Computational Approaches to Pronoun Resolution*

Carlos Augusto Prolo**
PUCRS

Abstract: In this paper I review and analyze four proposals for pronoun reference resolution: Hobbs' naive approach [5], an approach based on centering by Brennan, Friedman and Pollard [1], Lappin and Leass' salience method [8], and Kennedy and Boguraev's flat-syntax version of the salience method [7]. I discuss to what extent the proposals can really claim to be computational and how this conformance to an algorithm and other related notions should guide our interpretation of published results about their quantitative performance. Their general coverage of the phenomenon and possibilities of being improved are treated in order to suggest about long term life of the approaches.

1 Introduction

Anaphora is roughly speaking[1] the phenomenon of making in a discourse abbreviated references to entities that have been directly or indirectly introduced by a previous expression. The expression used to make the abbreviated reference is called the *anaphor* and the previous expression the *antecedent*. In this paper I restrict myself to the cases where the anaphor is a pronoun. The concept is illustrated in (1). The occurrence of the pronoun *he* makes an anaphoric reference to the discourse entity introduced by the noun phrase *Carlos*. Hence *he* is the anaphor and *Carlos* is the antecedent. There are two occurrences of the possessive pronoun *its*, the first has as antecedent the paper, and the second, as well as the occurrence of *it*, refers to the introduction of the paper.

---

* This critical review was actually written in 1997 while I was at the Department of Computer and Information Science. University of Pensylvania.

** Faculdade de Informática. PUCRS. Porto Alegre, Brasil. prolo@inf.pucrs.br

1 By using "roughly" I intend also to follow a common conservative position with the definition of this phenomenon not completely understood to the moment.

(1) *Carlos is writing a paper. Right now* ***he*** *is rewriting* ***its*** *introduction because* ***he*** *is not happy with* ***its*** *current form.* ***It*** *should be clear enough so that people not familiar with the area can understand at least the topic the paper is about.*

By *pronoun resolution* I mean the process of finding for each pronoun its antecedent, and the rules that govern the choice of the antecedent by the hearer/listener are the central point of this problem that the proposals here discussed try to capture.

The term *utterance* stands for the use of a sentence or a fragment of sentence in a context. A noun phrase when uttered in a discourse is able to bring to mind the cognitive representation of a *discourse entity* that is part of what is being talked about. Because of this characteristic of being able to refer to an entity it is also called a *referential expression*. A referential expression may introduce a new discourse entity or make reference (e.g. anaphoric reference) to a previously introduced one.

Traditionally an anaphor and its antecedent are said to *corefer*, because they would be evoking the same discourse entity, and hence anaphora would be a mechanism characterized by *coreference*. Two objections can be raised against the use of this term. One is due to the fact that in traditional semantics, reference is used (in extensional semantics) to mean the relation between a referential expression and an entity in a real world represented by the expression. Since we want to stay in a more abstract level, where entities are represented by their properties, and also where non-physical entities as well as imaginary ones can be represented, Sidner [10] proposes the term *specification* for the representations at this level, restating that the representation of the anaphor *co-specify* (instead of corefer) with the representation of the antecedent.

The second objection is that both coreference and this first definition of co-specification are not adequate enough to cope with anaphora since anaphor and antecedent do not always represent the same entity. In (2) the interpretation for *they* is strongly related to the previous occurrence of *a monster Harley 1200*, and we still say that the latter is the antecedent of the former. But in this case, the specifications associated with these two referential expressions are certainly not the same, i.e. neither they corefer in a strict sense, nor their specifications co-specify in Sidner's definition. However Sidner herself loosen her definition and uses the term co-specification to mean the relation existing between anaphor and antecedent (see note 1 in [10]). This seems to be the widespread usage of the expression in the literature. In this paper I will use both co-specification and the traditional coreference with this latter extended meaning.

(2) *My neighbor has a monster Harley 1200.*
They are really huge but gas-efficient bikes.

Strictly speaking, in anaphora, as defined above, the antecedent comes before the anaphor in the discourse. There is a similar process where the abbreviated expression comes before the referential expressions to which it corefers. This process, called *cataphora*, is exemplified in (3) where *he* is coreferential with *Paul*. Together anaphora and cataphora would be called *endophora* (Hirst [4]). Because anaphora is much more widely used and has been given more attention than cataphora it ended up taking the place of endophora in much of the literature and I will here follow the tendency to include cataphora as a special case of anaphora (for some, "backward" anaphora).

