University of Essex

Underspecification and Anaphora: Theoretical Issues and Preliminary Evidence

Massimo Poesio
University of Essex

Patrick Sturt
University of Edinburgh

Ron Artstein
University of Essex

Ruth Filik
University of Glasgow

Contents

Underspecification and Anaphora: Theoretical Issues and Preliminary Evidence

Massimo Poesio
University of Essex

Patrick Sturt
University of Edinburgh

Ron Artstein
University of Essex

Ruth Filik
University of Glasgow

Abstract   Much experimental work in psycholinguistics suggests that fully specified syntactic and semantic interpretations are obtained incrementally. The finding that intepretation takes place incrementally is very robust and underlies our own view of sentence processing as well; however, most of this work tends to test very simple interpretive judgments, and using materials which have very clean-cut interpretations, which makes the view expressed above more questionable when applied to semantic interpretation. This paper discusses a class of anaphoric expressions that do not appear to have a clear antecedent, using both corpus analysis and psychological experiments. We argue that these cases of anaphora are similar to cases of lexical polysemy, and propose an explicit semantic representation for such cases.

1   Introduction

There is very strong evidence that most language interpretation processes take place quickly and incrementally (Marslen-Wilson, 1975; Swinney, 1979; Frazier, 1987; Tanenhaus et al., 1995). Most such papers also suggest that the interpretation thus derived is fully disambiguated; however, this further conclusion tends to be based on very simplified views about ambiguity and interpretation. The starting point for the work presented in this paper are the results by Frazier and Rayner (1990), indicating the need to distinguish between two types of ambiguity—homonymy and polysemy. Frazier and Rayner discuss evidence suggesting that polysemous expressions are interpreted differently from homonymous expressions: when people process homonymous expressions, there is evidence for processing difficulty when subsequent text indicates the need for a reanalysis of the initially preferred interpretation. However, this pattern is not found for polysemous expressions. The first goal of the research project discussed in this paper is to carry out similar experiments for another type of semantic interpretation process, anaphora. We show evidence suggesting that in the case of anaphoric interpretation, as well, a distinction can be drawn between two types of 'ambiguity,' and that in one of these cases at least, the final interpretation may not be fully specified, but only 'good enough' in the sense of Ferreira et al. (2002). A second goal of our research project is to go further than Frazier and Rayner, or Ferreira and colleagues, in specifying what a 'good enough interpretation' may be; we make some preliminary theoretical suggestions in this respect. This paper builds on previous work (Poesio et al., 2001), but goes further in at least two directions: first of all, we undertook a large

1

scale experiment of ambiguity annotation, confirming that indeed these anaphoric expressions can be identified as ambiguous, and secondly, online studies were carried out.

The structure of the paper is as follows. We begin by discussing the previous papers just mentioned suggesting that initial interpretations may not be fully disambiguated (Frazier and Rayner, 1990; Ferreira et al., 2002), as well as our own preliminary work. In our own preliminary work we use corpora and corpus annotation to identify 'interesting' expressions from the point of view of semantic interpretation, then use behavioral studies to investigate the interpretation of these cases in detail. In Section 3 we report evidence from corpora about a class of anaphoric expressions that we call MEREOLOGICAL REFERENCES, which appear to be ambiguous in a different way from other types of anaphoric expressions. In Section 4 we show that these anaphoric expressions behave differently from other anaphoric expressions also from the point of view of offline studies. Finally, in Section 5, we consider what the initial 'good enough' interpretation of these expressions may be.

2   Background

Much experimental work in psycholinguistics suggests that fully specified syntactic and semantic interpretations are obtained incrementally (Marslen-Wilson, 1975; Swinney, 1979; Frazier, 1987; Tanenhaus et al., 1995). This finding is very robust and underlies our own view of interpretation as well. However, most of this work tends to test very simple interpretive judgments, and using materials which have very clean-cut interpretations. The work reviewed in this section, by contrast, suggests that when more complicated judgments are tested, this view of interpretation may need to be revised.

2.1   Frazier and Rayner: Polysemy vs Homonymy

Work on lexical access (Swinney, 1979; Simpson, 1994) tends to focus on HOMONYMY—cases of lexical ambiguity like *pitcher* or *records*, in which the two interpretations of a word are clearly distinct (often also from an etymological perspective), and can be argued to be associated with distinct lexical entries. However, Frazier and Rayner (1990) compared the interpretation of such words with the interpretation of POLYSEMOUS words such as *newspaper*, whose senses are closely interrelated,[1] and found that these two types of lexical ambiguity could be experimentally separated. Whereas garden paths could be observed when the initially preferred interpretation of an homonym such as *record* was disconfirmed by subsequent context, as in (1d), no garden paths were observed for polysemous words like *newspaper* in (2d).

