Centering Theory in Spanish: Coding Manual[*]

Loreley Hadic Zabala and Maite Taboada
lmhadic@sfu.ca, mtaboada@sfu.ca
Simon Fraser University

Current version: June 6, 2006

0. Introduction

This is a manual for coding Centering Theory (Grosz et al., 1995) in Spanish. The manual is still under revision. The coding is being done on two sets of corpora:

- ISL corpus. A set of task-oriented dialogues in which participants try to find a date where they can meet. Distributed by the Interactive Systems Lab at Carnegie Mellon University. Transcription conventions for this corpus can be found in Appendix A.

- CallHome corpus. Spontaneous telephone conversations, distributed by the Linguistics Data Consortium at the University of Pennsylvania. Information about this corpus can be obtained from the LDC.

This manual provides guidelines for how to segment discourse (Section 1), what to include in the list of forward-looking centers (Section 2), and how to rank the list (Section 3). In Section 4, we list some unresolved issues.

1. Utterance segmentation

1.1 Utterance

In this section, we discuss how to segment discourse into utterances. Besides general segmentation of coordinated and subordinated clauses, we discuss how to treat some spoken language phenomena, such as false starts.

In general, an utterance U is a tensed clause. Because we are analyzing telephone conversations, a turn may be a clause or it may be not. For those cases in which the turn is not a clause, a turn is considered an utterance if it contains entities.

The first pass in segmentation is to break the speech into intonation units. For the ISL corpus, an utterance U is defined as an intonation unit marked by either {period}, {quest} or {seos} (see Appendix A for details on transcription). Note that {comma}, unless it is followed by {seos}, does not define an utterance.

In the example below, (1c.) corresponds to the beginning of a turn by a different speaker. However, even though (1c.) is not a tensed clause, it is treated as an utterance because it contains entities, it is followed by {comma} {seos}, and it does not seem to belong to the following utterance.

[*] This work is supported by the Ministry of Science and Technology of Spain and the Xunta de Galicia, under project MCYT-FEDER BFF2002-

1

a. fvgc:	así	que	si	Ø	te	viene	bien
	so	that	if	nullpro:3SG	OBJ:2SG	go:3SG.PRES	well

	de	diez	a	doce {comma}
	from	ten	to	twelve

‘So if (it) is good for you from ten to twelve’
Cf: fsnm (te), 10-12
Cb: 0

b.este	Ø	est-á	bien {period} #key_click# {seos}
eh	nullpro:3SG	be -3SG:PRES	well

‘then, (it) is good’
Cf: 10-12 (zero)
Cb: 10-12
Transition: CONTINUE

c. fsnm:	**perfecto {period} {seos}**	**diez**	**a**	**doce**	**el**	**veintitrés {comma}{seos}**
	perfect	ten	to	twelve SG: 10		

3 fcba_08_02: /h#/ **bueno {period} {seos}** el

We are, for the time being, considering the first model of intrasentential Centering as our general model of Centering. That is, each clausal unit (segmented as described below) is a Centering unit. We believe this is the most appropriate model for spoken discourse. Exceptions are those mentioned by Kameyama: reported speech and non-report complements, where the reported part is embedded in the same Centering unit as the reporting unit (see below). These are to be processed differently: the embedded part becomes a segment and undergoes Centering analysis, but is not considered an update unit for the following clause. This is the approach taken by Suri and McCoy (1994) for processing main-subordinate clauses pairs (“X because Y”). We do not believe that approach is appropriat [illegible]

b.	**y**	Æ	**ten -és**	**acceso,**
	and	nullpro:2SG	have-2SG:PRES	access

‘And (you) have access’
Cf: B (nullpro), internet (acceso)
Cb: B
Transition: SMOOTH SHIFT

c.	**yo**	**ten -go**	**también**	**acceso,**
	I	have-1SG:PRES	too	access

‘I also have access’
Cf: A (yo), internet (acceso)
Cb: internet
Transition: ROUGH SHIFT

d. ∅ access

b. **y** **(yo** **teng -o)** **otro -s** **laburo-s**
and (I have-1SG:PRES) other-MASC:PL job -MASC:PL
‘and (I have) other jobs’
Cf: B (ellipsis), laburos
Cb: B

1.2.4 Tenseless adjuncts: Tenseless clausal and phrasal adjuncts belong to the same utterance unit as the immediately superordinate clause (Tenseless adjunct hypothesis, TlessAdj). In example (11), the tenseless adjunct ‘para enganchar todo’ does not constitute a center-updating unit and belongs in the same unit with the main clause.

