# Cataphora, backgrounding and accessibility in discourse

Radoslava Trnavac [a,*], Maite Taboada [b]

[a] *Simon Fraser University, Department of Linguistics/Cognitive Science Program, Robert C. Brown (RCB) Building 8206, 8888 University Dr., Burnaby, B.C. V5A 1S6, Canada*
[b] *Simon Fraser University, Department of Linguistics, Robert C. Brown (RCB) Building 8206, 8888 University Dr., Burnaby, B.C. V5A 1S6, Canada*

## Abstract

We examine discourse factors that are involved in the occurrence of intra/inter-sentential cataphora. On the basis of a corpus analysis of natural language, we test two cognitive theories that attempt to explain the phenomenon of cataphora: clausal backgrounding (Harris and Bates, 2002) and Accessibility Theory (Ariel, 1990). In the first part of the paper, we investigate the presence of cataphoric *he*, *she, it* and *they* in a corpus annotated with nucleus-satellite labels (Mann and Thompson, 1988), which are interpreted to be discourse correlates of the foreground-background distinction. The analysis shows that cataphora cannot be restricted to backgrounded parts of texts, and that backgrounding as an explanation for the occurrence of cataphora cannot be applied at the discourse level. In the second part of the paper, we investigate a cognitively related phenomenon to backgrounding: accessibility and its influence on cataphora. We demonstrate that in different conditions accessibility parameters such as Givenness, Distance and Unity do not show a clear influence (especially when an instance of cataphora and its antecedent are in different sentences), while Saliency and (non)Competition do play a role in the presence of cataphora (both intra- and inter-sententially).
© 2015 Elsevier B.V. All rights reserved.

## 1. Problem characterization

The general function of pronouns is to facilitate rapid access to the current discourse topic (Ariel, 1990; Givón, 1983; Harris and Bates, 2002). In cases of backward anaphora or cataphora, however, a pronoun is used before the referent has been introduced. Cataphora is a relatively rare phenomenon, and it has specific restrictions. Typically, cataphora occurs in a syntactically subordinate clause.[1] For instance, in Example (1a) the pronoun cannot be coreferential with the subsequent noun, but in (1b) it can, because in (1b), the cataphor occurs in a subordinate adjunct clause.

(1)   a.   *__He__ ate the cake when __the Smurf__ was in the box. (From Crain (1991), cited in Harris and Bates, 2002.)
      b.   When __he__ was in the box, __the Smurf__ ate the cake.

---

* Corresponding author. Tel.: +1 778 986 1822.
  *E-mail address:* rtrnavac@sfu.ca (R. Trnavac).
[1] The term 'cataphora' in this paper is used for the pronouns *he, she, it, they*, when they linearly precede their antecedents (including false starts, repairs and right dislocations). We do not discuss here the more general phenomenon of cataphora, as described, for instance, through cataphoric devices in Gernsbacher and Jescheniak (1995).

Most of the earlier accounts have suggested that pronouns cannot precede their referents when they are the subject of the main clause (the (a) sentence in the example above), but may when the pronoun appears in a syntactically subordinate clause (the (b) sentence). On the other side, Carden (1982)[2] claims that syntactic restrictions are only symptoms of pragmatic restrictions and explores the view that an NP can appear 'to the right' of a coreferential pronoun if it answers some need for more semantic information, for instance, to re-identify a prior referent that is distant, to avoid ambiguity, to emphasize the nature of the referent, or to reintroduce the referent as topic.

Carden (1982) describes the phenomenon of backwards anaphora, or cataphora (the term that we will use throughout the paper) in the following way[3]:

$NP_1 \ldots Pro_1 \ldots NP_2$
$Pro_1 \ldots NP_1$

According to Carden (1982), in the first case, we need to test whether $NP_1$ or $NP_2$ is the antecedent for the pronoun. While in some instances the sequence Pro . . . NP is a true cataphoric relation, in others it is a re-identification of a previously mentioned referent (Bolinger, 1979), which may be considered cataphora, and which we later on will refer to as repeated mention cataphora.

In the second case proposed by Carden, the sequence represents the first appearance of the referent in the discourse. This type of cataphora is common in journalistic discourse and in fiction (van Hoek, 1997), as we see in Example (2).[4,5]

(2)　　Once **<u>it</u>** is finished, **this new high school in Riverside, California**, will serve a mostly white, mostly middle class community. [BN]

Many cases of cataphora, such as in (1), are part of the same sentence. Other cases, however, involve reference across sentences, as can be seen in (3).

(3)　　You know **<u>she</u>** wasn't going to claim me. **My sister** had uh, an apartment in NJ and I remember there was, she had a roommate, young girl which hindsight is 20/20 she used to say she was going to kidnap in the middle of the night, put me in bed with her. [OANC]

A possible explanation for inter-sentential cataphora, as in (3), is that the pronoun appears in a pragmatically subordinate structure, that is, a clause or sentence that is in a pragmatic relation to the clause or sentence where the antecedent appears, and where the relation is one of pragmatic subordination or backgrounding (Harris and Bates, 2002). In this paper, we attempt to establish that parallelism between intra- and inter-sentential cataphora by examining examples drawn from corpora. In addition, we apply Ariel's (1990) Accessibility Theory, and characterize instances of cataphora in terms of the discourse factors that Ariel suggests are involved in anaphoric relations. We also propose that accessibility and backgrounding are cognitively related phenomena. The goal of this paper is to answer the question posed by the examples presented so far: What discourse features trigger the occurrence of intra/inter-sentential cataphora? The paper also presents a partial refutation of the above mentioned two linguistic theories that explain licensing of cataphora—backgrounding (Harris and Bates, 2002) and Accessibility Theory (Ariel, 1990). Our approach is original in that the two theories are applied for the first time, and in a corpus-based study.

The paper proceeds as follows: In Section 2, we briefly present some well-established syntactic and discourse theories that attempt to explain the phenomenon of cataphora. In Section 3, we outline the discourse theory of coherence relations of Mann and Thompson (1988), which we use to characterize backgrounding in discourse. Section 4 describes the corpora that we used in our analysis and describes the analysis based on the theory of clausal backgrounding (Harris and Bates, 2002). In Section 5 we present the distribution of cataphora with regard to backgrounding. In Section 6 we connect the phenomenon of backgrounding and accessibility and describe the four parameters of Accessibility Theory as they are presented in Ariel (1990, 2001). Section 7 demonstrates the analysis of our corpus according to the parameters of Accessibility Theory. Finally, Section 8 presents a discussion of the analysis results, and Section 9, conclusions.

---

[2] See more on the syntactic versus pragmatic approach to cataphora in Schlenker (2005).

[3] Carden (1982) also discusses cases with a quantifier: $Pro_1 \cdots Q\ NP_1$. We are not concerned with such cases here.

[4] Corpus sources are provided in square brackets after the examples: OANC (Open American National Corpus), BN (Broadcast News), RST (RST Discourse Treebank). See Section 4.

[5] In the examples, the cataphoric pronoun is indicated with bold and underlining. The referent is set in bold.

## 2. Approaches to cataphora

In this section we present a very brief overview of the literature on cataphora that is relevant for our analysis.

Generative syntax suggests that syntactic structure determines the coreference patterns in (1) above (Langacker, 1969; Lasnik, 1976; Reinhart, 1981, 1983; Ross, 1969; and others). According to this position, pronouns can generally refer only to referents that are higher up in the phrase structure diagram.[6] Principle C of Chomsky's Binding Theory (Chomsky, 1981) specifies that a definite description or a proper name cannot appear in the scope of (i.e., it cannot be c-commanded by) a coreferring expression. From the perspective of the Binding Principles linear order matters only to the extent that it changes the structural relation between the elements. Reinhart (1983: 42) argues that when two NPs or an NP and a pronoun are *not* in the domain of each other, c-command does not apply. Whether they are coreferential or not depends on pragmatic, rather than syntactic (sentence-level) considerations, but such considerations are not explored further.

