EVALUATING DISCOURSE PROCESSING ALGORITHMS

Marilyn A. Walker

Hewlett Packard Laboratories
Filton Rd., Bristol, England BS12 6QZ, U.K.
& University of Pennsylvania
lyn%lwalker@hplb.hpl.hp.com

Abstract

In order to take steps towards establishing a methodology for evaluating Natural Language systems, we conducted a case study. We attempt to evaluate two different approaches to anaphoric processing in discourse by comparing the accuracy and coverage of two published algorithms for finding the co-specifiers of pronouns in naturally occurring texts and dialogues. We present the quantitative results of hand-simulating these algorithms, but this analysis naturally gives rise to both a qualitative evaluation and recommendations for performing such evaluations in general. We illustrate the general difficulties encountered with quantitative evaluation. These are problems with: (a) allowing for underlying assumptions, (b) determining how to handle underspecifications, and (c) evaluating the contribution of false positives and error chaining.

1 Introduction

In the course of developing natural language interfaces, computational linguists are often in the position of evaluating different theoretical approaches to the analysis of natural language (NL). They might want to (a) evaluate and improve on a current system, (b) add a capability to a system that it didn't previously have, (c) combine modules from different systems.

Consider the goal of adding a discourse component to a system, or evaluating and improving one that is already in place. A discourse module might combine theories on, e.g., centering or local focusing [GJW83, Sid79], global focus [Gro77], coherence relations[Hob85], event reference [Web86], intonational structure [PH87], system vs. user beliefs [Pol86], plan or intent recognition or production [Coh78, AP86, SI81], control[WS88], or complex syntactic structures [Pri85]. How might one evaluate the relative contributions of each of these factors or compare two approaches to the same problem?

In order to take steps towards establishing a methodology for doing this type of comparison, we conducted a case study. We attempt to evaluate two different approaches to anaphoric processing in discourse by comparing the accuracy and coverage of two published algorithms for finding the co-specifiers of pronouns in naturally occurring texts and dialogues[Hob76b, BFP87]. Thus there are two parts to this paper: we present the quantitative results of hand-simulating these algorithms (henceforth Hobbs algorithm and BFP algorithm), but this analysis naturally gives rise to both a qualitative evaluation and recommendations for performing such evaluations in general. We illustrate the general difficulties encountered with quantitative evaluation. These are problems with: (a) allowing for underlying assumptions, (b) determining how to handle underspecifications, and (c) evaluating the contribution of false positives and error chaining.

Although both algorithms are part of theories of discourse that posit the interaction of the algorithm with an inference or intentional component, we will not use reasoning in tandem with the algorithm's operation. We have made this choice because we want to be able to analyse the performance of the algorithms across different domains. We focus on the linguistic basis of these approaches, using only selectional restrictions, so that our analysis is independent of the vagaries of a particular knowledge representation. Thus what we are evaluating is the extent to which these algorithms suffice to narrow the search of an inference component[1]. This analysis gives us

[1] But note the definition of success in section 2.1.

some indication of the contribution of syntactic constraints, task structure and global focus to anaphoric processing.

The data on which we compare the algorithms are important if we are to evaluate claims of generality. If we look at types of NL input, one clear division is between textual and interactive input. A related, though not identical factor is whether the language being analysed is produced by more than one person, although this distinction may be conflated in textual material such as novels that contain reported conversations. Within two-person interactive dialogues, there are the task-oriented master-slave type, where all the expertise and hence much of the initiative, rests with one person. In other two-person dialogues, both parties may contribute discourse entities to the conversation on a more equal basis. Other factors of interest are whether the dialogues are human-to-human or human-to-computer, as well as the modality of communication, e.g. spoken or typed, since some researchers have indicated that dialogues, and particularly uses of reference within them, vary along these dimensions [Coh84, Tho80, GSBC86, DJ89, WS89].

We analyse the performance of the algorithms on three types of data. Two of the samples are those that Hobbs used when developing his algorithm. One is an excerpt from a novel and the other a sample of journalistic writing. The remaining sample is a set of 5 human-human, keyboard-mediated, task-oriented dialogues about the assembly of a plastic water pump [Coh84]. This covers only a subset of the above types. Obviously it would be instructive to conduct a similar analysis on other textual types.

2 Quantitative Evaluation–Black Box

2.1 The Algorithms

When embarking on such a comparison, it would be convenient to assume that the inputs to the algorithms are identical and compare their outputs. Unfortunately since researchers do not even agree on which phenomena can be explained syntactically and which semantically, the boundaries between two modules are rarely the same in NL systems. In this case the BFP centering algorithm and Hobbs algorithm both make ASSUMPTIONS about other system components. These are, in some sense, a further specification of the operation of the algorithms that must be made in order to hand-simulate the algorithms. There are two major sets of assumptions, based on discourse segmentation and syntactic representation. We attempt to make these explicit for each algorithm and pinpoint where the algorithms might behave differently were these assumptions not well-founded.

