Ambiguity in Anaphora Workshop Proceedings, pages 13–21
ESSLLI 2006, Málaga, Spain, 7–11 August 2006
Edited by Ron Artstein and Massimo Poesio
http://cswww.essex.ac.uk/anaphora/

On the (Ir)relevance of Psycholinguistics for Anaphora Resolution

Lucas Champollion
champoll [at] ling.upenn.edu

Abstract

Psycholinguistic experiments show that pronouns tend to be resolved differently depending on whether they occur in main or subordinate clauses. If a pronoun in a subordinate clause has more than one potential antecedent in the main clause, then the pronoun tends to refer to the antecedent which has a certain thematic role (depending on the verb and on the subordinating conjunction). In contrast, pronouns in main clauses tend to refer back to the subject of the previous main clause, and this tendency is not affected by any verbs or conjunctions. In natural language processing, these findings have recently led to a proposal that pronoun resolution systems should have a split architecture, i.e. that they should use different mechanisms for pronoun resolution in the two cases.

With the help of two parsed and coreference-annotated corpora, this paper estimates the impact of the split-architecture proposal. The findings of this work are as follows: (1) Subject pronouns in authentic texts behave the same way in main and subordinate clauses. (2) The number of sentences in which a split architecture would behave differently than a system that treats both cases the same way is close to zero. Therefore, a separate treatment of resolution within and across units is unlikely to improve the performance of any system. This result casts a doubt on the split-architecture proposal, and more generally on approaches that directly incorporate psycholinguistic results into performance-oriented algorithms for anaphora resolution without assessing the relative importance of the phenomena that underlie them.

1 Introduction

Both in natural language processing and in psycholinguistics, the resolution of pronouns has long been a center of attention.

Computational approaches have ranged from purely syntax-oriented treatments (Hobbs, 1978) to work in the framework of centering theory (Joshi and Kuhn (1979); Kehler (1997); Joshi et al. (to appear)) to analyses based on statistical methods (e.g. Ge, Hale, and Charniak, 1998) and genetic programming (Orasan et al., 2000).

Psycholinguists have studied the processes involved in human anaphora resolution. Seemingly contradictory results were obtained in experiments by Stevenson et al. (2000) and Hudson-D'Zmura and Tanenhaus (1998). Stevenson et al. (2000) reports that the choice of verbs in a clause affect the interpretation of pronouns in subsequent subordinate clauses. Certain arguments of the verb (depending on the verb and on the conjunction between the clauses) are more likely to act as antecedents than others. This can be seen in the following minimal pair:

(1) a. Ken$_i$ admired Geoff so he$_i$...
b. Ken impressed Geoff$_j$ so he$_j$...

Although the pronoun *he* is ambiguous in these sentence fragments, subjects of a sentence completion experiment preferred to resolve it in both cases as indicated, that is, they coindexed it with the experiencer of the verb. Thus, this experiment seems to show that antecedents are preferred based on their thematic roles and not on their subjecthood.

Hudson-D'Zmura and Tanenhaus (1998) report experiments that seem at first sight to contradict this finding. When participants were presented with the following sentences (without indexings) and asked to judge the continuations for naturalness, they strongly preferred the subject interpretation shown in (2a).

(2) Max_i despises Ross_j.

a. He_i always gives Ross_j a hard time.

b. He_j always gives Max_i a hard time.

Crucially, this tendency was unaffected by the thematic role of the subject. In the previous example, the subject was the experiencer, while in the following example, it is the agent. Yet participants still favored the subject interpretation, shown in (3a).

(3) Jack_i apologized profusely to Josh_j.

a. He_i had been rude to Josh_j yesterday.

b. He_j had been offended by Jack_i's comment.

To the extent that this experiment is comparable with the previous one, it seems to show exactly the opposite tendency: that antecedents at least of subjects are preferred based on their subjecthood (a preference known as subject-to-subject parallelism) and not on their thematic roles.