(1)   a.   *After they were scratched, the records were carefully guarded.*

       b.   *After the political takeover, the records were carefully guarded.*

       c.   *The records were carefully guarded after they were scratched.*

       d.   *The records were carefully guarded after the political takeover.*

(2)   a.   *Lying in the rain, the newspaper was destroyed.*

       b.   *Managing advertising so poorly, the newspaper was destroyed.*

       c.   *Unfortunately the newspaper was destroyed, lying in the rain.*

       d.   *Unfortunately the newspaper was destroyed, managing advertising so poorly.*

---

[1]According to WordNet 2.1, *newspaper* has four senses. The two senses considered by Frazier and Rayner are WordNet sense 2, "a business firm that publishes newspapers", and 3, "the physical object that is the product of a newspaper publisher". Sense 1 is "a daily or weekly publication on folded sheets; contains news and articles and advertisements" and Sense 4 is "cheap paper made from wood pulp and used for printing newspapers".

Although Frazier and Rayner do not explain what the initial interpretation of *newspaper* may be, their study clearly suggests that in cases where such a 'preliminary' interpretation may be available, the requirement that all phrases are immediately interpreted may be weakened; Frazier and Rayner call this the IMMEDIATE PARTIAL INTERPRETATION HYPOTHESIS. They also suggest that this initial partial interpretation in some sense 'covers' the disambiguated interpretations.[2]

2.2   Good enough syntactic representations

Ferreira et al. (2002) discuss several experiments challenging "...the assumption that utterance interpretations are compositionally built up from words clustered into hierarchically organized constituents". In one such experiment, Duffy et al. (1989) measured naming times for the final word in sentences such as those in (3), finding that the naming times for the word *cocktails* were significantly slower in (3b) than in (3a) (in which *bartender* serves as a prime for *cocktails*). However, they also found no difference between the naming times for (3a) and (3c).

(3)   a.   The boy watched the *bartender* serve the <u>cocktails</u>.

   b.   The boy saw that the *person* liked the <u>cocktails</u>.

   c.   The boy who watched the *bartender* served the <u>cocktails</u>.

According to Ferreira et al., this suggests that priming, one task usually taken to operate at the 'conceptual' level, in fact either operates at a level in which syntactic structure is not realized, or at the very least is not affected by syntactic structure.[3] Further evidence questioning the extent to which semantic interpretation is affected by syntactic structure was reported in a second series of experiments cited by Ferreira et al. In one study, Christianson et al. (2001) were concerned with the question of whether after reanalysis people delete the initial incorrect interpretation. They tested this by presenting their subjects with garden path sentences such as (4):

(4)   While Anna dressed the baby played in the crib.

and afterwards asking them questions such as *Did the baby play in the crib?* (which should be answered positively if *the baby* is successfully re-interpreted as the subject of *played*) and *Did Anna dress the baby* (which, should be answered negatively if *dressed* is successfully re-interpreted as an intransitive verb). What Christianson et al. found is that both questions tended to be answered positively, compared with unambiguous controls. This shows that part of the final interpretation persisted from the initial misanalysis, and was not consistent with the globally correct analysis of the sentence. Therefore, either the participants did not always complete the syntactic reanalysis of the sentence, or they did not always complete the change of interpretation following syntactic reanalysis.

On the basis of these and other experiments, one could argue that semantic interpretation may not always be entirely parasitic on syntactic structure, or perhaps that subjects sometimes only derive incomplete syntactic representations, which are 'good enough' for their purposes (Ferreira et al.'s conclusion). No specific hypotheses concerning the form of these representations are made by Ferreira et al.

---

[2]Any lexicographer will point out that matters are much more complicated, and that the distinction between homonymy and polysemy is sometimes very difficult to draw (Pinkal, 1995; Lyons, 1995). We take this point to provide further support for our main point, that the findings about incremental interpretation may be based on an overly simplified view of interpretation.

[3]However, see Morris (1994) for contrasting results using eye fixation times as a measure of priming.

2.3   Underspecification and anaphoric reference

Anaphoric reference is generally viewed as a case of H-AMBIGUITY (Pinkal, 1995; Poesio, 1999): the type of ambiguity in which the alternative interpretations of an expression are semantically distinct, as in the case of homonymy. Accordingly, most evidence on anaphoric interpretation suggests that it is quick and incremental, both in the case of definite descriptions (Tanenhaus et al., 1995) and in the case of pronouns and reflexives (Arnold et al., 2000; Sturt, 2003) as would be predicted by Frazier and Rayner's Immediate Partial Interpretation Hypothesis, although there is some evidence for a delay in processing of pronouns (Corbett and Chang, 1983; Garrod et al., 1994). However, the class of ambiguous anaphoric reference investigated in this project do not appear to suffer from some of the problems caused by ambiguity, resembling in this the cases of lexical access studied by Frazier and Rayner. In addition, it makes intuitive sense that these expressions may only be assigned a 'good enough' interpretation, like the cases reported by Ferreira et al.

