The discourse functions of Italian subjects: a centering approach

Barbara Di Eugenio
Computational Linguistics
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
dieugeni@andrew.cmu.edu

Abstract

This paper examines the discourse functions that different types of subjects perform in Italian within the centering framework (Grosz et al., 1995). I build on my previous work (Di Eugenio, 1990) that accounted for the alternation of null and strong pronouns in subject position. I extend my previous analysis in several ways: for example, I refine the notion of CONTINUE and discuss the centering functions of full NPs.

1 Introduction

Interpreting referential expressions is important for any large coverage NL system; while such systems do exist for Italian, e.g. (Stock et al., 1993; Lombardo and Lesmo, 1994), to my knowledge not much attention has been devoted to the interpretation of Italian referential expressions. Some exceptions are (Samek-Lodovici and Strapparava, 1990), that discusses interpretation of referential expressions within dialogues to access a videodisc on Italian art; (Not and Zancanaro, 1995), that adopts a systemic grammar approach (Halliday, 1976); and (Di Eugenio, 1990), which uses centering theory (Grosz et al., 1995) to account for the alternation of null and strong subjects.

In this paper, I build on and expand (Di Eugenio, 1990) in several ways. First, I reanalyze the hypotheses I proposed earlier with respect to a corpus of naturally occurring data:[1] I show that those hypotheses are basically supported, and that when they aren't an elegant explanation can be found by looking at a two member sequence of centering transitions rather than at just one transition. Second, I extend my previous analysis by also discussing the centering functions of full NPs in subject position, and some occurrences of pronouns unaccounted for by centering.

2 Centering theory

Centering theory (Grosz et al., 1986; Brennan et al., 1987; Grosz et al., 1995) models local coherence in discourse: it keeps track of how local focus varies from one utterance to the next. Centering postulates that:[2]

- Each utterance U_n has associated with it a set of discourse entities, the FORWARD-LOOKING CENTERS or Cfs. The Cf list is ranked according to discourse salience.
- The BACKWARD-LOOKING CENTER, or Cb, is the member of the Cf list that U_n most centrally concerns, and that links U_n to the previous discourse.
- Finally, the PREFERRED CENTER, or Cp, is the highest ranked member of the Cf list. The Cp represents a prediction about the Cb of the following utterance.

Transitions between two adjacent utterances U_{n-1} and U_n can be characterized as a function of *looking backward* — whether Cb(U_n) is the same as Cb(U_{n-1}) — and of *looking forward* — whether Cb(U_n) is the same as Cp(U_n). Table 1 illustrates the four transitions that are defined according to these constraints. (Brennan et al., 1987) proposes a default ordering on transitions which correlates with discourse coherence: CONTINUE is preferred to RETAIN is preferred to SMOOTH-SHIFT is pre-

[1] The examples in (Di Eugenio, 1990) were constructed.

[2] The version of centering I present here is from (Brennan et al., 1987).

	$\mathrm{Cb}(\mathrm{U}_n) = \mathrm{Cb}(\mathrm{U}_{n-1})$	$\mathrm{Cb}(\mathrm{U}_n) \neq \mathrm{Cb}(\mathrm{U}_{n-1})$
$\mathrm{Cb}(\mathrm{U}_n) = \mathrm{Cp}(\mathrm{U}_n)$	CONTINUE	SMOOTH-SHIFT
$\mathrm{Cb}(\mathrm{U}_n) \neq \mathrm{Cp}(\mathrm{U}_n)$	RETAIN	ROUGH-SHIFT

Table 1: Centering Transitions

ferred to ROUGH-SHIFT.[3]

The saliency ordering on the Cf list, which is generally equated with grammatical function, for Western languages is SUBJECT > OBJECT2 > OBJECT > OTHERS, where OTHERS includes prepositional phrases and adjuncts. (Kameyama, 1985) was the first to point out that for languages such as Japanese empathy and topic marking affect the Cf ordering, and proposed the following ranking

(1) **empathy** > SUBJECT > OBJECT2 >
OBJECT > OTHERS

I follow (Turan, 1995) in adopting (1) also for Western languages. Turan argues that a notion analogous to empathy arises in Western languages as well: e.g. with perception verbs, it is the experiencer, which is often in object position, rather than the grammatical subject, that should be ranked higher.