(3) *Was he younger, Paul would be able to climb that tree.*

The kind of pronouns that makes sense to talk about for pronoun resolution are the *specific* or *definite* pronouns. The approaches discussed in this paper deals with the subset of *central* pronouns, consisting of the *reflexives* (e. g. herself), *reciprocals* (e. g. each other), *personal* (he, him) and *possessive* (his) pronouns. Because they have the same locality constraints, I will refer to reflexives and reciprocals together as *reflexives* (also *lexical anaphors*, according to [8,7]), while using *non-reflexive* (*non-lexical anaphors* in [8,7]) for the other two categories (personal and possessive pronouns).[2]

To resolve a pronoun is actually more complicated that finding an adequate noun phrase in the discourse to be the antecedent. It may require finding out that there is no antecedent. In English sometimes a pronoun does not represent a discourse entity. The usages of "it" in (4) are known as pleonastic or expletive. They do not point to an entity, they are just there to fill the requirement in English that sentences have to have an overt subject. In other situations the pronoun may corefer with something else than a referential expression introduced by a noun phrase, and in order to understand their antecedent we have to nominalize some other kind of phrase, as in (5) where *it* refers to the whole fact expressed by the initial clause. Finally in (6)[3] it is really an art to carve out the antecedent of *they* which is not explicit in the text. These are typical problems that have to be considered by the computational methods, some of them fortunately of infrequent occurrence.

---

[2] For an account of demonstrative pronouns see [10]. For a complete taxonomy and description of pronouns see [9].

[3] From Sidner [10].

(4) *It is snowing but it is time go home.*

(5) *I didn't say anything wrong in the last two minutes and you know it.*

(6) *I went to a concert last night.*
They played Beethoven's ninth.

I finally mention in passing that the approaches presented here do not take into account the precise semantic relation between the anaphor and the antecedent (or between the specifications they introduce). For instance, in Sidner's (7) and (8) there is clearly a semantic distinction between the interpretations of the antecedents (*a vegomatic* and *TWA 384*) uttered in the context of the first sentence of each fragment, and the posterior reference using the pronoun *it*. This involves the concepts of specific versus non-specific interpretation, prototype versus instance. However, all the approaches would get the pronouns bound correctly in this examples. For a more extensive discussion of anaphora and related concepts I suggest the reading of Hirst([4]) and Sidner([10]).

(7) *Sally wanted to buy a vegomatic.*
She had seen it advertised on TV.

(8) *TWA 384 was so bumpy this Sunday I almost got sick.*
It usually is a very smooth flight.

I consider in this paper four approaches for pronoun resolution: Hobbs' naive approach [5], the approach based on centering due to Brennan, Friedman and Pollard [1], Lappin and Leass' salience method [8] and Kennedy and Boguraev's version of Lappin and Leass' method relying on approximate configurational relations from a flat morpho-syntactic analysis [7]. In section 2 it is given an overview of the approaches with some contextual information needed to understand them (e.g., a brief introduction to centering to understand Brennan, Friedman and Pollard's approach). I also identify the general assumptions they make on the input and pre- and co-processing modules.

I decided to split the description of the approaches into two very distinct parts: the morpho-syntactic filter and the core of the approach. The morpho-syntactic filters presented in section 3 are variations of some more or less fixed general principles that have been studied for a long time, giving locality constraints for the binding of pronouns to antecedents. Hence in Section 4 I focus

specifically on the core of the approaches. One objective of this factorization is that one can read it assuming the existence of a filter in any of its theoretical formulations or pragmatic implementations.[4] In both sections 3 and 4 the algorithms presented are critically analyzed.

Section 5 contains some published results about the individual or comparative performance of the algorithms that gives a short-time perspective in their evaluation.

Most of the readers of Hobbs [5] and Sidner [10] would be amazed to see the variety of problems they present while discussing their methods of solving definite anaphora. I particularly was equally surprised to see how their methods decrease in practical feasibility as they try to cover more complicated cases.[5] It is essential for a computational method either that it be presented as an algorithm (an unambiguous description containing a finite number of mechanically executable steps) or that one can easily see how to find such algorithm. A discussion of these two fundamental aspects are presented in Sections 6 and 7. In Section 6 I discuss to what extent the proposals are really computational approaches, i.e., algorithms, as well as some other subtle aspects very important for the interpretation of results as those of section 5, and which are frequently not given the due importance by the reader.[6] Chapter 7 has a general coverage analysis which helps to see how far the approaches are of solving the problem.

Finally in Section 8 I briefly touch the aspect of improvability, to give an idea of the long-term perspective for the approaches. I conclude in Section 9.

2 Overview of the approaches

Hobbs' naive method[7] is very simple in conception. He presents an algorithm that, given a pronoun, visits the nodes of the parse trees in a pre-determined order starting at the pronoun,

---

4 For instance factorizing the filter from Hobbs' algorithm results in a much cleaner presentation of his method of ordering the choices for antecedents to see why he called it naive.

5 This is not to be seen as a criticism. For instance Hobbs start by presenting its naive algorithm and proceeds by showing its flaws in coverage. Both Hobbs and Sidner give a very good insight on how difficult it is to get an algorithm for definitely solving the problem of anaphora resolution.