Poesio et al. (2001) argued that in dialogue corpora, a fairly powerful test can be used to find candidates for underspecification: look for anaphoric expressions that (i) don't seem to have a preferred antecedent, yet (ii) do not result in the listener signalling a misunderstanding. Poesio et al. carried out an analysis of the anaphoric expressions in the TRAINS corpus collected at the University of Rochester (Gross et al., 1993), identifying several classes of expressions passing this test. We concentrate here on what Poesio et al. called the MEREOLOGICAL cases. An example of this class is the pronoun *it* in utterance 5.1 in the fragment in (5).

```
(5)   3.1      M: can we .. kindly hook up
      3.2       : uh
      3.3       : engine E2 to the boxcar at .. Elmira
      4.1      S: ok
      5.1      M: +and+ send it to Corning
      5.2       : as soon as possible please
      6.1      S: okay
                  [2sec]
      7.1      M: do let me know when it gets there
      8.1      S: okay it'll /
      8.2       : it should get there at 2 AM
      9.1      M: great
      9.2       : uh can you give the
      9.3       : manager at Corning instructions that
      9.4       : as soon as it arrives
      9.5       : it should be filled with oranges
     10.1      S: okay
     10.2       : then we can get that filled
```

In this example, it's not clear whether the pronoun *it* in 5.1 refers to *the engine E2* which has been hooked up to *the boxcar at Elmira*, to the boxcar itself, or indeed whether that matters. It's only at utterance 9.5 that we get evidence that *it* probably referred to *the boxcar at Elmira*, since it is only boxcars that can be filled with oranges; yet, if anything, focusing theories would predict engine E2 to be the antecedent, since engine E2 is the direct object, the THEME, and comes first (Sidner, 1979; Stevenson et al., 1994; Grosz et al., 1995; Pearson et al., 2001). This didn't only happen with the verb *hook up*, but more generally whenever two objects were put together, e.g., by loading a cargo into some sort of transport vehicle, as in the following example:

```
(6)   26.1   S: okay
      27.1   M: so then we'll
      27.2    : ... we'll be in a position to
      27.3    : load the orange juice into the tanker car
      27.4    : ... and send that off
```

There are important similarities and differences between these cases of anaphoric ambiguity and the cases studied by Frazier and Rayner. Just like the cases of lexical polysemy studied by Frazier and Rayner (and unlike the cases of lexical homonymy), mereological pronouns can receive an interpretation that 'covers' both possible antecedents: the 'train' formed by the engine and the boxcar in (5), the 'tanker car full of orange juice' in (6). As said above, this is not normally the case with anaphoric expressions, so we would expect to find a contrast between mereological pronouns and cases in which no such 'merged' interpretation has been created. However, there is also an important difference: whereas in the cases studied by Frazier and Rayner the existence of a superinterpretation was part of lexical knowledge, in the case of mereological interpretation the merged interpretation is the result of actions discussed in the text.[4]

Mereological pronouns are also an interesting case of expression for which only a 'good enough' interpretation may be reached, for in this case, we can make a fairly precise hypothesis concerning what this interpretation may be. From the point of view of the plan, once the engine and the boxcar are hooked up, or once the orange juice is loaded in the tanker car, the two objects will move together. So strictly speaking, all that the listener needs to do in this case is to somehow restrict the interpretation to the object obtained by joining together the two possible antecedents and its parts: that will be 'good enough' for her purposes. (This idea is made more precise in Section 5.)

3   Corpus evidence

Poesio et al. (2001) only carried out informal studies of the mereological cases. We carried out a more systematic analysis of the data in the corpus, through an annotation experiment with many naïve subjects (Poesio and Artstein, 2005). In this study, our annotators can mark ambiguity by indicating multiple antecedents for an anaphoric expression, to reveal whether indeed subjects perceived the mereological cases as ambiguous. In addition to this *explicit* marking of ambiguity, we also assessed *implicit* ambiguity, where the pattern of annotation by multiple subjects revealed that an item was ambiguous even when the annotators were not aware of this themselves.

3.1   Methods

Coding Scheme.   The coding manual used in this experiment is based on the approach to anaphoric annotation developed in MATE (Poesio et al., 1999) and GNOME (Poesio, 2004). The task and instructions

---

[4] Note that this type of pronoun use cannot be viewed as an example of vagueness, at least not according to the standard vagueness tests (Lakoff, 1970; Zwicky and Sadock, 1975). Whereas the ellipsis test, for example, would suggest that *a glove* in (7a) is indeterminate, because it's possible for John to have lost its left glove and for Bill to have lost its right one, the same test applied to *it* in (7b) suggests that this expression is ambiguous, in that it's not possible to interpret *Then, John should check if IT gets to Bath in time, and Bill should too* as meaning that John should check that the engine gets to Bath in time, whereas Bill should check that the boxcar gets there in time.

(7)   a.   *John lost a glove, and Bill did too*

      b.   *Let's hook the engine to the boxcar.*
           Then, John should check if IT gets to Bath in time, and Bill should too

were simplified, the primary simplification being that we did not annotate bridging references; on the other hand, we added instructions for marking multiple antecedents for ambiguous anaphoric expressions, and a simple way for marking discourse deixis.

