
            and I asked how much it would be
            and I gave her the money right there

A majority of the disagreements (72 in total) were due to entities missing from the Cf list, and those were mostly temporal and propositional references (which, most likely, would not affect the calculation of the Cb for the following utterance). Finally, 14 disagreements had to do with the ranking, in particular with possessives, and when it came to deciding whether a third person plural pronoun was impersonal (thus ranked last) or not. Agreement in the Spanish conversation is similar: 76.13 % (252 out of 331 utterances that had a Cf list), and again most of the disagreements (53) were less important entities missing from the list. Table 3 summarizes the figures for Cf ranking. The first row shows the total

Table 3. *Summary of disagreements in Cf ranking*

	English	Spanish
Coder 3 agreements/	314/406	252/331
Coders 1 + 2 utterances	(77.34 %)	(76.13 %)
Disagreements:		
Different entities	6	5
Missing entities	72	53
Different ranking	14	21

Table 4. *Summary of disagreements in topic assignment*

	English	Spanish
Coder 3 agreements/	281/338	240/332
Coders 1+2 utterances	(83.14%)	(72.29%)
Disagreements:		
One assigned, other did not	19	30
Different topics	38	62

number of utterances that Coders 1 and 2 (the two authors) had agreed upon (406 for English), and for how many of those Coder 3 also had the same Cf ranking (314, that is, 77.34% in English). The next three rows show a breakdown of the disagreements into different types.

The agreement on topic assignment is summarized in Table 4 below. The coders agreed 83.14% of the time in English (281 out of 338 utterances). Of the 57 utterances where the two sets of coders (the two authors vs. the third coder) did not agree, 19 (33.3%) were disagreements in whether there was a topic or not: one coder found a topic for the utterance; the other did not. The rest of the disagreements were about which particular entity was the topic. In the Spanish conversation, agreement was 72.29% (240 out of 332 utterances), and a similar percentage of the disagreements (32.61%) were due to a disagreement over whether a topic could be established, not about the topic itself.

We believe that the levels of agreement are high enough to warrant a comparison among the segmentation methods. The comparison performed for this paper is based on the initial coding done by the two authors, with some changes after the reliability study with the third coder that we have reported on in this section. It is clear to us now that most disagreements can be solved with further training and annotation experience. We have annotated a further five conversations in each language (Taboada 2008), and we find that the process is more streamlined and the instructions easier to follow, especially since we update our coding manual to reflect current practices (Hadic Zabala and Taboada 2006). We believe that the level of disagreement in the data for this paper is low enough that the results of the comparison will be sufficiently informative and reliable for us to choose one of the segmentation methods for all future coding.

6. Results

We created Centering structures for each of the four segmentation methods. Tables 4 and 5 present the transitions for each method. The percent-

Table 5. *Transitions for each method, in English*

	Continue	Retain	Smooth Shift	Rough Shift	Utterances
Kameyama	42.67%	11.28%	11.99%	2.84%	1,268
Suri and McCoy	42.51%	11.36%	11.75%	2.84%	1,268
Miltsakaki	42.04%	11.79%	12.30%	3.12%	1,187
Poesio	36.74%	7.45%	7.32%	1.66%	1,625

Table 6. *Transitions for each method, in Spanish*

	Continue	Retain	Smooth Shift	Rough Shift	Utterances
Kameyama	45.51%	12.01%	10.55%	3.61%	1,024
Suri and McCoy	45.41%	11.72%	10.74%	3.52%	1,024
Miltsakaki	45.39%	12.16%	11.84%	3.88%	954
Poesio	42.24%	9.80%	7.76%	2.19%	1,276

ages are with respect to the total number of utterances considered. They do not add up to 100, because the rest are utterances with an empty Cb.

In general, Centering's Rule 2 (that CONTINUE transitions are more frequent than RETAIN, and those more than SMOOTH and ROUGH SHIFT; see Section 3) holds. In the corpus, most of the non-zero transitions are CONTINUE, and SMOOTH SHIFT is preferred to ROUGH SHIFT.[25] In English, however, there is a slightly higher number of SMOOTH SHIFT than RETAIN for all methods except Poesio. Although Rule 2 has been confirmed in other corpus-based analyses (Taboada 2002), it is possible that the preference of RETAIN over SMOOTH SHIFT is fragile (Kibble 2001).

6.1. Characteristic 1: Empty Cbs

The numbers of empty backward-looking centers for each language are shown in Table 7.

Table 7. *Empty backward-looking centers*

	English	Spanish
Kameyama	396 (31.23%)	290 (28.32%)
Suri and McCoy	400 (31.55%)	293 (28.61%)
Miltsakaki	365 (30.75%)	255 (26.73%)
Poesio	761 (46.83%)	485 (38.01%)

The worst method in this aspect is Poesio, which produces 46.83 % of empty Cbs in English, and 38.01 % in Spanish (in relation to the total number of utterances for each method). As we increase the length of utterances, the number of zero Cbs decreases. Kameyama has more zero Cbs than Suri and McCoy, and the latter more than Miltsakaki. It is obvious that the more entities included in an utterance, the more the chances that one of those will have a connection to the previous utterance. Contrary to what we expected, linking main clauses and ignoring intervening adjunct clauses in Suri and McCoy did not result in a lower number of empty Cbs. In Example (26), the antecedent for the zero first person plural in (26c) is to be found in (26b). If we ignore (26b), the method produces an empty Cb for (26c).

