# On the Information Status of Antecedents: Referring Expressions Compared

*Iker Zulaica-Hernández and Javier Gutiérrez-Rexach*
The Ohio State University

## Abstract

The differences in use among referring expressions have been explained on the basis of the information status or the cognitive status of their antecedents. Thus, for example, it has been proposed that highly accessible referents (in the current focus of attention of the discourse participants) license the use of personal pronouns while banning the use of demonstratives. This paper compares the referential properties of Spanish demonstrative expressions and the neuter personal pronoun through the study of a Spanish corpus. Our hypothesis is that these two referring expressions are very similar, if not identical, regarding the information status of their antecedents. We will argue that the difference between these expressions lies in that demonstratives actively contribute to information structure by marking topic or subtopic shifts in discourse, whereas speakers use neuter personal pronouns to refer to established topics.

## 1 Introduction

Over the last decades, corpus-based research has turned out to be of great importance in helping provide adequate solutions to many theoretical issues in different linguistic fields. A number of corpus studies have been conducted on the phenomenon of discourse deixis[1] achieving outstanding advances in the comprehension of the mechanisms that govern the class of referential chains that arise in discourse. Some of these studies have put the focus on providing an adequate annotation scheme for discourse deixis, other studies are focused on the quantitative part and most of them combine the two perspectives. For example, Poesio and Artstein (2008) present their annotation scheme for the ARRAU[2] corpus and tackle important questions like the referential ambiguity of certain expressions in discourse-anaphora patterns. Regarding the reliability tests on discourse deixis, these authors point out the following: "for discourse deixis we found that annotators agreed on the general textual regions that evoke the referents, though they often disagreed on the exact boundaries, resulting in agreement of around $\alpha = 0.55$" (2008:1171).

Dipper and Zinsmeister's work (2009) focuses on German and provides rigorous annotation guidelines to determine the semantic type of anaphor and antecedent. The authors justify their semantic annotation due to the idiosyncrasy of antecedents in discourse deixis, that is, the anaphoric link cannot be resolved through grammatical restrictions. Navarretta and Olsen's study (2008) is an extension of the MATE/GNOME co-reference annotation scheme (Poesio, 2004) that accounts for abstract anaphora in Danish and Italian[3]. Besides annotating the type of clausal antecedent, the semantic type of the referent (events, states, fact-like entities, etc.)

1 For an overview of discourse deixis, see the general studies by Asher (1993), Byron (2004), Fox (1987), Webber (1979) or the studies on Spanish demonstratives by Gutiérrez-Rexach and Zulaica-Hernández (2007) and Zulaica-Hernández (2008).

2 The home page of the ARRAU project is http://cswww.essex.ac.uk/Research/nle/arrau.

3 The English home page of the DAD project is http://www.cst.dk/dad.

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena", pages 173-186, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

they also annotate anaphoric distance, measured in terms of clauses in between the anaphor and the antecedent. Navarretta and Olsen draw important conclusions on the differences between Italian and Danish abstract anaphora and on how some of the proposals made by Gundel et al. (2004) concerning the relationship between antecedent and pronoun types do not hold for Italian.

Recasens (2008) conducted a corpus study of the discourse-deictic properties of Spanish and Catalan expressions, including demonstratives, based on the annotated corpus AnCora[4]. In her study, the author tests whether Webber's ideas on discourse deixis also hold for Catalan and Spanish. Besides the importance of her quantitative study, one of the most significant points in Recasen's paper concerns the intrinsic difficulty to clearly delimit the exact boundaries of the antecedents in cases of discourse deixis. Thus, she appeals to Webber's (1988, 1991) ideas on the unspecificity of the antecedent and to Poesio et al.'s (2006) theory on the underspecification of anaphora (*The Justified Sloppiness Hypothesis*) that, in essence, postulates that certain ambiguous anaphoric expressions may be left unresolved or simply not fully specified in the right context.

Other studies have placed the focus on the analysis of referring expressions and information status across languages (Bosch et al., 2003; Carminati, 2000; Kaiser and Trueswell, 2005; Kameyama, 1999; Navarretta, 2005, 2007; Sturgeon, 2008; Vieira et al., 2002). Although there is no total consensus when the referential properties of demonstratives and personal pronouns are compared, the most widely accepted thesis is that antecedents of demonstratives are most commonly non-topical whereas personal pronouns commonly have topical elements as their antecedents. Topichood is assumed to be dependent on syntactic configurations; namely, highly prominent positions (subject) are topical whereas less prominent syntactic positions (object, adjunct) are non-topical.