Finally, centering provides an interesting framework for studying the functions of pronouns, as the observation that the Cb is often deleted or pronominalized can be stated as the following rule:

Rule 1 *If some element of Cf(U_{n-1}) is realized as a pronoun in U_n, then so is Cb(U_n).*

This rule has been computationally interpreted to individuate the Cb. If U_n has:

- a single pronoun, that is $\mathrm{Cb}(\mathrm{U}_n)$;
- zero or more than one pronoun, $\mathrm{Cb}(\mathrm{U}_n)$ is:
  - $\mathrm{Cb}(\mathrm{U}_{n-1})$ if $\mathrm{Cb}(\mathrm{U}_{n-1})$ is realized in U_n;
  - otherwise the highest ranked $\mathrm{Cf}(\mathrm{U}_{n-1})$ which is realized in U_n.

Let's apply centering to the constructed example in (2). In (2a) Cb = ? because the Cb of a segment initial utterance is left unspecified; in (2b) the Cb is *John*, as it is the only pronoun, and also the only entity belonging to the Cf list of (2a) realized in (2b).

(2a) **John** *is a nice guy.*
`Cb = ?   Cf = [John]`

(2b) **He** *met* **Mary** *yesterday.*
`Cb = John, Cf = [John > Mary]`

(2c) i. **He** *likes* **her**. (CONTINUE)
`Cb = John, Cf = [John > Mary]`
ii. **She** *likes* **him**. (RETAIN)
`Cb = John, Cf = [Mary > John]`
iii. **She** *was with* **Lucy**. (SMOOTH-SHIFT)
`Cb = Mary, Cf = [Mary > Lucy]`
iv. **Lucy** *was with* **her**. (ROUGH-SHIFT)
`Cb = Mary, Cf = [Lucy > Mary]`

In (2c).i we have a CONTINUE, as its Cb is *John* (the highest entity on the Cf list of (2b)), and so is its Cp. In (2c).ii, the Cb is still *John* as in (2c).i, but the Cp now is *Mary*, thus we have a RETAIN. In both (2c).iii and (2c).iv the Cb is *Mary* (the only entity belonging to the Cf list in (2b) that is realized): as *Mary* is also the Cp in (2c).iii, a SMOOTH-SHIFT occurs. Instead, as *Lucy* is the Cp in (2c).iv, a ROUGH-SHIFT occurs.

Centering theory has appealing traits from both cognitive and computational points of view. From a cognitive perspective, it explains certain phenomena of local discourse coherence (e.g. pronominal "garden paths"), and is supported by psycholinguistic experiments (Gordon et al., 1993). Computationally, it is a simple mechanism, and thus it has been the basis for simple algorithms for anaphora resolution (Brennan et al., 1987).

Much work still remains to be done on centering. For example, most development so far has been based on simple constructed examples: to apply centering to real text, issues such as how possessives and subordinate clauses affect referring expression resolution must be addressed. This paper is a contribution in that direction.

3 The Italian pronominal system

Italian has two pronominal systems (Calabrese, 1986): weak pronouns, that must always be cliticized to the verb (e.g. lo, le, gli - respectively him, accusative; them, feminine accusative or her, dative; him, dative), and strong pronouns (lui, lei,

[3] (Grosz et al., 1986; Grosz et al., 1995) propose that the ordering on transitions pertains to *sequences* of transitions rather than to single transitions.

loro - respectively he or him; she or her; they or them).[4] The null subject is considered part of the system of weak pronouns.

Weak and strong pronouns are often in complementary distribution, as strong pronouns have to be used in prepositional phrases, e.g. per lui, *for him*. However, this syntactic alternation doesn't apply in subject position. The choice of null versus strong pronoun depends on pragmatic factors; the centering explanation offered in (Di Eugenio, 1990) goes as follows:

(3a) Typically, a null subject signals a CONTINUE, and a strong pronoun a RETAIN or a SHIFT.

(3b) A null subject can be felicitously used in cases of RETAIN or SHIFT if in U_n the syntactic context up to and including the verbal form(s) carrying tense and / or agreement forces the null subject to refer to a particular referent and not to $\mathrm{Cb}(U_{n-1})$.