(26)   a. *bueno, mandan   todos los impuestos de la   zona*
            well,   send.3PL.PRES all   the taxes   of the area
            'Well, (they) send all the taxes for the area'
        b. *para que los   cobremos*
            so   that CL.3PL collect.2PL.SUBJ
            'so that (we) can collect them'
        c. *viste,   entonces bueno, funcionamos,   eh,*
            you.see, then   well,   work.2SG.PRES uh
            'you see, then, well, (we) manage, uh'

We compared the four methods according to each characteristic, to see if there were statistically significant differences among them. We used ANOVA, with a randomized complete block design, where each conversation is a block (Montgomery 2005). All tests were done at the $\alpha = 0.05$ level. The results showed a statistically significant difference (p-value $<$ 0.01) in mean levels among the methods both for English ($F_{3,11} =$ 178.46) and for Spanish ($F_{3,10} = 118.95$).[26] This just tells us that there is some difference among the methods, but not which one or ones are different. A Least Squares Means test using Tukey's HSD (Tukey 1953) shows that Poesio is significantly different from the other methods. We re-run the model, leaving out the conversations coded according to Poesio's method, and this time we found no statistically significant difference in English. In Spanish, on the other hand, Miltsakaki proved to be different from Kameyama and Suri and McCoy ($F_{2,8} = 7.68$, $p < 0.05$).

6.2. Characteristic 2: Topic = Cb

Table 8 displays the number of Cbs that coincided with what we assigned as topic for the utterance. The percentages are with respect to the total number of utterances for the method. Recall that a number of utterances

Table 8. *Coincidence of topic with Cb*

	English	Spanish
Kameyama	714 (56.31%)	570 (55.66%)
Suri and McCoy	712 (56.12%)	567 (55.37%)
Miltsakaki	666 (56.11%)	540 (56.60%)
Poesio	721 (44.37%)	596 (46.71%)

were left out of this analysis, because we could not establish a clear topic for them (see Table 2 above).

The results are close, with the exception of Poesio. In Spanish, Miltsakaki's method is the best, but in English it is Kameyama's, by a small margin. The ANOVA tests show differences in both languages ($p <$ 0.001), due to Poesio (English $F_{3,10} = 230.37$; Spanish $F_{3,12} = 54.69$). Once Poesio was removed, there was no difference in English, but a small one in Spanish ($F_{2,8} = 7.69$, $p < 0.05$).

In many cases, there was a topic, but not a Cb. Those cases were excluded from the results that we are considering, since we are analyzing only utterances with a Cb. In Example (3), speaker A starts a new topic, dealing with construction, and that was assigned as the topic for her utterance. Since there were no entities connecting the utterance by speaker B and that by speaker A, the second utterance had no Cb. It would have been possible to include *construction* as part of *what is new*, as an indirectly-realized entity. We thought that would be beyond indirect realization for this example.

(3)   B: but uh but anyway so what else is new you know
       A: well let's see our construction has started

Of the cases that had a Cb, Cb and topic did not coincide most often because the topic was a general discourse topic, but it had not been mentioned in the immediately preceding utterance, and therefore it was not part of the preceding Cf list. In Example (4), the two speakers have been discussing Gary, a friend of B's, or possibly her partner. Gary has had an offer for a teaching job in a black community. Gary has been the topic all along, but in (28a), he is not mentioned. The Cb for (28a) is empty (the previous entity did not include a reference to *it*, the community or the school), and the topic is *community*. Utterance (28b) is ignored for Centering purposes, since it is a backchannel. When we come to process (28c), *community* is the Cb, because it is the only entity that (28a) and (28c) have in common. However, we both determined that the

topic for (28c) is *Gary*. In this example, then, topic and Cb do not coincide.

(4) B: a. eh yeah it's uh it's like a eh it's just about a totally black community now
     A: b. no kidding
     B: c. and anyway the more he hears about the community the more he hears …

Space precludes a full analysis of each case where topic and Cb did not coincide, but we would like to point out that, in the best case, only a bit over half of the Cbs were sentence topics. This is possibly a result of the type of data under analysis: Conversations between relatives and friends are bound to include entities and topics that do not need to be specified in the discourse, but that are available as referents. Eckert and Strube (2000) mention this as a reason to use the Switchboard corpus for analysis. Switchboard (Godfrey and Holliman 1993) is a collection of telephone conversations between strangers that were assigned a common topic to discuss. In Switchboard, presumably, there are fewer shared assumptions that the observer does not have access to.

In order to better understand the data, we present the percentages of utterances that had a topic, for each method, in Table 9. (Note that figures for the Suri and McCoy method are exactly the same as those for Kameyama.) The first column for each language displays the raw number of utterances with a Cb. The second column shows the percentage with respect to all utterances in the corpus. The third column contains the percentage of utterances with a topic, with respect to utterances that had a Cb. This is a more informative percentage, since utterances that had no Cb are also likely to present more difficulties in topic assignment. In general, we see that around 50% of all utterances had an easily identifiable topic, and around 70% of utterances with a Cb had a topic

Table 9. *Utterances with a topic, and percentage with respect to all utterances and to utterances with a Cb*

	English			Spanish		
	Utts. with topic	All utts.	Utts. with a Cb	Utts. with topic	All utts.	Utts. with a Cb
Kameyama	868	50.80%	68.51%	709	53.94%	69.23%
Suri and McCoy	868	50.80%	68.51%	709	53.94%	69.23%
Miltsakaki	823	50.80%	69.33%	676	54.25%	70.86%
Poesio	847	40.10%	52.12%	745	46.88%	58.38%

assigned to them. The low number of utterances with an easily identifiable number could be explained using the argument by Eckert and Strube mentioned above: Topics were difficult to identify given the amount of common ground and presuppositions shared by the speakers.

6.3. Characteristic 3: Cheap and expensive transitions

We calculated the number of cheap and expensive transitions in relation to the total number of utterances for each method (Table 10). The percentages of cheap and expensive for each method do not add up to 100%, because of utterances with empty Cbs.