The aim of this paper is to compare the referential properties of Spanish demonstratives and the neuter personal pronoun *lo* ('it') as elements that participate in discourse anaphora and discourse deixis patterns. Following previous work on the information status of referring expressions (Prince, 1981b; Ariel, 1988, 1990; Gundel et al, 1993; Hegarty et al, 2003; Poesio and Modjeska, 2005), we put the main focus in checking whether there are significant differences in the referring behavior of these two Spanish linguistic expressions and whether these differences, if any, may have a bearing on the information status of their referents. With this purpose, we have conducted a corpus study where we have tested two factors that can help us distinguish these two elements, namely, the textual distance of the antecedent[5] and the morphological type of the antecedent. The corpus used in this study is the CREA corpus[6]. Contrary to what is argued by proponents of the *Accessibility Scale* (Ariel, 1988) or the *Givenness Hierarchy* (Gundel et al., 1993), our hypothesis is that Spanish demonstratives (determiners and pronouns) and the neuter personal pronoun

4 The AnCora corpora – Annotated Corpora for Catalan and Spanish (Taulé et al. 2008) – consist of two corpora of 500,000 words for Catalan (AnCora-Ca) and Spanish (AnCora-Es). The corpora are accessible from http//clic.ub.edu/ancor.

5 Referential distance has already been considered as a factor possibly influencing the degree of accessibility of different referring expressions (see Maes & Noordman 1995 and Ariel 2001 for discussion on this topic).

6 The home page of the CREA corpus is http://corpus.rae.es/creanet.html.

*lo* (‘it’) do not differ in their basic referring capabilities or the information status of their antecedents. Rather, we will argue that the difference between these elements lies in that speakers use demonstratives to mark topic or subtopic shifts in the discourse. This is accomplished by focusing the “hearer's attentional state” on specific discourse referents. By using this strategy, speakers would make hearers aware of a change in the general or local topic at a certain point in discourse. On the other hand, the main function of the neuter personal pronoun is to refer to topics already established in discourse or, in other words, to maintain topic continuity.

## 2 Information Status: Accessibility and the Current Focus of Attention

Different hierarchical scales have been proposed to account for the different distribution shown by the range of referring expressions across languages. Thus, for example, Prince (1981b) was the first to propose a hierarchy for discourse entities called the *Scale of Familiarity*, which is based on three main factors: predictability, saliency and the common knowledge shared by the speaker and addressee.

In Ariel’s (1988) *Accessibility Scale*, the notion of accessibility is defined as the relative ease with which the addressee can identify the referent of a referring expression or, alternatively, the ease with which the addressee can retrieve the intended referent from memory. According to the scale, demonstratives occupy an intermediate position in terms of the degree of accessibility they confer to their referents. As the scale clearly indicates, the less informative (null) forms (gaps, PRO, etc.) occupy the highest position, that is, they are high accessibility markers. Unstressed and cliticized pronouns also occupy a high position in the scale.

Gundel et al.’s (1993) *Givenness Hierarchy* is an implicational hierarchy of cognitive states and linguistic forms aimed to resolve the different anaphoric behavior of pronominal and non-pronominal anaphors. According to this hierarchy, the referents of demonstratives have either activated or familiar status but never in focus, whereas the referent of a neuter personal pronoun always has the status in focus. Being *activated* for a referent means that, at a given point in the discourse, there must be a representation of the referent in short-term memory. On the other hand, being *in focus* means that the referent is not only in short-term memory but also at the current center of attention. As the authors pointed out, at a given discourse point, entities *in focus* are the partially ordered subset of *activated* entities that are more likely to be the topic in subsequent discourse.

Gundel et al. (2005) analyzed the behavior of English demonstratives ‘this/that’ and the unstressed pronoun ‘it’ in the Santa Barbara corpus of spoken American English. These authors observed that demonstrative anaphors were used to refer to abstract entities in 85% of the analyzed cases, whereas only 15% of the cases were anaphorically referred to with the pronoun ‘it’. They explained this fact by assuming that material introduced in clauses (e.g. clausally introduced entities like propositions or events, which are typical antecedents for demonstrative anaphors) is *activated* compared to material introduced via noun phrases in prominent syntactic positions, which is more likely to be *in focus*.

Poesio and Modjeska (2005) tried to make the cognitive notions from the Givenness Hierarchy (i.e., *in focus*, *activated*) and short-term memory, more precise.

They primarily adopt the computational approach to anaphora resolution of *Centering Theory* (Grosz, Joshi & Weinstein, 1995) and follow previous findings on that field to better define these notions. Poesio and Modjeska annotated the corpus GNOME for this purpose and tested the following hypothesis regarding the speaker's non-preference to use This-NP's to refer to *in focus* entities:

- This-NPs are preferentially used to refer to entities other than the $CB(U_i)$, the CB of the utterance containing the This-NP.
- They are used to refer to entities other than the $CB(U_{i-1})$, the CB of the previous utterance.
- They are used to refer to entities other than $CP(U_{i-1})$, the most highly ranked entity of the previous utterance.

