Volume 42, Issue 7,
July 2010
ISSN 0378-2166
ELSEVIER
Editors-in-Chief
JONATHAN CULPEPER
JACOB L. MEY
Founding Editor
HARTMUT HABERLAND
Editors
CLAUDIA CAFFI
NEAL R. NORRICK
THEODOSSIA-SOULA PAVLIDOU
Assistant Editor
LEO HOYE
journal of
PRAGMATICS
An Interdisciplinary Journal
of Language Studies
Special Issue Section: Linguistic and Cognitive aspects of Reference

Available online at www.sciencedirect.com

ScienceDirect

Journal of Pragmatics 42 (2010) 1816–1828

journal of PRAGMATICS

www.elsevier.com/locate/pragma

Subjects and topics in conversation

Maite Taboada [*], Loreley Wiesemann

Department of Linguistics, Simon Fraser University, 8888 University Dr., Burnaby, BC, V5A 1S6, Canada

Received 23 May 2008; received in revised form 1 April 2009; accepted 4 April 2009

Abstract

This paper presents an examination of the expression of subject and topic in conversation, using parameters established by Centering Theory. Based on a corpus analysis of English and Spanish casual conversations, we show (i) under what circumstances the subject and the topic of an utterance coincide ('topic' is defined, according to Centering, as the backward-looking center of an utterance); and (ii) what referring expressions are used to encode subject and topic. We found that the Cb is realized as subject most of the time (80% of the utterances in English and 73% in Spanish), showing that topichood and subjecthood tend to be assigned to the same entity. The preferred realization for the Cb in those cases is a pronoun in English and a zero pronoun in Spanish.
© 2009 Elsevier B.V. All rights reserved.

Keywords: Topic; Subject; Conversation; Anaphora; Centering Theory

1. Introduction

In this paper, we are concerned with the notions of subjecthood and topichood. Since they are both ways of encoding a salient entity in the discourse, we are interested in cases where the topic is realized as a subject, and also in the cases where it is not.

Definitions for these two concepts abound. The notion of subject is traditionally associated with salience. Among the many criteria that Keenan (1976) proposes for a universal definition of subject are that subjects tend to be topics, ''i.e. they identify what the speaker is talking about'' (Keenan, 1976:318); and that they are ''highly referential''. Givón (1983) points out that subject was first considered a grammaticalized topic, but that was soon discovered to be an insufficient definition, since there are sentences in which there seem to be two clearly distinguishable NPs, one a topic and the other one a subject (*John, we saw him yesterday*). He then proposes different degrees of topicality, and a divorce between subject and topic. Other criteria for subjecthood in Keenan (1976) include aspects such as verb agreement, case, autonomy of reference, and coreference across clause boundaries. In our study, verb agreement was the main criterion for subject identification.

The definition of topic is more elusive than that of subject. According to one view, topic is what ''sets a spatial, temporal, or individual framework within which the main predication holds'' (Chafe, 1976:50). This is in contrast to subject, which, broadly, is ''what we are talking about'' (Chafe, 1976:43). Li and Thompson (1976:464) say that ''[t]he topic is the 'center of attention'; it announces the theme of the discourse''. For our definition, we turned to a theory that provided a very specific set of parameters to identify the topic of an utterance. In Centering Theory (see section 2),

[*] Corresponding author. Tel.: +1 778 782 5585; fax: +1 778 782 5659.
E-mail addresses: mtaboada@sfu.ca (M. Taboada), lmhadic@sfu.ca (L. Wiesemann).

0378-2166/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.pragma.2009.04.009

topic is the backward-looking center of an utterance, the most salient entity from the previous utterance that is present in the current utterance. Subject, then, is defined within the clause or utterance, but the definition of topic relies on context.

Centering Theory aims at accounting for coherence in discourse by explaining how speakers and hearers maintain the focus of attention in discourse. It is concerned with how both global and local discourse structure have an influence on the expressions used to refer to entities that are in the participants' focus of attention. Those entities are commonly known as centers of attention, hence the name Centering. Centering has its origins within computational linguistics, and it has been applied to anaphora resolution (Brennan et al., 1987), the generation of referring expressions (Kibble and Power, 2004), and the computation of coherence in discourse (Miltsakaki and Kukich, 2004), among other applications. It has always been connected to efforts into the study of salience (Stevenson, 2002), topic tracking (Miltsakaki, 2002), discourse structure (Walker, 1998, 2000) and coherence in general (Hurewitz, 1998).

In this paper, we explore the relationship between subjects and topics, using Centering Theory as a tool. The application of Centering Theory and some of its constructs may sound technical, but our choice is motivated because Centering allows for precise definitions of the concepts that are of interest to us, subject and topic. As Poesio et al. (2004a,b) observe, the basic tenets of Centering Theory are present in most research in discourse: Utterances tend to be mainly about one entity; some discourse entities are more salient than others; utterances that concern themselves with the same entity form more coherent discourse segments; and pronouns tend to refer to the entity most in focus.

The paper is organized as follows: in section 2, we provide a brief introduction to Centering Theory. Ours is a corpus-based study, and the data for the study is discussed in section 3. Section 4 outlines the general results of the analysis, in terms of types and frequency of transitions. Section 5 delves into the relationships between subjects and topics, and examines cases where they are and where they are not the same, discussing contexts for each. Finally, section 6 provides conclusions.

2. Centering Theory

There are a number of good introductions to Centering (Grosz et al., 1995; Walker et al., 1998b), including summaries presented in our previous work (Taboada, 2008; Taboada and Hadic Zabala, 2008). Here, we provide a brief description, outlining the basic concepts needed to understand the rest of the paper. For a more detailed explanation and examples, see the works cited above.

Centering is concerned with centers of attention, semantic entities that are part of the discourse model of each utterance in a discourse segment. For each utterance, Centering proposes that there is 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 or highest-ranked member of 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.