Table 10. *Cheap and expensive transitions*

	English		Spanish	
	Cheap	Exp.	Cheap	Exp.
Kameyama	658 (51.89%)	213 (16.80%)	564 (55.08%)	170 (16.60%)
Suri and McCoy	660 (52.05%)	207 (16.32%)	556 (54.30%)	175 (17.09%)
Miltsakaki	615 (51.81%)	208 (17.52%)	536 (56.18%)	163 (17.09%)
Poesio	696 (42.83%)	169 (10.40%)	619 (48.51%)	172 (13.48%)

We performed tests of significance based on the percentage of cheap transitions, showing a difference ($p < 0.01$) across methods and languages (English $F_{3,12} = 32.98$; Spanish $F_{3,11} = 61.23$). As with the other methods, we excluded Poesio from the analysis, which showed that there was no difference in mean levels of the response for English, but a significant one in Spanish ($F_{2,8} = 6.40$, $p < 0.05$). This means that Poesio is significantly different from the other methods, and no other method is different in English. In Spanish, there was a different among the three remaining methods. The change from the previous analyses is that, in this case, the Spanish difference was between Suri and McCoy and Miltsakaki. We can see in the table that Miltsakaki has more cheap transitions than any of the other methods, but a very similar percentage of expensive ones. There was no significant difference when performing the test on the percentages of expensive transitions.

The percentages of cheap and expensive transitions that we found contradict Poesio et al.'s (2004) results, who found many more expensive transitions, and expensive transition pairs (82.23% expensive transitions, albeit in a particular instantiation of Centering, which does not exactly correspond to our settings). It is also surprising that our results are the opposite of the English CallHome analysis by Byron and Stent (1998). This could be because they used different segmentation and Cf ranking

methods, but we believe the most likely reason for the discrepancy is that the lexical cohesive ties give us a higher number of realizations. That means that we have more non-empty Cbs, and more transitions, than other studies. We carried out a replication study of exactly the same three conversations that Byron and Stent analyzed, and found that, using our segmentation and realization methods, we have more cheap than expensive transitions. This is confirmation that different interpretations of the Centering parameters potentially yield very different results. It also means that our approach to segmentation is better than Byron and Stent's, if the preference for cheap over expensive transitions is a criterion.

6.4. Characteristic 4: Pronoun antecedents

We calculated the number of pronouns that had an antecedent in the same utterance. For this purpose, we excluded first and second pronouns, because they are deictic, and thus not covered by a Centering-based pronoun resolution method. We also excluded relative pronouns. The list included: zero, clitics, third person, demonstrative, and (independent) possessive pronouns. The ideal method is one that gives us fewer instances of antecedents in the same utterance as the pronoun that refers to them, since Centering does not help solve that type of anaphora, but only anaphora across utterances. Results are presented in Table 11.

Table 11. *Pronoun antecedents inside the utterance*

	English	Spanish
Kameyama	60 (8.01%)	68 (11.76%)
Suri and McCoy	60 (8.01%)	68 (11.76%)
Miltsakaki	87 (11.61%)	89 (15.40%)
Poesio	13 (1.73%)	13 (2.25%)
n	749	578

From the percentages in the table, the worst method is Miltsakaki. This is to be expected, since her utterances include all adjunct clauses, and therefore more antecedents are found within the same utterance as the pronouns that refer to them. This is the only test that shows an advantage for Poesio. We also performed tests of significance in this method, but their results are questionable, due to the small numbers of pronouns per conversation. The results do show that Poesio is, again, significantly different (English $F_{3,12} = 23.52$. $p < 0.01$; Spanish $F_{3,12} =$

33.90, $p < 0.01$). Once Poesio is removed, Miltsakaki is also different from the other two (English $F_{2,8} = 16.08$, $p < 0.01$; Spanish $F_{2,8} = 36.97$, $p < 0.01$), performing worse.

7. Discussion

In summary, we found that the differences among methods did not show a clear advantage for any of them. The only clear result is that the more fine-grained Poesio/GNOME approach does not offer enough benefits. The only measure where it performs better is in the number of pronouns with antecedents inside their utterance. The statistical results show that it is different from all the other methods, performing worse: a higher number of empty Cbs, and fewer Cbs that are also topics. Its only advantage is that, since units are so small, most pronouns have antecedents outside their utterance.

There were no conclusive results from the statistical tests once the Poesio method was removed from the running. We still have three different methods that seem to perform quite similarly. We decided to weigh in different considerations about each method, and to proceed in elimination fashion, until we arrived at one that we found satisfactory in most aspects. After Poesio, Suri and McCoy was eliminated next for a number of reasons: it resulted in fewer Cbs that coincided with our definition of topic in English (as compared to Kameyama); it had more empty Cbs; it had fewer Cbs that coincided with the topic of the utterance; and it had fewer cheap transitions in English. Although Suri and McCoy specifically addressed pronoun antecedents, we found some counterexamples, as shown in (5). The antecedent for the zero object pronoun in (29d) is in the preceding adjunct clause, which would be ignored in the Cf list of (29d).

(5)   a. *Hay   hay   cosas que yo puedo   hacer*
        there.is there.is things that I   can.1SG.PRES do.INF
        'There are, there are things that I can do,'

     b. *pero hay   muchas cosas   que   no*
        but   there.is many   things that not
        'but there are many things that (I) can't'

     c. *porque   necesito   instrumentos de precisión*
        because need.1SG.PRES instruments   of precision
        'because (I) need precision instruments'

     d. *y   yo no   tengo*
        and I   not have.1SG.PRES
        'and I don't have (any).'

The decision boiled down to either Miltsakaki or Kameyama. We believe that there are good arguments for pursuing either. Poesio et al. (2004) found that the sentence seemed to be the best unit for Centering purposes. We agree, especially when applied to Spanish. However, they also point out that it is dangerous to equate unit of analysis with the sentence for all computational applications, or for all discourse analysis. This would mean a discrepancy between Centering and other analyses such as RST (Mann and Thompson 1987), which most often uses the finite clause as basic unit.