According to functionalist approaches to pronominal reference, the main function of pronouns is to refer to discourse entities that are highly accessible in working memory (Ariel, 1990; Garnham, 1987; Givón, 1983; Gordon and Hendrick, 1997; Gundel et al., 1993; Prince, 1981; van Hoek, 1997). Consider the example in (4):

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

This example presents a conflict in accessibility status: *he* must be highly accessible (pronoun in subject position); *John* must be a new concept (proper name). Therefore, *he* and *John* cannot refer to the same entity. Harris and Bates (2002) point out that foregrounding/backgrounding is a crucial factor in determining coreference interpretations. Backgrounding refers to a dependency relation between two structures (clauses or sentences). The most frequent backgrounding devices are syntactic subordination (Bolinger, 1979; Hopper, 1979; Matthiessen and Thompson, 1988), pragmatic subordination, and imperfective aspect (Hopper, 1979). Syntactic subordination reflects the same structural constraints captured through c-command for subordinate clauses. In pragmatic subordination, the pronoun occurs in the pragmatically non-dominant clause, while the antecedent is part of the dominant clause (see McCray, 1980; see also Cristofaro, 2005, on differences between syntactic and pragmatic subordination), as in (5), where the second clause can be considered to be dominant because it represents the informational focus for the speaker.

(5)     **He** lied to me, and **John** was my friend.

One of the main goals of this paper is to answer the question whether all instances of cataphora can be accounted for through backgrounding in discourse. Backgrounding (Hopper, 1979) is relatively straightforward to determine in complex sentences (main-subordinate structures), but the question is how to define it in discourse. We propose to determine backgrounding at the discourse level by mapping foregrounding and backgrounding to the nucleus-satellite structures of a text, as defined within Rhetorical Structure Theory (Mann and Thompson, 1988). In the following two sections, we will address the issue of how cataphora and backgrounding interact, starting with a definition of nucleus and satellite in terms of coherence relations.

## 3. Coherence relations

In this study, in order to test the interaction between cataphora and backgrounding, we make use of constructs defined within Rhetorical Structure Theory (RST), which distinguishes between nuclear and satellite components in discourse relations (Mann and Thompson, 1988). The information presented in the nucleus is what the author or speaker considers to be the most important information, whereas satellites contain additional, secondary or supplementary information. Clauses, but also entire sentences and paragraphs of a text, may be linked as nuclei and satellites. The linkage between them produces the effect of coherence in discourse. The deletion of nuclear information will make the discourse less coherent, the remaining parts becoming more difficult to comprehend. The deletion of satellite information will make the discourse less explicit, but the remaining information should still be coherent.

The basic units of analysis in RST are clauses, sentences or phrases, i.e., syntactic units. The nucleus-satellite distinction at the discourse level is grammaticalized in many languages into the main-subordinate distinction at the sentence level. According to Mann and Thompson (1988) and Matthiessen and Thompson (1988), we can postulate a direct mapping between subordinate clauses and satellites of coherence relations. If we take this hypothesis as a starting

---

[6] For contrasts in cataphora among different clause types in cross-linguistic research see Reuland and Avrutin (2010).

Background

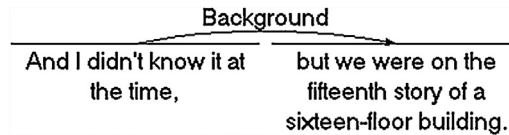| And I didn't know it at the time, | but we were on the fifteenth story of a sixteen-floor building. |

Fig. 1. RST analysis for Example (6).

point, then satellites roughly correspond to backgrounded material in discourse. For example, in (6), the clause containing the pronoun *it* acts as background for what is to follow. The first clause is a satellite to the clause that contains the antecedent (the fact that we were on the fifteenth story). We enclose each unit in square brackets, and use the subscripts S and N for satellite and nucleus, respectively. Fig. 1 is a schematic representation, where the arrow indicates subordination, from satellite to nucleus. Note that the status of a unit as nucleus or satellite does not necessarily correspond to its syntactic status. A syntactically subordinate clause may be presented as more important in the discourse. Additionally, nuclei and satellites may be composed of multiple units, which in turn contain additional relations, in a hierarchical recursive structure.

(6)      [And I didn't know **it** at the time,]$_S$ [**but we were on the fifteenth story of a sixteen-floor building**.]$_N$ [OANC]

Space precludes a full explanation of discourse units in RST, and the analysis process, but more detail is available in the original publication (Mann and Thompson, 1988), in a more recent review (Taboada and Mann, 2006a, 2006b), and on the RST web site (Mann and Taboada, 2015).

In the following section, we present our corpus analysis, whereby we analyzed instances of cataphora to determine whether the pronoun occurs in the backgrounded part of a coherence relation (either a clause-level or a sentence-level satellite).

## 4. Data analysis: corpus and methodology

Our corpus includes transcripts of spontaneous and non-spontaneous speech, in addition to written material. Carden (1982) points out that cataphora is less frequent in speech than in writing, but it is also the case that it has been less frequently studied in speech. This is why we decided to examine instances of cataphora in spontaneous and non-spontaneous speech, as well as in a written corpus, for comparison. The majority of our data comes from the Open American National Corpus (http://americannationalcorpus.org/OANC/), which has 3.2 million words of spoken language (face-to-face and telephone conversations). The second corpus that we used is the English Broadcast News (Alabiso et al., 1998), with English radio and television news broadcasts (200,000 words). The written corpus that we selected for comparison, and because it is already annotated for RST relations, is the RST Discourse Treebank (Carlson et al., 2002), which contains a collection of Wall Street Journal articles from the Penn Treebank (176,000 words).

In each of the corpora, we searched for instances of third person pronouns (*he, she, it, they*) and their morphological variants (e.g., *him, his* for *he*) using WordSmith (Scott, 2012), and extracted all instances, as shown in Table 1.

Not all of the instances in Table 1 are of course cataphoric. To determine whether a pronoun is cataphoric or not, a significant amount of the context needs to be considered, sometimes the entire text or conversation. Given that such analysis needs to be performed manually, we restricted the number of examples we could examine. In Table 2, we show the number of instances that we examined in the three corpora, to check whether they were cataphoric or not. These correspond, for OANC and BN, to the first *x* instances (e.g., the first 1000 instances of *he* in the OANC). The RST corpus contained much fewer pronouns, and thus we analyzed all of them. The two authors distributed the data and labelled

Table 1
Total instances of *he, she, it, they* in the corpora.

|  | Open American National Corpus (OANC) | Broadcast News (BN) | RST Discourse Treebank (RST) | Total |
| --- | --- | --- | --- | --- |
| Number of words | 3.2 m | 200,000 | 176,000 | ~3.6 m |
| Total instances of *he* | 19,676 | 6553 | 1124 | 27,253 |
| Total instances of *she* | 12,397 | 2350 | 362 | 15,109 |
| Total instances of *it* | 62,179 | 1372 | 924 | 64,475 |
| Total instances of *they* | 59,089 | 8209 | 982 | 68,280 |

Table 2
Instances of *he, she, it, they* analyzed.

|  | OANC | BN | RST | Total |
|---|---|---|---|---|
| Instances of *he* analyzed | 1000 | 500 | 1124 | 2624 |
| Instances of *she* analyzed | 1000 | 500 | 362 | 1862 |
| Instances of *it* analyzed | 2979 | 765 | 924 | 4668 |
| Instances of *they* analyzed | 1000 | 500 | 982 | 2482 |
| Total pronouns analyzed | **5979** | **2265** | **3392** | **11,636** |
| Total instances of first mention cataphora | **47** | **8** | **1** | **57** |

pronouns independently, but we consulted with each other in cases where one of us was not sure as to the cataphoric status of the pronoun in question. In the extraction of examples we disregarded examples with non-anaphoric instances of *it* (pleonastic *it*).[7] Particularly difficult were cases of false starts (which we included) and formulaic expressions (*believe it or not; it just dawned on me*). The latter can be considered constructions (Goldberg, 1995), but we tended to include them, because the pronouns that they contain seem to have referential status.

An important point with regard to methodology is the fact that, for the spoken data, we are dealing with transcripts of recorded speech. We did not check the transcriptions against the original recordings. In particular for clause and sentence boundaries, we rely on the boundaries that seem to be determined by the punctuation in the transcripts. This is important when determining whether the cataphora occurs within the same clause/sentence or across clauses and sentences, but in all cases, we used the transcripts as the source.

Altogether, we examined 11,636 pronouns, of which only 57 could be clearly determined to be cataphoric. Cataphora is indeed a rare phenomenon in naturally-occurring discourse, with few of the 'classic' cases that linguists have argued over for decades (e.g., Example 1 in this paper). Halliday and Matthiessen (2014: 625) do point out that inter-clausal cataphora is rare, with the exception of *structural cataphora*, where the referent appears immediately after the pronoun, as is the case with some relative clauses (*Good things come to those who wait*, where the referent for *those* is provided in the relative clause *who wait*[8]).