The evidence for (3b) provided in (Di Eugenio, 1990) derived, among others, from modal and control verb constructions, in which clitics may be cliticized to the infinitival complement of the higher verb or may climb in front of the higher verb. When the clitic climbs, certain pronominal "garden path" effects, deriving from a wrong interpretation initially assigned to the null subject and later retracted, are avoided.

4 Italian subjects in discourse

4.1 The corpus

The corpus amounts to about 25 pages of text, and 12,000 words; it is composed of excerpts from two books (von Arnim, 1989; Fallaci, 1989), a letter (Mila, 1993), a posting on the Italian bulletin board (SCI, 1994), a short story (Nichetti, 1993), and three articles from two newspapers (del Buono, 1993; Pagetti, 1993; La Nazione, 1994). The excerpts are of different lengths, with the excerpts from the two books being the longest.

Texts were chosen to cover a variety of contemporary written Italian prose, from formal (newspaper articles about politics and literature), to informal (posting on the Italian bulletin board), and according to the following criteria: a) minimal direct speech, which has not been addressed in centering yet; b) prose that describes situations involving several animate referents, because strong pronouns can refer only to animate referents.

Table 2 shows the distribution of animate third person subjects partitioned into: full NPs — the numbers in parentheses refer to possessive NPs; strong pronouns; null subjects — I counted only those whose antecedents are not determined by contraindexing constraints (Chomsky, 1981).; other anaphors (e.g. tutte, all_{fem}) — they won't be analyzed in this paper.

4.2 Issues

When applying centering to real text, one realizes that many issues have not been solved yet. I will comment here on how deictics, possessives, and subordinate clauses affect centering.

Deictics such as *I*, *you*, etc. The problem is whether they are part of the Cf list or not. I follow (Walker, 1993) in assuming that deictics are always available as part of global focus, and therefore are outside centering.

Possessives. Table 3 includes a category marked *possessive*, which refers to full NPs that include a possessive adjective referring to an animate entity, such as *i suoi sforzi* — *his efforts*.

The problem is how possessives affect Cb computation and Cf ordering. While Cb computation does not appear to be affected by a possessive, that behaves like a pronoun, the Cf ranking needs to be modified. An NP of type *possessive* refers to two entities, the possessor P_{or} and the possessed P_{ed}. P_{ed} corresponds to the full NP, and thus its position in Cf is determined by the NP's grammatical function; as regards P_{or}, my working heuristics is to rank it as immediately preceding P_{ed} if P_{ed} is inanimate, as immediately following P_{ed} if P_{ed} is animate. Such heuristics appears to work, but needs to be rigorously tested.

Subordinates. Another important issue, that has not been extensively addressed yet — but see (Kameyama, 1997; Suri and McCoy, 1993) — is how to deal with complex sentences that include coordinates and subordinates. The questions that arise concern whether there are independent Cb's and Cf lists for every clause; if not, how the Cb of the complex sentence is computed, and how semantic entities appearing in different clauses are ordered on the global Cf list.

In this paper, I will loosely adopt Kameyama's proposal (1997) that sentences containing conjuncts and tensed adjuncts are broken down into a linear sequence of centering "units", while tense-

[4] Lui, lei, loro are the oblique forms of the strong system, while the nominative forms are respectively egli, ella, essi/e: in current Italian the latter forms are rarely used as the oblique forms have replaced them in subject position — in my corpus there are only four occurrences of these nominative forms, and they all occur in the same article (Pagetti, 1993).

Text	Total	Full NPs		Strong	Zero	Other
(von Arnim, 1989)	111	45	(11)	23	36	7
(Fallaci, 1989)	17	6	(0)	2	9	0
(Mila, 1993)	8	1	(0)	2	4	1
(SCI, 1994)	18	7	(1)	0	7	4
(Nichetti, 1993)	40	26	(1)	1	13	0
(del Buono, 1993)	36	28	(6)	1	6	1
(Pagetti, 1993)	22	19	(6)	3	0	0
(La Nazione, 1994)	35	27	(4)	1	5	2
Total	287	159	(29)	33	80	15

Table 2: Animate 3rd person subjects

less adjuncts don't generate independent centering units[5].