In addition, there may be a number of UNDERSPECIFICATIONS in the descriptions of the algorithms. These often arise because theories that attempt to categorize naturally occurring data and algorithms based on them will always be prey to previously unencountered examples. For example, since the BFP salience hierarchy for discourse entities is based on grammatical relation, an implicit assumption is that an utterance only has one subject. However the novel *Wheels* has many examples of reported dialogue such as *She continued, unperturbed, "Mr. Vale quotes the Bible about air pollution."* One might wonder whether the subject is *She* or *Mr. Vale*. In some cases, the algorithm might need to be further specificied in order to be able to process any of the data, whereas in others they may just highlight where the algorithm needs to be modified (see section 3.2). In general we count underspecifications as failures.

Finally, it may not be clear what the DEFINITION OF SUCCESS is. In particular it is not clear what to do in those cases where an algorithm produces multiple or partial interpretations. In this situation a system might flag the utterance as ambiguous and draw in support from other discourse components. This arises in the present analysis for two reasons: (1) the constraints given by [GJW86] do not always allow one to choose a preferred interpretation, (2) the BFP algorithm proposes equally ranked interpretations in parallel. This doesn't happen with the Hobbs algorithm because it proposes interpretations in a sequential manner, one at a time. We chose to count as a failure those situations in which the BFP algorithm only reduces the number of possible interpretations, but Hobbs algorithm stops with a correct interpretation. This ignores the fact that Hobbs may have rejected a number of interpretations before stopping. We also have not needed to make a decision on how to score an algorithm that only finds one interpretation for an utterance that humans find ambiguous.

2.1.1 Centering algorithm

The centering algorithm as defined by Brennan, Friedman and Pollard, (BFP algorithm), is derived from a set of rules and constraints put forth by Grosz,

Joshi and Weinstein [GJW83, GJW86]. We shall not reproduce this algorithm here (See [BFP87]). There are two main structures in the centering algorithm, the CB, the BACKWARD LOOKING CENTER, which is what the discourse is 'about', and an ordered list, CF, of FORWARD LOOKING CENTERS, which are the discourse entities available to the next utterance for pronominalization. The centering framework predicts that in a local coherent stretch of dialogue, speakers will prefer to CONTINUE talking about the same discourse entity, that the CB will be the highest ranked entity of the previous utterance's forward centers that is realized in the current utterance, and that if anything is pronominalized the CB must be.

In the centering framework, the order of the forward-centers list is intended to reflect the salience of discourse entities. The BFP algorithm orders this list by grammatical relation of the complements of the main verb, i.e. first the subject, then object, then indirect object, then other subcategorized-for complements, then noun phrases found in adjunct clauses. This captures the intuition that subjects are more salient than other discourse entities.

The BFP algorithm added linguistic constraints on CONTRA-INDEXING to the centering framework. These constraints are exemplified by the fact that, in the sentence *he likes him*, the entity cospecified by *he* cannot be the same as that cospecified by *him*. We say that *he* and *him* are CONTRA-INDEXED. The BFP algorithm depends on semantic processing to precompute these constraints, since they are derived from the syntactic structure, and depend on some notion of c-command[Rei76]. The other assumption that is dependent on syntax is that the the representations of discourse entities can be marked with the grammatical function through which they were realized, e.g. subject.

The BFP algorithm assumes that some other mechanism can structure both written texts and task-oriented dialogues into hierarchical segments. The present concern is not with whether there might be a grammar of discourse that determines this structure, or whether it is derived from the cues that cooperative speakers give hearers to aid in processing. Since centering is a local phenomenon and is intended to operate within a segment, we needed to deduce a segmental structure in order to analyse the data. Speaker's intentions, task structure, cue words like *O.K. now..*, intonational properties of utterances, coherence relations, the scoping of modal operators, and mechanisms for shifting control between discourse participants have all been proposed as ways of determining discourse segmentation [Gro77, GS86, Rei85, PH87, HL87, Hob78, Hob85, Rob88, WS88]. Here, we use a combination of orthography, anaphora distribution, cue words and task structure. The rules are:

- In published texts, a paragraph is a new segment unless the first sentence has a pronoun in subject position or a pronoun where none of the preceding sentence-internal noun phrases match its syntactic features.
- In the task-oriented dialogues, the action PICK-UP marks task boundaries hence segment boundaries. Cue words like *next, then,* and *now* also mark segment boundaries. These will usually co-occur but either one is sufficient for marking a segment boundary.

BFP never state that cospecifiers for pronouns within the same segment are preferred over those in previous segments, but this is an implicit assumption, since this line of research is derived from Sidner's work on local focusing. Segment initial utterances therefore are the only situation where the BFP algorithm will prefer a within-sentence noun phrase as the cospecifier of a pronoun.