Contrary to what was suggested by Carden (1982), our results show that the written RST corpus has significantly fewer instances of cataphora than the spoken language corpora. This may be an artefact of the specific genre within spoken and written language that each corpus represents, but it is certainly interesting that cataphora does not seem frequent at all in our written data.

Examples (7) and (8) show the most frequent patterns of cataphoric reference, which are cases of noun and proposition reference, respectively. In some cases, such as (7), right dislocation also seems to be at play (Ziv, 1994).

(7)   When did they close that place? Nineteen, I forget. Nineteen sixty something. Yeah, um-- I mean a whole, is it an island? Absolutely. And the whole thing's a prison? Yeah. Absolutely! Well, **they** had warden's houses, **all the officers, all the guards**; they all lived there. [OANC]

(8)   And I did not know **it** at the time, but **the hotel was built on six-foot rollers**. [OANC]

We should point out that many examples in our corpus involve the pronoun *it*. Out of 57 instances of first mention cataphora, only 5 are the pronouns *he* or *his*, 1 the pronoun *she* and 7 *they*. All the other instances of first mention cataphora are expressed with the pronoun *it*. This is partly because, in the OANC, we examined more instances of *it* than of the other pronouns. But we also believe that *it* is frequent because it captures both noun and proposition reference, and enters into semi-fixed expressions that make use of cataphora to introduce long referents (such as *you won't believe it, but. . .*). In contrast to the pattern with *it* that is frequently discussed in the linguistic and syntactic literature on cataphora, such as in (1a) and (1b) from Crain (1991), the one that is particularly prominent in our corpus is represented by example (3) where the pronoun and the referent are in separate sentences.

In the next section, we describe the distribution of our data in terms of the correlation between cataphora and backgrounding in discourse.

---

[7] Following Quirk et al. (1985), Lappin and Leass (1994) and Mitkov (2002), instances of pleonastic *it* included the following cases: *it* appearing with modal adjectives (*it is dangerous*), with cognitive verbs (*it is believed that*), weather predicates (*it is sunny*) or in cleft constructions (*It was him who made an offer*). For this reason, even if that kind of *it* was part of Elaboration relation, it was not included in our data. In the linguistic literature, there is no clear-cut consensus regarding non-anaphoric uses of *it* (see Mitkov, 2002; Morgan, 1968). For instance, Lee-Goldman (2011, 2012) claims that syntactic, semantic and information-structural characteristics of weather predicates, extrapositions and clefts indicate that such instances of *it* can be considered referential. For the role of *it* in extrapositions, see also Couper-Kuhlen and Thompson (2008).

[8] Example from Halliday and Matthiessen; not from our data.

## 5. Data analysis: distribution of cataphora with regard to backgrounding

Harris and Bates (2002) proposed that backgrounding is the factor which permits a pronoun to deviate from its normal role of referring backwards to an established discourse entity. Cataphora is possible in backgrounded clauses because listeners/readers are aware of foregrounding/backgrounding structures and adjust accordingly, allowing for a pronoun to appear before its antecedent if the pronoun is in a backgrounded structure.

Harris and Bates (2002) discuss backgrounding at the clausal level, examining mainly clauses with imperfective aspect, which Hopper (1979) proposed are backgrounded. This is a generalization of the c-command constraint for subordinated clauses. Harris and Bates, however, do not apply their analysis beyond the confines of the sentence, and thus do not account for cataphora occurring across clauses. We propose to extend their analysis by drawing a parallel between satellites in Rhetorical Structure Theory analyses of text and backgrounded material in discourse. Matthiessen and Thompson (1988) proposed that main-subordinate structures at the sentence level are the result of grammaticalization of nucleus-satellite structures at the discourse level. Veins Theory (Cristea et al., 1998) also postulates that referents in general are to be found in nuclei.

In this section, we test the hypothesis that there is a correlation between first mention cataphora and backgrounding at the discourse level by analyzing satellites as backgrounded structures. If backgrounding is the factor which determines the distribution of cataphora in discourse, then all instances of cataphora should be restricted to the satellite position (the backgrounded element in discourse).[9]

In order to determine nucleus-satellite structures in our corpus, we used RST analyses of the clauses or sentences containing cataphora, and their surrounding context. For the RST corpus, we simply used the analyses provided by the annotators who created the corpus. Full details of the corpus analysis are provided in Marcu (1999) and of the corpus in Carlson et al. (2002). For the other two corpora, the authors performed RST analyses of a few sentences surrounding the pronoun, and determined whether the pronoun was in a nucleus or a satellite.

As mentioned above, we found, among 11,636 pronouns, a total of 57 instances of first mention cataphora (where the pronoun is the first instance of the entity in the text). Out of those 57 instances, 35 cases are found within the nucleus, and 21 instances within the satellite, thus with a preference for the majority of our pronouns appearing in the nucleus of an RST relation. Our initial hypothesis, that cataphora would occur in backgrounded units, is not supported by the data at all.

The fact that cataphora is found in both positions (nuclei and satellites) confirms the proposal made by Matthiessen and Thompson (1988) that backward pronominalization cannot be a criterion for hypotaxis or subordination. Cataphora can occur in both main and subordinate discourse structures and, at the sentence level, it can appear in both main and subordinate clauses. In addition, we observe that cataphora does not seem to be restricted to any specific type of discourse relation. For instance, the cataphora in Example (9) occurs in the nucleus of an Elaboration relation.

(9)  [I think that racism is uh inherent in the American uh life and system]ₙ [and to the point where **they** don't believe that they are racist]ₙ, [and they being **the American people**.]ₛ [OANC]

A relatively frequent pattern that we found is cataphoric *it* with a proposition reference in Elaboration discourse relations.[10] In those instances, longer material is usually postponed (in a separate sentence), and first introduced with a pronoun (van Hoek, 1997). Example (10) illustrates this use of *it* in the nucleus part of an Elaboration relation.

(10)  And I was just fascinated with the big city: San Francisco life, so many cars and, and the day that we were going the shuttle bus was taking us back to um, the airport: LAX. We were going to L.A. [Well, we got on the shuttle bus and **it** just dawned on me, you know.]ₙ [And I said, and I told him, ''**You know, I've been here three days and as many police cars as I've seen**,'' I said, you know, ''**I've not seen a single accident**.'']ₛ And he was like, ''Oh, don't say that, don't say that... [OANC]

Based on our data, backgrounding seems to be insufficient to account for all cases of a cataphoric pronoun in discourse. Thus, in order to explain the phenomenon in a more complete manner, additional factors must be included. In the next section, we discuss accessibility as a phenomenon that is related to backgrounding and that has an influence on the occurrence of cataphora.

---

[9] In this part of the paper we focus only on first mention cataphora and nucleus-satellite structures. The parameters of intra/intersentential cataphora and first mention versus continuous cataphora become relevant as we test predictions of Accessibility Theory (Ariel, 1990).

[10] Out of 57 instances of cataphora, we found 25 instances of cataphoric *it* with a proposition reference in Elaboration relations.

## 6. Backgrounding and accessibility

The phenomenon of clausal backgrounding can be related to accessibility and experimental work made within the Structure Building Framework (Gernsbacher, 1990; Gernsbacher et al., 1989). According to Gernsbacher et al. (1989), dependent (backgrounded) clauses with forward referring pronouns are highly accessible structures, since they serve as a foundation for representation of the whole sentence. They also argue that there is an implicit correlation between backgrounding and the accessibility of the dependent substructures. In the rest of the paper, we test claims which suggest that the process of referring is also largely dependent on accessibility of the referent itself (Ariel, 1990; Gernsbacher and Jescheniak, 1995; Givón, 1983; Gundel et al., 1993).