The statistical tests showed that Miltsakaki sometimes behaved significantly differently from the Kameyama and Suri and McCoy methods, and only in Spanish. It was a better method in terms of having fewer empty Cbs, and in that the Cbs it identified coincided with the topic of the sentence more often. The cheap/expensive transition distinction showed that Miltsakaki in Spanish had more cheap transitions, but also more expensive ones than the other methods (due to differences in the number of empty Cbs it produced). It was the worst method for inter-utterance anaphora resolution, since many more pronouns had antecedents in the same utterance. (Recall, however, that the statistical tests for the pronoun measure are not as reliable, due to small numbers of pronouns.)

Statistics failing to help us make a clear decision, we have decided to pursue a Kameyama-style analysis. The most important reason is that we found that it was easier to segment spoken discourse when the basic unit was the finite clause.

Another recently completed study lends support to this choice. Hadic Zabala (2007) investigated the segmentation of relative clauses in Centering Theory for two types of Spanish texts, 10 telephone conversations extracted from CallHome and a fragment from the novel *Cien Años de Soledad* (*One Hundred Years of Solitude*), by Gabriel García Márquez. The texts were segmented following two of the approaches discussed in this paper: Kameyama and Miltsakaki. In the Kameyama-style segmentation, relative clauses were said to constitute an embedded utterance, that is, a separate unit from the main clause that contained them, but not accessible to the following utterance. In the Miltsakaki segmentation, relative clauses were included with their main clauses in the same utterance. Centering transitions (CONTINUE, RETAIN, SMOOTH SHIFT, ROUGH SHIFT, as well as cheap and expensive) were computed for all utterances in both types of segmentation. The transitions for utterances containing a relative clause as well as those of the utterances following relative clauses were selected for analysis. Among other findings concerning the distinction between restrictive and non-restrictive relative clauses, the results of the study showed a significantly better perform-

ance of the Kameyama approach over the Miltsakaki approach for both types of relative clauses. The better performance of the Kameyama approach was specifically due to a significant difference in the number of transitions with zero Cbs, which show no cohesion between two utterances.

The choice of a Kameyama-style analysis is not without problems. Many cases are difficult to process for more than one approach. In (30), an adjunct clause intervenes between matrix clause and its complement. The pronoun this refers cataphorically to the complement clause. Under the Kameyama approach, the adjunct clause should be in a different utterance. The question is where to place it. It could be processed after the matrix + complement clause, in which case we would alter the sequential order. Or it could be processed after *this*, in which case we would have to create a new segment for the complement clause. The latter analysis raises the question of which would be the previous utterance for the complement clause: the matrix or the adjunct clause. Similarly, Suri and McCoy do not account for adjunct clauses embedded in the middle of an utterance. We decided to segment this example into three units, and to process each linearly.

(30)   I didn't realize this [until I got to San Francisco] that the people in Chicago they took the wrong ticket from me.

8. Conclusions

We have presented a comparison of four different segmentation methods for Centering Theory. The measures for comparison were: number of empty Cbs; Cb and topic agreement; number of cheap versus expensive transitions; and presence of pronouns and antecedents within the same utterance.

The results are that a method that segments down to each individual clause (the Poesio method) is clearly not the best method for Centering-based segmentation. The differences among the other three methods considered are small, and often not statistically significant. After examining those differences, we propose to follow a Kameyama-style segmentation, that is, one where the finite clause is the unit of analysis. An alternative is the full sentence segmentation, in situations where consistency in unit size across different types of analysis is not an issue.

A different line of research in anaphora resolution uses general discourse structure to estimate which clauses are most accessible for a pronoun under consideration. This research is based on an assumption already in Fox (1987) that the choice of a particular referring expression for an entity depends on the distance between the mention of the entity

and its antecedent. That distance is not linear, but organized around rhetorical structure, which could be represented in the relations proposed in Rhetorical Structure Theory (Mann and Thompson 1988; Taboada and Mann 2006a, 2006b). For example, the work carried out within Veins Theory (Cristea et al. 2000; Ide and Cristea =, 2000) emphasizes the importance of hierarchical units for the disambiguation of anaphora. One example that employs RST specifically is the work of Tetreault and Allen (2003), who tested whether reference could be solved more easily if discourse structure were taken into account. The initial results were not encouraging, but more recent work (Tetreault 2005) suggests that discourse structure does improve the success of reference resolution methods. Chiarcos and Krasavina (2005a =,; 2005b) are also exploring this issue.

A long-term goal of our work is to combine Centering-based analyses with RST-based analyses. It is likely that the use of rhetorical structure will help in the disambiguation of anaphora. Given such a goal, the choice of a clausal unit for analysis is even more justified: Most RST analyses segment discourse into clauses (with the exception of complement clauses).

Stenström (2006), in her review of the C-ORAL-ROM corpus (Cresti and Moneglia 2005), points out that the annotators of the corpus chose 'utterance' as unit of analysis, "a concept that has no precise linguistic definition" (Stenström 2006: 249). We believe that any time a unit is selected, whichever its name, a very precise definition has to be provided for it. The word utterance may be vague, but it can be equipped with a precise definition for corpus annotation. The definition is crucial, as we have discussed, in Centering-based annotation. This paper details our efforts at fully spelling out what the definition should be for our annotation. We concluded that the finite clause should be our unit of analysis and, since 'finite clause' still stands to be further defined, we also specify how to deal with phenomena such as noun and relative clauses, false starts and backchannels. To make the annotation task easier, we have prepared a coding manual that addresses the most common issues in segmentation (Hadic Zabala and Taboada 2006).