11 porque Ø tene-mos un -o -s (( )) por.ahí

i. **Ø no me bronc -a. </DA>**

they tend to occur with 1st person subjects (for comprehensive list, see Thompson 2002: 138). Following Thompson (2002: 136), these CTPs and their clausal complements will be analyzed as monoclausal utterances. In other words, the clausal complements of epistemic, evidential or evaluative CTPs do not constitute embedded segments. CTPs express the epistemic/evidential/evaluative stance of the speaker towards the information contained in the complement clause, and could be substituted by modals or adverbs (Thompson, 2002: 132). The analysis of these clauses is a flat analysis, i.e., as if there was no embedding. The subject of the CTP is typically the first entity in the Cf list.

Examples (14) and (15) illustrate this type of construction. In (14d.) the verb *creo* 'believe' creates an epistemic frame for the clause that follows. It is the speaker's belief that his friend ended her relationship with her boyfriend in England.

14 a."B" una amiga que dejó la escuela, le entró la locura y se fue con su novio que estudia medicina en [PAUSE] el Medical College o algo así, de Inglaterra, y se largó con él
'A friend who left school, went crazy and left with her boyfriend who studies medicine at the English Medical College, or something like that, in England, and she went with him.'

b."A" Mmm </DA>

c."B" y	este, </DA>	y	∅	se	fue	a
and	uh	and	nullpro:3SG	3SG:RFL	go: 3SG:PAST	to
	la	aventura,				
	the: FEM:SG	adventure				

'And, uh, she went looking for adventure'

mm </DA>

15 a.

b.∅ Se compr-ó algun-as cosa -s de est -a follows. In terms of the Cf-list, the ranking of the entities in the false start with respect to the entities in the repaired speech proceeds linearly. Note however, that only the false starts that contain entities are taken into account. This is illustrated in example (20). In (20a.), *te*, a pronoun referring to the addressee, becomes part of the Cf list. In (20b), there are no entities in the false start (marked with angled brackets), and therefore there is nothing to include in the Cf list.

20 a.fmcs_01_11: *pause*

bueno {period} {seos} <	**te {seos}** > /mm/	entonces	
well	2SG:OBJ	then	
∅	qued -amos	así {period} {seos}	
nullpro:1PL	arrange-1PL:PRES	so/like.this	

‘Well, <**you**> then (we) agree on this’
Cf: **fmgl (te)**, nosotras
Cb: 0
Transition: NONE

b.

por favor	no	∅	te	olvid -es	
please	not	nullpro.2SG	2SG:RFL	forget-2SG:PRES	
de	tra -er	tod-os	l -os	legajo-s {period}	
of	bring-INF	all -MASC:PL	the-MASC:PL	file -PL	
/h#/ <	**para**	**pod -er**	**este**>	para	ten -er
	to	be.able-INF	eh	to	have-INF
tod-a	l -a	información	a	mano	
all -FEM:SG	the-FEM:-	to ndTj 51.75 0 mano TD 0 Tc 0.5411 Tw (mano) Tj 3875 0 f 0 T			

e backward 0 3 6 3c - 0 . 1 F 0

b. nadie	l -a	co- </DA>
nobody	OBJ-FEM:SG	knw-

‘Nobody (knows) her’
Cf: nadie, Mónica Martínez (la)
Cb: Mónica Martínez
Transition: RETAIN

c. "A" un-a	much-	**un-a**	**muchacha**
a -FEM:SG	gir-	a -FEM:SG	girl

‘**A gir- a girl**’
Cf: Mónica Martínez (muchacha)
Cb: Mónica Martínez
Transition: CONTINUE

que	nac	-ió	en	Camiri [PAUSE]	Cochabamba,	Bolivia </DA>
that	be.born	-3SG:PRET	in	Camiri	Cochabamba	Bolivia