Ariel (1990) proposes that there are two types of cataphora, each allowed under different circumstances. Dependency,[11] which usually implies the presence of a backgrounded element, is only crucial when the antecedent is a new entity, what she terms first mention cataphora. When the referent is repeated (an already given entity), dependency is not needed at all. Ariel makes these hypotheses regarding only intra-sentential cataphora. Our data includes examples of both intra- and inter-sentential cataphora. Based on Ariel's proposal, we hypothesize that the occurrence of cataphora in the nucleus part of a relation is explained by the two parameters of Accessibility Theory: (i) Unity (low cohesion) and (ii) Givenness: first mention vs. continuation of discourse referent. In order to test this hypothesis, we need to empirically show that cataphora within the nucleus tends to refer to a continuing discourse referent and that the linkage[12] between the discourse unit where it is positioned and the discourse unit of its antecedent is looser[13] than in first mention cataphora, which is supposed to appear only in the satellite.[14]

To test the above, we apply the parameters of Accessibility both to the foregrounded and backgrounded parts of discourse (nuclei and satellites in coherence relations) with a special focus on the difference in behaviour between first and repeated mention cataphora.

### 6.1. Accessibility Theory and cataphora

Within research on anaphora there is a general consensus that the more reduced an anaphoric expression is, the more salient/accessible its antecedent has to be. Accessibility has been linked to the referent's discourse status, namely the role an entity has played in the preceding discourse (Arnold, 2010). Ariel (1990) suggests that the referential form is a marker for the discourse status of the referent, which helps listeners identify the location of the referent in their mental representation. A hierarchy of explicitness of referential expressions (Almor and Nair, 2007; Ariel, 1990; Arnold, 1998, 2010; Givón, 1983; Gundel et al., 1993) ranges from semantically rich expressions, such as full NPs, which signal less salient/accessible referents, to demonstratives, pronouns or zero pronouns, which reflect referents that are more salient/accessible. Ariel (1990) proposes that multiple factors can influence Accessibility, such as Old/New information (Givenness),[15] Distance, Competition, Saliency, and Unity. Similar to that, although under various nomenclatures, what is known as Preferred Argument Structure (Du Bois, 2003; and related work by Givón, 2001; Chafe, 1994) has shown that subjects, pronoun distribution and old/new information strongly correlate. This leads one to predict that saliency, topics, subjects, old information, and pronominal cataphors are not independent factors.[16] Indeed, as the results of this study demonstrate, there is a correlation between saliency and first mention cataphora, and similar results with repeated mention cataphora.

In the next few paragraphs we describe each of the factors that influence Accessibility as proposed by Ariel (1990), after which we test them with respect to cataphora.

**Old/New information (Givenness)**: Speakers usually reserve pronouns for referents that have already been evoked in the discourse. By contrast, ''new'' entities must be introduced with descriptions or names (Prince, 1981). However, with first mention cataphora, new information can be introduced through a pronominal form.

---

[11] The notion of *dependency* is explained in the subsection on the Accessibility parameter Unity in 6.1.

[12] This is a link between the main and the subordinate parts of sentences (or larger units of discourse) which is realized through discourse markers (signals of particular discourse relations).

[13] The looser link is between two independent structures of discourse (two main clauses) in multinuclear relations such as Contrast, List or Sequence, while the tighter link is between main and subordinate structures (nuclei and satellites).

[14] For Ariel (1990), while the distinction between nucleus and satellite would be relevant for first mention cases as related to the Unity factor, it would not be relevant for repeated mention anaphora cases. We apply our analysis regarding the distinction between nucleus and satellite structures to both first and repeated mention instances of cataphora.

[15] Ariel (1990) proposes that the concept of Givenness, which is a development of the Praguian notion of Old, should be replaced with the notion of Accessibility. However, in this paper we use the term Givenness in its original meaning.

[16] We thank an anonymous reviewer for this comment.

**Distance**: This factor refers to the distance between the antecedent and the anaphor. Pronouns in general favour a position where the antecedent occurs in a previous sentence (Ariel, 1990). Additionally, co-reference in intra-sentential position almost always requires the use of pronouns. In the case of cataphora, we would argue that distance plays a similar role, with the referent having to be close to the pronoun.

**Competition**: Speakers must distinguish the intended referent from potential competitors. As reported by Arnold and Griffin (2007) and Arnold (2010), speakers tend to use pronouns more than explicit forms if there is only one gender-matched character in the context.

**Saliency**: The notion of saliency in Ariel (1990) relates to the linguistic category of topic. Topicality has been hypothesized to increase the likelihood of using pronouns (e.g., Ariel, 1990; Broadbent, 1973; Gernsbacher, 1990). The accessibility of sentential topics correlates with evidence that grammatical subjecthood confers accessibility, as subjects are considered a topical position (cf. Chafe, 1976). Topics can be identified at the sentential (Reinhart, 1981) or at the discourse level for a given discourse segment (e.g., Grosz and Sidner, 1986; Kehler, 2004). One identifying aspect of topichood is that (sentence) topics can be defined in terms of the ''aboutness'' condition (Reinhart, 1981). The topic of an utterance is what the assertion expressed by that utterance is about, given a particular situation. The notion of discourse topic can be defined in an analogous fashion to that of sentence topic, only at the level of sets of propositions rather than a single proposition. Following Reinhart (1981), who proposed that cataphora is restricted to sentence topics, we analyzed our data in terms of sentence topics. There are several practical diagnostics to identify sentence topics. Topics often coincide with grammatical subjects in English, especially when the subject is in canonical position (Cowles et al., 2007; Givón, 1983; Gundel et al., 1993; Lambrecht, 1994; Reinhart, 1981). They must be referring expressions, and cannot be expletive subjects (Cowles et al., 2007; Lambrecht, 1994). In addition, following Taboada and Hadic Zabala (2008), we used four criteria: (i) we determined what question the sentence was answering (Gundel, 1977); (ii) we used the *as for* test (Gundel, 1977); (iii) the *say about X that S* test (Reinhart, 1981); and (iv) the pseudo-cleft test (Cohan, 2000).

**Unity**: This notion refers to the degree of connectivity between the segment where the antecedent is positioned and the segment containing the cataphor. When sentence level anaphora is involved, the question of Unity is related to the degree of connectivity between different sentential components. This is illustrated in the following example from Ariel (1990: 132).

(11)    a.    Because **Noga** cannot resist sweets, <u>**she**</u> bought a whole load of them.
         b.    Because <u>**she**</u> cannot resist them, **Noga** bought a whole load of sweets.

According to Ariel, and also following Gernsbacher et al. (1989), the sentence in (b) reflects a stronger clause dependency, since one has to delay the interpretation of the two pronouns in it (*she* and *them*) until the second clause has been processed. Ariel hypothesizes that the more dependent the clause is, the higher the Accessibility marker chosen as anaphor (see also Green, 2014). If the two adjacent clauses are anaphorically related, anaphor choice will vary with the clause relations. Adjacent independent sentences, conjoined sentences, and matrix plus embedded clauses might use different options. The more independent clauses may even use a proper name (Low accessibility marker). Conjoined sentences would typically use pronouns, while embedded clauses may use a zero (e.g., infinitivals in English). As pointed out by Ariel and based on Silverstein (1976) and Foley and van Valin (1984), the differences between various languages can be much more fine-grained than such a classification may suggest.

For the purpose of this paper, we will use the basic distinction between two types of Unity: (1) between clauses in independent sentences with inter-sentential cataphora, and (2) between intra-sentential clauses with intra-sentential cataphora. Based on our data, we created the following hierarchy of Unity relations: (1) Unity within the same clause, (2) the same sentence (syntactic and pragmatic subordination), (3) between clauses in different sentences. The backgrounding factor is related to the second level of Unity relations.

In the analysis that follows, we test the behaviour of two types of cataphora as compared to four Accessibility parameters—Distance, Competition, Saliency and Unity for cataphora that occurs in both the nucleus and the satellite of coherence relations. Fig. 2 summarizes the parameters of the analysis. A square brackets represents an exclusive choice (e.g., cataphora may be found either in the nucleus or in the satellite), whereas the curly bracket indicates that all four parameters are considered.

## 7. Data analysis in terms of accessibility theory

### 7.1. First mention cataphora in the nucleus

In this part of our paper, we analyze instances of first mention cataphora against the four parameters of Accessibility.
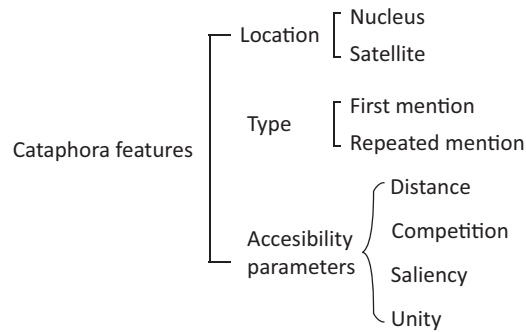
Fig. 2. Features of cataphora analyzed.