One way of making sense of this contradiction is to assume that there is a distinction between pronoun resolution within a main clause and its subordinate clauses on the one hand and pronoun resolution across main clauses on the other hand. A sentence-completion experiment by Miltsakaki (2002), henceforth Miltsakaki, confirms this hypothesis (see also Miltsakaki, 2003). She reports a strong preference for the following coindexings:

(4) a. The groom_i hit the best man_j violently. However, he_i...

b. The groom_i hit the best man_j violently although he_j...

This minimal pair exhibits subject-to-subject parallelism in the case of two main clauses (4a), but not in the case of a main and a subordinate clause (4b). In the latter case, the main clause verb *hit* seems to focus its experiencer *the best man* and thereby to make it more likely to be an antecedent. (From here on, following Miltsakaki, I will refer to pronouns which are located in a dependent (subordinate) clause with respect to the clause of their antecedent as *intrasentential*. Pronouns whose antecedent is located in a different main clause will be called *intersentential*. This case includes clausal conjunction. For example, in the sentence "John loves Mary and she loves him", both pronouns are intersentential, because their antecedent is located in a different main clause.)

2 Miltsakaki's model

Miltsakaki's anaphora resolution architecture models this split behavior. In her model, entities inside main clauses are ranked in two different ways: according to grammatical function for the

purposes of resolution across main clauses, and according to semantic focusing preferences for resolution within a main clause and its subordinate clauses. These preferences are determined lexically by the main verb and by discourse connectives (i.e. subordinating conjunctions). Whenever a pronoun occurs inside a subordinate clause and there is no compatible potential antecedent inside that clause (for example because the pronoun is a subject), the pronoun is resolved to the most highly ranked available candidate, if any, inside the main clause as defined by semantic focusing preferences. In a second step, pronouns that could not be resolved so far are matched against candidates from the previous discourse unit, this time ranked according to grammatical function. Thus, the preferred reading in (2) as well as in (4a) will be obtained because pronouns will first be resolved to the subject (highest ranked in terms of grammatical function), whereas in the case of (4b), the pronoun is resolved to the expression *the best man* (highest ranked in terms of semantic focus).

Both the tendency of certain verbs to promote one of their arguments as a potential antecedent and the tendency for subject pronouns to resolve to subject antecedents have been previously observed in the literature on anaphora resolution.

- In the natural language processing (NLP) literature, Mitkov (1997) observes the tendency of certain verbs to promote their objects, which he calls “verb preference”. In contrast to Miltsakaki, however, he regards this as only one of several preference factors influencing pronoun resolution. — Psycholinguistic studies of activation of antecedents by *implicit causality verbs*, a closely related phenomenon, are cited in McDonald and MacWhinney (1995).

- The subject-to-subject parallelism is explicitly modeled in several centering-based algorithms, such as Left-to-Right Centering (Tetreault, 1999) or the RAFT/RAPR algorithm (Suri and McCoy, 1994). Again, Mitkov (1997) cites this as just one of several factors. — In psycholinguistics, the tendency of the first element of a sentence (which, in English, generally coincides with the subject) to influence subsequent pronoun resolution has been called the "advantage of first mention" (see McDonald and MacWhinney (1995) for references).

However, Miltsakaki is the first to suggest, based on the dichotomy in the findings described above, that a separate treatment of intra- and intersentential resolution is appropriate for the purpose of anaphora resolution.

While the results of the psycholinguistic experiments described above are beyond doubt, it is less clear how, if at all, the performance of a pronoun resolution algorithm would be improved by redesigning it along the lines of Miltsakaki’s proposal. In other words, the reported effects are undoubtedly true, but are they relevant to natural language processing? More generally, how can and should results obtained in psycholinguistics be incorporated in the development of applications whose goal is to improve system performance?

In the remainder of this paper, I investigate the relevance of Miltsakaki’s two-way model in two corpus-based experiments. It is shown that pronouns found in corpora do not exhibit the kind of split behavior you would expect from them based on Miltsakaki’s model. Furthermore, the corpora are analyzed in detail to show that the number of sentences in which semantic focusing properties can apply is very low and thus Miltsakaki’s algorithm is not going to increase the performance of anaphora resolution algorithms more than marginally. I conclude by a brief discussion of the role of psycholinguistic results for anaphora resolution in natural language processing.