In Centering Theory it is assumed that new discourse entities (forward-looking centers or CFs) introduced by each utterance are ranked based on information status. The forward-looking centers of $U_n$ only depend on the expressions that constitute that utterance; they are not constrained by features of any previous utterance in the segment. The most highly ranked entity of the forward-looking set is called the CP (the preferred center). The CB (the backward-looking center of an utterance $U_n$) is Centering's equivalent of the notion of topic or focus. The backward-looking center of $U_i$ connects with one of the forward-looking centers of $U_{i-1}$. The $CB(U_i)$, the backward-looking center of utterance $U_i$, is the highest ranked element of $CF(U_{i-1})$, i.e. the CP of $U_i$.

The authors propose a general hypothesis regarding the speaker's preference to use This-NPs for reference to *activated* (*active* in their own terminology) discourse entities. An entity is *active* if:

- It is in the visual situation; or
- it is a CF of the previous utterance; or
- it is part of the implicit linguistic focus. They only considered as part of the implicit focus those entities that can be *constructed* out of the previous utterance. An entity can be constructed out of an utterance if: A) It is a plural object whose elements or subsets have been explicitly mentioned in that utterance; or B) It is an abstract entity introduced by that utterance. They consider two types of abstract entities:
  i. Propositions
  ii. Types[7]

7 For Poesio and Modjeska (2005), types are those cases of generic reference that have concrete objects as instances.

They got the following results in terms of distribution:

| Class | Number (%) |
|---|---|
| Anaphora | 45 (40%) |
| Visual Deixis | 28 (25%) |
| Discourse Deixis | 19 (17%) |
| Type | 9 (8%) |
| Plurals | 1 |
| Ellipsis | 1 |
| Time | 1 |
| Unsure | 5 |
| Disagreement | 3 |
| Total | 112 |

Table 1. Distribution of This-NPs (Poesio and Modjeska, 2005)

With respect to the correlation between focus and This-NPs, they found the following principal results:

- 8-11 violations to the hypothesis that a This-NP is used to refer to entities other than the CB($U_{i-1}$) were found, which is therefore verified by 90%-93% of This-NPs.
- The hypothesis that This-NPs are used to refer to entities other than CP($U_{i-1}$) is verified by 75-80% of This-NPs.
- The hypothesis that This-NPs are used to refer to entities other than CB($U_i$) is verified by 61-65% of This-NPs.

So the hypothesis that received more empirical support is the following: This-NPs are used to refer to entities which are active but not the backward-looking center of the previous utterance. Based on these results and an in-depth study of the violation cases they proposed the version that leads to the fewest number of violations of Grice's *Maxim of Quantity* (1989):

**The This-NP Hypothesis**: This-NPs are used to refer to entities, which are *active* in the sense specified above. However, pronouns should be preferred to This-NPs for entities other than CB($U_{i-1}$).

In a series of papers, Hegarty (2003, 2006) and Hegarty et al. (2001, 2003) studied abstract object anaphora from a semantic perspective. Generally speaking, all these studies coincide in that clausally introduced entities are more commonly referred to with a demonstrative pronoun hence indicating that the cognitive status of the entities is activated. There is an important point to be made regarding the theoretical appropriateness of the cognitive statuses as reflected in the Givenness Hierarchy. Hegarty indicates that an entity will be *in focus* only if it has been mentioned by a

nominal expression in a prominent syntactic argument position earlier in the utterance or in the previous utterance; a supposition which is compatible with *Centering Theory* and results in the experimental psycholinguistic literature. On the other hand, peripherally introduced entities, including those introduced by less prominent nominal expressions and by clauses, will be *activated* upon their introduction, placed in working memory within the field of attention, but never at the center of attention.

As we have seen so far and concerning English data, there appears to be consensus on the information and cognitive status of the entities referred to with demonstratives and the weak pronoun 'it', especially when reference to abstract entities is involved. Thus, speakers would use demonstratives to refer to *activated* entities (or *active* in Poesio and Modjeska's terminology), which rank lower than *in focus* entities regarding their cognitive and information status. Unlike demonstratives, the pronoun 'it' would be strongly preferred for reference to entities in the current focus of attention, i.e. *in-focus*. But there are reasons to believe that these findings cannot be extrapolated to all languages. For example, Navarretta (2008) found language-specific results for Danish and Italian regarding the referential behavior of demonstratives and personal pronouns. She found that the most frequently used abstract anaphor in Danish is the ambiguous *det* ('it/this/that') and her data indicate that the anaphors *det* and *dette* (demonstrative 'this') are used with all antecedent types and to make reference to all sorts of referents. Also, personal pronouns are also used in Danish with clausal antecedents. Regarding Italian, Navarretta found that zero anaphors and personal pronouns are often used in this language in contexts where demonstrative pronouns occur in English. Also, zero anaphors are the most frequently used pronouns to refer to propositions (let us remind that the referents of zero pronouns are *in focus* in the Givenness Hierarchy). These data indicate important cross-linguistic differences in the referential behavior of referring expressions and/or the information status of abstract referents. Our Spanish data appear to point in a similar direction. We present our findings in the next sections.