‘who was born in Camiri, Cochabamba, Bolivia.’
Cf: Mónica Martínez (que), Camiri, Cochabamba, Bolivia
Cb: Mónica Martínez
Transition: CONTINUE

2.2 Synonyms and near synonyms (when they have the same reference)

In examples (22a.) and (22c.) below, the words *picture* and *icon* have the same reference and are used as synonyms.

22 a. B: if someone could send me the %um **the blessed virgin picture**
Cf: someone, B (me), picture
Cb: B
Transition: ROUGH SHIFT

that I have in my room
Cf: B (I), picture (that), B (my), room
Cb: B
Transition: CONTINUE

b. A: okay

c. B: **the icon** that's next to that gold %uh cross
Cf: icon, gold cross
Cb: icon
Transition: SMOOTH SHIFT

that I have
Cf: B (I), cross (that)
Cb: cross
Transition: ROUGH SHIFT

2.3 Superordinate

23	a. "B"	Sí, </DA>	además	Ø	no	te	dijeron	que	tipo
		yes	also	nullpro:3PL	not	2SG:OBJ	say:PRET :3PL	what	type

de	**ganado**, </DA>
of	cattle

‘And also, (they) didn’t tell you what type of **animal**.”
Cf: 3pl (nullpro), A (te), ganado
Cb: ganado
Transition: RETAIN

b. a.lo.mejor	Ø	son	**topo-s**,	o -- </DA>
maybe	nullpro:3PL	be:3PL:PRES	mole-PL	or

‘Maybe (they) are **moles**.’
Cf: ganado (topos)
Cb: ganado
Transition: CONTINUE

2.4 Inclusive relation

24	a. "A"	Y	l -os,	mija,	y	l -os
		and	the-MASC:PL	my.daughter	and	the-MASC:PL

niñ -it -os	qué	tal	est-án. </DA>
kid –DIM-MASC:PL	what	¿?	be -3PL:PRES

‘And the, dear, and **the kids** how are they?’
Cf: niñitos
Cb: 0
Transition: NONE

b. "B"	Bien, </DA>	**Samuel**	ayer	se	ca-yó	en
	well	Samuel	yesterday	3SG:RFL	fall-3SG:PRET	in

l -a	pisci-	afuera	de	l -a	piscina, </DA>
the-FEM:SG	swim-	outside	of	the-FEM:SG	swimingpool

‘Well, yesterday **Samuel** fell in the swim- outside the swimingpool.’
Cf: Samuel (one of niñitos), piscina
Cb: Samuel
Transition: CONTINUE

2.5 Part – whole

25 a. I mean there was **trees** down
Cf: trees
Cb: 0
Transition: NONE

b. there was **branches** all over
Cf: trees (branches)
Cb: trees
Transition: CONTINUE

3. Cf –Ranking

3.1 Ranking criterion

The most important aspect of adapting Centering Theory to a new language is to determine the ordering of the Cf list, what Cote (1998) calls the *Cf template* for a language.

We mainly follow grammatical relations as the basis for ordering the Cf list in Spanish, therefore Subjects are ranked higher than Objects, whether they appear as full pronouns (26), or as null pronouns.

26	como	**vos**	**me**	has	dicho	en
	like	2SG:SUBJ	1SG:OBJ 2			

28	/h#/ **me**	viene		mejor	el	jueves {comma} {seos}	
	1SG	come:3SG:PRES		better	the:MASC:SG	Thursday	

‘Thursday is better for **me**’
Cf: mphb (me), jueves
Cb: 0

29	/h#/ *pause*	este /ls/	qué.tal	para	**ti** {comma} *pause*		
		so	how	for	2SG		
		del		quince	a -l	diecinueve {period} {seos}	
		from.the:MASC:SG		fifteenth	to-the:MASC:SG	nineteenth	

‘How is it for **you** from the fifteenth to the nineteenth?’
Cf: meba (tí), del 15 al 19
Cb: 0

Empathy also includes verbs with clausal grammatical subjects, but with an animate experiencer, or person from whose point of view the statement is to be interpreted. In (30), the experiencer is in a prepositional phrase (*para mí,* ‘for me’). We believe the experiencer should be ranked higher than either the clause as a whole that has the function of subject (*juntarme con vos ese día,* ‘to get together with you that day’), , or any of the entities included in that clause.