3 Experiments

In order to assess the extent to which overall pronoun resolution could be improved by Miltsakaki's proposal, two parsed and manually coreference-annotated corpora from different domains were used. Both corpora are subsets of the Penn Treebank (Marcus et al., 1993). The first corpus is a collection of the Wall Street Journal articles 1 to 199 (about 94,100 words in total); the second corpus consists of three fictional texts taken from the Brown corpus section of the Penn Treebank (about 8,700 words). From the standpoint of anaphora resolution, the major difference between the two corpora is the abundance of *he* and *she* tokens in the fictional corpus. By contrast, *it* is the most frequently mentioned pronoun in the Wall Street Journal corpus.

A list of all pronoun-antecedent pairs from the corpora was automatically extracted. In case the pronoun had multiple antecedents (i.e. when the corresponding discourse entity had been mentioned several times), only the last one (the closest to the pronoun) was recorded as its antecedent. Each pronoun or antecedent extracted from the Wall Street Journal subcorpus was identified as subject or nonsubject. (Due to poor quality of annotation, subjects could not be identified in the Brown subcorpus.) Finally, for each pronoun-antecedent pair it was recorded whether it was located in the same clause, in two different clauses but in the same sentence, or in two different sentences.

(*Technical note.* Two coindexed expressions occurring in the same sentence were counted as interclausal if and only if there existed an SBAR node that dominated one but not both of them. In the Penn Treebank annotation, SBAR nodes identify most finite subordinate clauses. I did not consider other subordinate clauses. This is in part guided by Miltsakaki's working assumption (p.c.) that only the boundaries of *finite* subordinate clauses should be relevant for her algorithm. Moreover, among those subordinate clauses that are not identified by SBAR nodes, many are instances of inverted direct speech (identifiable by SINV nodes), and like many others, Miltsakaki's algorithm does not make provisions for resolution in connection with quoted texts.)

3.1 Experiment 1

In a first experiment, I determined how often a subject pronoun refers to an antecedent that is also a subject, depending on whether they are inter- or intrasentential. (More precisely, I determined how often the nearest clause that mentions the referent of a pronoun mentions it at least in the subject. This formulation filters out irrelevant factors such as parentheticals: For example, we regard the sentence "[John Doe]$_i$, [vice chairman of XYZ Co.]$_i$, announced that he$_i$ would resign" as a case of (intrasentential) subject-to-subject parallelism even though the subject of the main clause is strictly speaking not the *closest* antecedent of the pronoun.)

Since Miltsakaki suggests a separate treatment for intra- and intersentential pronoun resolution, a difference in behavior would indicate that her separate treatment is on the right track. More precisely, since the intersentential component of Miltsakaki's algorithm tries to resolve subjects to subjects but the intrasentential component does not have that preference, her approach would be validated if intersentential pronouns in subject position resolved more often to subjects than intrasentential pronouns did.

For this experiment, the set of pronouns was restricted to *he, she, it* and *they*, since these are the only pronouns that can occur as subjects of finite clauses and this experiment dealt specifically at subject pronouns. (In addition, *I, we, you* and possessive pronouns were excluded because it is not clear whether Miltsakaki intends these pronouns to fall into the scope

of her algorithm.) The results of this experiment are shown in Table 1. Again, note that the Brown subcorpus could not be included in this experiment.

Intersentential subject pronouns	
Antecedent is a subject:	682 instances (85%)
Antecedent is not a subject:	123 instances (15%)
Intrasentential subject pronouns	
Antecedent is a subject:	179 instances (87%)
Antecedent is not a subject:	26 instances (13%)

Table 1: *A comparison of pronouns which are subjects with respect to subjecthood of their antecedents.*

As can be seen, subject pronouns uniformly tend to resolve to the subject of the closest clause that mention their discourse referent at least once. There is no significant difference between intra- and intersentential pronouns. This result is in contrast with Miltsakaki's split model of pronoun resolution, and it suggests that at least for applications dealing with newspaper-style texts, it may not be necessary to treat intra- and intersentential pronouns differently as Miltsakaki suggests, since most of them tend to be resolved to subject position anyway.