2.1.2 Hobbs' algorithm

The Hobbs algorithm is based on searching for a pronoun's co-specifier in the syntactic parse tree of input sentences [Hob76b]. We reproduce this algorithm in full in the appendix along with an example. Hobbs algorithm operates on one sentence at a time, but the structure of previous sentences in the discourse is available. It is stated in terms of searches on parse trees. When looking for an intrasentential antecedent, these searches are conducted in a left-to-right, breadth-first manner. However, when looking for a pronoun's antecedent within a sentence, it will go sequentially further and further up the tree to the left of the pronoun, and that failing will look in the previous sentence. Hobbs does not assume a segmentation of discourse structure in this algorithm; the algorithm will go back arbitrarily far in the text to find an antecedent. In more recent work, Hobbs uses the notion of COHERENCE RELATIONS to structure the discourse [HM87].

The order by which Hobbs' algorithm traverses the parse tree is the closest thing in his framework to predictions about which discourse entities are salient. In the main it prefers co-specifiers for pronouns that

are within the same sentence, and also ones that are closer to the pronoun in the sentence. This amounts to a claim that different discourse entities are salient, depending on the position of a pronoun in a sentence. When seeking an intersentential co-specification, Hobbs algorithm searches the parse tree of the previous utterance breadth-first, from left to right. This predicts that entities realized in subject position are more salient, since even if an adjunct clause linearly precedes the main subject, any noun phrases within it will be deeper in the parse tree. This also means that objects and indirect objects will be among the first possible antecedents found, and in general that the depth of syntactic embedding is an important determiner of discourse prominence.

Turning to the assumptions about syntax, we note that Hobbs assumes that one can produce the correct syntactic structure for an utterance, with all adjunct phrases attached at the proper point of the parse tree. In addition, in order to obey linguistic constraints on coreference, the algorithm depends on the existence of a $\overline{N}$ parse tree node, which denotes a noun phrase without its determiner (See the example in the Appendix). Hobbs algorithm procedurally encodes contra-indexing constraints by skipping over NP nodes whose $\overline{N}$ node dominates the part of the parse tree in which the pronoun is found, which means that he cannot guarantee that two contra-indexed pronouns will not choose the same NP as a co-specifier.

Hobbs also assumes that his algorithm can somehow collect discourse entities mentioned alone into sets as co-specifiers of plural anaphors. Hobbs discusses at length other assumptions that he makes about the capabilities of an interpretive process that operates before the algorithm [Hob76b]. This includes such things as being able to recover syntactically recoverable omitted text, such as elided verb phrases, and the identities of the speakers and hearers in a dialogue.

2.1.3 Summary

A major component of any discourse algorithm is the prediction of which entities are salient, even though all the factors that contribute to the salience of a discourse entity have not been identified [Pri81, Pri85, BF83, HTD86]. So an obvious question is when the two algorithms actually make different predictions. The main difference is that the choice of a co-specifier for a pronoun in the Hobbs algorithm depends in part on the position of that pronoun in the sentence. In the centering framework, no matter what criteria one uses to order the forward-centers list, pronouns take the most salient entities as antecedents, irrespective of that pronoun's position. Hobbs ordering of entities from a previous utterance varies from BFP in that possessors come before case-marked objects and indirect objects, and there may be some other differences as well but none of them were relevant to the analysis that follows.

The effects of some of the assumptions are measurable and we will attempt to specify exactly what these effects are, however some are not, e.g. we cannot measure the effect of Hobbs' syntax assumption since it is difficult to say how likely one is to get the wrong parse. We adopt the set collection assumption for both algorithms as well as the ability to recover the identity of speakers and hearers in dialogue.

2.2 Quantitative Results of the Algorithms

The texts on which the algorithms are analysed are the first chapter of Arthur Hailey's novel *Wheels*, and the July 7, 1975 edition of Newsweek. The sentences in *Wheels* are short and simple with long sequences consisting of reported conversation, so it is similar to a conversational text. The articles from *Newsweek* are typical of journalistic writing. For each text, the first 100 occurrences of singular and plural third-person pronouns were used to test the performance of the algorithms. The task-dialogues contain a total of 81 uses of *it* and no other pronouns except for *I* and *you*. In the figures below note that possessives like *his* are counted along with *he* and that accusatives like *him* and *her* are counted as *he* and *she*[2].

	N	Hobbs	BFP
Wheels	100	88	90
Newsweek	100	89	79
Tasks	81	51	49

Figure 1: Number correct for both algorithms for Wheels, Newsweek and Task Dialogues

We performed three analyses on the quantitative results. A comparison of the two algorithms on each data set individually and an overall analysis on the three data sets combined revealed no significant differences in the performance of the two algorithms

[2]Hobbs reports his algorithm's performance and the examples it fails on in [Hob76b, Hob76a]. The numbers reported here vary slightly from those. This is probably due to a discrepancy in exactly what the data-set consisted of.

($\chi^2 = 3.25$, not significant). In addition for each algorithm alone we tested whether there were significant differences in performance for different textual types. Both of the algorithms performed significantly worse on the task dialogues ($\chi^2 = 22.05$ for Hobbs, $\chi^2 = 21.55$ for BFP, $p < 0.05$).