### 7.1.1. Distance

In all 35 cases of first mention cataphora in the nucleus we find that cataphora and antecedent were either in the same (''false'' starts) or adjacent clause (within the same or adjacent sentence). This is illustrated in (12) and (13).

Anaphora and antecedent in the same clause (false start)

(12)    I know there are degrees of modernness. I'm talking about modern in the 19th, or actually 20th century-type of reading, of novels in literature. They deal very much with realism, **they** don't like, **the authors** don't like the Cinderella endings, the Cinderella type of, of storyline. The 'Father Knows Best''-types of families, you know. [OANC].

Anaphora and antecedent in adjacent clauses (adjacent sentence)

(13)    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]

### 7.1.2. Competition

In the majority of cases of first mention cataphora, just one referent is linked to the cataphoric element. There were three instances where two antecedents were competing for the referent position. Two of these examples were with cataphoric *it*.

### 7.1.3. Saliency

All 35 examples of first mention cataphora in the nucleus are topics.

### 7.1.4. Unity

As was mentioned earlier in this section, there are two types of Unity that we were looking for in our data: (1) between clauses in independent sentences in case of inter-sentential cataphora (see example 13); and (2) between intra-sentential clauses with intra-sentential cataphora (see example 12). We classified each instance of cataphora based on the following parameters: syntactic subordination (adverbial clauses), pragmatic subordination (see example 3), coordination, and distance between the pronoun and the antecedent when they are in different sentences (in terms of number of clauses). Syntactic subordination in the nucleus includes instances in which cataphora is in the main clause of the sentence. Pragmatic subordination is represented with instances of coordinate constructions that have the conjunction *but*, when pragmatically one sentence is subordinated to the other (Mittwoch, 1983). Table 3 presents results for first mention cataphora in the nucleus. We can see that, out of the 35 instances of first mention cataphora in the nucleus, 14 were instances of inter-sentential cataphora.

The table does not include 11 instances of false starts in which the first segment with cataphora is repeated, in part or in whole, in the next segment, but repeated with a noun instead of a pronoun.

Table 3
Unity parameter: distribution of first mention cataphora in the nucleus.

|         | Syntactic subordination | Pragmatic subordination | Coordination | Different sentences |
|---------|-------------------------|-------------------------|--------------|---------------------|
| Nucleus | 9                       | 1                       | –            | 14                  |

The table also demonstrates that in the majority of cases the syntactic bond between the cataphoric element in the nucleus and its antecedent is absent, since the antecedent usually appears in the next clause of a new sentence. Consider the example in (14), where the antecedent appears in the following clause.[17]

(14)     **They** lived across the street from us in PA. **That's my dad's brother and uh, his family** and um, one of his daughters call over there and she says, "We got a bird stuck in the house and it won't get out." (Distance: next clause) [OANC]

### 7.2. Repeated mention cataphora in the nucleus

If the pronoun is already part of the discourse context, the question is why a full NP is used after the pronoun. According to Ariel (1990), what determines whether cataphora is acceptable or not is distance and low/high cohesion (Unity). When the referent is already a given entity, there is relatively low cohesion (Unity), and dependency of the clause with the pronoun on the antecedent clause is not needed at all. This allows for the use of a full nominal (a Low Accessibility Marker). In the following examples from corpora we analyze sentences with repeated mention cataphora against the four parameters of Accessibility. For this part of the analysis, we extracted further examples. We analyzed 1900 instances of the same pronouns (*he, she, it, they*) in the OANC, and found 88 examples of repeated mention cataphora, 40 of them in the nucleus. Those were further analyzed according to the parameters of Accessibility Theory.

### 7.2.1. Distance
Of the 40 examples of repeated mention cataphora in the nucleus seven had the cataphor and the secondary antecedent separated by one or more clauses. One of those seven examples is illustrated in (15).

(15)     It was, when the first one arrived, **the first soldier**, and he said, "Er, ma'am, make," he said, "some four, six rations," he said, "because plenty of soldiers have come, and I'm going to inform the other houses." I stayed, making the rations of food when a group of soldiers arrives, and when I was, fixing the food, in the kitchen, because of the sun, because of the reflection of the sun, I saw, a weapon pointing at me, so that was when I said to him, "What happened?" And **he** said, "Don't play." **The soldier** said, "don't play," he said, "innocent," he said, in that house, I, our house is huge, and I stayed locked up, because just me and my son lived, my baby son, tiny, and me, and well, that's all. [OANC]

### 7.2.2. Competition
Repeated mention entities in the majority of cases have one candidate for the position of referent. Just three examples out of 40 have two competitors for the referent. Cataphora with *it* shows an interesting pattern: The referent is either reiterated as a secondary antecedent in the form of a noun phrase or it is elaborated in a separate proposition(s). The latter is illustrated in (16). The event of setting the mailbox and newspaper on fire was previously mentioned in the text.

(16)     It was uh, me, a guy name Jeff Teague, and Craig Barsley. . . . And uh, you know after a little while we ended up talking him into **it** and uh, so we all snuck . . . And uh, I don't know who thought of **it** I don't know if it was me or Jeff or Craig but uh, we uh, we had some cigarettes on us and **we were smoking trying to be the big rebels you know and we had a lighter.** And uh, one of us thought of **the idea of sticking a newspaper in the mailbox and setting the mailbox and newspaper on fire.** [OANC]

### 7.2.3. Saliency
All 88 instances with repeated mention entities are topics.

### 7.2.4. Unity
The same criteria of Unity are applied to repeated mention entities. The results are presented in Table 4, where we show that out of the 40 examples of repeated mention cataphora in the nucleus, 18 correspond to inter-sentential cataphora.

The distribution of different types of Unity partially depends on the pronoun itself. In 18 cases with *it* there is no syntactic bond between the cataphor and the antecedent, because they are in different sentences. The rest of the pronouns (seven

---

[17] The notion of clause and sentence boundary is fuzzy in transcribed spoken language. In Example (14), for instance, the unit that includes "that's my dad's brother. . ." could be considered a parenthetical, with the main unit continuing at "one of his daughters". We have relied on the transcriptions to determine clause and sentence boundaries, but it is possible that the transcriptions constitute an interpretation of boundaries.

Table 4
Unity parameter: distribution of repeated mention cataphora in the nucleus.

|  | Same clause | Syntactic subordination | Pragmatic subordination | Coordination | Different sentences |
|---|---|---|---|---|---|
| Nucleus | 4 | 4 | 1 | 6 | 18 |

of them) were found with false starts, coordinate constructions and syntactic subordination. An additional category with repeated mention entities that did not exist with the first mention entities are pronouns occurring within the same clause, as illustrated with the following example, with an antecedent expressed as the subject of the sentence and a possessive pronoun in the adjunct part of the same sentence.

(17)     In **his** Saturday radio address, t**he President** repeated his call [breath] for a federal ban on human cloning. [OANC]

### 7.3. First mention cataphora in the satellite

In our original data collection of first mention cataphora, we found 21 instances in the satellite part of coherence relations.

#### 7.3.1. Distance
Two out of 21 instances were separated by more than one clause. The rest of the examples are within the same or adjacent clause in the same or adjacent sentence. Example (18) illustrates first mention cataphora in the satellite position in adjacent clauses:

(18)     Although **he** called current market conditions "highly competitive," **Mr. LaMothe, Kellogg's chairman and chief executive officer,** forecast an earnings increase for the full year. [RST]

#### 7.3.2. Competition
In all 21 instances of first mention cataphora in the satellite there is only one candidate for the status of referent.

#### 7.3.3. Saliency
All instances of first mention cataphora in the satellite are topics.

#### 7.3.4. Unity
The results for the Unity parameter are illustrated in Table 5. For first mention cataphora in the satellite, out of 21 examples, six of them are inter-sentential.
As expected, cataphora in the satellite position is frequently found with the subordination Unity parameter (syntactic and pragmatic subordination). Distribution of cataphora in the clauses of adjacent sentences is not as high as in the previous cases of first/repeated mention cataphora in the nucleus position. The two instances of cataphora are found in the same clause. Their antecedents were cases of cataphora expressed with the possessive pronoun in the sentence adjunct and the noun phrase filling in the position of antecedent.