Transition types are 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 topics. Cb_i and Cp_i refer to 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 of the current utterance is the same as the Cb of the previous utterance. Transitions are one explanation for how coherence is achieved: A text that maintains the same centers is perceived as more coherent. Transitions are ranked, according to how difficult it is for the hearer or reader to process the transition from one utterance to the next. CONTINUE transitions are considered to pose the least processing demands, followed by RETAIN, SMOOTH SHIFT and ROUGH SHIFT. This is known as Rule 2 of Centering.

There are different sets of transitions used in the literature. Most common is the set that includes CONTINUE, RETAIN, SMOOTH SHIFT and ROUGH SHIFT. In our case, we have also adopted the ESTABLISH transition proposed by Kameyama (1986)

Table 1
Centering transitions.

	$\mathrm{Cb}_{i-1} = \varnothing$ and $\mathrm{Cb}_i \neq \varnothing$	$\mathrm{Cb}_{i-1} \neq \varnothing$ and $\mathrm{Cb}_i = \varnothing$	$\mathrm{Cb}_{i-1} = \varnothing$ and $\mathrm{Cb}_i = \varnothing$	$\mathrm{Cb}_i = \mathrm{Cb}_{i-1}$	$\mathrm{Cb}_i \neq \mathrm{Cb}_{i-1}$
$\mathrm{Cb}_i = \mathrm{Cp}_i$	ESTABLISH	ZERO	NULL	CONTINUE	SMOOTH SHIFT
$\mathrm{Cb}_i \neq \mathrm{Cp}_i$				RETAIN	ROUGH SHIFT

and NULL and ZERO, suggested in Poesio et al. (2004a,b). They all occur in situations where there is an empty Cb. ESTABLISH has been separated from CONTINUE and RETAIN (where some researchers classify it). This distinguishes it from transitions that maintain a topic, rather than introduce it. ZERO and NULL transitions are, to a certain extent, undesirable, since they signal a lack of connection across utterances and are, in turn, a new subcategorization of what other researchers term ‘‘no-Cb transitions’’. A ZERO transition indicates the abandonment of a topic, the inverse of ESTABLISH, since the Cb of the previous utterance is not repeated in the current utterance, and no other entity picks up that role. A NULL transition indicates an entire sequence of at least two utterances without a connection (in the Centering sense).

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

For each U_i in a discourse segment D consisting of utterances $U_1, \ldots, U_m$, 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. In its most strict interpretation, it only states that if anything in the utterance is pronominalized, then the Cb must also be a pronoun. In more general terms, 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 allowed in the same utterance, but the most salient entity must be realized by the least marked referring expression.

As we indicated above, the main principles of Centering are meant to capture general insights common to most discourse research. They encapsulate principles of degrees of salience (Gundel et al., 1993); hierarchies that feature the pronoun as the optimal form of expression for the most salient entity in the discourse (Ariel, 1990; Givón, 1983; Gundel et al., 1993); and coherence maintained through repetition and co-reference (Halliday and Hasan, 1976).

3. Data analysis

We used two corpora from the CallHome collection, a series of spontaneous telephone conversations between relatives and friends, available in several languages. We analyzed 20 conversations, 10 in English and 10 in Spanish (Kingsbury et al., 1997; Wheatley, 1996), and coded them according to the parameters of Centering Theory. A few contentious issues, such as the segmentation of embedded clauses, the application of Centering to spoken language, and the ranking of entities in Spanish, were coded according to the methodology described in our previous work (Hadic Zabala and Taboada, 2006; Taboada, 2005, 2008; Taboada and Hadic Zabala, 2008). We provide a brief summary here.

Many aspects of Centering Theory are open to interpretation (Poesio et al., 2004a,b). In our work, we have found that segmenting discourse into what Centering calls *utterances* necessitates a very precise set of instructions for how to deal with embedded and complex clauses, finite or not. We discuss this issue at length elsewhere (Taboada and Hadic Zabala, 2008). The conclusion of our research is that a Centering utterance is a finite clause, whether subordinate or not. This applies to cases where a large unit has to be broken down into its component clauses; there also exist instances of utterances being smaller than a finite clause, such as phrases occurring independently from other utterances, and which contain entities that can form a Cf list.

The ordering of the Cf list (Cf template) is also subject to interpretation. In English, researchers often follow grammatical relations, with subjects ranked higher than objects (Brennan et al., 1987; Grosz et al., 1995; Walker et al., 1998a). Our proposal of a template for Spanish is presented in (1), inspired by Di Eugenio's (1998) for Italian. The template is based on grammatical relations, with two exceptions: (i) the experiencer in psychological predicates is the highest-ranked entity, even though it is not the subject[1]; and (ii) animate indirect objects are ranked higher than direct objects.

$$\text{Experiencer} > \text{Sub}_j > \text{Animate IOb}_j > \text{DOb}_j > \text{Other} > \text{Impersonal/Arbitrary pronouns} \qquad (1)$$

We also need to decide on a treatment for inferable entities. To populate the Cf list, indirect realization of entities was permitted. We used the relations identified by Halliday and Hasan (1976) as lexical cohesion: synonymy, hyponymy, and superordinate, but not collocation or antonymy, which do not necessarily involve reference to the same entity. This is what Fais (2004) calls cohesive transitions.

[1] This we do for several reasons that have been correlated with topichood and salience. The experiencer tends to be animate, a criterion for Stevenson (2002); and it tends to be pre-verbal, a criterion for Diderichsen and Elming (2005), among others.

Table 2
Transitions in the data.

	English		Spanish	
CONTINUE	806	29.33%	910	34.85%
ESTABLISH	546	19.87%	472	18.08%
RETAIN	159	5.79%	189	7.24%
SMOOTH SHIFT	287	10.44%	242	9.27%
ROUGH SHIFT	72	2.62%	65	2.49%
ZERO	503	18.30%	439	16.81%
NULL	365	13.28%	284	10.88%
NO CB	10	0.36%	10	0.38%
Total	2748		2611	

So far, we have described how we code standard Centering parameters in the data. In addition to the usual Centering parameters (Cf list, Cb and transition type for each utterance), we coded the subject and its realization. Subject is defined as the grammatical subject, based on verb agreement, or the logical subject in existential clauses. Since we were interested in the cases where the subject and the Cb did not coincide, we investigated, for those cases, how far in the conversation the previous mention of both was. This distance was counted in number of utterances.

