Cataphoric *it* and backgrounding from the point of view of coherence relations*

Maite Taboada and Radoslava Trnavac

Simon Fraser University
Burnaby, BC – Canada

mtaboada@sfu.ca, rtrnavac@sfu.ca

1 Problem characterization

In (1a) the pronoun cannot be coreferential with the subsequent noun, but in (1b) it can.

(1) a. ***He** finished breakfast before John went to school. (from Carden & Dieterich 1980, cited in Harris & Bates 2002)
b. After **he** finished breakfast, John went to school (he = John)

Earlier accounts have suggested that pronouns cannot precede their referents when the pronoun is the subject of the clause (1a), but may when the pronoun appears in a syntactically subordinate clause (1b).

Many cases, such as (1) are part of the same sentence. Other cases, however, involve reference across sentences.

(2) And I just remember that movie was, had a lot of hype about being really scary and suspenseful and it's nothing like these movies that are out today like Scream and, what is **it**? Nightmare on Elm Street whatever it is, where they're actual horror movies. [OANC][1]

2 Research questions

We address two related questions. First of all, we would like to know why in some cases a pronoun is allowed to refer ahead to a subsequent noun, as in (1b). Secondly, we are interested in the discourse factors involved in inter-sentential cataphora.

3 Previous approaches

Cataphora is a well-studied phenomenon, and a lot of the 'classical' work has suggested that syntactic structure determines coreference patterns (Langacker 1969; Lasnik 1976; Reinhart 1981, 1983; Ross 1969; and others). According to this position, pronouns can generally only

* Extended abstract of a paper presented at the 86th Annual Meeting of the Linguistic Society of America. Portland, OR. January 2012.

[1] Corpus sources: OANC (Open American National Corpus), BN (Broadcast News), RST (RST Discourse Treebank). See Section 5.

refer to referents that are higher than they are in the phase structure diagram. This obeys the c-command principle, where a pronoun must have its antecedent c-commanding it (Reinhart 1983). However, Reinhart herself (1983: 42) points out that when two NPs or an NP and a pronoun are *not* in the domain of each other, then c-command does not apply. Whether they are coreferential or not depends on pragmatic, rather than syntactic (sentence-level) considerations. In Examples like (2), the reference is across sentences, where c-command principles do not apply.

Functionalist approaches to pronominal reference (e.g., Ariel 1990, Garnham 1987; Givón 1983; Gordon & Hendrick 1997; Prince 1981) state that the main function of pronouns is to refer to discourse entities that are highly accessible in working memory. Thus, in an example like (3), there is a conflict in accessibility status: *he* must be highly accessible (pronoun in subject position), whereas *John* must be a new concept (proper name). Therefore, *he* and *John* cannot refer to the same entity.

(3) **He** finished breakfast before John went to school.

Within this line of argument, Harris and Bates (2002) argue that the foregrounding-backgrounding distinction is a crucial factor in determining coreference interpretations. In (4) coreference is possible although c-command is violated, because the first clause is backgrounded.

(4) **He** had been staring at the control panel for over an hour when Jack received a message from his commander.

Backgrounding can be achieved through different means. One is the use of syntactic/pragmatic subordination (Bolinger 1979, Hopper 1979, Matthiessen & Thompson 1988). This results in structural constraints similar to the c-command principle when subordinate clauses are involved. Another way of backgrounding information is through aspect, through the use of imperfective verb forms (Hopper 1979), such as the progressive in Example (4).

Backgrounding is relatively easy to determine in complex clauses (main-subordinate structures), but we are interested in cataphora across sentences, so we needed a way to determine backgrounding at the discourse level, which we found in the use of coherence relations in Rhetorical Structure Theory (Mann and Thompson 1988).

The fundamental question, then, that we try to address, is whether all instances of cataphora can be accounted for through backgrounding, whether at the clause or the discourse level, and whether we can identify backgrounded clauses through RST.

4 Coherence relations

In brief, Rhetorical Structure Theory is a theory of coherence (Mann & Thompson 1998) that proposes a limited set of relations, such as Cause, Concession, Condition or Elaboration. In RST, clauses, but also entire sentences and paragraphs are linked as main and secondary parts (nucleus and satellite). The notion of satellite roughly corresponds to backgrounded material (Matthiessen & Thompson 1988).

5 Data

We focused on the pronoun *it* in spontaneous and non-spontaneous speech. In the first stage of our corpus analysis, we extracted cases which fit two of the patterns proposed by Carden (1982): (1) NP_1 … Pro_1 … NP_2; and (2) Pro_1 … NP_1, and analyzed the first mention instances of cataphora with noun and proposition reference. We found 4,668 examples of *it* in all four corpora, out of which 45 were cataphoric. We disregarded sentences with pleonastic *it*. The following table illustrates the distribution of the pronoun in the nucleus/satellite positions in four corpora.

Table 1. Distribution of *it* in nucleus and satellites[2]

	Open American National Corpus	**Broadcast News**	**RST Discourse Treebank**	**Total**
Nucleus	21	3	4	28
Satellite	15	2	-	17

As Table 1 shows, cataphoric *it* is found both in the nucleus and satellite of relations which confirms the idea of Matthiessen & Thompson (1988) that backward pronominalization is not a criterion for hypotaxis. Backgrounding seems to be an insufficient factor to account for all the cases of cataphoric *it* since 21 out of 28 cases of first mention cataphora were found in the nucleus position.