(19)     In **his** review of "Saturday Night With Connie Chung," **Tom Shales**, the TV critic of the Washington Post and generally an admirer of CBS, wrote that while the show is "impressive,.. one has to wonder if this is the proper direction for a network news division to take." [BN]

One instance of false start is found with this position of cataphora.

Table 5
Unity parameter: distribution of the first mention cataphora in the satellite position.

|  | Same clause | Syntactic subordination | Pragmatic subordination | Coordination | Different sentences |
|---|---|---|---|---|---|
| Nucleus | 2 | 6 | 6 | – | 6 |

Table 6

Unity parameter: distribution of repeated cataphora in the satellite.

|  | Same clause | Syntactic subordination | Pragmatic subordination | Coordination | Different sentences |
|---|---|---|---|---|---|
| Satellite | 5 | 6 | – | 1 | 28 |

### 7.4. Repeated mention cataphora in the satellite

Out of the 88 instances of repeated mention cataphora, 48 cases were found in the satellite part of coherence relations.

#### 7.4.1. Distance
The instances of repeated mention cataphora in the satellite part of relations show the highest distance between the anaphor and its secondary antecedent when compared to all other cases of cataphora (16 out of 48 separated with more than one clause). One of such examples is illustrated in (20).
(Monika Lewinsky has been mentioned in the previous text)

(20)      So that's the reason **<u>she</u>** went to Portland. Yes. Yes. Yes. Yes. Yes. Yes. The tabloid Star. Like its mainstream competitors, [breath] it finds people who say they knew her. Were you stunned? [breath] This just couldn't be? [breath] Not the **Monica Lewinsky** that I knew? [breath] No. No, not s- not stunned. But it's you know, a lot could happen in five years. [BN]

#### 7.4.2. Competition
Competition for the position of the referent does not seem to be a relevant factor, since none of the examples show two candidates competing for the same referent.

#### 7.4.3. Saliency
As with previous instances of cataphora, cases of repeated mention cataphora in the satellite are all sentence topics.

#### 7.4.4. Unity
In this case, we find a somewhat different distribution of the unity parameter than with first mention cataphora in the satellite. The majority of cases have cataphora and the secondary antecedent positioned in different sentences, as we can see in Table 6: Out of 48 examples, 28 were cases of inter-sentential cataphora.
(Antecedent previously mentioned)

(21)      Well, **<u>he</u>** was almost, old enough to be her daddy. Oh, no. And he was, **my grandfather** was not a nice man.

Apart from the examples represented in Table 6, we found eight instances of false starts with cataphora (repeated mention) in satellites. As with first mention cataphora in the satellite, five instances that are found in the same clause with its antecedent are cases of cataphora expressed with the possessive pronoun in the adjunct part of the sentence, and the noun phrase filling in the position of its antecedent as in the sentence below:

(22)      In **<u>his</u>** memoirs, **Ambassador Seitz** says that Mrs. Kennedy Smith, who was then and still is ambassador to Ireland, was quote, too shallow to understand the past and too naive to anticipate the future.

## 8. Discussion

According to Harris and Bates (2002), cataphora is allowed in the backgrounded part of a sentence when backgrounding is achieved through subordination. In this paper we examined backgrounding at the discourse level. We analyzed both intra-sentential and inter-sentential instances of cataphora in three corpora of naturally-occurring discourse. We made use of the nucleus-satellite distinction which, as Matthiessen and Thompson (1988) suggest, directly corresponds to the syntactic phenomenon of subordination. We examined the presence of cataphoric *he*, *she*, *it* and *they* in data annotated with nucleus-satellite labels (roughly corresponding to main and subordinate clauses) and concluded that there is no correlation between cataphora and backgrounding at the discourse level, as cataphoric pronouns appear in both the nucleus and the satellite.

We then tested Ariel's (1990) hypothesis that only first mention cataphora requires a highly cohesive relation with the antecedent unit, and hence only occurs within adverbial subordinate clauses or pragmatically non-dominant clauses,

while continuing discourse referents require separation between the pronoun (local cataphora) and the antecedent in order for a Low Accessibility Marker (noun) to appear. In following and extending Ariel's proposal, we hypothesized that the nucleus tends to (i) include a continuing discourse referent and that (ii) the linkage between the nuclear discourse unit where the repeated mention cataphora is positioned and the discourse unit of its antecedent is looser than between first mention cataphora and its antecedent in the satellite position. We now summarize the results of the analysis from the previous section.

The main distinction between different types of cataphora seems to be focused on two Accessibility parameters, Distance and Unity. The results regarding the other two parameters, Competition and Saliency, are quite homogenous across the examples and this can be taken as a solid discourse pattern that characterizes cataphora. In terms of Competition, most cataphoric pronouns have one candidate for the position of referent. Since cataphoric devices[18] in general enhance the activation of the concepts that they mark and suppress activation of previously mentioned concepts (Gernsbacher and Jescheniak, 1995), the lack of competition for the position of referent might be an expected result. In terms of Saliency, all our examples are instances of sentence topics. They also tend to be discourse topics, because the same referent is repeated in subsequent sentences as a topic. As a result of our representation of the Unity parameter, we suggest that the other two parameters, Distance and Unity, go hand in hand: The tighter the connection between the units which contain the cataphoric pronoun and the antecedent, the shorter the distance between the two elements. Based on this observation, we discuss these two parameters together. We summarize our results starting with the hypothesis by Ariel (1990: 158), which states that, when the antecedent is newly introduced into the discourse, a speaker has to make sure that the pronoun clause is dependent on the antecedent clause so that the interpretation of the pronoun can rely on the material from the independent clause. However, when the entities form part of the discourse already, dependency is not needed at all.

The results of the corpus analysis do not fit the above hypothesis in a straightforward manner. For instance, a frequent pattern among new entities (especially with *it*) is that cataphora occurs in the nucleus (independent segment, the main clause of the sentence), while the referent is expressed with a proposition that elaborates the pronoun in the next sentence. There is no syntactic dependency between the unit that contains the cataphoric pronoun and the unit of the antecedent. The semantic dependency, however, is obvious: A pronoun needs elaboration. An example is presented in (23), where the bold portion in the second part of the example elaborates on the content of the pronoun *it*.[19]

(23)    It was about 7:30 and here come Richard. And that's, I'll never forget **it**. **He came and then he said,** "**Shh, what happened to your van?**" **I'm like,** "**Nice try, Richard but my van is fine. I've been going over there all night long**." [OANC]

At the same time, other results of the analysis regarding Unity and Distance partially confirm the above claim of Ariel. Some instances of repeated mention cataphora, both in the nucleus and in the satellite, show greater preference for looser Unity and higher Distance between the pronoun and the antecedent. In the nucleus, seven out of 40 cases of cataphoric pronouns are separated from the antecedent by one or more clauses. In the satellite position we have found 16 out 48 such cases. First mention cataphoric pronouns show similar behaviour in satellites: Two out of 21 pronouns have a separation between the pronoun and the antecedent of one or more clauses. However, this behaviour is not documented with first mention instances in the nucleus (see Table 7).

This conclusion is, of course, based on a limited data set. We cannot completely exclude the possibility that there are cases where the distance between the antecedent and the first mention pronoun in the nucleus is higher than one clause. In addition, first mention entities in the satellite position show a tendency towards tighter Unity between the unit that contains the cataphoric pronoun and the antecedent than repeated mention instances of anaphora in the satellite. Only six out of 21 instances of the first mention cataphora in the satellite are in separate sentences. Consider the example in (24), in which cataphora appears as first mention in the satellite of a Concession relation, and the antecedent, in the nucleus, is in an adjacent clause.

(24)    But I do remember that we stood and watched him go by and waved at him. [I don't remember whether **he** waved back or not]$_S$ [so I did get to see **President Roosevelt himself** and in person]$_N$ so that was a big thrill for me as a young person. [OANC]

---

[18] Cataphoric devices include, but are broader than, cataphoric pronouns (for more detailed information on cataphoric devices see Gernsbacher and Jescheniak (1995: 25).

[19] This type of cataphora can be characterized as a ''point of view'' case which has an ''egocentric'' pronoun and which points out to the perspective adopted by the speaker. It should be distinguished from discourse-initial types of cataphora that are typical of the media discourse (see Ariel's (1990) notes 9 and 12 on p. 155 and 157 respectively).