We provide a full description of the coding process and of a reliability study in a different paper (Taboada and Hadic Zabala, 2008). We compared our joint coding of one conversation in each language with the coding done by another researcher who followed our coding manual (Hadic Zabala and Taboada, 2006). Agreement in unit segmentation was 91.89% in English and 92.89% in Spanish; in Cf ranking the agreement was 77.34% for English and 76.13% for Spanish; for topic assignment the agreement figures were 83.14% (English) and 72.29% (Spanish).

The next section describes the results of our analysis, and section 5 discusses the distribution of topics and subjects in the data.

4. General results

The 20 conversations contained a total of 5359 utterances, 2748 in English and 2611 in Spanish. Table 2 shows the number of utterances, broken down into the types of transitions defined for this study.

We see that, in both languages, the most frequent transition by far is CONTINUE, followed by ESTABLISH, which, in some versions of the theory, is counted as a form of CONTINUE. Let us point out that the numbers are roughly the same in both languages. Despite many other differences between English and Spanish, the two sets of conversations have a similar distribution of Centering transitions. It seems to be the case that the genre and style of conversations are very similar, in terms of how the focus of attention is presented, across the two languages.

Our numbers are quite different from those presented in Poesio et al. (2004a,b), where the most frequent transition was NULL, followed by ESTABLISH, ZERO and CONTINUE. There are three reasons, we believe, for these differences: genre, mode and our application of Centering. With respect to genre, Poesio and colleagues analyzed written texts (museum descriptions and pharmaceutical leaflets), whereas we are studying casual speech among family and friends. Besides the genre differences, the mode (spoken vs. written) means that first and second person pronouns are part of our analysis. Finally, our application of Centering differs from that of Poesio and colleagues in that we allow a wider range of indirect realization, including bridging reference and entities in a lexical cohesion relation. This alone would account for a much lower number of transitions that involve a zero Cb in our case.

In both languages, a significant number of utterances still violate Constraint 1 of Centering (Brennan et al., 1987), that each utterance has a Cb. This refers to the sum of ZERO and NULL transitions, and does not concern utterances at the beginning of the conversation, which naturally do not have a Cb, or have one that is underspecified (to be established as the conversation proceeds).[2] Poesio et al. (2004a,b) found, in a corpus evaluation, that about 49% of their utterances had violated this Constraint. In our corpus, there were 878 utterances with a zero Cb (31.95%) in English, and 733 (28.07%) in Spanish.

[2] In Table 2, the 10 utterances for each language with no Cb correspond to the beginning of the transcripts. The transcripts are about 5-min transcriptions of a half-hour conversation, and begin at points where the conversation has been ongoing.

One valid reason for the presence of ZERO and NULL transitions is the beginning of a discourse segment. A new discourse segment may introduce new entities, and abandon those in the previous segment. We performed a rough calculation of how many times ZERO and NULL transitions introduced new segments. For that purpose, we looked at those two transition types, and every time they appeared, we determined whether the utterance was the beginning of a new discourse segment (whether at the top level or embedded). Discourse segments were defined as chunks of the conversation that deal with the same topic. These were large chunks, including embedded segments, typically longer than the more fine-grained segments of Grosz and Sidner (1986). In English, 746 out of 850 utterances with ZERO or NULL transitions continued the same discourse segment (87.8%), and 104 (12.2%) introduced a new discourse segment. In the Spanish conversations, the two transitions were inside a segment in 539 out of 699 cases (77.1%), and in 160 cases (22.9%), they signalled the beginning of a new segment. In some cases, it was difficult to determine which one was the case (new segment or same segment), and we discounted those cases. These findings leave a high number of discourse-segment-internal ZERO and NULL transitions unaccounted for.

Below there are two examples of ZERO and NULL transitions that do not coincide with the beginning of new discourse segments. In (2),[3] the speaker has been explaining that his son Benjamín saw a television program about Nostradamus. Benjamín then believed that the world was coming to an end. In (2b), the Cb is *Benjamín*, but in (2c), the Cf list only contains *mundo* (‘world’), so the Cb is empty and the transition ZERO. In (2e) the speaker returns to discussing Benjamín, and again the Cb is empty, and the transition NULL. The problem is the repetition of *y que el mundo se iba a acabar* (‘and that the world was coming to an end’) in (2c), which breaks the sequence with *Benjamín* as Cb. This is definitely not a new discourse segment, merely a repetition of something that has already been said.

(2) a. . . . pero Benjamín convencido de que el mundo se iba a acabar,
‘. . . but Benjamín (was) convinced that the world was coming to an end,’
b. no, porque todo lo que vio lo convenció, ¿no?
‘right, because everything he saw convinced him, right?’
c. y que el mundo se iba a acabar
‘and that the world was coming to an end’
e. y entonces en eso empezó a temblar, a temblar–
‘and then with that (he) began to tremble, to tremble–’

The next example contains a ZERO transition between utterance (3b), an answer to a question, and (3c), the next question, after a CONTINUE. (We assume an implied entity, *A*, in (3b), so that the utterance could be *I am coming back on the 19th*.) There is a clear parallel structure, question–answer, followed by a new question–answer. Since the answer to the first question and the second question do not have any entities in common, the Cb of (3c) is empty.

(3) B: a. yeah. When are you coming back?
A: b. The nineteenth.
B: c. Oh. When does school start?
A: d. The twenty-second.