## 3 Corpus: Methodology and Results

The CREA corpus is a large linguistic database (over 160 million words) comprising several language varieties, text types and genres. Corpus queries allow users to retrieve a text fragment, situating words in context. 50% of its sources are from Spain (45 million speakers), and 50% from Latin America (350+ million speakers). 90% of the words in CREA are from written sources, and only 10% from oral sources. The CREA corpus of Spanish is not annotated so we did the annotation manually and only for the cases analyzed. The size of the corpus and the high frequency of the expressions analysed (neuter personal pronoun and demonstratives) made it unfeasible to analyze all the occurrences found.

We have analysed a total number of 327 occurrences divided as follows: 120 occurrences of the neuter personal pronoun *lo* ('it') and 207 occurrences of demonstrative expressions. All the occurrences of neuter personal pronouns analysed ($n$ = 120) were divided into three groups corresponding to three different corpus searches: *lo entiendo* ('I understand it'), *lo necesito* ('I need it') and *lo tengo* ('I have

it'), so 40 occurrences per group were scrutinized. The reason for having analysed these particular combinations is twofold. On the one hand, this allowed us to discard other, non-referential uses of the personal pronoun in Spanish. On the other hand, these three groups would allow us to test not only referential distance but also the denotation of the antecedent and check whether it may possibly have a bearing on the cognitive status and different accessibility marking shown by the neuter personal pronoun. Thus, by using the predicates *entender* ('understand'), *necesitar* ('need') and *tener* ('have') we have tried to force different semantic readings for the antecedent. The predicate *entender* ('understand') would show a preference for higher order antecedents such as concepts, ideas or hypotheses rather than concrete, physical objects. Conversely, the verb *tener* in the expression *lo tengo* ('I have it') exhibits a preference for physical-object denoting antecedents as, under normal conditions, people have/own physical objects. The verb *necesitar* ('need') is intended to occupy an intermediate position in between the former two predicates. The aim overall was to obtain a sample ample enough to be able to draw some initial conclusions regarding the possible influence of antecedent denotation.

The first factor analysed was referential distance, that is, the distance between the anaphor and the antecedent. In order to check referential distance we segmented our examples into clauses. Our definition of a clause includes main and subordinate clauses, where the verbal phrase (VP) is taken as the clausal indicator. Thus, for example, two clauses joined with conjunction *y* ('and') count as two clauses and a main clause with a subordinate clause counts as two clauses as well. Obviously, we came across problematic cases like infinitival clauses (e.g. *Having a relationship is not in my plans for the moment*), which were also taken as a clause. The results are shown in Table 2.

| Anaphor | $CL_0$ | | $CL_1$ | | $CL_2$ | | $CL_3$ | | $CL_{\geq 4}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | # | % | # | % | # | % | # | % | # | % |
| *Lo necesito (I need it)* | 3 | 7.5 | 30 | 75.0 | 4 | 10.0 | 0 | 0 | 3 | 7.5 |
| *Lo entiendo (I understand it)* | 2 | 5.0 | 37 | 92.5 | 0 | 0 | 1 | 2.5 | 0 | 0 |
| *Lo tengo (I have it)* | 11 | 27.5 | 26 | 65.0 | 2 | 5.0 | 1 | 2.5 | 0 | 0 |
| Total | 16 | 13.33 | 93 | 77.5 | 6 | 5.0 | 2 | 1.66 | 3 | 2.5 |

Table 2: Referential distance for accusative personal pronoun

In total, we analyzed 207 occurrences of demonstratives ($n$ = 207). The results of this sample are shown in Tables 3 and 4. In the first place, we retrieved a sample of 50 adnominal demonstratives divided into two groups of 25 cases each: *este hecho* ('this fact') and *ese hombre* ('that man'). The reasons for having analyzed these particular NPs are the same that we explained for the neuter personal pronoun in the previous paragraph. With the NPs *hecho* and *hombre* we analyzed different denotations of the antecedent, namely, a higher order entity and a physical entity, respectively. A second corpus search consisted of 157 cases of demonstrative pronouns: 63 instances of demonstrative pronoun *esto* ('this'), 69 of *eso* ('that') and 25 of *aquello* ('that

further'). The disparity of the analyzed occurrences of demonstrative pronouns, in particular the low number of tokens for pronoun *aquello* (25), is due to the actual frequency of use of demonstratives in modern Spanish. Overall corpus figures show that pronominal demonstrative *aquello* has a very low frequency of use (6%) compared to the frequencies shown by *esto* and *eso*. Even between these two pronouns the differences are quite relevant (*eso*: 60%) and (*esto*: 34%). Nevertheless, overall figures vary when the frequency of use as demonstrative determiners is considered. Demonstrative determiner *ese* has a frequency of 30% whereas determiner *este* shows a percentage as high as 61%. Again, demonstrative determiner *aquel* shows a rather low frequency of use (9%).