Table 7
Separation between pronoun and antecedent of one or more clauses.

|  | Nucleus | Satellite |
| --- | --- | --- |
| First mention cataphora | – | 2/21 |
| Repeated mention cataphora | 7/40 | 16/48 |

The rest of the first mention instances in the satellite position (15 cases) are limited to subordination (syntactic or pragmatic), or show a pattern with both antecedent and pronoun in the same clause (all cases of higher Unity). The repeated instances of cataphora in the satellite exhibit crucially different behaviour: 28 instances out of 48 are with a pronominal element in a different sentence from the antecedent. In general, however, all the above groups except for the first mention entities in the satellite prefer for the cataphor and the antecedent to be in adjacent clauses of different sentences, which is a looser level of Unity. Even in cases of first mention cataphora in the nucleus, this type of Unity is quite high: In 14 out of 35 instances the pronoun and the antecedent are in different sentences. In summary, the results regarding the relation between level of Unity and Givenness do not provide a clear-cut distinction.

Our hypothesis about the correlation between the nucleus and continuous anaphora (repeated mention cataphora) can be rejected. The hypothesis that repeated mention instances in the nucleus position have a looser connection between the pronominal element and the antecedent than the first mention instances in the satellite position is partially proven (compare 18 out 40 instances of repeated mention cataphora in the nucleus, appearing in a different sentence to its antecedent, to six out of 21 instances of first mention cataphora with the same Unity parameter).

Our analysis provides a challenge for the following three hypotheses: (i) the hypothesis about a correlation between discourse backgrounding and cataphora, (ii) the hypothesis about a correlation between continuous cataphora and nucleus versus first mention cataphora and satellite, and (iii) the hypothesis about a tighter connection between first mention instances of cataphora and antecedent versus looser connection between continuous anaphora and its antecedent.

Based on a very limited set of data, we can conclude that there are conflicting linguistic conditions in which cataphora occurs in natural discourse. First, when a cataphoric pronoun and its antecedent are within one sentence (usually two adjacent clauses), clausal and discourse backgrounding is at work: Cataphora is in the satellite, while the antecedent is in the nucleus. There are, however, instances of cataphoric pronouns in the nucleus within one sentence as in example 25 (see also Table 4).

(25)     [**They** deal very much with realism, **they** don't like,]$_N$ [**the authors** don't like the Cinderella endings.]$_S$

Second, continuous anaphora occurs both in the nucleus and the satellite parts of coherence relations. The conflicting factor is that the same also applies to instances of first mention cataphora, that is, it occurs in both types of cataphora (first and repeated mention). Third, while continuous cataphora occurs more often in examples where the pronoun is separated from the antecedent by one or more clauses in different sentences, the significance of that result might not be high, since both first and repeated mention cataphora show tendencies[20] for the antecedent and the pronoun to be in adjacent clauses within different sentences (not a very high Unity level). In future research it would be interesting to examine whether linguists and psycholinguists are able to come to a consensus regarding whether there is one relevant unit (clause, sentence or larger discourse unit) for the distribution of full vs. attenuated forms (Ariel, 2001). We suspect that, in different conditions, different units are relevant.

So far the following conclusion is emerging from our data: the occurrence of cataphora cannot be mapped to a single linguistic factor. Backgrounding (or dependency) is one of the factors in combination with which cataphora occurs frequently, both at the clausal and discourse levels. However, it is neither a sufficient nor a necessary element in spoken discourse. It is not sufficient because it operates within the constraints of Accessibility, and it usually interacts with other parameters of Accessibility, such as Givenness, Saliency of the referent and short/long Distance between the referent and the cataphoric element. Backgrounding is not a necessary element either. As we have discussed, cataphora can occur in the nucleus (foregrounded) part. In this case, no special linguistic signal is present apart from intonation.[21] As for Accessibility, our data shows that its two most stable parameters related to cataphora are Saliency and Competition. As

---

[20] In 66 out of 144 instances of first and repeated mention cataphora the pronominal element and its antecedent are in different sentences (both adjacent and non-adjacent clauses). In 41 of those instances the cataphoric pronoun and the antecedent are in adjacent clauses of different sentences.

[21] Compare with the example in (5) earlier: *He lied to me, and John was my friend.*

suggested by Reinhart (1983), cataphora is always a sentence topic and usually there is no competition between two or more referents that are linked to cataphora. Since there are multiple factors in place which surround the occurrence of cataphora and that reflect its complexity, we tend to agree with Gernsbacher and Jescheniak (1995), who claim that the underlying mechanism of cataphoric devices is cognitive in nature. For instance, some instances of cataphora in our data were false starts, which signal that the speaker has a mental representation of the cataphoric referent as highly accessible in her mind. At some point in the discourse, however, the speaker realizes that the same referent might not be as accessible for the addressee. We consider fruitful the hypothesis of Gernsbacher et al. (1989), that different referring expressions enhance differently the accessibility of mental representations associated with them. As pointed out in Ariel (2001), and according to Gernsbacher et al. (1989), more explicit referring expressions boost the activation of their mental representations faster than higher accessibility markers. Also, the same accessibility markers code a specific current degree of accessibility, but at the same time they contribute to the opposite degree of future accessibility. In some instances, the speaker chooses to attend to her addressee's needs by choosing her accessibility marker in accordance with the current degree of accessibility (e.g., high marker of Accessibility) or she may choose to ensure the future high accessibility (by choosing an accessibility marker which is relatively low). Possibly, the speaker, by using forward referring pronouns (high accessibility markers) followed by the referent expressed with a noun or a proposition (low accessibility markers), signals current high accessibility, and at the same time also high future accessibility. This goes along the lines of the proposal of Gernsbacher and Jescheniak (1995) that cataphoric devices enhance the activation of concepts that they mark, suppress the activation of previously mentioned concepts and protect the concepts that they mark from being suppressed by subsequently mentioned concepts. It also suggests that factors that enable the occurrence of cataphora in natural discourse seem to be to a certain degree extralinguistic in nature.

## 9. Conclusion

After analyzing 57 instances of first mention cataphora and 88 instances of repeated mention cataphora (out of 11,636 instances of the third-person pronouns *he, she, they, it*), we demonstrate that backgrounding (or dependency) is an insufficient factor to explain the occurrence of cataphora in natural discourse. Cataphoric pronouns can be positioned both in the backgrounded and foregrounded parts of discourse. As for the Accessibility parameters of Givenness, Distance and Unity, we demonstrate that in different conditions they do not show a clear influence (especially when a cataphor and its antecedent are in different sentences). However, Saliency and Competition are two stable parameters—cataphoric pronouns are always sentence topics, while competition between two or more different referents for the role of cataphora is rare. Future research should include more data with a focus on distinguishing between various types of cataphora that may provide additional discourse patterns for the occurrence of this phenomenon.

## Acknowledgments

## References

Alabiso, Jennifer, MacIntyre, Robert, Graff, David, 1998. 1997 English Broadcast News Transcripts (HUB4), LDC98T28 [Corpus]. Linguistic Data Consortium, Philadelphia, PA.
Almor, Amit, Nair, Veena A., 2007. The form of referential expressions in discourse. Lang. Linguist. Compass 1 (1–2), 84–99.
Ariel, Mira, 1990. Accessing Noun Phrase Antecedents. Routledge, London.
Ariel, Mira, 2001. Accessibility theory: an overview. In: Sanders, Ted, Schilperoord, Joost, Spooren, Wilbert (Eds.), Text Representation: Linguistic and Psycholinguistic Aspects. John Benjamins, Amsterdam, pp. 29–88.
Arnold, Jennifer E., 1998. Reference Form and Discourse Patterns. Stanford University Thesis Type.
Arnold, Jennifer E., 2010. How speakers refer: the role of accessibility. Lang. Linguist. Compass 4 (4), 187–203.
Arnold, Jennifer E., Griffin, Zenzi M., 2007. The effect of additional characters on choice of referring expression: everyone counts. J. Mem. Lang. 56 (4), 521–536.
Bolinger, Dwight L., 1979. Pronouns in discourse. In: Givón, T. (Ed.), Syntax and Semantics, Vol. 12 Discourse and Syntax. Academic Press, New York, pp. 289–309.
Broadbent, Donald E., 1973. In Defence of Empirical Psychology. Methuen, London.
Carden, Guy., 1982. Backwards anaphora in discourse context. J. Linguist. 18 (2), 361–387.
Carlson, Lynn, Marcu, Daniel, Okurowski, Mary Ellen, 2002. RST Discourse Treebank, LDC2002T07 [Corpus]. Linguistic Data Consortium, Philadelphia, PA.