We might wonder with what confidence we should view these numbers. A significant factor that must be considered is the contribution of FALSE POSITIVES and ERROR CHAINING. A FALSE POSITIVE is when an algorithm gets the right answer for the wrong reason. A very simple example of this phenomena is illustrated by this sequence from one of the task dialogues.

Exp_1:	Now put IT in the pan of water.
Exp_2:	Stand IT up.
Exp_3:	Pump the little handle with the red cap on IT.
Cli_1.	ok
Exp_4.	Does IT work??

The first *it* in Exp_1 refers to *the pump*. Hobbs algorithm gets the right antecedent for *it* in Exp_3, which is *the little handle*, but then fails on *it* in Exp_4, whereas the BFP algorithm has *the pump* centered at Exp_1 and continues to select that as the antecedent for *it* throughout the text. This means BFP gets the wrong co-specifier in Exp_3 but this error allows it to get the correct co-specifier in Exp_4.

Another type of false positive example is *"Everybody and HIS brother suddenly wants to be the President's friend," said one aide.* Hobbs gets this correct as long as one is willing to accept that *Everybody* is really the antecedent of *his*. It seems to me that this might be an idiomatic use.

ERROR CHAINING refers to the fact that once an algorithm makes an error, other errors can result. Consider:

Cli_1:	Sorry no luck.
Exp_1:	I bet IT's the stupid red thing.
Exp_2:	Take IT out.
Cli_2:	Ok. IT is stuck.

In this example once an algorithm fails at Exp_1 it will fail on Exp_2 and Cli_2 as well since the choices of a cospecifier in the following examples are dependent on the choice in Exp_1.

It isn't possible to measure the effect of false positives, since in some sense they are subjective judgements. However one can and should measure the effects of error chaining, since reporting numbers that correct for error chaining is misleading, but if the error that produced the error chain can be corrected then the algorithm might show a significant improvement. In this analysis, error chains contributed 22 failures to Hobbs' algorithm and 19 failures to BFP.

3 Qualitative Evaluation–Glass Box

The numbers presented in the previous section are intuitively unsatisfying. They tell us nothing about what makes the algorithms more or less general, or how they might be improved. In addition, given the assumptions that we needed to make in order to produce them, one might wonder to what extent the data is a result of these assumptions. Figure 1 also fails to indicate whether the two algorithms missed the same examples or are covering a different set of phenomena, i.e. what the relative distribution of the successes and failures are. But having done the hand-simulation in order to produce such numbers, all of this information is available. In this section we will first discuss the relative importance of various factors that go into producing the numbers above, then discuss if the algorithms can be modified since the flexibility of a framework in allowing one to make modifications is an important dimension of evaluation.

3.1 Distributions

The figures 2, 3 and 4 show for each pronominal category, the distribution of successes and failures for both algorithms.

	Both	Neither	Hobbs only	BFP only
HE	66	1	1	7
SHE	6			
IT	6	3	3	
THEY	5	1	1	
Total	83	5	5	7

Figure 2: Distribution on Wheels

Since the main purpose of evaluation must be to improve the theory that we are evaluating, the most interesting cases are the ones on which the algorithms' performance varies and those that neither algorithm gets correct. We discuss these below.

	Both	Neither	Hobbs only	BFP only
HE	53		8	2
IT	11	5	4	1
THEY	13	3		
Total	77	8	12	3

Figure 3: Distribution on Newsweek

	Both	Neither	Hobbs only	BFP only
IT	48	29	3	1

Figure 4: Distribution on Task Dialogues

3.1.1 Both

In the *Wheels* data, 4 examples rest on the assumption that the identities of speakers and hearers is recoverable. For example in *The GM president smiled. "Except Henry will be damned forceful and the papers won't print all HIS language."*, getting the *his* correct here depends on knowing that it is the GM president speaking. Only 4 examples rest on being able to produce collections or discourse entities, and 2 of these occurred with an explicit instruction to the hearer to produce such a collection by using the phrase *them both.*

3.1.2 Hobbs only

There are 21 cases that Hobbs gets that BFP don't, and of these these a few classes stand out. In every case the relevant factor is Hobbs' preference for intrasentential co-specifiers.

One class, ($n = 3$), is exemplified by *Put the little black ring into the the large blue CAP with the hole in IT.* All three involved using the preposition *with* in a descriptive adjunct on a noun phrase. It may be that with-adjuncts are common in visual descriptions, since they were only found in our data in the task dialogues, and a quick inspection of Grosz's task-oriented dialogues revealed some as well[Deu74].