| Anaphor | $CL_0$ | | $CL_1$ | | $CL_2$ | | $CL_3$ | | $CL_{\geq 4}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | # | % | # | % | # | % | # | % | # | % |
| *Este hecho (this fact)* | 1 | 4.0 | 21 | 84.0 | 3 | 12.0 | 0 | 0 | 0 | 0 |
| *Este hombre (this man)* | 2 | 8.0 | 19 | 76.0 | 2 | 8.0 | 0 | 0 | 2 | 8.0 |
| Total | 3 | 6.0 | 40 | 80.0 | 5 | 10.0 | 0 | 0 | 2 | 4.0 |

Table 3: Referential distance for demonstrative determiners

Our sample of demonstrative pronouns was restricted to events as type of referents of demonstrative anaphors. In order to restrict the referential potential of demonstratives, we searched the corpus for expressions consisting of a combination of a demonstrative pronoun plus a typical predicate of events like *suceder* ('happen'), *ocurrir* ('occur') or *pasar* ('happen'); e.g. *eso sucedió* ... ('that happened...'), etc. This forces a specific denotation for the antecedent: events. The principal advantages of this strategy were to restrict the large number of demonstrative pronouns in the corpus and also eliminating potential exophoric (extra-textual) reference while having a denotation that is not particularly biased as for the morphological type of antecedent used to convey it (NP or clausal).

| Anaphor | $CL_0$ | | $CL_1$ | | $CL_2$ | | $CL_3$ | | $CL_{\geq 4}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | # | % | # | % | # | % | # | % | # | % |
| *Esto (this)* | 0 | 0 | 50 | 79.4 | 11 | 17.5 | 1 | 1.6 | 1 | 1.6 |
| *Eso (that)* | 0 | 0 | 54 | 78.3 | 5 | 7.2 | 7 | 10.1 | 3 | 4.3 |
| *Aquello (that further)* | 0 | 0 | 19 | 76.0 | 4 | 16.0 | 0 | 0 | 2 | 8.0 |
| Total | 0 | 0 | 123 | 79.0 | 20 | 12.0 | 8 | 5.0 | 6 | 4.0 |

Table 4: Referential distance for demonstrative pronouns

The second factor analyzed was the morphosyntactic type of the antecedent. We have included the total number of occurrences analyzed in this study (n = 327). We have considered two types: NP and Other (clausal). Within the type *Other (clausal)* we have also included infinitival clauses and other antecedents that expand beyond the clause (i.e. complex clauses or even larger text spans). The results of the study

involving antecedent type are shown in Table 5.

| Anaphor | NP | | Other (Clausal) | |
|---|---|---|---|---|
| | # | % | # | % |
| *Lo necesito (I need it)* | 20 | 50.0 | 20 | 50.0 |
| *Lo entiendo (I understand it)* | 13 | 32.5 | 27 | 67.5 |
| *Lo tengo (I have it)* | 33 | 82.5 | 7 | 17.5 |
| *Este hecho (this fact)* | 6 | 24.0 | 19 | 76.0 |
| *Este hombre (this man)* | 24 | 96.0 | 1 | 4.0 |
| *Esto (this)* | 6 | 9.5 | 57 | 90.5 |
| *Eso (that)* | 16 | 23.2 | 53 | 76.8 |
| *Aquello (that further)* | 5 | 20.0 | 20 | 80.0 |
| Total | 123 | 37.6 | 204 | 62.4 |

Table 5: Morphological type of antecedent

In general, clausal antecedents are widely preferred (62.4%) over NP antecedents (37.6%) when all referring expressions are taken together. When we analyze the expression types individually, we found the following frequencies: the neuter pronoun *lo* shows a slight preference for non-clausal antecedents (55%) over clausal ones (45%). Some individual differences appear to be based on the type of the predicate accompanying the personal pronoun or demonstrative determiner analyzed. For example, the expression *lo entiendo* ('I understand it') shows a strong preference for clausal antecedents over NPs (27 and 13 occurrences, respectively). Conversely, the neuter pronoun in the expression *lo tengo* ('I have/posses it') shows a strong preference for NP over clausal antecedents (33 and 7 occurrences, respectively). Demonstrative pronouns (*esto*, *eso* and *aquello*) show a strong preference for clausal antecedents (90%, 76% and 80%, respectively), whereas demonstrative determiners show opposite preferences depending on the noun involved in each particular expression: the NP *este hecho* ('this fact') shows a strong preference for clausal antecedents (76%), most likely due to the denotation of the noun, whereas the NP *este hombre* ('this man') shows an even stronger preference for NP antecedents (96%).