The previous are only some examples of the many cases of ZERO and NULL transitions. In both languages, about 30% of the utterances do not exhibit entity coherence at the local level, since they do not contain an entity from the previous utterance. We have shown in two examples that repetition and parallelism break the continuous flow of Cbs across utterances, albeit without disrupting overall coherence. A preliminary analysis suggests that discourse phenomena of this type account for some of the ZERO and NULL transitions. Other discourse phenomena that result in zero Cbs arise from characteristics of spoken language. This is the case with false starts and reported speech (the latter also present in writing). Examples of those can be found in our Coding Manual (Hadic Zabala and Taboada, 2006). In other cases, coherence may be maintained through relational coherence in the form of coherence relations (Poesio et al., 2004a,b). A more detailed analysis of the role of discourse phenomena and coherence relations is beyond the scope of this paper.

[3] We provide free translations for the Spanish examples, trying to keep the word order similar to the original. When a pronoun is null, we place it between parentheses in the translation.

Table 3
Subjects and Cbs, per transition, in English.

	Same	Different	No subject	No Cb
CONTINUE	729	12	65	0
ESTABLISH	388	128	30	0
RETAIN	3	152	4	0
SMOOTH SHIFT	258	4	25	0
ROUGH SHIFT	0	67	5	0
ZERO	0	0	0	503
NULL	0	0	0	365
Total (*n* = 2738)	1378 (50.33%)	363 (13.26%)	129 (4.71%)	868 (31.70%)

Table 4
Subjects and Cbs, per transition, in Spanish.

	Same	Different	No subject	No Cb
CONTINUE	699	70	141	0
ESTABLISH	277	122	73	0
RETAIN	12	166	11	0
SMOOTH SHIFT	174	20	48	0
ROUGH SHIFT	4	55	6	0
ZERO	0	0	0	439
NULL	0	0	0	284
Total (*n* = 2601)	1166 (44.83%)	433 (16.65%)	279 (10.73%)	723 (27.80%)

5. Subjects and topics

We are interested in how subjects and topics are realized in conversation. For each utterance in the conversations, we coded whether the subject and the Cb coincided, and what happened when they did not. Table 3 shows the results of that comparison for English, and Table 4 does the same for Spanish. In both cases, we excluded the initial utterance for each conversation that had no Cb.

Comparing the two languages, we can say that the distribution is similar, with subject and topic coinciding roughly half of the time. If we reinterpret the two tables, and take into account only the cases where an utterance has both a subject and a Cb, they are the same in about 80% of the cases in English, and 73% in Spanish. In other words, subjecthood and topichood are assigned to the same entity in most utterances.

The rest of the cases are utterances with either no subject or no Cb. The main difference between Spanish and English is a higher number of utterances with no subject in Spanish, and a lower number of utterances without a Cb. We have already discussed, in the previous section, why utterances may have no Cb. Utterances have no subject when they are fragments (such as noun and prepositional phrases) that are uttered as separate units, and contain entities that can form a Cf list. Even when a sentence with a finite predicate is present, the subject may be non-referential (such as English *it*). In Spanish, the count of utterances without a subject does not include cases of pro-drop, where the subject is recoverable through context and morphology, and therefore included in the Cf list.[4] Most of the no subject cases are fragments with inferred entities and without a finite verb, which seem to be more frequent in Spanish.

By transition, as it was to be expected, subject and Cb are the same most often in CONTINUE (90.5% of all CONTINUE transitions in English and 76.8% in Spanish) and ESTABLISH transitions (71.1% in English and 58.7% in Spanish). RETAIN and ROUGH SHIFT follow the opposite pattern. These patterns are true for both languages. There is a crucial difference between the two types of transitions classified according to whether the Cb equals the Cp, the first center in the list of

[4] The subject reference for pro-drop subjects in Spanish is easily recoverable in the vast majority of the cases. The verb shows agreement with the dropped subject, making the reference resolution no harder or easier than that of an English subject pronoun.

Table 5
Realization of Cb when it equals subject, in English.

	CONTINUE	ESTABLISH	RETAIN	SMOOTH SH.
Zero	69 (9.5%)	31 (8%)	–	10 (3.9%)
Pronoun	595 (81.6%)	303 (78.1%)	3 (100%)	187 (72.5%)
NP	29 (4%)	21 (5.4%)	–	18 (7%)
Other pronoun	36 (4.9%)	33 (8.5%)	–	43 (16.7%)
Total (*n* = 1378)	729	388	3	258

forward-looking centers. When they are the same (CONTINUE, some ESTABLISH, SMOOTH SHIFT), the subject is the Cb. When they are different (RETAIN, some ESTABLISH, ROUGH SHIFT), the subject does not encode the Cb. Since the highest-ranked entity in Cf (the Cp) tends to be the subject, this makes perfect sense. It validates Centering's distinction between CONTINUE/SMOOTH SHIFT and RETAIN/ROUGH SHIFT along salience lines.

We have mentioned in previous work (Taboada and Hadic Zabala, 2008), that Rule 2 of Centering (which states that there is a general ranking of preferred transition types: CONTINUE > RETAIN > SMOOTH SHIFT > ROUGH SHIFT) is fragile, and in corpus work, it has applied mostly to the two extremes. In other words, CONTINUE is always the most preferred transition, whereas ROUGH SHIFT is the least preferred, to the point of being non-existing in some studies. The preference of RETAIN over SMOOTH SHIFT does not always hold. Kibble (2001) characterizes Rule 2 as the formulation of two principles of discourse: cohesion and salience. Following cohesion, the same entity is repeated across utterances. Following salience, the most salient entity is realized as the subject. In CONTINUE transitions, both principles hold. In SMOOTH SHIFT, salience (the Cb is the subject) outranks cohesion (the current Cb is not the previous Cb). In RETAIN and SMOOTH SHIFT, only one holds at a time. That is why, we think, there is not a clear preference of one over the other. Kibble then proposes a different type of Rule 2, based on Strube and Hahn's (1999) cheap vs. expensive transitions. A cheap transition is one where the most salient entity of the previous utterance (the Cp) is realized as the Cb of the current utterance, capturing both salience and cohesion. The preference of cheap over expensive transitions, however, has not been consistently attested in data (just over 50% in Taboada and Hadic Zabala, 2008, but fewer cheap transitions in Poesio et al., 2004a,b and Byron and Stent, 1998). We believe that a four-way (or more) distinction seems to better represent both principles, and the continuum between obeying both or neither. We simply suggest that the strict hierarchy be changed to a relaxed ranking, with CONTINUE and ROUGH SHIFT at the edges, and with no ranking between RETAIN and SMOOTH SHIFT. With the addition of the dispreferred ZERO and NULL transitions, Rule 2 would be formulated as follows:

(4) CONTINUE > ESTABLISH > (RETAIN, SMOOTH SHIFT) > ROUGH SHIFT > ZERO > NULL

The next two sections discuss in more detail the cases where Cb and subject coincide, and where they do not, focusing on what referring expressions are used to encode them.