Another class, ($n = 7$), are possessives. In some cases the possessive co-specified with the subject of the sentence, e.g. *The SENATE took time from ITS paralyzing New Hampshire election debate to vote agreement*, and in others it was within a relative clause and co-specified with the subject of that clause, e.g. *The auto industry should be able to produce a totally safe, defect-free CAR that doesn't pollute ITS environment.*

Other cases seem to be syntactically marked subject matching with constructions that link two S clauses ($n = 8$). These are uses of *more-than* in e.g. *but Chamberlain grossed about $8.3 million more than HE could have made by selling on the home front.* There also are S-if-S cases, as in *Mondale said: "I think THE MAFIA would be broke if IT conducted all its business that way."* We also have subject matching in AS-AS examples as in *... and the resulting EXPOSURE to daylight has become as uncomfortable as IT was unaccustomed*, as well as in sentential complements, such as *But another liberal, Minnesota's Walter MONDALE, said HE had found a lot of incompetence in the agency's operations.* The fact that quite a few of these are also marked with *But* may be significant.

In terms of the possible effects that we noted earlier, the DEFINITION OF SUCCESS (see section 2.1 favors Hobbs ($n = 2$). Consider:

> K: Next take the red piece that is the smallest and insert it into the hole in the side of the large plastic tube. IT goes in the hole nearest the end with the engravings on IT.

The Hobbs algorithm will correctly choose *the end* as the antecedent for the second *it.* The BFP algorithm on the other hand will get two interpretations, one in which the second *it* co-specifies *the red piece* and one in which it co-specifies *the end.* They are both CONTINUING interpretations since the first *it* co-specifies the CB, but the constraints don't make a choice.

3.1.3 BFP only

All of the examples on which BFP succeed and Hobbs fails have to do with extended discussion of one discourse entity. For instance:

Exp_1:	Now take the blue cap with the two prongs sticking out (CB = blue cap)
Exp_2:	and fit the little piece of pink plastic on IT. Ok? (CB= blue cap)
Cli_1:	ok.
Exp_3:	Insert the rubber ring into that blue cap. (CB= blue cap)
Exp_4:	Now screw IT onto the cylinder.

On this example, Hobbs fails by choosing the co-specifier of *it* in Exp_4 to be *the rubber ring*, even

though the whole segment has been about *the blue cap.*

Another example from the novel *WHEELS* is given below. On this one Hobbs gets the first use of *he* but then misses the next four, as a result of missing the second one by choosing *a housekeeper* as the cospecifier for *HIS.*

> ..An executive vice-president of Ford was preparing to leave for Detroit Metropolitan Airport. HE had already breakfasted, alone. A housekeeper had brought a tray to HIS desk in the softly lighted study where, since 5 a.m., HE had been alternately reading memoranda (mostly on special blue stationery which Ford vice-presidents used in implementing policy) and dictating crisp instructions into a recording machine. HE had scarcely looked up, either as the mail arrived, or while eating, as HE accomplished in an hour what would have taken...

Since *an executive vice-president* is centered in the first sentence, and continued in each following sentence, the BFP algorithm will correctly choose the cospecifier.

3.1.4 Neither

Among the examples that neither algorithm gets correctly are 20 examples from the task dialogues of *it* referring to the global focus, the pump. In 15 cases, these shifts to global focus are marked syntactically with a cue word such as *Now*, and are not marked in 5 cases. Presumably they are felicitous since the pump is visually salient. Besides the global focus cases, pronominal references to entities that were not linguistically introduced are rare. The only other example is an implicit reference to 'the problem' of the pump not working:

Cli_1: Sorry no luck.
Exp_1: I bet IT's the stupid red thing.

We have only two examples of sentential or VP anaphora altogether, such as *Madam Chairwoman, said Colby at last, I am trying to run a secret intelligence service. IT was a forlorn hope.* Neither Hobbs algorithm nor BFP attempt to cover these examples.

Three of the examples are uses of *it* that seem to be lexicalized with certain verbs, e.g. *They hit IT off real well.* One can imagine these being treated as phrasal lexical items, and therefore not handled by an anaphoric processing component[AS89].

Most of the interchanges in the task dialogues consist of the client responding to commands with cues such as *O.K.* or *Ready* to let the expert know when they have completed a task. When both parties contribute discourse entities to the common ground, both algorithms may fail ($n = 4$).

Consider:

Exp_1: Now we have a little red piece left
Exp_2: and I don't know what to do with IT.
Cli_1: Well, there is a hole in the green plunger inside the cylinder.
Exp_3: I don't think IT goes in THERE.
Exp_4: I think IT may belong in the blue cap onto which you put the pink piece of plastic.

In Exp_3, one might claim that *it* and *there* are contraindexed, and that *there* can be properly resolved to *a hole*, so that *it* cannot be any of the noun phrases in the prepositional phrases that modify *a hole*, but whether any theory of contra-indexing actually give us this is questionable.