30	así	que	para	**mí**	ser-ía	imposible	
	so	that	for	1SG	be -PRES.COND	impossible	
		junt-ar -me		con	vos /h#/ /eh/ es -e		día /h#/
		join-3)0uney j 3vey empathy and T(ii)0uney ar19often place 2.1f -ejune Tj 15-10TD 0.-F45 0 TD 3 0 40imposibTw 0 Tc -0.v					

32	"A" ¡	Y	que	Ø	**se**	**l**	**-o**	d	-an! </DA>
		and							

Marta's letters)[3]. Thus, in Example (36), *una de Marta* refers to one (letter) from Marta. Since Marta is animate, it is ranked higher.

3.4 Relative pronouns

Relative pronouns should be ranked according to the role of the pronoun in the relative clause Subj>obj>etc., for the purpose of computing the Cf list. However, Poesio et al. (to appear) have shown that relative pronouns are not affected by Rule 1 of Centering, i.e. the Cb need not be a pronoun when a non-

b. **Ø**	Tiene-s	que	present-ar	un -o	cada	año.
imp.2SG	have -2SG:PRES	that	present -INF	one-MASC:SG	every	year

‘(**You**) have to take one every year’
Cf:examen (uno), cada año, **imp-tú (nullpro)**
Cb: examen
Transition: continue

It is interesting to note that this second person form is often used as an indirect form of reference to the speaker. In Example (39), the speaker is implying that he has to take one exam every year. The tú form might indicate simply that that’s the norm, and he is no exception. If we were to consider that the second person form has some reference to the speaker, its ranking in the Cf list would have to change. For the time being, however, we are considering it as a type of impersonal form.

3.5.2 Impersonal third person plural

Third person plural can be used impersonally when the speaker does not include him/herself or the hearer in the reference (Butt and Benjamin 1994:374). As above, impersonal third person plurals are included in the list of forward-looking centers, but they are ranked low (arbitrary plural pronouns).

40	"A" ¡	Y	que	**Ø**	se	lo	d -an! </DA>
		and	that	nullpro:3pl	3sg	obj:3sg	give- 3pl:pres

‘And it was given to her’
Cf: hermana (se), programa (lo), **imp-3pl (nullpro)**
Cb: hermana

3.5.3 Impersonal *se*

García (1975:24) identifies three impersonal *se* constructions:

- Impersonal sentence containing an inanimate nominal that is not the logical subject: *Se quemó el dulce* ‘The jam was burnt’ / ‘Someone burnt the jam’;
- Impersonal sentence containing an animate nominal preceded by *a*, in which case no subject is available for *se* to refer to: *Se fusiló a los prisioneros* ‘The prisoners were shot’/’Someone shot the prisoners’;
- Impersonal sentence containing no nominal: *Se vive mejor en España* ‘One lives better in Sp se

41	Ya	**se**	te	oye	muy	bien. </DA>
	already	imp.3SG	OBJ:2SG	hear: 3SG	very	well

‘You already sound very well’
Cf: B (te), **imp-se**
Cb: 0

3.6 Subjects and predicates of verb to be (*ser* & *estar*)

The verb to be functions as a linking verb, so subjects and predicates (nominal and adjectival) of the verb to be are co-referential and only need to be listed once in the Cf-list.