Materials. The data used for this coding experiment are the first edition of the TRAINS corpus collected at the University of Rochester (Gross et al., 1993). This corpus consists of transcripts of dialogues between two humans. One of the humans plays the 'manager' of a railway company, with the aim of developing a plan to achieve a transportation goal (delivering a certain amount of goods at a given town by a given deadline). The other participant in the dialogue plays a 'system', and his[5] role is to help managers develop this plan by providing them with required information such as timetables and equipment availability. The text annotated in the experiment was dialogue 3.2 from the corpus. This dialogue contains 16 instances of *it* (counting all tokens, whether referential, expletives, or of unclear use) and 16 instances of *that*. Subjects were trained on dialogue 3.1.

Annotation tools. The subjects entered their annotations directly on a computer, using the MMAX 2 annotation tool (Müller and Strube, 2003).[6]

Subjects. Eighteen paid subjects participated in the experiment, all students at the University of Essex, mostly undergraduates from the Departments of Psychology and Language and Linguistics. They were paid £30 for their participation.

Instructions. Subjects were instructed to go through the phrase markables in order, using MMAX 2's markable browser, and to assign each markable to one of the following four classes.

phrase: a markable referring to an object mentioned earlier in the dialogue;

segment: a markable referring to a plan, event, action, or fact discussed earlier in the dialogue;

place: one of the five railway stations in the 'TRAINS world'—Avon, Bath, Corning, Dansville, and Elmira—explicitly mentioned by name (this was done to save time, since pointers between explicit place names are not particularly interesting);

none: a markable that does not fit any of the above criteria, for instance one referring to a novel object, or a non-referential noun phrase.

For markables designated as phrase or segment, subjects were instructed to create a POINTER to the antecedent—a phrase markable or a dialogue turn. In case an expression was considered ambiguous, subjects were instructed to create more than one pointer. (Due to the limitations of the coding scheme they could only indicate ambiguity between two phrase antecedents or two discourse antecedents, not an ambiguity between a phrase antecedent and a discourse antecedent.)

3.2 Results

There was perfect agreement among the 18 annotators on 65 of the 148 markables in the dialogue (43.9%), and near perfect agreement (no more than two disagreeing coders) on another 14 markables

---

[5]The role of the system in all of the TRAINS 91 corpus was played by Derek Gross.
[6]Available from http://mmax.eml-research.de/

(9.5%)—in total, there were no real disagreements on 53.4% of markables. Of the remaining markables, 67 (45.3%) were annotated with at least two antecedents chosen by at least two coders each, which we take as a sign of implicit ambiguity (the remaining two items were assigned a single label by 14 or 15 coders, and distinct labels by each of the remaining coders). Of the 67 implicitly ambiguous markables, 24 (16.2% of the total number of markables) were marked as explicitly ambiguous by at least one annotator (overall there were 131 explicit ambiguity markings by all coders). In addition, 27 of these markables (18.2%) revealed an implicit ambiguity between a discourse-old and a discourse-new interpretation—an ambiguity that coders were not able to mark explicitly in our scheme. These results are summarized in Table 1.

	First Half	Second half	Total
Number of markables	72	76	148
Perfect agreement	27 (37.5%)	38 (50.0%)	65 (43.9%)
Almost perfect	10 (13.9%)	4 (5.3%)	14 (9.5%)
Ambiguous (total)	35 (48.6%)	32 (42.1%)	67 (45.3%)
Explicit ambiguity	18 (25.0%)	6 (7.9%)	24 (16.2%)
Old/new ambiguity	8 (11.1%)	19 (25.0%)	27 (18.2%)

Table 1: Coder judgments concerning ambiguity in TRAINS 91 Dialogue 3.2

The results of the experiment clearly show that annotators are able to identify ambiguity and mark it: of the 18 annotators, only 3 did not mark any item as explicitly ambiguous. However, ambiguity is seldom noticed by many coders. Of the 24 items marked as ambiguous by at least two coders (38 by at least one coder), only 6 were marked as ambiguous by seven coders or more, and 5 of those were mereological cases. The distribution of items according to the number of coders that marked them as ambiguous is shown in Figure 1.
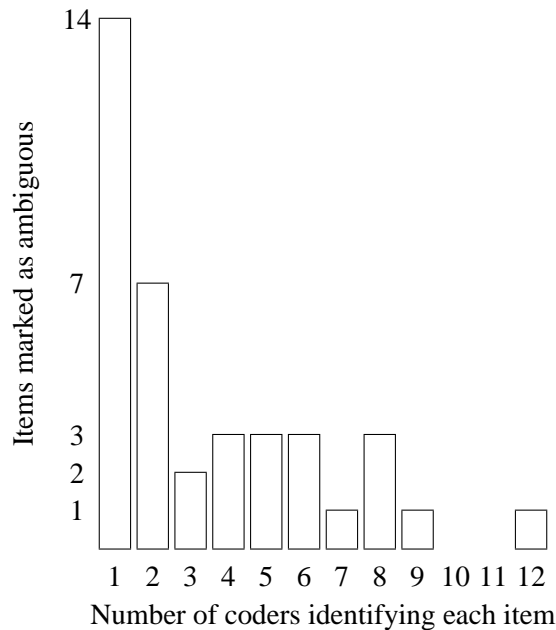