We hope to have shown that defining units of analysis is not only important for Centering research, or for anaphora studies, but for any study that involves corpus coding and analysis. Practically any corpus-based research needs to consider the unit issue. If one needs to calculate the frequency of any type of phenomena (taboo words, idioms, cohesive devices, collocations, etc.) per unit of analysis, the definition of unit of analysis is crucial.

Acknowledgements

Bionotes

Maite Taboada is Associate Professor in the Department of Linguistics at Simon Fraser University, in Canada. She works in the areas of discourse analysis, systemic functional linguistics and computational linguistics, concentrating on Centering Theory, coherence relations and subjectivity in text.

   Loreley Hadic Zabala is a PhD student in the Department of Linguistics at Simon Fraser University. Her areas of research include discourse analysis, second language acquisition and computational linguistics. In her dissertation, she investigates different approaches to the segmentation of text at the local level of discourse.

Notes

1. http://www.ldc.upenn.edu
2. The exact numbers differ, depending on the segmentation method used, as we shall explain in Section 6.
3. Not everybody agrees that a general definition of units is desirable. Ford (2004) advocates definitions suited to the context and the particular characteristics of the interaction and the participant. We believe that this is not feasible in large-scale corpus annotation of the type described here.
4. http://www.ldc.upenn.edu/Projects/MDE/
5. http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf
6. Most of this section, including the analysis of Example (1), is taken from Taboada (2008).
7. From J.K. Rowling (2003) *Harry Potter and the Order of the Phoenix.* Vancouver: Raincoast Books (p. 8).
8. The term 'utterance' has not been properly defined yet. For now, let us assume it corresponds to a sentence. A discourse segment is a portion of discourse that has a particular purpose, different from that of adjacent segments (Grosz and Sidner 1986).
9. Small capitals indicate that the list contains entities, not their linguistic realization. The reference to Dudley is conveyed by two different referring expressions: *their son* and *Dudley*.
10. Centering transitions are just one explanation for coherence. A text can be coherent without repeating or referring to the same entities (Brown and Yule 1983: 195−199; Poesio et al. 2000).
11. Grosz, Joshi and Weinstein (1995) is a revised version of a paper in circulation since 1986.

12. Abbreviations used in the glosses: 1/2/3 − first/second/third person; CL − clitic; SG − singular; PL − plural; FEM − feminine; MASC − masculine; PRES − present; INF − infinitive; SUBJ − subjunctive; SE − clitic with several functions (here, mostly a marker of passive or middle voice).

13. Technically, the unit introduced by *because* is a clause, not a sentence. We preserve some of the terminology and the abbreviations used by Suri and McCoy.

14. Miltsakaki does not explicitly discuss coordinated clauses within the same sentence. We have decided to group them as an utterance in our interpretation of her method. She does discuss, however, an example similar to (16), but considers it an instance of a subordinate conjunction in a consequence relation (*I had just been to the bank, so I had money*). Her discussion pertains to the fact that *so* in that example does not behave as expected from a subordinating conjunction.

15. In this example we have only provided a free translation into English, and not a full gloss. We believe this is sufficient to follow the structure.

16. The GNOME texts were written descriptions of museum artifacts, and written medical leaflets created for patients, all in English.

17. The comparisons carried out by Byron and Stent did not directly involve segmentation issues; they addressed other problems in conversation (first and second person pronouns, false starts). Their definition of topic was "the annotator's intuition of what the utterance is 'about'." (Byron and Stent 1998: 1476).

18. According to Grosz, Joshi and Weinstein (1995), Joshi and Kuhn (1979) and Joshi and Weinstein (1981) were the ones to introduce the notions of "forward-looking" and "backward-looking".

19. Typically, an utterance at the beginning of a discourse segment would have an empty Cb, and thus we cannot expect all utterances to have a Cb. Walker (1998) presents evidence that this is not always the case: some new segments carry over centers from the previous utterance; and some segment-internal utterances have empty Cbs.

20. Topic is a most elusive concept. The definition above is a general one, but some instances of what Gundel and Fretheim call topic would necessarily not be Cbs. A new topic can be introduced in an utterance that also discusses a previously mentioned entity: *John likes beans. As for Mary, she hates them.* The topic of the second sentence is *Mary.* The Cb is *beans.* In this paper, we are trying to have a general description of topic. We understand that not all topics are Cbs; however, we expect that a majority will be. We thank Nancy Hedberg for pointing this out to us, and for the example.

21. Thanks to Nancy Hedberg for helping us select the tests.

22. As described below, we performed a reliability test with a third coder. She happened to be a speaker of Japanese, and also used a -wa test for topic assignment. She translated some of the difficult sentences into Japanese, and considered which entity was most likely to have the topic marker -wa. That entity was then the topic in the utterance.

23. For the purposes of the annotation, we excluded such utterances out of our comparison with the Cb. This does not mean that they are all-focus utterances or thetic sentences (Kuroda 1972; Vallduví 1990). It simply means that we could not decide what the topic was, if there was one.

24. Thanks to Mayo Kudo for carrying out this comparison analysis.

25. Both CONTINUE and RETAIN include transitions where the previous Cb was empty (i. e., center continuation and center establishment are under CONTINUE).

26. In this test, and some of the ones reported below, it was often the case that one or more conversations were outliers, and so they were excluded from the analysis.