4.3 Centering Transitions

Table 3 illustrates the distribution of referring expressions with respect to centering transitions. The number of full NPs in Table 3 is about half their number in Table 2: in fact, full NPs often introduce entities new to the discourse, in which case centering does not apply.

Table 3 includes two columns that don't refer to centering transitions. The column labeled CENT-EST encodes referring expressions that don't refer to a member of $\mathrm{Cf}(\mathrm{U}_{n-1})$, but to an entity available in the discourse. While such transitions do not belong to centering, that models how centers change from one centering unit to the next, they constitute referential usages of pronouns that need to be explained. I call these transitions CENT-ESTAB, for CENTER ESTABLISHMENT, because such references appear to establish the new center of local discourse. Finally, OTHER includes e.g. expressions that build a set out of $\mathrm{Cb}(\mathrm{U}_{n-1})$ and some other entity, such as *sia lui che sua moglie* — *both him and his wife*. It is not clear how to deal with these constructions within the centering framework, and thus, I have left them unanalyzed for the time being.

The results are as follows. Null subjects are, not surprisingly, the most frequently used expression — 58% — for CONTINUE's; the difference between null subjects and all the other referring expressions is also statistically significant ($\chi^2 = 7.128$, $p < 0.01$).[6] Vice versa, CONTINUE's account for 70% of null subjects. However, even full NPs can be used for CONTINUE's — such usages accounts for 16% of CONTINUE's, and for 20% of full NPs. Also, 12% of CONTINUE's are encoded by means of possessive NP's, and vice versa, 41% of possessive NP's are used for CONTINUE's.

The situation for RETAIN's and SHIFT's is not very clear, as none of the four categories of referring expressions is predominant. All these SHIFT's are actually SMOOTH-SHIFT's, i.e., there are no ROUGH-SHIFT's at all. This is not surprising for null subjects, that are never used for ROUGH-SHIFT (Turan, 1995), however it is puzzling for full NPs. Apparently the Italian writers I selected adhere to the default ranking of transitions, in which ROUGH-SHIFT's are the least preferred.

A significant difference in the usages of the four referring expressions regards CENT-EST. In this case, full NP's are used 59% of the times, and the difference between full NP's, and all the other expressions is significant ($\chi^2 = 8.88$, $p < 0.01$).

I will now focus on the contrast between zeros and strong pronouns, in order to assess the strategies proposed in (3). Initially, (3a) — zeros used for CONTINUE, strong pronouns for RETAIN and SHIFT — appeared not to be supported, not even as regards the preference for null subjects for CONTINUE: given the numbers in Table 3, the difference between zeros and strong pronouns used for CONTINUE is **not** significant ($\chi^2 = 2.436$, $p < 0.20$). This finding puzzled me, because the usage of null subjects for CONTINUE seems to be a robust cross-linguistic phenomenon: it occurs in languages as diverse as Japanese (Kameyama, 1985; Walker et al., 1994; Shima, 1995) and Turkish (Turan, 1995).

The puzzle can be solved by examining the transition preceding the CONTINUE in question. Table 4 shows the different possible transitions in U_n, that precedes U_{n+1} in which a CONTINUE occurs. The configuration in which a CONTINUE is preceded by a RETAIN, which I call RET-CONT, differs from the other two because of the constraint $\mathrm{Cp}(\mathrm{U}_n) \neq \mathrm{Cb}(\mathrm{U}_n)$ in the RETAIN. This in a sense predicts that the center will shift: but in a RET-

[5] The situation for complements is more complicated, and space prevents me from discussing it.

[6] χ^2 test results are reported here more as a source of suggestive evidence than as strong indicators, as the observations in the corpus, which come from only 8 authors, are not totally independent.