Figure 1: Distribution of ambiguity markings

The dialogue used in this study contains 8 instances of *it* or *that* (about 25% of the total number of tokens of these types) that, according to our own judgment, fit the 'mereological reference' pattern discussed in Section 2.3. These cases were all classified as explicitly ambiguous by at least 2 coders.[7] We can see an example of an explicitly marked ambiguous 'mereological' pronoun in utterance unit 3.1 of the following fragment, where the two *it* pronouns display the same type of ambiguity already seen in (5): they could refer to engine E2, the boxcar, or both.

```
(8)   1.4    M: first thing ˍI'dˍ like you to do
      1.5       is send engine E2 off with a boxcar to Corning
                to pick up oranges
      1.6       uh as soon as possible
      2.1    S: okay [6 sec]
      3.1    M: and while it's there it should pick up the tanker
```

All of our subjects considered the two *it* pronouns to be a 'phrase' reference. The first pronoun was marked by 9 coders as explicitly ambiguous between engine E2 and the boxcar, by 6 coders as unambiguous and referring to engine E2, and by 3 coders as unambiguous and referring to the boxcar. The second pronoun was marked by 8 coders as ambiguous between engine E2 and the boxcar—2 did so directly, and 6 marked the previous, ambiguous pronoun as an antecedent; the remaining coders marked it as unambiguous—8 as referring to engine E2, and 2 as referring to the boxcar. Thus, even without explicit marking of ambiguity, we would be able to infer, implicitly, that these pronouns are ambiguous.

In addition, the dialogue contained two examples of anaphoric references to a mereological structure which was not the result of an action performed in the plan, but was pre-existing. These two examples are shown in the following fragment.

```
(9)  18.6    S: it turns out that the boxcar at Elmira
     18.7       has a bad wheel
     18.8       and they're .. gonna start fixing that at midnight
     18.9       but it won't be ready until 8
     19.1    M: oh what a pain in the butt
```

Just as in the mereological cases, it is not clear to us whether the pronouns *that* and *it* in utterance units 18.8 and 18.9 refer to the boxcar or to its bad wheel. However, these cases were treated differently by our annotators. None of our annotators identified the pronoun *that* as ambiguous: 15 marked it as referring to the wheel, and one as referring to the boxcar (two others marked it with a discourse antecedent). The pronoun *it* was marked as ambiguous by a single annotator; 11 marked it as referring to the boxcar alone, and 6 to the wheel alone.

In summary, this study shows that mereological references are fairly common, at least in this type of dialogue, and that in each case, alternative interpretations are indeed available: in fact, no particular preference can be found among the interpretations annotated. We also found that even with limited training it is possible to get coders to identify many of these ambiguous cases, which suggests that a systematic study of this phenomenon is possible. What the study does *not* tell us is whether these mereological cases are somehow problematic during interpretation, and how they are interpreted; this is the task of behavioral studies, to which we now turn.

---

[7]The dialogue exhibits two other types of anaphoric ambiguity: ambiguity in discourse deictic references to actions / plans, and ambiguity between discourse new / discourse old interpretations.

4 Psychological evidence

4.1 Offline evidence

Poesio et al. (2001) tested whether sentences that contain a potentially ambiguous anaphoric reference are easier to process when the two potential antecedents are part of a single mereological structure which makes the possible interpretations equivalent for the purposes of the plan, compared to a condition where the two potential antecedents are not joined. Poesio et al. tested this hypothesis using the offline Magnitude Estimation technique proposed in Bard et al. (1996). They asked participants to judge whether 'MEREOLOGY' sentences such as (10a), which depicts *the engine* and *the boxcar* being attached together, are 'more acceptable' (in that less ambiguous) than 'NON-MEREOLOGY' sentences like (10b), in which *the engine* and *the boxcar* are not attached together:

(10) a. The engineer hooked up the engine to the boxcar and sent it to London.

b. The engineer separated the engine from the boxcar and sent it to London.

The experiment was run using WebExp, a software package for running experiments on the Web developed at the Universities of Edinburgh and Saarbruecken. A significant effect of MEREOLOGY was found, such that mereological sentences like (10a) were judged to be reliably more acceptable than non-mereology sentences like (10b). The participants were also asked to judge the acceptability of the initial parts of the sentences, to ensure that any differences in acceptability were not due to uncontrolled features of the sentences before the anaphoric section:

(11) a. The engineer hooked up the engine to the boxcar.

b. The engineer separated the engine from the boxcar.

It was found that the shorter control sentences like (11a) and (11b) did not differ in acceptability, leading to an interaction between mereology and sentence length. Therefore Poesio et al. (2001) concluded that the differences between (10a) and (10b) were indeed due to the anaphoric content of the sentences, and they interpreted the difference in terms of a perceived difficulty associated with the referential ambiguity in the non-mereological case (10b). The lack of such perceived difficulty in (10a) was interpreted as evidence for underspecification.