## 4 Discussion

After having performed the *Chi-Square* test to check the statistical significance of our results, we can draw the following conclusions from the data presented in Tables 2–5. As far as referential distance is concerned, when we group the three categories together (i.e. personal pronoun, demonstrative determiners and demonstrative pronouns) the distribution observed is highly significant ($X^2$=29.999 (df = 8), p <0.0005); all three categories show a strong tendency to find their antecedents in the clause immediately preceding the anaphor ($CL_1$). Total frequencies are very similar for all three types of referring expressions: 77% (personal pronoun *lo*), 80%

(demonstrative determiners) and 79% (demonstrative pronouns). With only minor exceptions, a general tendency is observed that can be stated as follows: The higher the textual distance between the antecedent and the anaphor, the lower the frequency of occurrence of an antecedent-anaphor pattern.

As far as referential distance is concerned, our data show that all three anaphors show a strong preference to find their antecedent in CL1, so there appear to be no differences between demonstratives and the neuter personal pronoun in this respect. The personal pronoun *lo* though shows a somewhat significant rate of co-occurrence with antecedents in $CL_0$ (the clause containing the anaphor.) This is mainly due to a somewhat frequent Spanish construction that combines a demonstrative with the referential neuter personal pronoun *lo*. An example from the corpus is shown in (1).

(1) El ser humano es una bestia. Eso lo se hace años.
'The human being is a beast. I have known that since long.'

In this study, the occurrences of the neuter pronoun in this particular configuration have been included in the $CL_0$ group. The demonstrative pronoun *eso* refers back to the antecedent in the previous clause and the pronoun *lo*, in turn, has the demonstrative as antecedent. The relevance of this construction lies in the ability of both referring expressions to co-occur within the same clause and refer to the same discourse entity while having different morphological antecedents. Notice that the demonstrative pronoun occupies a highly prominent position within the sentence (subject), which is most commonly filled with topical elements. Also, the antecedent expression is highly salient as regards processing effort, recency of mention or memory retrieval, since it is introduced by the utterance closest to the demonstrative. In addition, the antecedent is not a subject or an object but a whole proposition. All these factors together lead us to suggest that the antecedent of the demonstrative is, contrary to expectations, a topical antecedent (e.g. either the general discourse topic or a local subtopic). Notice how the demonstrative in discourse (1) is immediately followed by the personal pronoun *lo*, which is commonly assumed to refer to highly topical entities. This co-occurrence is not obligatory, as is manifested by the ability of the pronoun *lo* to appear without the demonstrative in the same type of construction. This is shown in (2):

(2) Es algo incómodo revisar tu trabajo, pero (eso) ya lo tengo asumido.
'It is somewhat uncomfortable to revise your own work, but I have already accepted that.'

As regards the morphological type of the antecedent, we observed some highly significant distributions. When total figures for all three referring expressions are considered, we get extremely few demonstrative pronouns referring to NP antecedents ($X^2$=54.0238 (df = 2), p <0.0005). This is not surprising as demonstrative pronouns are most commonly used to refer to abstract entities in Spanish, so what this figure indicates is that abstract entities are most usually conveyed via clausal antecedents. Also, when we consider all demonstratives (determiners and pronouns together) and the neuter pronoun ($X^2$=24.4165 (df = 1), p <0.0005) we still get a very

strong preference for clausal antecedents over NPs (62.4% and 37.6%, respectively). We get similar frequencies when the neuter personal pronoun and demonstrative pronouns are compared (66.4% of clausal antecedents and 33.6% of NP antecedents, respectively) with a highly significant statistical significance ($X^2$=43.5814 (df = 1), p <0.0005). Although the neuter personal pronoun shows a slight preference for NP over clausal antecedents (55% and 45%, respectively), these figures are somewhat surprising, given that we did not expect to find so many cases of clausal antecedents for the neuter personal pronoun.

In view of these data, it appears that referential distance will not help us to discriminate among the referring properties of the expressions analyzed in this study. Let us recall that overall figures indicate that the preferred location of the antecedent is $CL_1$ for all the expressions involved. If we consider recency of mention as a factor to explain the information status of an antecedent, then we can conclude that there are no significant differences in the information status (i.e. in focus vs. activated) of the entity referred to with a neuter personal pronoun or a demonstrative expression. On the other hand, although figures indicate that demonstratives show a strong preference over the neuter personal pronoun for clausal antecedents, our data also show a high number of cases of clausal antecedents with the personal pronoun *lo* (55% and 45% of NP and clausal antecedents for the neuter personal pronoun, respectively, for n = 120).