The main factor seems to be that even though Exp_1 is not syntactically a question, *the little red piece* is the focus of a question, and as such is in focus despite the fact that the syntactic construction *there is* supposedly focuses *a hole in the green plunger* ...[Sid79]. These examples suggest that a questioned entity is left focused until the point in the dialogue at which the question is resolved. The fact that *well* has been noted as a marker of response to questions supports this analysis[Sch87]. Thus the relevant factor here may be the switching of control among discourse participants [WS88]. These mixed-initiative features make these sequences inherently different than text.

3.2 Modifiability

Task structure in the pump dialogues is an important factor especially as it relates to the use of global focus. Twenty of the cases on which both algorithms fail are references to *the pump*, which is the global focus. We can include a global focus in the centering framework, as a separate notion from the current C_B. This means that in the 15 out of 20 cases where the shift to global focus is identifiably marked with a cue-word such as *now*, the segment rules will allow BFP to get the global focus examples.

BFP can add the VP and the S onto the end of the

forward centers list, as Sidner does in her algorithm for local focusing [Sid79]. This lets BFP get the two examples of event anaphora. Hobbs discusses the fact that his algorithm cannot be modified to get event anaphora in [Hob76b].

Another interesting fact is that in every case in which Hobbs' algorithm gets the correct co-specifier and BFP didn't, the relevant factor is Hobbs' preference for intrasentential co-specifiers. One view on these cases may be that these are not discourse anaphora, but there seems to be no principled way to make this distinction. However, Carter has proposed some extensions to Sidner's algorithm for local focusing that seem to be relevant here(chap. 6, [Car87]). He argues that intra-sentential candidates (ISCs) should be preferred over candidates from the previous utterance, ONLY in the cases where no discourse center has been established or the discourse center is rejected for syntactic or selectional reasons. He then uses Hobbs algorithm to produce an ordering of these ISCs. This is compatible with the centering framework since it is underspecified as to whether one should always choose to establish a discourse center with a co-specifier from a previous utterance. If we adopt Carter's rule into the centering framework, we find that of the 21 cases that Hobbs gets that BFP don't, in 7 cases there is no discourse center established, and in another 4 the current center can be rejected on the basis of syntactic or sortal information. Of these Carter's rule clearly gets 5, and another 3 seem to rest on whether one might want to establish a discourse entity from a previous utterance. Since the addition of this constraint does not allow BFP to get any examples that neither algorithm got, it seems that this combination is a way of making the best out of both algorithms.

The addition of these modifications changes the quantitative results. See the Figure 5.

	N	Hobbs	BFP
Wheels	100	88	93
Newsweek	100	89	84
Tasks	81	51	64

Figure 5: Number correct for both algorithms after Modifications, for Wheels, Newsweek and Task Dialogues

However, the statistical analyses still show that there is no significant difference in the performance of the algorithms in general. It is also still the case that the performance of each algorithm significantly varies depending on the data. The only significant difference as a result of the modifications is that the BFP algorithm now performs significantly better on the pump dialogues alone ($\chi^2 = 4.31, p < .05$).

4 Conclusion

We can benefit in two ways from performing such evaluations: (a) we get general results on a methodology for doing evaluation, (b) we discover ways we can improve current theories. A split of evaluation efforts into quantitative versus qualitative is incoherent. We cannot trust the results of a quantitative evaluation without doing a considerable amount of qualitative analyses and we should perform our qualitative analyses on those components that make a significant contribution to the quantitative results; we need to be able to measure the effect of various factors. These measurements must be made by doing comparisons at the data level.

In terms of general results, we have identified some factors that make evaluations of this type more complicated and which might lead us to evaluate solely quantitative results with care. These are: (a) To decide how to evaluate UNDERSPECIFICATIONS and the contribution of ASSUMPTIONS, and (b) To determine the effects of FALSE POSITIVES and ERROR CHAINING. We advocate an approach in which the contribution of each underspecification and assumption is tabulated as well as the effect of error chains. If a principled way could be found to identify false positives, their effect should be reported as well as part of any quantitative evaluation.

In addition, we have taken a few steps towards determining the relative importance of different factors to the successful operation of discourse modules. The percent of successes that both algorithms get indicates that syntax has a strong influence, and that at the very least we can reduce the amount of inference required. In 59% to 82% of the cases both algorithms get the correct result. This probably means that in a large number of cases there was no potential conflict of co-specifiers. In addition, this analysis has shown, that at least for task-oriented dialogues global focus is a significant factor, and in general discourse structure is more important in the task dialogues. However simple devices such as cue words may go a long way toward determining this structure.

Finally, we should note that doing evaluations such as this allows us to determine the GENERALITY of our

approaches. Since the performance of both Hobbs and BFP varies according to the type of the text, and in fact was significantly worse on the task dialogues than on the texts, we might question how their performance would vary on other inputs. An annotated corpus comprising some of the various NL input types such as those I discussed in the introduction would go a long way towards giving us a basis against which we could evaluate the generality of our theories.