The off-line acceptability experiment left at least two questions open: how do the patterns of acceptability map on to on-line processing, and what was the interpretation assigned to the anaphoric expression?

4.2 Online evidence

Filik et al. (2005) conducted an eye-tracking study to examine the on-line processing of anaphoric references to mereological structures. Experimental items were as in (12):

(12) a. **Mereology/singular**
There were many delays. / The railwayman hooked up / the engine and the boxcar, / and sent it / quickly/ to the central station.

b. **Mereology/plural**
There were many delays. / The railwayman hooked up / the engine and the boxcar, / and sent them / quickly/ to the central station.

c. **Neutral/singular**
There were many delays. / The railwayman saw / the engine and the boxcar, / and sent it / quickly/ to the central station.

d.   **Neutral/plural**

There were many delays. / The railwayman saw / the engine and the boxcar, / and sent them / quickly/ to the central station.

The experimental design involved a manipulation of whether the main predicate was mereology-constructing (*hooked up*) or neutral (*saw*). The experiment also manipulated whether the subsequent pronoun was singular *it* or plural *them*. Previous work has found that a singular pronoun causes difficulty when it is forced to refer to a single conjunct in a coordinated noun phrase (the so-called "conjunction cost") while a plural pronoun is relatively easy in such cases (Albrecht and Clifton, 1998; Moxey et al., 2004). Therefore more processing difficulty is expected following the pronoun in (12c) than in (12d), as these are the two neutral conditions, involving non-mereological predicates. From the point of view of underspecification, the interesting contrast is between (12a) and (12b), which involve mereological predicates. Here, the prediction is that the mereological structure makes alternative interpretations for a singular pronoun available, for the reasons discussed above. This should eliminate the conjunction cost, making (12a) relatively easy to process compared with (12c).

Eye-movement data from the region containing the pronoun (*and sent it*) showed that singular reference was indeed easier for mereology-constructing than neutral sentences: there were more initial regressions from this region in the singular neutral condition than the singular mereology condition. A similar pattern of effects was found in the final region of the sentence (*to the central station*) in regression path time (the time taken to "go past" the region). It was also the case that plural reference was equally easy to process for mereology constructing and neutral sentences. One possible reason for this is that the use of the conjoined noun phrase *the engine and the boxcar* may automatically make a plural reference object available (Moxey et al., 2004), facilitating subsequent reference through a plural pronoun, whether a mereology constructing predicate is used or not.

Filik et al. further investigated this issue using the TEXT CHANGE DETECTION paradigm (Sanford and Sturt, 2002). In this procedure, participants read a piece of text at their own pace. This piece of text is then shown for a second time, with a possible change to one of the words. The task for the participant is to signal whether or not they noticed this change, and if so, which word it is that has changed. The likelihood of noticing a change may reflect the level of specification applied to the readers' mental representation of this word. If fewer changes are detected this may indicate that readers have underspecified their representations in some way. In the current change detection study, participants read sentences based on those used in the eye-tracking study described above. In these sentences, the plural pronoun 'them' appeared in the first display of the text, and was changed to the singular pronoun 'it' for the second presentation. Results showed that readers noticed fewer changes from 'them' to 'it' in mereology-constructing than neutral sentences, suggesting that readers may have underspecified the referent of the pronoun in mereology-constructing cases, supporting the results of the eye-tracking study.

A questionnaire study was also carried out to investigate the more specific hypothesis discussed in Section 5—see below.

5   The Justified Sloppiness Hypothesis

Up until now, we have discussed evidence suggesting that cases of anaphoric reference to structured entities such as those discussed above do indeed behave differently from other cases of anaphoric ambiguity. We have hinted that this may be because, as in the cases of lexical polysemy, an interpretation that supersedes all alternatives appears to be available, and that assigning this interpretation to the pronoun may be 'good enough' for the purposes of the plan. In this section we will attempt to

characterize in a more systematic fashion the possible interpretations of these mereological pronouns, borrowing some notation from the theory of plurals and parts proposed by Link (1983).

We will write $oj \oplus tc$ to indicate the object that has oj (e.g. orange juice) and tc (e.g. tanker car) as subparts, and $a \lhd b$ to say that a is a mereological part of b. With this notation, we can formalize the first and most obvious property of examples (5) and (6): namely, that actions like *hooking up* and *loading* are performed that create a new object $a \oplus b$ out of the potential antecedents a and b (e.g., $oj \oplus tc$ in (6)).

The second property of these examples is that four interpretations for the pronominal expression are possible. The complete list of the possible interpretations of *that* in (6), 27.4 is as follows:

1. the orange juice, oj;

2. the tanker car, tc;

3. the composite object formed by loading the orange juice in the tanker car, $oj \oplus tc$;

4. or an indeterminate $x \lhd (oj \oplus tc)$

This latter interpretation ($x \lhd (oj \oplus tc)$) is what has been called a P-UNDERSPECIFIED INTERPRETATION in Poesio (1999)—i.e., a 'disjunctive' interpretation that 'covers' all of the alternative interpretations, similar to those proposed for certain cases of lexical polysemy in Copestake and Briscoe (1995). We will hypothesize below that the existence of such an underspecified interpretation may be a further important property of these contexts. Either this interpretation or interpretation 3, the composite object, may be what Frazier and Rayner had in mind when they talked of an 'interpretation that covers all alternatives'.