Type	Total	CONTINUE	RETAIN	SHIFT	CENT-EST	OTHER
zero	80	56	4	6	12	2
strong	33	13	3	5	11	1
NP	81	17	11	7	44	2
poss.	25	11	5	1	8	0
Total	219	97	23	19	75	5

Table 3: Distribution of centering transitions

	CONTINUE	RETAIN	SHIFT
U_n	$Cb_n = Cb_{n-1}$ $Cp_n = Cb_n$	$Cb_n = Cb_{n-1}$ $\mathbf{Cp_n \neq Cb_n}$	$Cb_n \neq Cb_{n-1}$ $Cp_n = Cb_n$
U_{n+1}	$Cb_{n+1} = Cb_n$ $Cp_{n+1} = Cb_{n+1}$		

Table 4: Transitions preceding a CONTINUE

Type	Total	CONT-CONT+ SHIFT-CONT	RET-CONT
zero	56	51	5
strong	13	7	6
Total	69	58	11

Table 5: Pronoun occurrences for RET-CONT

CONT such prediction is not fulfilled. As Table 5 shows, this has some consequences on the usage of null and strong pronouns. Compared to strong pronouns, null subjects are used 87% of the times for CONT-CONT and SHIFT-CONT taken together and only 45% of the times for RET-CONT, and the puzzle discussed above is explained. In fact, in the case of CONT-CONT and SHIFT-CONT, there is a significant difference between zeros and strong pronouns, $\chi^2 = 6.279$, $p < 0.02$. Instead, in the case of RET-CONT, there is no significant difference, $\chi^2 = 2.986$, $p < 0.10$.[7] Fig. 1 presents two examples of RET-CONT, one in (4c) realized with a strong pronoun, the second in (4e) realized with a null subject. In the utterance preceding (4a), Cb = Irais and Cf = [Irais].

(4a) Φ *Incominceró a ricondurre il* **suo pensiero** *sui* **suoi doveri** *chiedendo***le** *ogni giorno*
(I) will start to bring **her thoughts** back to **her duties** by asking **her** every day
`Cf:[Irais > I's thoughts, I's duties],`
`Cb:Irais, continue`

(4b) *come sta* **suo marito.**
how **her husband** is.
`Cf:[husband > Irais], Cb:Irais, retain`

(4c) *Non è che* **lei gli** *voglia granché bene,*
It's not the case that **she** cares much about **him**
`Cf:[Irais > husband], Cb:Irais, continue`

(4d) *perché* **lui** *non corre ad aprir***le** *la porta*
because **he** doesn't run to open the door **for her**
`Cf:[husband > Irais], Cb:Irais, retain`

(4e) *ogni volta che* Φ *si alza per lasciare la stanza;*
whenever **(she)** gets up to leave the room.
`Cf:[Irais], Cb:Irais, continue`

Figure 1: Examples of RET-CONT

As far as RETAIN's and SHIFT's go, the numbers are both too small to draw any conclusion, and they don't seem to identify any preferred usage for strong pronouns, contrary to what claimed by (3a); also in the case of CENT-EST there doesn't seem to be any significant difference in usage. A topic for future work is to verify whether there are any factors affecting the choice between null and strong pronouns in these cases, especially because null subjects used for SHIFT or for CENT-EST sometimes result in a slightly less coherent discourse.

The second part of the claim, (3b) — a null subject can be used if U_n provides syntactic clues that force the null subject not to refer to $Cb(U_{n-1})$ — is supported; however, given the small numbers (four RETAIN's and six SHIFT's) this conclusion can just be tentative. The most frequent clue is agreement in gender and / or number.

5 Conclusions

In this paper, I examined the referring functions that different types of subjects perform in Italian within the centering framework. I built on the analysis presented in (Di Eugenio, 1990), and extended it in several directions: first, I used a corpus of really occurring examples; second, I included phenomena such as possessives and subordinate clauses; third, I refined the notion of

[7] Also (Turan, 1995) independently noticed the existence of RET-CONT's, and reports results similar to mine.

CONTINUE by pointing out the peculiarity of RET-CONT's; fourth, I included full NPs; fifth, I illustrated a type of pronominal usage, CENT-EST, outside the purview of centering.

Future work includes further analysis of a somewhat surprising finding from the current study, i.e. that NP's encoding CONTINUE's are not so rare. It is worth while to examine the data further, to see under which conditions a full NP is licensed to encode a CONTINUE. I also want to collect more RET-CONT's, RETAIN's, and SMOOTH-SHIFT's to refine the analysis presented in this paper. Finally, another topic of research is CENT-EST, even if it is outside the centering framework, and under what conditions zeros are used to encode it.

References

Susan Brennan, Marilyn Walker Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *Proc. 25th Meeting, Association for Computational Linguistics*, pages 155–162.