## 5 Conclusions

A widely accepted thesis concerning the information status of referring expressions is that antecedents of demonstratives are most commonly non-topical whereas personal pronouns are commonly anteceded by topical elements. Topichood is commonly assumed to be dependent on syntactic configurations, that is, highly prominent positions (i.e. subject) are topical whereas less prominent syntactic positions (i.e. object or adjunct) are non-topical. When information status is defined in cognitive terms, it is commonly assumed that the referents of demonstratives occupy a lower position in terms of cognitive accessibility (activated) whereas personal pronouns mark their referents as highly accessible (in focus).

As regards Spanish in discourse anaphora/deixis uses, our main hypothesis is that demonstratives and the neuter personal pronoun do not differ much in the way they refer to discourse entities. We have studied two factors in a corpus of Spanish, which are directly related to the referential properties of these elements: distance of the antecedent and morphological type of the antecedent. The data indicate that antecedent distance is not a distinguishing factor as all the expressions analysed showed a strong preference to find their antecedents in the clause that is the closest to the anaphor. Thus, if we consider antecedent distance as a factor having an effect on the information status of the antecedent (i.e. most recent antecedents are more accessible than antecedents located at a greater distance), then demonstratives and the neuter personal pronoun show a very similar behavior. On the other hand, the resulting figures show that demonstrative pronouns have a clear preference for clausal antecedents over NP ones but we also found a significant number of cases of the neuter personal pronouns with clausal antecedents. This may be due to the

denotation of the referent involved, since the referents of clausal antecedents are most commonly abstract entities such as propositions, facts, events, etc.
The main empirical conclusions can be summarized as follows:

- Demonstratives and the neuter personal pronoun alike show a strong preference to find their antecedents in the clause closest to the anaphor ($CL_1$).
- Demonstratives and the neuter personal pronoun alike can refer to abstract entities (propositions, facts, etc.)
- Demonstratives show a stronger preference for clausal antecedents but the neuter personal pronoun shows a high number of clausal antecedents (45% out of total number of cases analysed of the neuter personal pronoun.)

In many respects, the results from our study coincide with previous cross-linguistic research in the sense that different languages show language-specific characteristics in the way they realize abstract pronominal anaphora. In particular, our data resemble the findings by Fraurud (1992) who found no differences between the Swedish anaphor *det* (ambiguous 'it/that/this') and the demonstrative anaphor *detta* ('this') in abstract reference; or the findings by Navarretta (2008) on Danish and Italian mentioned earlier. In our view, our data appear to confirm our initial hypothesis that the main role of Spanish demonstratives in discourse-anaphora/deixis uses involves marking (sub)-topic shifts in discourse. This procedure should be conceived as an instruction on the part of the speaker for the addressee to focus on a particular discourse entity and with a precise communicative intention, i.e., making her interlocutor aware that a topic shift is taking place or a new local subtopic has been introduced. In terms of cognitive or information status, demonstratives are devices used by speakers to bring entities into the current focus of attention. We defend that this focusing property closely resembles that of demonstratives in deixis proper (i.e. use of demonstratives to point to physical entities) or nuclear pitch accent in phonological focus marking. Additional support in favor of our hypothesis comes from a Spanish specific construction consisting of a demonstrative anaphor and a neuter personal pronoun both co-occurring within the same clause, next to one another and co-referential *eso lo* ('that it'); see examples (1) and (2).

## References


Mira Ariel. Referring and accessibility. *Journal of Linguistics*, 24(1): 65-87, 1988.

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

Mira Ariel. Accessibility theory: an overview. In T. Sanders, J. Schilperoord and W. Spooren, *Text Representation: Linguistic and Psycholinguistic Aspects*. Amsterdam: John Benjamins, pages 29-89, 2001.

Nicholas Asher. *Reference to Abstract Objects in Discourse*. Dortrecht: Kluwer, 1993.

Peter Bosch, Tom Rozario, Yufan Zhao. Demonstrative pronouns and personal pronouns. German *der* vs. *er*. *Proceedings of EACL2003, Workshop on the Computational Treatment of Anaphora*, 2003.

Donna Byron. Resolving pronominal reference to abstract entities. Technical report 815, University of Rochester, 2004.

Maria Nella Carminati. *The Processing of Italian Subject Pronouns*. PhD thesis. University of Massachussets, 2002.

Stefanie Dipper and Heike Zinsmeister. Annotating discourse anaphora. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP 2009, pages 166-169, 2009.

Barbara A. Fox. *Discourse Structure and Anaphora*. Cambridge, UK: Cambridge University Press, 1987.

Kari Fraurud. *Processing noun phrases in natural discourse*. PhD Dissertation, Stockholm University, 1992.

Herbert P. Grice. Presupposition and conversational implicature. In H. P. Grice (ed), *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press, pages 269-283, 1989.

Barbara J. Grosz, Scott Weinstein and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2): 203-225, 1995.