42	a. no,	∅	l -a	conoc-ieras, </DA>
	no	nullpro:2SG	OBJ-FEM:SG	know –2SG:IMP:SUBJ

‘If (you) knew her’
Cf: A (nullpro), maestra (la)
Cb: A
Transition: SMOOTH SHIFT

b. **Ø**	es	**un**	**cuerazo.** </DA>
nullpro:3SG	be:3SG:PRES	a:MASC:SG	good.body

‘(She) has a good body.’
Cf: **maestra (nullpro) = cuerazo**
Cb: maestra
Transition: SMOOTH SHIFT

It is possible to have only a predicate (elliptical subject and predicator) in an utterance. In these cases, since the predicate is co-referential with the elliptical subject of the elliptical predicator, we include the subject in the list of forward-looking centers.

43	a. "A" Ay, a	poco	si	Ø	est-ás	hinchad-o. </DA>
	at	little	if	nullpro:2SG	be -2SG:PRES	swollen-MASC:SG

‘If anything, (you) are swollen.’
Cf: B (nullpro)
Cb: 0
Transition: NONE

b."B"	Sí,	sí. </DA>
	yes	yes

‘Yes, yes.’

c.No,	**hinchad-o**,	no,	no. </DA>
no	swollen –MASC:SG	no	no

‘Not swollen,’
Cf: **B**
Cb: B
Transition: CONTINUE

d. **Llen-o**	de	granitos,	no,	este </DA>
full –MASC:SG	of	zits	no	eh

‘full of zits.’
Cf: **B**, granitos
Cb: B
Transition: CONTINUE

4. Unresolved issues

4.1 Speech directed to a third party not in conversation

One issue to be resolved is the treatment of utterances directed to a third party not in conversation.

Following Kameyama (1998:107), the reported unit in (44d.), *que más o menos*, constitutes an embedded segment, unaccessible to centering. The reporting unit, *Æ dice*, however, is not an embedded segment and is accessible to centering. A Centering analysis shows that the speech directed to a third party must be included in the analysis since it contains the antecedent for the null pronoun.

Analysis including speech directed to a third party

44	a. ¿	Te	lat -e	que	como	quince? </DA>
		OBJ:2SG	beat-3SG:PRES	that	like	fifteen

‘Does fifteen (minutes) sound about right?
Cf: B (te), quince minutos
Cb: 0
Transition: NONE

b. "B"	Pues	no	sé	yo </DA>
	well	not	know:1SG:PRES	I

‘Well, I don’t know.’
Cf: B (yo)
Cb: B
Transition: CONTINUE

c. **// Ø llev -amos como quince ‘W Tf 0 .75 0 w (CONTINUE)>**

b. "B"	Pues	no	sé	yo </DA>
	well	not	know:1SG:PRES	I

‘Well, I don’t know.’
Cf: B (yo)
Cb: B
Transition: CONTINUE

c. // Ø	llev	-amos	como	quince	minutos,	mamá? // </DA>
nullpro:1PL	be.talking-1PL:PRES		like	fifteen	minutes	mom?

‘Mom, have we been talking for fifteen minutes?’

d. Ø	dice	que	más	o	menos </DA>
nullpro:3SG	say: 3SG:PRES	that	more	or	less

‘(She) says that (we have been talking for about fifteen minutes) more or less.’
Cf: mamá (nullpro)
Cb: 0
Transition: NONE

4.2 Pronouns referring to discourse segments

A second unresolved issues concerns the use of pronouns to refer to discourse segments, and how to deal with it within Centering Theory. The following example illustrates such use of pronouns. In (46c.), the demonstrative eso (‘that’) refers to the consequences of e-mail use that have been described in the two previous utterances. It is unclear how to list such “entities” as forward-looking centers.