5.1. Subject and topic are the same

In this section, we examine cases where the Cb or topic of the sentence is also encoded as the grammatical subject of the sentence, focusing on its linguistic realization in terms of referring expressions.

Tables 5 and 6 provide a summary, broken down by transition. In English (Table 5), the trend is that if the Cb is the subject, then it will likely be encoded as a pronoun. This is the case regardless of the type of transition. In Spanish, the preferred realization is a zero pronoun, with a full pronoun a distant second. Noun phrases are used rarely across all transitions. Other pronouns include demonstratives, wh-pronouns, indefinite pronouns and, in a handful of cases in Spanish, the impersonal *se*.[5] This is consistent with theories that equate salience with minimal form of expression (Ariel, 1990; Givón, 1983; Gundel et al., 1993).

Both CONTINUE and ESTABLISH transitions show very high percentages of either pronouns or zero pronouns, as is to be expected. Although many of those are first and second person pronouns, a great deal are also third person.

[5] For example, *no se puede creer las cosas*, 'one cannot believe (those) things'.

Table 6
Realization of Cb when it equals subject, in Spanish.

	CONTINUE	ESTABLISH	RETAIN	SMOOTH SH.	ROUGH SH.
Zero	552 (79%)	184 (66.4%)	8 (66.7%)	118 (67.8%)	2 (50%)
Pronoun	86 (12.3%)	36 (13%)	–	20 (11.5%)	–
NP	26 (3.7%)	26 (9.4%)	3 (25%)	17 (9.8%)	–
Other pronoun	35 (5%)	31 (11.2%)	1 (8.3%)	19 (10.9%)	2 (50%)
Total ($n = 1166$)	699	277	12	174	4

Pronouns other than personal pronouns are also used in CONTINUE transitions. One such case is (5), where the Cb in (5c) is *almond tea with honey and milk*, realized as a demonstrative pronoun, *that*, because the speaker is emphasizing the subject, and picking up on the realization also of *that* in (5b).

(5) a. um and for my favorite beverage I put down of course almond tea with honey and milk
b. and so she read that
c. because that was what was there.

As can be seen in the table, in SMOOTH SHIFT transitions pronouns in English and zero pronouns in Spanish are the majority of the Cb realizations. Example (6) shows a Spanish example of a SMOOTH SHIFT transition, between (6b) and (6d),[6] but since the Cb is the speaker and therefore salient, a zero pronoun is sufficient to determine reference.

(6) B: a. está bueno eso
'that's good.'
Cb: that (systems and programming)
A: b. y acá estoy practicando, viste, con lenguajes de programación y demás
'and here I'm practicing, you see, with programming languages and such.'
Cb: languages
B: c. sí
'yes'
A: d. y estoy re ocupado con eso ahora.
'and (I)'m really busy with that now.'
Cb: A

There are cases, however, of SMOOTH SHIFT with NPs and other pronouns, in both languages. In the next example, there is a smooth shift from (7b) to (7c), since the center switches from the speaker to the ring. In (7c), the ring is realized as a demonstrative pronoun, probably because of the false start. The speaker started out the utterance with *Mike*, but then switched to *the ring* as the center of attention.

(7) a. but uh I think it really be like the ring
b. I stopped wearing the ring a—
c. Mike that's going to be Mikey's too.

The following is another example of a SMOOTH SHIFT with a demonstrative pronoun as Cb. This case is particularly noteworthy, because it could be interpreted as a violation of Rule 1, namely that if the utterance contains a pronoun, then the Cb should also be a pronoun. We did not compute all the cases of violations of Rule 1, but there are few, and most of them, like Example (8), not clear-cut. In the example, the Cb is the demonstrative pronoun *that* (the mistake), but the Cf list also contains a personal pronoun (*me*). Strictly speaking, Rule 1

[6] (6c) is a backchannel, and not part of the Centering analysis.

only mentions pronouns, and does not postulate a ranking among them, but *that* can be argued to be more marked than *me*.[7]

(8) a. . . .she she called Chicago
b. and she recognized the mistake [they made].
c. Well that caused great confusion for me

5.2. *Subject and topic are not the same*

The fact that the subject and the Cb of an utterance are not the same means that a very salient entity in the utterance (Cb) is not encoded as the grammatical subject of the utterance, a grammaticalized way of encoding salience. We identified three main reasons why this happens. First of all, some utterances encode a subject that contains very little information, such as impersonal subjects. There are also situations where empathy plays a role, especially in Spanish. Another reason for the separation of Cb and subject is that the current utterance is positioning a new entity as prominent (subject), while still keeping a connection to the previous utterance via the Cb. We explore these issues in this section.

The first case we will discuss is a straightforward one: Impersonal subjects and expletives such as *it* and *there* in English typically do not become Cbs in a sentence. Equivalent expressions exist in Spanish, typically realized in pro-drop form. A related case is that where an agent is briefly introduced in the conversation, but never made explicit, and the topic clearly remains something else. In (9), speaker A is describing, in great detail, her fertility treatment. Speaker B reminds her that the conversation is being recorded. Speaker A then switches from talking about herself, including the *I don't care* statement, to introducing a new referent in the discourse, *they*, while still keeping herself as a center of attention (*me*). The conversation is still clearly about A; *they* has not become a topic, but it is the subject of the sentence, a reference to the person or persons doing the recording, or listening to it. Such instances of *they* and the impersonal *you* explain some of subjects introduced in the conversation as discourse new, but realized as pronouns.