References

Austin, J. L.
  1962   *How to Do Things with Words.* Harvard: Harvard University Press.
Bäcklund, Ingegerd
  1992   Theme in English telephone conversation. *Language Sciences* 14(4), 545−564.
Bardovi-Harlig, Kathleen
  1992   A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly* 26(2), 390−395.
Barnwell, David
  1988   Some comments on T-unit research. *System* 16(2), 187−192.
Barzilay, Regina and Mirella Lapata
  2005   Modeling local coherence: An entity-based approach, *Proceedings of the 43rd Meeting of the Association for Computational Linguistics.* Ann Arbor, MI, 141−148.
Brennan, Susan E., Marilyn W. Friedman and Carl J. Pollard
  1987   A Centering approach to pronouns, *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL-87).* Stanford, CA, USA, 155−162.
Brown, Gillian and George Yule
  1983   *Discourse Analysis.* Cambridge: Cambridge University Press.
Byron, Donna K. and Amanda Stent
  1998   A preliminary model of Centering in dialog, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-98).* Montréal, Canada, 1475−1477.
Carletta, Jean
  1996   Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2), 249−254.
Chafe, Wallace
  1994   *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing.* Chicago: University of Chicago Press.
Chafe, Wallace (ed.).
  1980   *The Pear Stories: Cognitive, Cultural and Narrative Aspects of Narrative Production.* Norwood, NJ: Ablex.
Chiarcos, Christian and Olga Krasavina
  2005a   Rhetorical distance revisited: A parametrized approach, *Proceedings of Workshop in Constraints in Discourse.* Dortmund, Germany.
Chiarcos, Christian and Olga Krasavina
  2005b   Rhetorical distance revisited: A pilot study, *Proceedings of Corpus Linguistics 2005.* Birmingham, UK.
Chiarcos, Christian and Manfred Stede
  2004   Salience-driven text planning, *Proceedings of the Third International Conference on Natural Language Generation.* Careys Manor, Hampshire.
Coelho, Carl A.
  2002   Story narratives of adults with closed head injury and non-brain-injured adults: Influence of socioeconomic status, elicitation task, and executive functioning. *Journal of Speech, Language, and Hearing Research* 45(6), 1232−1248.

Cohan, Jocelyn
  2000   The Realization and Function of Focus in Spoken English. Unpublished
            Ph.D. dissertation, University of Texas at Austin.
Cresti, Emanuela and Massimo Moneglia (eds.)
  2005   *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Lan-
            guages.* Amsterdam/Philadelphia: John Benjamins.
Cristea, Dan, Nancy Ide, Daniel Marcu and Valentin Tablan
  2000   Discourse structure and coreference: An empirical study, *The 18th In-
            ternational Conference on Computational Linguistics (COLING'00)*.
            Saarbrücken, Germany, 208−214.
Eckert, Miriam and Michael Strube
  1999   Resolving discourse deictic anaphora in dialogues, *Proceedings of the 9th
            Conference of the European Chapter of the Association for Computational
            Linguistics (EACL-99)*. Bergen, Norway, 37−44.
  2000   Dialogue acts, synchronizing units, and anaphora resolution. *Journal of
            Semantics* 17, 51−89.
Eggins, Suzanne and James R. Martin
  1997   Genres and registers of discourse. In Dijk, Teun A. van (ed.), *Discourse
            as Structure and Process. Discourse Studies: A Multidisciplinary Introduc-
            tion.* London: Sage, 230−256.
Eggins, Suzanne and Diana Slade
  1997   *Analysing Casual Conversation.* London: Cassell.
Ford, Cecilia
  2004   Contingency and units in interaction. *Discourse Studies* 6(1), 27−52.
Ford, Cecilia, Barbara A. Fox and Sandra A. Thompson
  1996   Practices in the construction of turns: The 'TCU' revisited. *Pragmatics*
            6(3), 427−454.
Ford, Cecilia and Sandra A Thompson
  1996   Interactional units in conversation: Syntactic, intonational, and prag-
            matic resources for the management of turns. In Ochs, Elinor, Emman-
            uel Schegloff and Sandra A Thompson (eds.), *Interaction and Grammar.*
            Cambridge: Cambridge University Press, 134−184.
Fox, Barbara A.
  1987   *Discourse Structure and Anaphora: Written and Conversational English.*
            Cambridge: Cambridge University Press.
Fries, Peter H.
  1994   On Theme, Rheme and discourse goals. In Coulthard, Malcolm (ed.),
            Advances in Written Text Analysis. London: Routledge, 229−249.
Gaies, Stephen J.
  1980   T-Unit analysis in second language research: Applications, problems and
            limitations. *TESOL Quarterly* 14(1), 53−60.
GNOME
  2000   *GNOME Project Final Report.* Edinburgh: University of Edinburgh.
Godfrey, John J. and Ed Holliman
  1993   Switchboard-1 Transcripts, LDC93T4 [Corpus]. Philadelphia, PA: Lin-
            guistic Data Consortium.
Gordon, Peter C., Barbara J. Grosz and Laura A. Gilliom
  1993   Pronouns, names, and the Centering of attention in discourse. *Cognitive
            Science* 17(3), 311−347.