Andrea Calabrese. 1986. PRONOMINA - Some properties of the Italian pronominal system. In N. Fukui, T. Rapoport, and E. Sagey, editors, *MIT Working papers in Linguistics. Papers in Theoretical Linguistics. Vol. 8.*

Noam Chomsky. 1981. *Lectures on Government and Binding.* Foris, Dordrecht.

Oreste del Buono. 1993. Avanti a tutto Metz contro la retorica fascista. *La Stampa, TuttoLibri*, December.

Barbara Di Eugenio. 1990. Centering Theory and the Italian Pronominal System. In *Proceedings 13th International Conference on Computational Linguistics (COLING 90)*, pages 270–275.

Oriana Fallaci. 1989. *Penelope alla guerra.* Bollati Boringhieri, Torino.

Peter Gordon, Barbara Grosz, and Laura Gilliom. 1993. Pronouns, Names, and the Centering of Attention in Discourse. *Cognitive Science*, 17:311–347.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1986. Towards a computational theory of discourse interpretation. Unpublished manuscript.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225.

M. A. K Halliday. 1976. *System and Function in Language.* London: Oxford University Press. Edited by G. R. Kress.

Megumi Kameyama. 1985. *Zero anaphora: the case of Japanese.* Ph.D. thesis, Stanford University.

Megumi Kameyama. 1997. Intrasentential Centering: A Case Study. To appear in *Centering in Discourse*, Ellen Prince, Aravind Joshi, Lyn Walker editors, Oxford University Press.

La Nazione. 1994. Il declino di Ross Perot: la nazione piu' capitalista del mondo sa di non essere un'azienda, March.

Vincenzo Lombardo and Leonardo Lesmo. 1994. A compact syntactic representation. In C. Martin-Vide, editor, *Current Issues in Mathematical Linguistics*, pages 191–200. Elsevier Science B.V.

Massimo Mila. 1993. Letter to his mother. *(Reprinted in) La Stampa, TuttoLibri*, December.

Maurizio Nichetti. 1993. La tv delle formiche. *Comix*, 84.

Elena Not and Massimo Zancanaro. 1995. The double nature of anaphora. a discussion with a flavour of systemic linguistics. In Wolfgang Höppner and Helmut Horacek, editors, *Principles of Natural Language Generation. Papers from a Dagstuhl-Seminar.* Technical Report Duisburg Universität, Bericht Nr. SI-12,2/95.

Carlo Pagetti. 1993. Dick l'illuminato. *La Stampa, TuttoLibri*, December.

Vieri Samek-Lodovici and Carlo Strapparava. 1990. Identifying Noun Phrase References: the Topic Module of the AlFresco System. In *Proceedings of ECAI 90, Ninth European Conference on Artificial Intelligence*, pages 573–578.

SCI. 1994. Posting on the soc.culture.italian electronic newsgroup, March.

Kaori Shima. 1995. Anaphora Resolution in Japanese: a Centering Approach. Master's project, Carnegie Mellon University, May.

Oliviero Stock, Giuseppe Carenini, Federico Cecconi, Enrico Franconi, Alberto Lavelli, Bernardo Magnini, Federico Pianesi, Marco Ponzi, Vieri Samek-Lodovici, and Carlo Strapparava. 1993. AlFresco: Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration. In Mark T. Maybury, editor, *Intelligent Multimedia Interfaces.* The MIT Press.

Linda Z. Suri and Kathleen F. McCoy. 1993. Focusing and Pronoun Resolution in Particular Kinds of Complex Sentences. In *Proceedings of Workshop on Centering, University of Pennsylvania.*

Ümit Deniz Turan. 1995. *Subject and Object in Turkish Discourse: a Centering Analysis.* Ph.D. thesis, University of Pennsylvania.

Elizabeth von Arnim. 1989. *Il giardino di Elizabeth.* Bollati Boringhieri, Torino.

Marilyn Walker, Masayo Iida, and Sharon Cote. 1994. Japanese Discourse and the Process of Centering. *Computational Linguistics*, 20(2):193–231.

Marilyn Walker. 1993. Initial Contexts and Shifting Centers. In *Proceedings of Workshop on Centering, University of Pennsylvania.*