(9) B: a. wait a second
b. this is being recorded, Michelle
A: c. I don't care
d. they don't know me
B: e. oh okay
A: [laughter]

Altogether, expletives and this type of impersonal subjects account for a total of 5.23% (19/363) subjects that were different from the Cb in English and 12.9% (56/433) in Spanish. Figures in Spanish are higher because Spanish tends to use the impersonal *se* construction, which is often used to demote or defocus a human entity (García, 1975; Hadic Zabala and Taboada, 2006).

Secondly, empathy affects the ranking of entities in Spanish, as we mentioned in section 3. Empathy takes precedence over grammatical function, which means that the subject may not be the Cb. (10b) has nada ('nothing') as subject, but the speaker (*me*) as Cb, and as highest-ranked entity in the Cf list. Clearly, the Cb is the speaker throughout the segment, complaining about mail not arriving after she moved.

(10) a. lo que pasa es que como me he cambiado de casa
'what happens is that since I moved house'
b. no no me llega nada
'nothing nothing arrives for me' ('no mail reaches me')
c. y no sé nada
'and I don't know anything'

[7] A further complication in this case is that the other pronoun, *me*, is second person. Theories that propose hierarchies of referring expressions do not usually discuss first and second person pronouns (e.g., Gundel et al., 1993).

A third reason why the Cb is not the subject concerns topic progression. In these cases, an entity is reintroduced in the conversation as subject, while another entity is still the link to the previous utterance, the Cb. For example, in (11), *Fatima* (or *going to Fatima*) is a new topic, introduced in the first utterance as an oblique. It then becomes the subject of the following utterance, captured in a demonstrative pronoun. However, the Cb is still speaker B and somebody else (*you guys*).

(11) B: a. and eh next year we should be going to Fatima.
A: b. oh wow that'll be a big break for you guys then.

In general terms, we were interested in exploring whether subjects that were not Cbs were completely new entities, or whether they had already been introduced in the conversation and, if so, whether they were in a salient position. To this end, we annotated these two factors.

First of all, we annotated, for each instance of the subject and the Cb being different, whether the subject had been mentioned already, and how far back, i.e., the referential distance (Givón, 1983). Following Givón, we coded few referents past 20 clauses preceding. In Spanish, about 70% of the referents for subjects that were not Cbs had appeared earlier in the conversation, including referents to the participants. In English, the percentage is a bit over 75%. Most referents are to be found in the immediately preceding utterances, four at most. In cases where the referent appeared much earlier than four or so clauses, we can say that it is part of the global focus of the discourse, rather than the local focus for the current utterance. By global focus we mean the entire conversational context, which includes entities that have been discussed in previous discourse segments (Grosz and Sidner, 1986; Poesio et al., 2004a,b, 2006; Walker, 1998). The few non-Cb subjects that have not been previously mentioned in the conversation seem to be hearer-old (Prince, 1981), even when they are discourse-new.

The common pattern for a local referent is similar to the one found in Example (12). The referent is introduced as an object or oblique, *Mica* in (12c). Then it may be mentioned again, but still not in subject position: In (12d), it is a direct object pronoun (*him*). And, finally, it becomes a subject in (12e). However, since (12e) still keeps a reference to speaker A, which was highly ranked in (12d), speaker A is the Cb of (12e).

(12) B: a. w– we played pool with them like on Saturday night.
A: b. uh-huh.
B: c. with her and Anders and some guy named Mica.
A: d. oh that's oh I know him.
e. He's my my friend Medea's brother.

This pattern follows the Preferred Argument Structure (PAS) proposed by Du Bois (1987), who found that clauses typically have at most one full NP, and that NP is rarely the subject of an intransitive clause. More specifically, there tends to be at most one new argument per clause, introduced often as subject of an intransitive or object in a transitive clause. In (12) above, *Mica* is first a full NP, then a pronoun object, and it only becomes a pronoun subject in the third mention of the referent. Similarly, in a Centering analysis, Brennan (1995) found that a new entity introduced as a NP object should be mentioned as a full NP before being pronominalized.

For each subject that had been mentioned before, we also annotated the syntactic role that its referent previously had in the conversation. The majority of them had subject or object roles previously (65% in Spanish, 72% in English), including objects of preposition. Many of those are references to the participants in the conversation, which can be considered part of the global focus. But some are local referents, within the current discourse segment. The following example discusses e-mail. The referent is first introduced in the conversation within a discussion about how to get news. Then speaker A starts a long segment, explaining to B how e-mail works, including references to modems, accounts, and universities. The word *correo* ('mail') is mentioned again in (13b), in subject position, with A as the Cb. Then it is a subject again in (13g), this time a null subject.[8]

[8] In this example there are three relative clauses, enclosed in square brackets. We process relative clauses and calculate Centering structures for them, but they are ignored for the computation of the following utterance, following Kameyama (1998).

(13) A. Aquí me llegan las noticias por el correo electrónico.
'Here I get news through electronic mail.'
[. . . 46 utterances . . .]
A: a. Yo de mi casa llamo para esta universidad [donde está mi cuenta,]
'I call from home to this university [where I have my account,]'
b. todo el correo me llega a esta universidad, a mi cuenta de esta universidad.
'all mail arrives at this university, in my account at this university.'
B: c. Okey, okey.
A: d. Entiendes, con tu, con tu computadora en casa, tú te conectas a la universidad.
'You see, with your, with your computer at home, you connect to the university.'
e. [a donde está tu cuenta,] [a donde está todo,]
'[where you have your account,] [where everything is,]'
f. no necesitas ni el disco duro.
'(you) don't even need the hard drive.'
B: g. Ah, no me llega aquí, sino a la universidad
'Ah, (it) doesn't arrive here, but at the university.'