[2] Corpora used in this study:
- Open American National Corpus (http://americannationalcorpus.org/OANC/)
  - 3.2 m words of spoken language (face-to-face and telephone)
- English Broadcast News (Alabiso et al. 1998)
  - Radio and television news broadcasts; 200,000 words
- RST Discourse Treebank (Carlson et al. 2002)
  - Collection of Wall Street Journal articles from the Penn Treebank; 176,000 words

6 Analysis

We suggest that additional factors that make cataphora as a first mention possible in the nucleus of coherence relations are Distance and Unity between the pronoun and the antecedent (Accessibility Theory, Ariel 1990)[3]. The unity parameter for anaphoric marking refers to the degree of connectivity between various sentential components in which anaphora and antecedent are found (see Bolinger 1979 and Ariel 1990). Higher dependency between these components suggests the use of higher Accessibility markers (pronouns), while lower dependency requires the use of lower Accessibility markers (nouns).

In the second stage of our analysis, we decided to compare first mention instances of the pronoun with instances of repeated reidentification in terms of distribution based on Distance and Unity factors. For each instance of *it*, we annotated: syntactic subordination (adverbial clauses), pragmatic subordination (*but* clauses), coordinate construction, and distance between *it* and the antecedent (in number of clauses). The results of this analysis show the following:

1. In all 28 first mention cases of *it* in the nucleus, the distance between pronoun and antecedent was one clause. First mention entities are frequently complements of cognition and propositional attitude verbs, in fact implicit complement clauses.
2. In 400 instances with reidentification reference, we found 32 examples of continuous anaphora, 14 in the nucleus. In 8 of those, we find cases where unity is not high – 8 out of 14 were separated by more than one clause.
3. Most of the examples with *it* in the nucleus were Elaborations in which longer material is postponed, and first introduced with a pronoun (van Hoek 1997).

7 Conclusions

Backgrounding does not completely capture the constraints involved in cataphora. At the discourse level, cataphora occurs both in the nucleus and satellite (foregrounded and backgrounded parts). When cataphoric *it* occurs within the nucleus, it requires higher unity and shorter distance between the pronoun and the antecedent. Instances of *it* in the nucleus tend to have a longer, heavier antecedent that is therefore postponed. Instances of reidentification reference do not require high unity or low distance.

In our future work we will analyze more instances of *it* and other types of pronouns and referring expressions.

[3] We focused our analysis on two out of four parameters of Accessibility Theory (Distance, Unity, Competition, Salience).

References

Alabiso, Jennifer, Robert MacIntyre and David Graff (1998*). 1997 English Broadcast News Transcripts (HUB4)* [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Ariel, Mira (1990). *Accessing Noun Phrase Antecedents.* London/New York: Routledge.

Bolinger, Dwight (1979). Pronouns in discourse. In *Syntax and Semantics: Discourse and Syntax*, T. Givón (Ed.), 289-309. New York/London: Academic Press.

Carden, Guy (1982). Backwards anaphora in discourse context. *Journal of Linguistics, 18* (2), 361-387.

Carden, Guy and Thomas Dieterich (1980). Introspection, observation and experiment: An example where experiment pays off. *Journal of the Philosophy of Science Association*, 2, 583-597.

Carlson, Lynn, Daniel Marcu and Mary Ellen Okurowski (2002*). RST Discourse Treebank, LDC2002T07* [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Givón, Talmy (1983). *Topic Continuity in Discourse: A Quantitative Cross-Linguistic Study.*Amsterdam: John Benjamins.

Garnham, Alan (1987). Understanding anaphora. In A. W. Ellis (Ed.). *Progress in the Psychology of Language, Vol. 3.* Hillsdale, NJ: Erlbaum.

Halliday, Michael A.K. and Ruqaiya Hasan (1976) *Cohesion in English*. London: Longman.

Harris, L. Catherine, and Elizabeth A. Bates (2002). Clausal backgrounding and pronominal reference: A functionalist approach to c-command. *Language and Cognitive Processes* 17(3):237-269.

van Hoek, Karen (1997). *Anaphora and Conceptual Structure*. Chicago: University of Chicago Press.

Hopper, Paul (1979). Aspect and foregrounding in discourse. In T. Givón (Ed.), *Syntax and Semantics: Discourse and Syntax* (pp. 213–241). New York: Academic Press.

Langacker, Ronald (1969). On pronominalization and the chain of command. In W. Reibel and S. Schane (Eds.), *Modern Studies in English.* Englewood Cliffs, NJ: Prentice Hall.

Lasnik, Howard (1976). Remarks on coreference. *Linguistic Analysis 2,* 1-22.

Mann, William, and Thompson, Sandra (1988). Rhetorical structure theory: Towards a functional theory of text organization. *Text* 8 (3): 243-281.

Matthiessen, Christian and Thompson, Sandra (1988). The structure of discourse and “subordination”. In *Clause Combining in Discourse and Grammar*, John Haiman and Sandra Thompson (Eds), 275-329. Amsterdam: Benjamins.

Prince, Ellen (1981). Toward a taxonomy of given-new information. In P. Cole, (Ed.), *Radical Pragmatics.* New York: Academic Press.

Reinhart, Tanya (1981). Definite NP anaphora and c-command domains. *Linguistic Inquiry, 12,* 4, 605-635.

Reinhart, Tanya (1983). *Anaphora and Semantic Interpretation*. London: Croom Helm.

Reppen, Randi, Nancy Ide and Keith Suderman (2005). *American National Corpus (ANC) Second Release, LDC2005T35* [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Ross, John R. (1969). On the cyclic nature of English pronominalization. In W. Reibel and S. Schane (Eds.), *Modern studies in English.* Englewood Cliffs, NJ: Prentice Hall.