Grosz, Barbara J., Aravind K. Joshi and Scott Weinstein
   1983   Providing a unified account of definite noun phrases in discourse, *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL-83)*. Cambridge, Massachusetts, USA, 44−50.
   1995   Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics* 21(2), 203−225.
Grosz, Barbara J. and Candace L. Sidner
   1986   Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3), 175−204.
Gundel, Jeanette K.
   1977   *Role of Topic and Comment in Linguistic Theory*. Bloomington: Indiana University Linguistics Club.
Gundel, Jeanette K. and Thorstein Fretheim
   2004   Topic and focus. In Horn, Laurence and Gregory Ward (eds.), *The Handbook of Pragmatics*. Malden, Mass: Blackwell, 175−196.
Hadic Zabala, Loreley
   2007   The segmentation of Spanish relative clauses: A pilot study. Paper presented at the Conference on Interdisciplinary Approaches to Relative Clauses. Cambridge, UK.
Hadic Zabala, Loreley and Maite Taboada
   2006   *Centering Theory in Spanish: Coding Manual*. Unpublished manuscript, Simon Fraser University. Available from: http://www.sfu.ca/~mtaboada.
Halliday, Michael A. K. and Ruqaiya Hasan
   1976   *Cohesion in English*. London: Longman.
Hearst, Marti
   1994   Multi-paragraph segmentation of expository text, *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*. Las Cruces, New Mexico, 9−16.
Henschel, Renate, Hua Cheng and Massimo Poesio
   2000   Pronominalization revisited, *Proceedings of COLING 2000*. Saarbrücken, 306−312.
Hitzeman, Janet and Massimo Poesio
   1998   Long distance pronominalisation and global focus, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL-98/ COLING-98)*. Montréal, Canada, 550−556.
Houtkoop, Hanneke and Harrie Mazeland
   1985   Turns and discourse units in everyday conversation. *Journal of Pragmatics* 9, 595− 619.
Hunt, Kellogg W.
   1977   Early blooming and late blooming syntactic structures. In Cooper, Charles R. and Lee Odell (eds.), *Evaluating Writing: Describing, Measuring and Judging*. Urbana, Il: NCTE, 91−104.
Hurewitz, Felicia
   1998   A quantitative look at discourse coherence. In Walker, Marilyn A., Aravind K. Joshi and Ellen F. Prince (eds.), *Centering Theory in Discourse*. Oxford: Clarendon, 273− 291.
Ide, Nancy and Dan Cristea
   2000   A hierarchical account of referential accessibility, *Proceedings of the 38th Meeting of the Association for Computational Linguistics*. Hong Kong, 416−424.

Joshi, Aravind K and Steve Kuhn
  1979   Centered logic: The role of entity centered sentence representation in natural language inferencing, *Proceedings of the 6th International Joint Conference on Artificial Intelligence*. Tokyo, Japan, 435−439.
Joshi, Aravind K and Scott Weinstein
  1981   Control of inference: Role of some aspects of discourse structure-centering, *Proceedings of the 7th International Joint Conference on Artificial Intelligence*. Vancouver, Canada, 385−387.
Kameyama, Megumi
  1986a   Japanese zero pronominal binding: Where syntax and discourse meet. In Poser, William J. (ed.), *Papers from the Second International Workshop on Japanese Syntax*. Stanford, CA: CSLI, 47−74.
  1986b   A property-sharing constraint in Centering, *Proceedings of the 24th Annual Meeting of Association for Computational Linguistics (ACL-86)*. New York, USA, 200−206.
  1998   Intrasentential Centering: A case study. In Walker, Marilyn A., Aravind K. Joshi and Ellen F. Prince (eds.), *Centering Theory in Discourse*. Oxford: Clarendon, 89−112.
Karamanis, Nikiforos
  2003   Entity Coherence for Descriptive Text Structuring. Unpublished Ph.D. dissertation, University of Edinburgh.
Kehler, Andrew
  1997   Current theories of Centering for pronoun interpretation. *Computational Linguistics* 23(3), 467−475.
Kibble, Rodger
  1999   Cb or not Cb? Centering Theory applied to NLG, *Proceedings of the ACL Workshop on Discourse/Dialogue Structure and Reference*. University of Maryland, 72−81.
  2001   A reformulation of Rule 2 of Centering. *Computational Linguistics* 27(4), 579−587.
Kim, Harksoo, Jeong-Mi Cho and Jungyun Seo
  1999   Anaphora resolution using an extended Centering algorithm in a multi-modal dialogue system, *Proceedings of the Workshop on the Relation of Discourse/Dialogue Structure and Reference (held in conjunction with the 37th Annual Meeting of the Association for Computational Linguistics ACL-99)*, 21−28.
Kingsbury, Paul, Stephanie Strassel, Cynthia McLemore and Robert McIntyre
  1997   CallHome American English Transcripts, LDC97T14 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.
Klecan-Aker, Joan S. and Beth Lopez
  1985   A comparison of T-units and cohesive ties used by first and third grade children. *Language and speech* 28(3), 307−315.
Kuroda, S.-Y.
  1972   The categorical and the thetic judgment: Evidence from Japanese syntax. *Foundations of Language* 9(2), 153−185.
Larsen-Freeman, Diane
  1978   An ESL index of development. *TESOL Quarterly* 12(4), 439−448.
Mann, William C. and Sandra A. Thompson
  1987   *Rhetorical Structure Theory: A Theory of Text Organization* (No. ISI/RS-87−190). Marina del Rey, CA: Information Sciences Institute.
  1988   Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3), 243−281.

Miltsakaki, Eleni
  2002   Toward an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics* 28(3), 319−355.
Miltsakaki, Eleni and Karen Kukich
  2004   Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering* 10(1), 25−55.
Montgomery, Douglas C.
  2005   *Design and Analysis of Experiments* (6th ed.). New York: John Wiley and Sons.
Mosegaard Hanse, Maj-Britt
  1998   *The Function of Discourse Particles: A Study with Special Reference to Spoken Standard French.* Amsterdam and Philadelphia: John Benjamins.
O'Donnell, Roy C.
  1976   A critique of some indices of syntactic maturity. *Research in the Teaching of English* 10(1), 31−38.
Passonneau, Rebecca
  1998   Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In Walker, Marilyn A., Aravind K. Joshi and Ellen F. Prince (eds.), *Centering Theory in Discourse*. Oxford: Clarendon, 327−358.
Passonneau, Rebecca and Diane J. Litman
  1997   Discourse segmentation by human and automated means. *Computational Linguistics* 23(1), 103−139.
Poesio, Massimo
  2000   *The GNOME annotation scheme manual* (Technical Report). Edinburgh: University of Edinburgh. (Available from: http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/).
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, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*. Hong Kong, 400−407.
Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio and Janet Hitzeman
  2004   Centering: A parametric theory and its instantiations. *Computational Linguistics* 30(3), 309−363.
Reed, Vicki A., Verity MacMillan and Sharynne McLeod
  2001   Elucidating the effects of different definitions of "utterance" on selected syntactic measures of older children's language samples. *Asia Pacific Journal of Speech, Language, and Hearing* 6(1), 39−45.
Reinhart, Tanya
  1981   Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27, 53−94.
Restrepo, Maria Adelaida and Vera F. Gutierrez-Clellen
  2001   Article use in Spanish-speaking children with Specific Language Impairment. *Journal of child language* 28(2), 433−452.
Sacks, Harvey, Emmanuel Schegloff and Gail Jefferson
  1974   A simplest systematics for the organization of turn-taking in conversation. *Language* 50, 696−735.
Schegloff, Emmanuel
  1996   Turn-organization: One intersection of grammar and interaction. In Ochs, Elinor, Emmanuel Schegloff and Sandra A. Thompson (eds.), *Interaction and Grammar*. Cambridge: Cambridge University Press, 52−133.