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

Chafe, Wallace, 1994. Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. The University of Chicago Press, Chicago.

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

Cohan, Jocelyn, 2000. The Realization and Function of Focus in Spoken English. University of Texas at Austin Thesis Type.

Couper-Kuhlen, Elizabeth, Thompson, Sandra A., 2008. On assessing situations and events in conversation: 'extraposition' and its relatives. Discourse Stud. 10 (4), 443–467.

Cowles, H. Wind, Walenski, Matthew, Kluender, Robert, 2007. Linguistic and cognitive prominence in anaphor resolution: topic, contrastive focus and pronouns. Topoi 26, 3–18.

Crain, Stephen, 1991. Language acquisition in the absence of experience. Behav. Brain Sci. 14, 597–650.

Cristea, Dan, Ide, Nancy, Romary, Laurent, 1998. Veins theory: a model of global discourse cohesion and coherence. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL-98/COLING-98), pp. 281–285.

Cristofaro, Sonia, 2005. Subordination. Oxford University Press, Oxford.

Du Bois, John, 2003. Argument structure. In: Du Bois, John, Kumpf, Lorriane, Ashby, William (Eds.), Preferred Argument Structure: Grammar as Architecture for Function. Benjamins, Amsterdam, pp. 11–60.

Foley, William A., van Valin Jr., Robert D., 1984. Functional Syntax and Universal Grammar. Cambridge University Press, Cambridge.

Garnham, Alan., 1987. Understanding anaphora. In: Ellis, Andrew W. (Ed.), Progress in the Psychology of Language. Lawrence Erlbaum, Hillsdale, NJ, pp. 253–300.

Gernsbacher, Morton Ann., 1990. Language Comprehension as Structure Building. Erlbaum, Hillsdale, NJ.

Gernsbacher, Morton Ann, Jescheniak, Jörg D., 1995. Cataphoric devices in spoken discourse. Cogn. Psychol. 29, 24–58.

Gernsbacher, Morton Ann, Hargreaves, David J., Beeman, Mark, 1989. Building and accessing clausal representations: the advantage of first mention versus the advantage of clause recency. J. Mem. Lang. 28 (6), 735–755.

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.

Givón, Talmy, 2001. Syntax: An Introduction, vol I. John Benjamins, Amsterdam/Philadelphia.

Goldberg, Adele, 1995. Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago.

Gordon, Peter C., Hendrick, Randall, 1997. Intuitive knowledge of linguistic coreference. Cognition 62, 325–370.

Green, Clarence, 2014. On the relationship between clause combination, grammatical hierarchy and discourse-pragmatic coherence. Funct. Lang. 21 (3), 297–332.

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

Gundel, Jeanette K., 1977. Role of Topic and Comment in Linguistic Theory. Indiana University Linguistics Club, Bloomington.

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

Halliday, Michael A.K., Matthiessen, Christian, 2014. Halliday's Introduction to Functional Grammar, 4th ed. Arnold, London.

Harris, Catherine L., Bates, Elizabeth A., 2002. Clausal backgrounding and pronominal reference: a functionalist approach to c-command. Lang. Cogn. Process. 17 (3), 237–269.

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

Kehler, Andrew, 2004. Discourse topics, sentence topics, and coherence. Theor. Linguist. 30 (2–3), 227–240.

Lambrecht, Knud, 1994. Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents. Cambridge University Press, Cambridge.

Langacker, Ronald W., 1969. On pronominalization and the chain of command. In: Reibel, David, Schane, Sanford (Eds.), Modern Studies in English: Readings in Transformational Grammar. Prentice Hall, Englewood Cliffs, NJ, pp. 160–200.

Lappin, Shalom, Leass, Herbert J., 1994. An algorithm for pronominal anaphora resolution. Comput. Linguist. 20 (4), 535–561.

Lasnik, Howard, 1976. Remarks on coreference. Linguist. Anal. 2, 1–22.

Lee-Goldman, Russell, 2011. Context in Constructions (PhD dissertation). University of California, Berkeley.

Lee-Goldman, Russell, 2012. A story-prefacing resource in conversational English. In: Paper presented at the 11th Conceptual Structure, Discourse and Language Conference, Vancouver, Canada.

Mann, William C., Taboada, Maite, 2015. RST Web Site. From http://www.sfu.ca/rst

Mann, William C., Thompson, Sandra A., 1988. Rhetorical structure theory: toward a functional theory of text organization. Text 8 (3), 243–281.

Marcu, Daniel, 1999. Instructions for Manually Annotating the Discourse Structures of Texts. Marina del Rey, California Unpublished manuscript.

Matthiessen, Christian M.I.M., Thompson, Sandra A., 1988. The structure of discourse and subordination. In: Haiman, John, Thompson, Sandra A. (Eds.), Clause Combining in Discourse and Grammar. John Benjamins, Amsterdam and Philadelphia, pp. 275–329.

McCray, Alexa, 1980. The semantics of backward anaphora. In: Jensen, John T. (Ed.), Cahiers Linguistiques d'Ottawa 9 (Proceedings of the Tenth Annual Meeting of the North Eastern Linguistic Society), Vol. 9. University of Ottawa, Ottawa, pp. 329–343.

Mitkov, Ruslan, 2002. Anaphora Resolution. Pearson, London.

Mittwoch, Anita, 1983. Backward anaphora and discourse structure. J. Pragmat. 7, 129–139.

Morgan, Jerry L., 1968. Some strange aspects of it. In: Paper presented at the Papers from the 4th Regional Meeting of the Chicago Linguistic Society, Chicago.

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

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, Svartvik, Jan, 1985. A Comprehensive Grammar of the English Language. Longman, London.

Reinhart, Tanya, 1981. Pragmatics and linguistics: an analysis of sentence topics. Philosophica 27, 53–94.

Reinhart, Tanya, 1983. Anaphora and Semantic Interpretation. Croom Helm, London.
Reuland, Eric, Avrutin, Sergey, 2010. Interpretive dependencies: the case of backward anaphora. Linguist. Anal. 34 (3–4), 271–315.
Ross, John R., 1969. On the cyclic nature of English pronominalization. In: Reibel, David, Schane, Sanford (Eds.), Modern Studies in English:
    Readings in Transformational Grammar.. Prentice Hall, Englewood Cliffs, NJ.
Schlenker, Philippe, 2005. Minimize restrictors! (Notes on definite descriptions, Condition C and epithets) In: Maier, Emar, Bary, Corien, Huitink,
    Janneke (Eds.), Proceedings of Sinn und Bedeutung 2004. Radboud University Nijmegen, Nijmegen, pp. 385–416.
Scott, Mike, 2012. WordSmith Tools Version 6. Lexical Analysis Software, Liverpool.
Silverstein, Michael, 1976. Hierarchy of features and ergativity. In: Dixon, R.M.W. (Ed.), Grammatical Categories in Australian Languages.
    Australian Institute of Aboriginal Studies, Canberra, pp. 112–171.
Taboada, Maite, Hadic Zabala, Loreley, 2008. Deciding on units of analysis within centering theory. Corpus Linguist. Linguist. Theory 4 (1), 63–108.
Taboada, Maite, Mann, William C., 2006a. Applications of rhetorical structure theory. Discourse Stud. 8 (4), 567–588.
Taboada, Maite, Mann, William C., 2006b. Rhetorical structure theory: looking back and moving ahead. Discourse Stud. 8 (3), 423–459.
van Hoek, Karen, 1997. Anaphora and Conceptual Structure. University of Chicago Press, Chicago.
Ziv, Yael, 1994. Left and right dislocations: discourse functions and anaphora. J. Pragmat. 22, 629–645.

**Radoslava Trnavac** is a Lecturer in the Cognitive Science Programme (Simon Fraser University, Canada). She works in the areas of discourse analysis. Her current research addresses the relational and referential types of coherence and evaluation in texts.

**Maite Taboada** is Professor of Linguistics at Simon Fraser University (Canada). She works in the areas of discourse analysis, systemic functional linguistics and computational linguistics. Ongoing research includes opinion and sentiment in text, coherence in multimodal documents, and cataphoric relations.