In this experiment, *all* subject pronouns in the Wall Street Journal corpus have been considered, as opposed to considering only pronouns in contexts that are comparable to the psycholinguistic experiments described above. Therefore, while the result does show a tendency to foreground certain discourse referents by placing them in subject position again and again, it neither confirms nor disconfirms a possible tendency of subject pronouns to refer back to subjects as opposed to other potential antecedents.

For example, it may be the case that most of the intersentential pronouns resolve trivially to the subject of the main clause, as there is simply no other grammatically compatible candidate available in the clause. Under this hypothesis, semantic focusing preferences would hardly ever have to be applied, and their application would be undetectable from the previous experiment. Another possibility is that semantic focusing preferences do apply often, but they happen to rank the subject highest in most cases. Since semantic focusing preferences are determined by lexical properties of individual verbs, it would then be conceivable that in other text genres more verbs are used which tend to rank the subject lower. Thus, the question arises how often the application of semantic focusing preferences would really make a difference, as opposed to a straightforward resolution procedure operating without focusing preferences.

3.2 Experiment 2

To answer this question, the following experiment was carried out. For each of the two corpora, those anaphoric pronouns were isolated for which semantic focusing properties actually had a chance of picking out one potential antecedent over another. Miltsakaki claims that this is only the case for pronouns in subordinate clauses whose closest antecedent occurred in the clause on which that clause was dependent. The corpora were filtered accordingly.

For this experiment, nonsubject pronouns were included, since Miltsakaki intends the split resolution algorithm to apply to those as well. Again, since it is not clear if *I, we, you* and possessive pronouns are supposed to fall in the scope of her algorithm, they were excluded. This means that the following pronouns were considered: *he, she, it, him* and objective case *her* (as opposed to possessive case *her*, which is exemplified by *Mary is looking for her purse*.)

All sentences in which there was only one grammatically compatible potential antecedent

in the relevant clause (and so semantic focusing preferences could trivially not apply) were eliminated. Furthermore, all sentences were removed in which there was more than one compatible potential antecedent but they were not assigned different thematic roles by the verb and so its focusing preferences could not have been used to disambiguate. This was the case when two potential antecedents were nested, for example. The sorting and removal had to be performed semi-automatically, as person, number, and binding constraints are not annotated in the corpora.

In this way, only those sentences were retained in which Miltsakaki's algorithm would actually apply semantic focusing preferences during resolution. The result is as follows. Seven sentences in total, containing eight relevant pronouns, were found. All of them were found in the Wall Street Journal corpus. By comparison, this corpus contains 846 instances of the pronouns under consideration. This means that on all except seven sentences (99% of the corpus), Miltsakaki's split-architecture resolution algorithm will yield exactly the same results as an otherwise equal algorithm that ignores semantic focusing preferences. In other words, if the corpus is representative, then the maximum improvement we can expect from switching to an architecture that takes semantic focusing preferences into account is around one percent. Since focusing preferences are likely to often coincide with subject-to-subject parallelism, the real number is likely to be even lower.

(Two reviewers are worried that we might be drawing conclusions based on a very small amount of evidence, i.e. seven examples, and that more examples should first be collected. But the very fact that almost no examples could be found is the main result of this experiment.)

A note of caution: The corpora are annotated for coreference, but not for potential antecedents. If a discourse referent is referred to at least twice (for example, by two definite descriptions, or by a definite description and a pronoun), then the two referents are marked as coreferential and can be assumed to be potential antecedents to other pronouns. But discourse referents which are only mentioned once are not marked as potential antecedents. They are merely marked as NP (noun phrase). However, it is not possible to simply assume that the set of NPs and the set of potential antecedents are equal, because not all entities which the corpus annotation considers NPs are potential antecedents:

(5) John does not see a car$_i$. #It$_i$ is blue. (von Heusinger, 2000)

(6) #Today$_i$'s notes will be posted online after it$_i$ is over.