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

Jeanette K. Gundel, Nancy Hedberg and Ron Zacharski. Demonstrative pronouns in natural discourse. In A. Branco, T. McEnery and R. Mitkov (eds.), *Proceedings of DAARC 2004*, pages 81-86, 2004.

Jeanette K. Gundel, Nancy Hedberg and Ron Zacharski. Pronouns without NP antecedents: how do we know when a pronoun is referential? In A. Branco, T. McEnery & R. Mitkov (eds.), *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*. Amsterdam: John Benjamins, pages 351-365, 2005.

Javier Gutiérrez-Rexach and Iker Zulaica-Hernández. Abstract reference and neuter demonstratives in Spanish. In *Proceedings of DAARC 2007*, pages 25-30, 2007.

Michael Hegarty, Jeanette K. Gundel and Kaja Borthen. Information structure and the accessibility of clausally introduced referents. *Theoretical Linguistics*, 27: 163-186, 2001.

Michael Hegarty. Semantic types of abstract entities. *Lingua*, 113: 891-927, 2003.

Michael Hegarty, Jeanette K. Gundel and Kaja Borthen. Cognitive status, information structure and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, 12 (3): 281-299, 2003.

Michael Hegarty. Type shifting of entities in discourse. In K. von Heusinger and K. Turner (eds.), *Where Semantics meets Pragmatics, Current Research in the Semantics/Pragmatics Interface*, 16. Amsterdam, Elsevier, pages 111-128, 2006.

Elsi Kaiser and John C. Trueswell. Investigating the interpretation of pronouns and demonstratives in Finnish: going beyond salience. In E. Gibson and N. Pearlmutter (eds.), *The Processing and Acquisition of Reference.* Cambridge, MA, MIT Press, 2005.

Megumi Kameyama. Stressed and unstressed pronouns: complementary preferences. In P. Bosch and R. van der Sandt (eds.), *Focus, Linguistic, Cognitive and Computational Perspectives*. Cambridge: Cambridge University Press, pages 306-321, 1999.

Alfons A. Maes and Leo G. M. Noordman. Demonstrative nominal anaphors: a case of nonidentificational markedness. *Linguistics,* 33: 255-282, 1995.

Costanza Navarretta. Combining information structure and centering-based models of salience for resolving Danish intersentential pronominal anaphora. In A. Branco, T. McEnery and R. Mitkov (eds.) *Anaphora Processing. Linguistic, Cognitive and Computational Modeling*. Amsterdam: John Benjamins, pages 329-350, 2005.

Costanza Navarretta. A contrastive analysis of abstract anaphora in Danish, English and Italian. In A. Branco, T. McEnery, R. Mitkov and F. Silva (eds.) *Proceedings of DAARC 2007 - 6th Discourse Anaphora and Anaphora Resolution Colloquium*, pages 103-109, 2007.

Costanza Navarretta. Pronominal types and abstract reference in the Danish and Italian DAD corpora. In C. Johansson (Ed.), *Proceedings of the Second Workshop on Anaphora Resolution*, pages 63-71, 2008.

Costanza Navarretta and Sussi Olsen. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC '08)*, pages 2046-2052, 2008.

Massimo Poesio. The MATE/GNOME proposals for anaphoric annotation, revisited. In

*Proceedings of the 5th SIGDIAL Workshop*, pages 154-162, 2004.
Massimo Poesio and Natalia N. Modjeska. Focus, activation and this-noun phrases. In A. Branco, T. McEnery and R. Mitkov (eds.), *Anaphora Processing*. John Benjamins, pages 429-442, 2005.
Massimo Poesio, Patrick Sturt, Ron Artstein and Ruth Filik. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*, 42: 157-175.
Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC '08)*, pages 1170-1174, 2008.
Ellen F. Prince. Toward a taxonomy of given-new information. In P. Cole (Ed), *Radical Pragmatics*. New York: Academic Press, pages 223-256, 1981.
Marta Recasens. Discourse deixis and coreference: evidence from AnCora. In C. Johansson (Ed), *Proceedings of the Second Workshop on Anaphora Resolution (WAR II)*. NEALT Proceedings Series Vol. 2, pages 73-82, 2008.
Anne Sturgeon. Topic and demonstrative pronouns in Czech. In G. Zybatow, L. Szuchsich, U. Junghans and R. Meyer (eds.), *Formal description of Slavic languages*. Berlin: Peter Lang, 2008.
Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang and Gabriel Othero. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium*, pages 233-238, 2002.
Bonnie L. Webber. *A Formal Approach to Discourse Anaphora*. Garland, New York, 1979.
Bonnie L.Webber. Discourse deixis: reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 113-122, 1988.
Bonnie L.Webber. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2): 107-135, 1991.
Iker Zulaica-Hernández. *Demonstrative pronouns in Spanish: a discourse based study*. PhD dissertation, The Ohio State University, 2008.