As in Example (12), we see confirmation of Brennan's (1995) findings that a new NP object is only pronominalized after a second mention as full NP. The entity *correo* is first an oblique NP, then a subject NP in (13b), and finally a zero pronoun in (13g). Similar principles probably apply to both Cbs and subjects: They have to be either hearer-old, or introduced first in object position, before they can become subjects or Cbs.

6. Conclusions

We have presented the results of a corpus analysis of spontaneous conversations from the point of view of Centering Theory. We used Centering as a way to distinguish between subjects (defined through verb agreement mostly) and topics (characterized as the backward-looking center).

This is a preliminary examination of the function and factors that affect choice of subject, topic, and referring expression. Using Centering Theory, we have been able to show that the majority of Cbs are also subjects, 80% of the time in English and 73% in Spanish. This means that, at least in our corpus, there is a strong preference for subjects that are topics, i.e., for conflating subjecthood and topichood. This is certainly not surprising; it confirms previous research that assigns salience to subjects, and it validates the claim of Centering that the Cb realizes the most salient entity. The tendencies were quite similar for both English and Spanish.

With respect to the frequency of transitions, the two languages had a similar distribution, with CONTINUE and ESTABLISH transitions being the most frequent, and ROUGH SHIFTS the least. RETAIN transitions were less numerous than SMOOTH SHIFT. In the paper, we discuss how this is related to our definition of ESTABLISH. We also point out that the absolute ranking of transitions in Rule 2 of Centering (CONTINUE > RETAIN > SMOOTH SHIFT > ROUGH SHIFT) may be better expressed as a relative ranking, with CONTINUE and ROUGH SHIFT at the edges, and no preference between RETAIN and SMOOTH SHIFT. With the addition of ESTABLISH as the most preferred after CONTINUE, the new ranking is: CONTINUE > ESTABLISH > (RETAIN, SMOOTH SHIFT) > ROUGH SHIFT > ZERO > NULL. Given that transitions capture the two principles of cohesion and salience, there needs to be a ranking of transitions that adhere to both, only one, or neither.

In terms of realization, when the Cb is also the subject, the preferred realization is pronoun in English and null pronoun in Spanish, regardless of the type of transition. This is also consistent with previous research: The most salient entity is expressed through the least marked referring expression available in the language.

When the Cb is not the subject, we find two different entities that are somehow prominent in the utterance. This is due to different factors. In some cases, the subject is impersonal, and thus more 'given' in the context than 'salient'. In Spanish, since empathy is part of the ranking of the Cf list, entities that are part of the point of view are often not subjects. As a result, there is a disassociation between the Cb, the point of view entity, and the subject, usually inanimate. This is the case in some constructions with psychological verbs (*Me gusta*, 'It pleases me/I like it'). Most other cases of Cb not being realized as the subject were instances of an entity becoming the subject and thus increasing its prominence, while another entity continued to provide a link to the previous utterance.

The Centering analysis revealed a large number of utterances in violation of Constraint 1, that each utterance have a Cb. Some of those corresponded to beginning of discourse segments, but the majority were segment-internal. Phenomena that played a role in the presence of zero Cbs were discourse-related, such as parallelism or reported speech, or associated with spoken language, such as false starts and repetitions. Absence of a Cb does not mean a lack of coherence in the conversations. Entity realization, the aspect of coherence that Centering deals with, is not the only factor responsible for discourse coherence. Relational coherence (in the form of coherence relations) and certain expectations derived from genre also play a role in the perception of coherence. For instance, the genre of the conversations, a sort of ''updating casual conversation'', where speakers update each other about recent events in their lives, means that they make frequent reference to family and friends that are hearer-old (Prince, 1981). The participants share a large amount of background knowledge that allows them to introduce new entities in the discourse without an explicit link to entities currently in the discourse.

Violations of Constraint 1 have been reported in the literature and, in fact, we have lower numbers than other studies. This is not necessarily a shortcoming of Centering Theory. It may be that the Constraint is not such, but merely a preference. Centering, which is mostly concerned with local coherence, captures local links between one utterance and the next. Other links across utterances are global in nature, and involve the conversational context.

What does this tell us about topichood and subjecthood in discourse? We have found that topics, in our case defined as the Cb of the utterance, tend to be subjects. In cases where subject and topic differ, the reasons are varied, but they all relate to discourse flow. Topichood is not only about mentioning and repeating entities, but about presenting those in a way that fits the context and the purpose of the discourse.

Acknowledgements

This work was supported by Social Sciences and Humanities Research Council of Canada (Standard Research Grant 410-2006-1009) and by Simon Fraser University, under a Discovery Parks grant and a President's Research Grant to Maite Taboada. Loreley Wiesemann was supported by a SSHRC CGS Doctoral Scholarship (767-2006-2048). We would like to thank Nancy Hedberg for many helpful discussions and feedback.

References

Ariel, Mira, 1990. Accessing Noun Phrase Antecedents. Routledge, London.

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

Brennan, Susan E., Friedman, Marilyn W., Pollard, Carl J., 1987. A Centering approach to pronouns. In: Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL-87), Stanford, CA, USA, pp. 155–162.

Byron, Donna K., Stent, Amanda, 1998. A preliminary model of Centering in dialog. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-98), Montréal, Canada, pp. 1475–1477.

Chafe, Wallace, 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In: Li, C.N. (Ed.), Subjects and Topics. Academic Press, New York, pp. 26–56.

Di Eugenio, Barbara, 1998. Centering in Italian. In: Walker, M.A., Joshi, A.K., Prince, E.F. (Eds.), Centering Theory in Discourse. Clarendon, Oxford, pp. 115–137.

Diderichsen, Philip, Elming, Jakob, 2005. A corpus-based approach to topic in Danish dialog. In: Proceedings of ACL Student Research Workshop. pp. 109–114.