For this reason, only those NPs which are annotated as coreferential could safely be considered potential antecedents for the purposes of this experiment. Therefore, an unknown number of "critical sentences" (sentences that contain a pronoun for which semantic focusing preferences will have the chance to apply) may have been missed. Identifying all potential antecedents in the corpus by hand is beyond the scope of this paper. Unfortunately, I am not aware of any parsed corpora in which all potential antecedents, as opposed to just the coreferential noun phrases, have been annotated and on which the experiment could be therefore carried out with greater accuracy.

An upper bound on this error has been estimated by eyeballing the data. An informal count performed on a 500-word sample, taken at random from one Wall-Street-Journal file, identified 114 potential antecedents, of which 81 (71%) were annotated as coreferential with some other noun phrase and were therefore visible to this experiment. If we assume the number of "critical sentences" to be roughly proportional to the number of potential antecedents, it can therefore be cautiously estimated that the number of "critical sentences" that were missed by the above

procedure is less than one third of the actual number. In other words, if the estimate is correct, there might perhaps be around 11 or 12 such sentences in the corpus, but not a lot more.

Taken together with the results of the previous experiment, this suggests that least in the newspaper and fictional genres, focusing preferences are of limited relevance as a factor in pronoun resolution.

4 Conclusion

While psycholinguistic experiments may seem to suggest that two different mechanisms are responsible for pronoun resolution within and across sentences, these experiments have considered a type of sentence that appears to occur very rarely in actual texts, if at all. The algorithm described in Miltsakaki (2002) may model human preferences in anaphora resolution acccurately when running on a certain set of restricted cases, but its most important feature – a separate treatment of intra- and intersentential pronouns – is unlikely to result in a significant improvement in performance.

It has to be stressed that these results are preliminary and a larger corpus study would be necessary to confirm them, preferably one that involves corpora from more varied domains than those used here. Note, too, that not all potential antecedents have been annotated in the corpora used here. Therefore, as explained above, the number of sentences in which semantic focusing preferences apply may be larger than what is reported here. However, if it were much larger, we would expect this to result in a marked difference in the frequency of subject-to-subject parallelism within as opposed to across clauses. This has been shown not to hold for the present corpus.

The identification of (abstract) thematic roles is a hard task. It is notoriously difficult to assign thematic roles consistently even by hand, let alone to build a system that identifies them (see e.g. Thompson et al., 2003). The focusing preferences of verbs could perhaps be restated in terms of more easily identifiable features such as grammatical function. However, no wide-coverage investigation has yet been carried out on whether focusing preferences are predictable from more easily obtainable features. For this reason, it is a welcome result that a unified treatment of pronouns within and across sentences seems possible. The simple heuristics of subject-to-subject resolution suggested by Miltsakaki (2002) and others for intersentential pronouns can likely be applied to intrasentential cases without any significant loss of accuracy.

The present study can of course not answer the general question of psycholinguistic results for preferences in anaphora resolution should be handled in performance-oriented algorithms. While psycholinguistics can bring attention to hitherto unknown anaphora resolution *factors* in the sense of Mitkov (1997), evaluating actual resolution systems is the only way to know how the best use can be made of these factors. As an example, see Mitkov (1997) for a comparative evaluation of two approaches based on the same set of factors.

Nevertheless, the present work has shown one way of how psycholinguistic results, specifically the ones cited, should *not* be handled: Directly implementing the semantic focusing preferences of verbs, as proposed by Miltsakaki, can require resources that are difficult to obtain (such as automatic semantic role labeling), while it seems unlikely that the overall performance will be affected at all. Speaking generally, corpus studies like the present one are a convenient tool of estimating the impact that a new anaphora resolution factor is likely to have on overall performance while avoiding the need and cost of implementing and evaluating the factor in an actual system.