Scott, Cheryl M. and Jennifer Windsor
　2000　　General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research* 43(2), 324−339.
Searle, John
　1969　　*Speech Acts.* Cambridge: Cambridge University Press.
Selting, Margret
　2000　　The construction of units in conversational talk. *Language in Society* 29, 477−517.
Stenström, Anna-Brita
　2006　　A step forward in multilingual corpus research (Review article). *International Journal of Corpus Linguistics* 11(2), 245−254.
Stolke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema and Marie Meteer
　2000　　Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26(3), 339−373.
Strube, Michael
　1998　　Never look back: An alternative to Centering, *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-98/ACL-98).* Montréal, Canada, 1251−1257.
Strube, Michael and Udo Hahn
　1996　　Functional Centering, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96).* Santa Cruz, CA, USA, 270−277.
　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.
Taboada, Maite
　2002　　Centering and pronominal reference: In dialogue, in Spanish. In Bos, Johan, Mary Ellen Foster and Colin Matheson (eds.), *Proceedings of the 6th Workshop on the Semantics and Pragmatics of Dialogue (EDILOG-2002),* 177−184.
　2004a　*Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish.* Amsterdam/Philadelphia: John Benjamins.
Taboada, Maite
　2004b　The genre structure of bulletin board messages. *Text Technology* 13(2), 55−82.
　2005　　Anaphoric terms and focus of attention in English and Spanish. In Butler, Christopher, María de los Ángeles Gómez-González and Susana Doval (eds.), *The Dynamics of Language Use: Functional and Contrastive Perspectives.* Amsterdam/Philadelphia: John Benjamins, 195−216.
　2008　　Reference, centers and transitions in spoken Spanish. In Gundel, Jeanette K. and Nancy Hedberg (eds.), *Reference and Reference Processing.* Oxford: Oxford University Press, 176−215.
Taboada, Maite and William C. Mann
　2006a　Applications of Rhetorical Structure Theory. *Discourse Studies* 8 (4), 567−588.

2006b   Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies* 8(3), 423−459.

Tetreault, Joel R.
2001   A corpus-based evaluation of Centering and pronoun resolution. *Computational Linguistics* 27(4), 507−520.
2005   Decomposing discourse. In Branco, António, Tony McEnery and Ruslan Mitkov (eds.), *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*. Amsterdam/Philadelphia: John Benjamins, 73−95.

Tetreault, Joel R. and James Allen
2003   An empirical evaluation of pronoun resolution and clausal structure, *International Symposium on Reference Resolution and its Applications to Question Answering and Summarization*. Venice, Italy, 1−8.

Thompson, Sandra A.
2002   "Object complements" and conversation: Towards a realistic account. *Studies in Language* 26(1), 125−164.

Thompson, Sandra A. and Elizabeth Couper-Kuhlen
2005   The clause as a locus of grammar and interaction. *Discourse Studies* 7(4−5), 481−505.

Traum, David and Elizabeth Hinkelman
1992   Conversation acts in task-oriented spoken dialogue. *Computational Intelligence* 8(3), 575−599.

Tukey, J.W.
1953   *The problem of multiple comparisons*. Princeton, NJ: Princeton University.

Vallduví, Enric
1990   The Informational Component. Unpublished Ph.D. dissertation, University of Pennsylvania.

Walker, Marilyn A.
1998   Centering, anaphora resolution, and discourse structure. In Walker, Marilyn A., Aravind K. Joshi and Ellen F. Prince (eds.), *Centering Theory in Discourse*. Oxford: Clarendon, 401−435.

Walker, Marilyn A., Aravind K. Joshi and Ellen F. Prince
1998a   Centering in naturally occurring discourse: An overview. In Walker, Marilyn A., Aravind K. Joshi and Ellen F. Prince (eds.), *Centering Theory in Discourse*. Oxford: Clarendon, 1−28.

Walker, Marilyn A., Aravind K. Joshi and Ellen F. Prince (eds.)
1998b   *Centering Theory in Discourse*. Oxford: Clarendon.

Wheatley, Barbara
1996   CallHome Spanish Transcripts, LDC96T17 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Witte, Stephen P.
1980   The stability of T-unit length: A preliminary investigation. *Research in the Teaching of English* 14(1), 5−17.

Wolf, Florian, Edward Gibson and Timothy Desmet
2004   Coherence and pronoun resolution. *Language and Cognitive Processes* 19(6), 665−675.

Yamura, Mitsuko
1996   Zero Anaphora Resolution in Japanese: A Computational Model for Machine Translation. Unpublished Master's thesis, Georgetown University.

Yngve, Victor H.
  1970   On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistics Society.* Chicago: University of Chicago, 567−577.
Yuksel, Ozgur and Cem Bozsahin
  2002   Contextually appropriate reference generation. *Natural Language Engineering* 8(1), 69−89.