5 Acknowledgements

David Carter, Phil Cohen, Nick Haddock, Jerry Hobbs, Aravind Joshi, Don Knuth, Candy Sidner, Phil Stenton, Bonnie Webber, and Steve Whittaker have provided valuable insights toward this endeavor and critical comments on a multiplicity of earlier versions of this paper. Steve Whittaker advised me on the statistical analyses. I would like to thank Jerry Hobbs for encouraging me to do this in the first place.

References

[AP86] James F. Allen and C. Raymond Perrault. Analyzing intention in utterances. In Barbara J. Grosz, Karen Sparck Jones, and Bonnie Lynn Webber, editors, *Readings in Natural Language Processing*, pages 419–422, Morgan Kauffman, Los Altos, Ca., 1986.

[AS89] Anne Abeille and Yves Schabes. Parsing idioms in lexicalized tags. In *Proc. 27th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 161–65, 1989.

[BF83] Roger Brown and Deborah Fish. The psychological causality implicit in language. *Cognition*, 14:237–273, 1983.

[BFP87] Susan E. Brennan, Marilyn Walker Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proc. 25th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 155–162, Stanford University, Stanford, Ca., 1987.

[Car87] David M. Carter. *Interpreting Anaphors in Natural Language Texts.* Ellis Horwood, 1987.

[Coh78] Phillip R. Cohen. *On Knowing What to Say: Planning Speech Acts.* Technical Report 118, University of Toronto; Department of Computer Science, 1978.

[Coh84] Phillip R. Cohen. The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10:97–146, 1984.

[Deu74] Barbara Grosz Deutsch. Typescripts of task oriented dialogs. August 1974.

[DJ89] Nils Dahlback and Arne Jonsson. Empirical studies of discourse representations for natural language interfaces. In *Proc. 27th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 291–298, 1989.

[GJW83] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proc. 21st Annual Meeting of the ACL, Association of Computational Linguistics*, pages 44–50, Cambridge, MA, 1983.

[GJW86] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Towards a computational theory of discourse interpretation. 1986. Preliminary draft.

[Gro77] Barbara J. Grosz. *The Representation and Use of Focus in Dialogue Understanding.* Technical Report 151, SRI International, 333 Ravenswood Ave, Menlo Park, Ca. 94025, 1977.

[GS86] Barbara J. Grosz and Candace L. Sidner. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:pp. 175–204, 1986.

[GSBC86] Raymonde Guindon, P. Sladky, H. Brunner, and J. Conner. The structure of user-adviser dialogues: is there method in their madness? In *Proc. 24st Annual Meeting of the ACL, Association of Computational Linguistics*, pages 224–230, 1986.

[HL87] Julia Hirschberg and Diane Litman. Now lets talk about now: identifying cue phrases intonationally. In *Proc. 25th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 163–

171, Stanford University, Stanford, Ca., 1987.

[HM87] Jerry R. Hobbs and Paul Martin. *Local Pragmatics.* Technical Report, SRI International, 333 Ravenswood Ave., Menlo Park, Ca 94025, 1987.

[Hob76a] Jerry R. Hobbs. *A Computational Approach to Discourse Analysis.* Technical Report 76-2, Department of Computer Science, City College, City University of New York, 1976.

[Hob76b] Jerry R. Hobbs. *Pronoun Resolution.* Technical Report 76-1, Department of Computer Science, City College, City University of New York, 1976.

[Hob78] Jerry R. Hobbs. *Why is Discourse Coherent?* Technical Report 176, SRI International, 333 Ravenswood Ave., Menlo Park, Ca 94025, 1978.

[Hob85] Jerry R. Hobbs. *On the Coherence and Structure of Discourse.* Technical Report CSLI-85-37, Center for the Study of Language and Information, Ventura Hall, Stanford University, Stanford, CA 94305, 1985.

[HTD86] Susan B. Hudson, Michael K. Tanenhaus, and Gary S. Dell. *The effect of the discourse center on the local coherence of a discourse.* Technical Report, University of Rochester, 1986.

[PH87] Janet Pierrehumbert and Julia Hirschberg. The meaning of intonational contours in the interpretation of discourse. In *Proc. Symposium on Intentions and Plans in Communication and Discourse*, Monterey, Ca., 1987.

[Pol86] Martha Pollack. A model of plan inference that distinguishes between the beliefs of actors and observers. In *Proc. 24st Annual Meeting of the ACL, Association of Computational Linguistics*, pages 207–214, Columbia University, New York, N.Y, 1986.

[Pri81] Ellen F. Prince. Toward a taxonomy of given-new information. In *Radical Pragmatics*, Academic Press, 1981.

[Pri85] Ellen F. Prince. Fancy syntax and shared knowledge. *Journal of Pragmatics*, pp. 65–81, 1985.

[Rei76] T. Reinhart. *The Syntactic Domain of Anaphora.* PhD thesis, MIT, Cambridge Mass., 1976.

[Rei85] Rachel Reichman. *Getting Computers to Talk Like You and Me.* MIT Press, Cambridge, MA, 1985.