Finally, the results described here make it necessary to rethink what we believe to be the *function* of phenomena we model by abstractions like subject preference, centering rules,

or semantic focusing of thematic roles. They are sometimes (see e.g. Hudson-D'Zmura and Tanenhaus, 1998, for centering rules) seen as strategies that readers or listeners unconsciously apply in order to constrain the interpretation process and in this way control inferential complexity. That is, since the strategies highlight certain entities as more likely antecedents, pronoun resolution is made easier. But if the cases in which this can happen are in practice very rare, then the overall reduction in processing load is very small, and phenomena like semantic focusing would be very inefficient strategies. They should then perhaps better be assumed to be epiphenomena of some more general, unknown processes, and their relative importance should be reassessed (contra e.g. Stevenson et al., 2000, p. 226: "Pronoun resolution is primarily determined by focusing..."). This would essentially mean that why we are so good at real-time pronoun resolution, and how we manage to reduce our inferential load most of the time when we do it, becomes an open question again.

5 Acknowledgements

I would like to thank Eleni Miltsakaki for her extensive support of this project and for her patience in explaining the details of her proposal to me. This project would not have seen the light without her initiative. I am also indebted to her for suggesting the first experiment to me. I wish to thank John Trueswell for helpful discussion, and ESSLLI reviewers for their valuable comments on this paper.

For access to the two corpora that were used in the experiments, I am grateful to Tom Morton (WSJ corpus) and Joel Tetreault (fictional corpus). (Unfortunately, the corpora are not publicly available at the time of writing. Requests for any of the corpora may be sent to the author of this paper and will be forwarded accordingly.)

References

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora* 161–170.

Klaus von Heusinger. 2000. Anaphora, antecedents, and accessibility. In *Theoretical Linguistics* 26:75-93.

Jerry R. Hobbs. 1978. Resolving pronoun references. In *Lingua* 44: 311–338.

Susan Hudson-D'Zmura and Michael Tanenhaus. 1998. Assigning antecedents to ambiguous pronouns: The role of the center of attentian as a default assignment. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, 273–291. Clarendon Press, Oxford.

Aravind K. Joshi and Steven Kuhn. 1979. Centered logic: The role of entity centered sentence representation in natural language inferencing. In *Sixth International Joint Conference on Artifical Intelligence* 435–439. Tokyo.

Aravind K. Joshi, Rashmi Prasad, and Eleni Miltsakaki. To appear. Anaphora resolution: a centering approach. In *Encyclopedia of Language and Linguistics, 2nd edition.* Elsevier.

Andrew Kehler. 1997. Current theories of centering for pronoun interpretation: A critical evaluation. In *Computational Linguistics*, 23(3):467–475.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, 19(2):313–330.

Janet L. McDonald and Brian MacWhinney. 1995. The time course of anaphor resolution: Effects of implicit verb causality and gender. Journal of Memory and Language, 34, 543ś566

Eleni Miltsakaki. 2002. Toward an aposynthesis of topic continuity and intrasentential anaphora. In *Computational Linguistics*, 28(3):319–355.

Eleni Miltsakaki. 2003. The syntax-discourse interface: Effects of the main-subordinate distinction on attention structure. Doctoral dissertation, University of Pennsylvania.

Ruslan Mitkov. 1997. Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches. In *Proceedings of the ACL '07/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution.* 14–21. Madrid, Spain.

Constantin Orasan, Richard Evans, and Ruslan Mitkov. 2000. Enhancing preference-based anaphora resolution with genetic algorithms. In *Proceedings of NLP '2000*, 185–195. Patras, Greece.

Rosemary Stevenson, Alistair Knott, Jon Oberlander, and Sharon McDonald. 2000. Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. *Language and Cognitive Processes* 15(3):225–262.

Linda Z. Suri and Kathleen F. McCoy. 1994. RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences. In *Computational Linguistics*, 20(2):301–317.

Joel R. Tetreault. 1999. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting*, 602–605. University of Maryland, June. Association for Computational Linguistics.

Cynthia A. Thompson, Roger Levy, and Christopher D. Manning. 2003. A generative model for semantic role labeling. In: *Proceedings of the European Conference on Machine Learning (ECML)*, 397-408.