The third property that these examples have in common is that both in situations involving attaching two objects together and in situations involving loading objects into other objects, all of the alternative interpretations of the anaphoric expression are equivalent as far as the plan of moving these objects to a new location is concerned: after the two explicitly mentioned potential antecedents are joined, if one of them gets moved, the other one must be moved as well. E.g., in (6), 27.4, all interpretations of the instruction *send that off* will achieve the same result irrespective of how the pronoun is interpreted. We will write X ~ Y to indicate that interpretation X is equivalent to interpretation Y for the purpose of the plan: i.e., we write

$$oj \sim tc$$

to indicate that from the point of view of the plan, interpreting the pronoun *that* as referring to the orange juice or the tanker car are equivalent. Similarly, in the case of (5), we will write $e \sim b$ to say that from the point of view of the plan, the interpretation of *it* in which it refers to engine E2 and that in which it refers to the boxcar are equivalent. In fact, in this case, *all* four situations discussed above are equivalent in the sense just discussed: they are 'good enough' in the sense of Ferreira et al.

To summarize, the mereological references in the TRAINS dialogues have at least three aspects in common:

1. Both explicitly mentioned potential antecedents x and y are elements of an underlying mereological structure with summum $\sigma = x \oplus y$ which has been explicitly constructed (and made salient) in the dialogue ($\sigma = oj \oplus tc$ in (6));

2. The existence of this structure makes it possible to construct a p-underspecified interpretation

in which the anaphoric expression is interpreted as denoting an element z included in the mereological structure—i.e., part-of its summum σ:

$$
\begin{array}{|l|}
\hline
x \; y \; \sigma z \\
\hline
\ldots \\
\sigma = x \oplus y \\
z \lhd^* \sigma \\
\ldots \\
\hline
\end{array}
$$

3. All possible interpretations $(x, y, z, x \oplus y)$ are equivalent for the purposes of the plan.

The evidence discussed in the previous sections may now be viewed as providing support for the following generalization:

Justified Sloppiness Hypothesis Ambiguous anaphoric expressions are not perceived as infelicitous provided that Conditions 1–3 hold.

This may be because if these three conditions hold, the speaker's sloppiness in using an anaphoric expression in an ambiguous context is not problematic; we will therefore use the term JUSTIFIED SLOPPINESS to indicate cases such as those discussed in the previous section, and refer to the hypothesis above as the JUSTIFIED SLOPPINESS HYPOTHESIS.

Of course, the fact that a p-underspecified interpretation exists does not mean that the listener will adopt it as its final interpretation; however, this possibility is what makes these examples interesting from an underspecification perspective. Some evidence that either the SUM interpretation (*engine* $\oplus$ *boxcar*) or the underspecified representation $z \lhd^*$ *engine* $\oplus$ *boxcar* (where, in this context, z can be instantiated either as the engine or the boxcar, or the two combined) are provided by an offline questionnaire run by Filik et al. The questionnaire showed that people tend to prefer the SUM interpretation of *it* in the context of (12a), and that they only rarely chose *the engine* or *the boxcar* as the referent on its own. However, this asymmetry may be due to the use of the conjunction; as mentioned above, we already know that single conjuncts of coordinated noun phrases are relatively inaccessible to subsequent singular reference. If this preference holds also in the context of mereological predicates, it is possible that the SUM reference is the only element of the underspecified interpretation that remains salient in the context of conjunction.

6   Discussion

The evidence presented in this paper suggests, first of all, that linguists (computational and theoretical) and psychologists ought to pay attention to cases in which the interpretation of an ambiguous expression is not clear-cut, as these are fairly numerous, and subjects *do* find them ambiguous, at least implicitly. Secondly, our findings, particularly from behavioral studies, extend to anaphoric ambiguity what we take to be the main point made by Frazier and Rayner, namely, that not all types of ambiguity are alike; the existence of a 'super-interpretation' appears to make certain types of pronominal usages more acceptable. An interesting difference in the cases of anaphoric reference we are studying is that this 'super-interpretation' is made available through context instead of lexically, as in the cases studied by Frazier and Rayner.

One of the goals of this project is to make more specific hypotheses about what these 'super interpretations' or 'good enough interpretations' may be. We sketched out in some detail what we think

the alternatives might be, and discussed a several bits of evidence that, while not conclusive, we feel will shed some light on this point. Some evidence that either the SUM interpretation (*engine*⊕*boxcar*) or the underspecified representation $z \triangleleft^* engine \oplus boxcar$ (where, in this context, z can be instantiated either as the engine or the boxcar, or the two combined) is provided by the offline questionnaire run by Filik et al. (2005).