[Rob88] Craige Roberts. *Modal Subordination and Pronominal Anaphora in Discourse.* Technical Report No. 127, CSLI, May,1988. Also to appear in Linguistics and Philosophy.

[Sch87] Deborah Schiffrin. *Discourse Markers.* Cambridge University Press, 1987.

[SI81] Candace Sidner and David Israel. Recognizing intended meaning and speakers plans. In *Proc. International Joint Conference on Artificial Intelligence*, pages 203–208, Vancouver, BC, Canada, 1981.

[Sid79] Candace L. Sidner. *Toward a computational theory of definite anaphora comprehension in English.* Technical Report AI-TR-537, MIT, 1979.

[Tho80] Bozena Henisz Thompson. Linguistic analysis of natural language communication with computers. In *COLING80: Proc. 8th International Conference on Computational Linguistics. Tokyo*, pages 190–201, 1980.

[Web86] Bonnie Lynn Webber. *Two Steps Closer to Event Reference.* Technical Report MS-CIS-86-74, Linc Lab 42, Department of Computer and Information Science, University of Pennsylvania, 1986.

[WS88] Steve Whittaker and Phil Stenton. Cues and control in expert client dialogues. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, 1988.

[WS89] Steve Whittaker and Phil Stenton. User studies and the design of natural language systems. In *Proc. 27th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 116–123, 1989.

A The Hobbs algorithm

The algorithm and an example is reproduced below. In it, NP denotes NOUN PHRASE and S denotes SENTENCE.

1. Begin at the NP node immediately dominating the pronoun in the parse tree of S.
2. Go up the tree until you encounter an NP or S node. Call this node X, and call the path used to reach it p.
3. Traverse all branches below node X to the left of path p in a left-to-right breadth-first fashion. Propose as the antecedent any NP node encountered that has an NP or S node on the path from it to X.
4. If X is not the highest S node in the sentence, continue to step 5. Otherwise traverse the surface parse trees of previous sentences in the text in reverse chronological order until an acceptable antecedent is found; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as the antecedent.
5. From node X, go up the tree to the first NP or S node encountered. Call this new node X, and call the path traversed to reach it p.
6. If X is an NP node and if the path p to X did not pass through the $\overline{N}$ node that X immediately dominates, propose X as the antecedent.
7. Traverse all branches below node X to the left of path p in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered. Propose any NP or S node encountered as the antecedent.
8. Go to step 4.

The purpose of steps 2 and 3 is to observe the contra-indexing constraints. Let us consider a simple conversational sequence.

U_1: Lyn's mom is a gardener.
U_2: Craige likes her.

We are trying to find the antecedent for *her* in the second utterance. Let us go through the algorithm step by step, using the parse trees for U_1 and U_2 in the figure.

1. NP_5 labels the starting point of step 1.

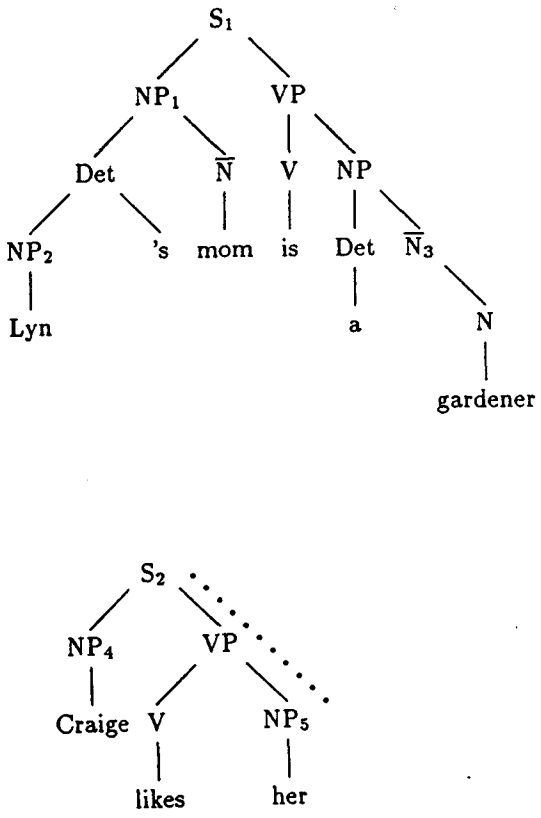

Figure 6: Parse Trees for U_1 and U_2

2. S_2 is called X. We mark the path p with a dotted line.
3. We traverse S_2 to the left of p. We encounter NP_4 but it does not have an NP or S node between it and X. This means that NP_4 is contra-indexed with NP_5. Note that if the structure corresponded to *Craige's mom likes her* then the NP for *Craige* would be an NP to the left of p that has an NP node between it and X, and *Craige* would be selected as the antecedent for *her*.
4. The node X is the highest S node in U_2, so we go to the previous sentence U_1. As we traverse the tree of U_1, the first NP we encounter is NP_1, so *Lyn's mom* is proposed as the antecedent for *her* and we are done.