46	a. B"	Porque	Ø	deja-s	de	escrib-ir =le	a
		because	imp:2SG	stop-2SG:PRES	of	write -INF=OBJ:3SG	to
				l -a	gente </DA>		
				the-FEM:SG	people		

‘Because (you) stop writing to people’
Cf: gente (-le), gente, imp-tú (nullpro)
Cb: 0
Transition: NONE

b. y	además	Ø	no	guard-as	l -as	carta-s </DA>
and	also	imp:2SG	not	keep -2SG:PRES	the-FEM:PL	letter-PL

‘and (you) don’t keep the letters either’
Cf: cartas, imp-tú (nullpro)
Cb: imp-tú (nullpro)
Transition: RETAIN

c. "A"	Sí,	**es -o**	es	l -o	mal-o, </DA>
	yes	that-MASC:SG	be: 3SG:PRES	the-MASC:SG	bad-MASC:SG

‘Yes, **that** is the bad thing about it’
Cf: **[dejas de escribirle a la gente y además no guardas las cartas] (eso)**
Cb: 0
Transition: NONE

Appendix A: Transcription conventions for the ISL corpus

The transcripts include a number of conventions introduced by the transcriber. These include human and non-human noises, as explained below.

CATEGORY	BRACKET	
human noises	/…/	slashes
non-human noises	#…#	hash marks/pound sign
silences	*…*	asterisks
mispronunciations	[…]	square brackets (around whole word)
	(…)	parentheses (supply missing part of word or correct pronunciation of word, only inside square brackets)
transcriber comments	{…}	curly braces
accent	|…|	vertical bars/pipes
false starts	<…>	angled brackets

In addition, transcriber comments include intonation, marked with one of the following at the end of the corresponding section of speech.

{period}	Falling intonation
{comma}	Slightly rising intonation, continuation of idea, and not a question
{quest}	Marked rising intonation

These comments do not reflect, or are influenced by, sentence structure. The speaker may have the intonation of a statement whether he or she is, in fact, asking a question. He or she may have the intonation of a period after a collection of words that do not, in any way, resemble a grammatically correct or complete sentence.

References

Butt, J. & C. Benjamin (1994). A New Reference Grammar of Modern Spanish. Edward Arnold/Hodder Headline Group.

Byron, D. & A. Stent (1998). A Preliminary Model of Centering in Dialog. University of Rochester CS Department. R687. http://citeseer.nj.nec.com/article/byron98preliminary.html

Cote, S. (1998) Ranking Forward-Looking Centers. In M. Walker, A. Joshi & E. Prince, Eds. Centering Theory in Discourse. Oxford University Press, 55-69.

Di Eugenio, B. (1998). Centering in Italian. In M. Walker, A. Joshi & E. Prince, Eds. Centering Theory in Discourse. Oxford University Press, 115-137.

Eckert, M. & M. Strube (1999). Resolving Discourse Deictic Anaphora in Dialogues. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway, 37-44. http://citeseer.nj.nec.com/miriam99resolving.html

García, E. (1975). *The role of theory in linguistic analysis: the Spanish pronoun system.* Amsterdam and Oxford: North-Holland.

Grosz, B., A. Joshi and S. Weinstein (1995) Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 2 (21), 203-226.

Kameyama, M. (1998). Intrasentential Centering: A Case Study. In M. Walker, A. Joshi & E. Prince, Eds. *Centering Theory in Discourse*. Oxford University Press, 89-112.

Kuno, S. (1987). *Functional Syntax*. The University of Chicago Press.

Radford, A. (1997). *Syntactic Theory and the Structure of English: A Minimalist Approach.* Cambridge University Press.

Taboada, M. (2002a) Centering and Pronominal Reference: In Dialogue, In Spanish. *Proceedings, 6th Workshop on the Semantics and Pragmatics of Dialogue, EDILOG*. Edinburgh. September 2002: 177-184

Taboada, M. (2002b) Foco y pronominalización en la lengua hablada: Una primera aproximación. *Documentos de Español Actual* 3-4: 173-200.

Taboada, M. and L. Hadic Zabala (2004) What are the units of discourse structure? Segmenting discourse within Centering Theory. Manuscript. Simon Fraser University.

Thompson, S.A. (2002). “Object complements” and conversation. Towards a realistic account. *Studies in Language*, 26 (1), 125-164.

Walker, M. and E. Prince (1996) A bilateral approach to givenness: A hearer-status algorithm and a Centering algorithm. In Thorstein Fretheim and Jeanette Gundel (eds.) *Reference and Referent Accessibility.* Amsterdam: John Benjamins. 291-306.