Further research will examine in more detail the effect of mereological predicates in sentences with and without coordination. However, finding ways of discriminating between the 'underspecified' and the 'sum' reading is likely to prove difficult.

Given the available evidence, the Justified Sloppiness Hypothesis is the strongest generalization that we may draw at the moment. The Justified Sloppiness Hypothesis is weaker than Frazier and Rayner's Immediate Partial Interpretation Hypothesis (which is specifically an hypothesis about interpretation), but may still be applicable to cases other than mereological references: e.g., cases of reference to events and other abstract objects (see Poesio et al. (2001)).

Acknowledgments

References

Albrecht, J. and Clifton, C., J. (1998). Accessing singular antecedents in conjoined phrases. *Memory and Cognition*, 26(3):599–610.

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., and Trueswell, J. C. (2000). The immediate use of gender information: eyetracking evidence of the time-course of pronoun resolution. *Cognition*, 76:B13–B26.

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Christianson, K., Hollingworth, A., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42:368–407.

Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12(1):15–68. Special Issue on Lexical Semantics.

Corbett, A. and Chang, F. (1983). Pronoun disambiguating: Accessing potential antecedents. *Memory and Cognition*, 11:283–294.

Duffy, S. A., Henderson, J. M., and Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15:791–801.

Ferreira, F., Bailey, K. G. D., and Ferraro, V. (2002). Good enough representations in language comprehension. *Current Directions in Psychological Science*, 11:11–15.

Filik, R., Sanford, A., and Sturt, P. (2005). Anaphoric reference to structured entities. Manuscript in Preparation.

Frazier, L. (1987). Sentence processing: a tutorial review. In Coltheart, M., editor, *Attention and performance XII: The Psychology of Reading*, pages 559–586. Erlbaum, Hove.

Frazier, L. and Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29:181–200.

Garrod, S. C., Freudenthal, D., and Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, 32:1–30.

Gross, D., Allen, J., and Traum, D. (1993). The TRAINS 91 dialogues. TRAINS Technical Note 92-1, Computer Science Dept. University of Rochester.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202–225. (The paper originally appeared as an unpublished manuscript in 1986.).

Lakoff, G. P. (1970). A note on vagueness and ambiguity. *Linguistic Inquiry*, 1(3):357–359.

Link, G. (1983). The logical analysis of plurals and mass terms: A lattice- theoretical approach. In Bäuerle, R., Schwarze, C., and von Stechow, A., editors, *Meaning, Use and Interpretation of Language*, pages 302–323. Walter de Gruyter.

Lyons, J. (1995). *Linguistic Semantics*. Cambridge University Press, Cambridge.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science*, 189:226–228.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20:92–103.

Moxey, L. M., Sanford, A. J., Sturt, P., and Morrow, L. (2004). Constraints on the formation of plural reference objects: the influence of role, conjunction, and the type of description. *Journal of Memory and Language*, 51:346–364.

Müller, C. and Strube, M. (2003). Multi-level annotation in MMAX. In *Proc. of the 4th SIGDIAL*, pages 198–207.

Pearson, J., Stevenson, R., and Poesio, M. (2001). The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. In Poesio, M., editor, *Proc. of the First Workshop on Cognitively Plausible Models of Semantic Processing (SEMPRO)*. University of Edinburgh, HCRC.

Pinkal, M. (1995). *Logic and Lexicon*. D. Reidel, Dordrecht.

Poesio, M. (1999). Utterance processing and semantic underspecification. HCRC/RP 103, University of Edinburgh, HCRC.

Poesio, M. (2004). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*, Boston.

Poesio, M. and Artstein, R. (2005). Annotating (anaphoric) ambiguity. In *Proc. of the Corpus Linguistics Conference*, Birmingham.

Poesio, M., Bruneseaux, F., and Romary, L. (1999). The MATE meta-scheme for coreference in dialogues in multiple languages. In Walker, M., editor, *Proc. of the ACL Workshop on Standards and Tools for Discourse Tagging*, pages 65–74.

Poesio, M., Reyle, U., and Stevenson, R. (2001). Justified sloppiness in anaphoric reference. In Bunt, H. and Muskens, R., editors, *Computing Meaning 3*. Kluwer. To appear.

Sanford, A. J. and Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Science*, 6:382–386.

Sidner, C. L. (1979). *Towards a computational theory of definite anaphora comprehension in English discourse*. PhD thesis, MIT.

Simpson, G. B. (1994). Context and the processing of ambiguous words. In Gernsbacher, M. A., editor, *Handbook of Psycholinguistics*, chapter 10, pages 359–374. Academic Press.

Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus, and the representation of events. *Language and Cognitive Processes*, 9:519–548.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18:545–567.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Zwicky, A. and Sadock, J. (1975). Ambiguity tests and how to fail them. In Kimball, J., editor, *Syntax and Semantics 4*, pages 1–36. Academic Press, New York.

Other Technical Notes and theses from the Natural Language Engineering and Web Applications group are available electronically at