Du Bois, John W., 1987. The discourse basis of ergativity. Language 63 (4), 805–855.

Fais, Laurel, 2004. Inferable centers, Centering transitions and the notion of coherence. Computational Linguistics 30 (2), 119–150.

García, Erica, 1975. The Role of Theory in Linguistic Analysis: The Spanish Pronoun System. North-Holland, Amsterdam.

Givón, Talmy, 1983. Topic continuity in discourse: an introduction. In: Givón, T. (Ed.), Topic Continuity in Discourse: A Quantitative Cross-Language Study. John Benjamins, Amsterdam and Philadelphia, pp. 1–41.

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

Grosz, Barbara J., Joshi, Aravind K., Weinstein, Scott, 1995. Centering: a framework for modelling the local coherence of discourse. Computational Linguistics 21 (2), 203–225.

Gundel, Jeanette K., Hedberg, Nancy, Zacharski, Ron, 1993. Cognitive status and the form of referring expressions in discourse. Language 69, 274–307.

Hadic Zabala, Loreley, and Taboada, Maite, 2006. Centering Theory in Spanish: Coding Manual. Unpublished manuscript, Simon Fraser University. Available from: http://www.sfu.ca/~mtaboada.

Halliday, Michael A.K., Hasan, Ruqaiya, 1976. Cohesion in English. Longman, London.

Hurewitz, Felicia, 1998. A quantitative look at discourse coherence. In: Walker, M.A., Joshi, A.K., Prince, E.F. (Eds.), Centering Theory in Discourse. Clarendon, Oxford, pp. 273–291.

Kameyama, Megumi, 1986. A property-sharing constraint in Centering. In: Proceedings of the 24th Annual Meeting of Association for Computational Linguistics (ACL-86), New York, USA, pp. 200–206.

Kameyama, Megumi, 1998. Intrasentential Centering: a case study. In: Walker, M.A., Joshi, A.K., Prince, E.F. (Eds.), Centering Theory in Discourse. Clarendon, Oxford, pp. 89–112.

Keenan, Edward L., 1976. Towards a universal definition of ‘‘subject’’. In: Li, C.N. (Ed.), Subject and Topic. Academic Press, New York, pp. 303–333.

Kibble, Rodger, 2001. A reformulation of Rule 2 of Centering. Computational Linguistics 27 (4), 579–587.

Kibble, Rodger, Power, Richard, 2004. Optimizing referential coherence in text generation. Computational Linguistics 30 (4), 401–416.

Kingsbury, Paul, Strassel, Stephanie, McLemore, Cynthia, McIntyre, Robert, 1997. CallHome American English Transcripts LDC97T14 [Corpus]. Linguistic Data Consortium, Philadelphia, PA.

Li, Charles N., Thompson, Sandra A., 1976. Subject and topic: a new typology of language. In: Li, C.N. (Ed.), Subject and Topic. Academic Press, New York, pp. 457–489.

Miltsakaki, Eleni, 2002. Toward an aposynthesis of topic continuity and intrasentential anaphora. Computational Linguistics 28 (3), 319–355.

Miltsakaki, Eleni, Kukich, Karen, 2004. Evaluation of text coherence for electronic essay scoring systems. Natural Language Engineering 10 (1), 25–55.

Poesio, Massimo, Mehta, Rahul, Maroudas, Axel, Hitzeman, Janet, 2004a. Learning to resolve bridging references. In: Proceedings of ACL, Barcelona, Spain.

Poesio, Massimo, Stevenson, Rosemary, Di Eugenio, Barbara, Hitzeman, Janet, 2004b. Centering: a parametric theory and its instantiations. Computational Linguistics 30 (3), 309–363.

Poesio, Massimo, Patel, Amrita, Di Eugenio, Barbara, 2006. Discourse structure and anaphora in tutorial dialogues: an empirical analysis of two theories of global focus. Research on Language and Computation 4, 229–257.

Prince, Ellen F., 1981. Towards a taxonomy of given-new information. In: Cole, P. (Ed.), Radical Pragmatics. Academic Press, New York, pp. 223–255.

Stevenson, Rosemary, 2002. The role of salience in the production of referring expressions: a psycholinguistic perspective. In: van Deemter, K., Kibble, R. (Eds.), Information Sharing: Reference and Presupposition in Language Generation and Interpretation. CSLI Publications, Stanford, pp. 167–192.

Strube, Michael, Hahn, Udo, 1999. Functional Centering: grounding referential coherence in information structure. Computational Linguistics 25 (3), 309–344.

Taboada, Maite, 2005. Anaphoric terms and focus of attention in English and Spanish. In: Butler, C., Gómez-González, M.d.l.Á., Doval, S. (Eds.), The Dynamics of Language Use: Functional and Contrastive Perspectives. John Benjamins, Amsterdam and Philadelphia, pp. 195–216.

Taboada, Maite, 2008. Reference, centers and transitions in spoken Spanish. In: Gundel, J.K., Hedberg, N. (Eds.), Reference and Reference Processing. Oxford University Press, Oxford, pp. 176–215.

Taboada, Maite, Hadic Zabala, Loreley, 2008. Deciding on units of analysis within Centering Theory. Corpus Linguistics and Linguistic Theory 4 (1), 63–108.

Walker, Marilyn A., 1998. Centering, anaphora resolution, and discourse structure. In: Walker, M.A., Joshi, A.K., Prince, E.F. (Eds.), Centering Theory in Discourse. Clarendon, Oxford, pp. 401–435.

Walker, Marilyn A., 2000. Toward a model of the interaction of Centering with global discourse structure. Verbum 22.

Walker, Marilyn A., Joshi, Aravind K., Prince, Ellen F., 1998a. Centering in naturally occurring discourse: an overview. In: Walker, M.A., Joshi, A.K., Prince, E.F. (Eds.), Centering Theory in Discourse. Clarendon, Oxford, pp. 1–28.

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

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

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

Loreley Wiesemann 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.