6 And here is a point where the writers can blurry the capacity of judgment by the readers.

7 Naive is how he himself refers to the method he presented in the first half of [5].

searching for possible antecedents in the parse tree of the sentences while filtering the noun phrases contra-indexed to the pronoun. This straightforward search strategy over the parse tree despite seeming naive, actually behaves somewhat in accordance with certain generally observed rules of preferences for antecedents in English.

Hobbs' method assumes its input is a parse tree for each sentence and moreover this tree has feature values of gender and number that allow to filter also on a morphological agreement basis. Moreover he assumes without further details that the algorithm should be part of a larger interpretation process which also recovers syntactically recoverable omitted material and records coreference and non-coreference relations. A version dealing with selectional constraints is also considered but no much is said about it.

The second approach due to Brennan, Friedman and Pollard [1] is based on centering. Centering was introduced by Grosz, Joshi and Weinstein [2] as a framework for modeling the local component of the attentional state in discourse. They were concerned with the relations among the choice of referring expressions in discourse and the level of discourse coherence. They would say that discourse coherence is partly determined by different inference demands made to readers/hearers by different choices of referring expressions. And they would claim that what determines the different inference demands (and consequently affecting the levels of discourse coherence) caused by different choices of referring expressions is indeed the attentional state at the time the expressions are uttered.

Centering has been proposed for anaphora resolution on the basis that pronoun interpretation would be related to achieving low inference load demands from hearers, i.e., that given a pronoun and the attentional state at the moment it was uttered, the antecedent for the pronoun as interpreted by a hearer would be the referring expression which led to minimum inference demands.[8] Brennan, Friedman and Pollard in [1] present a proposal on this basis where they apply a modified version of the original

[8] It may appear redundant to mention the hearer as the interpreter of the pronoun since the notion of antecedent is obviously based to the interpretation hearers would make. Notice however that hearers sometimes fail to resolve correctly the pronoun in the first attempt leading them to backtrack to get the correct choice. Centering theory would say that the wrong choice would be related to the fact that it led to a relative low level of inference demand at the moment the pronoun was uttered.

proposal in [2][9] for resolving pronouns with the set of candidate antecedents previously filtered by a conventional morpho-syntactic filter. I give next a brief introduction to centering to allow for the understanding of the algorithm presented in the next sections.

In the centering framework it is proposed that each utterance U in a discourse DS is assigned a partial order C_f of discourse entities called the *forward looking centers* of U and all but the first utterance are assigned a *backward looking center* C_b. The C_b of an utterance is the confirmation of the center of the discourse at the transition from the previous utterance. The C_f presents a rank, in the form of a partial order, of the prominence of the discourse referents candidates to be the center of the discourse as the discourse moves to the next utterance. The more highly ranked an element of C_f (U_n) is, the more likely it is to be the C_b (U_{n+1}) and indeed they state that the most highly ranked element of C_f (U_n) that is realized in U_{n+1} is the C_b (U_{n+1}).

The forward-looking centers of an utterance U_n depend only on the expressions that constitute that utterance. Although intuitively we might think they should include contributions from all noun phrases, this is not true, e.g. negated noun phrases do not contribute with forward-looking centers. I have used the term "contributes" because not only the entity directly realized by an expression is considered but also entities indirectly realized. For example in the discourse (9) the set of forward looking centers of the second sentence includes not only the door of the house but the house itself. Thus we see the need of a powerful inference mechanism for computing realization. Finally, other kinds of phrases may contribute, such as verbal phrases or entire clauses, say, with the event conveyed by them, in a process of nominalization (recall (5)).

(9) *I bought a house yesterday*
The door was broken.

It is defined four types of transition relations across pairs of utterances U_n and U_{n+1}.[10]

---

[9] The original work by Grosz, Joshi and Weinstein was widely known since the beginning of the 80's and much of the work derived from them refer to a non-published manuscript dated 1986. [2] is the published revised version of this manuscript.

[10] I actually present here the extended transition cases proposed in [1]. The original centering proposal has merged the two cases of shifting in one.

1. **Center Continuation:** When $C_b(U_{n+1}) = C_b(U_n)$ is the most highly ranked element of $C_f(U_{n+1})$. This means that in the transition from the utterances not only the center was preserved but also it remained the most ranked candidate to be the center of the following (U_{n+2}) utterance.
2. **Center Retaining:** In this case the center is preserved from U_n to U_{n+1}, but is no longer the best candidate to be the center of the following utterance (U_{n+2}). This case happens when $C_b(U_{n+1}) = C_b(U_n)$ is not the most highly ranked element of $C_f(U_{n+1})$.
3. **Center Shifting-1:** $C_b(U_{n+1}) \neq C_b(U_n)$ and $C_b(U_{n+1})$ is the highest ranked element of $C_f(U_{n+1})$, i.e. the center was changed and the new center is likely to be preserved for the next utterance.
4. **Center Shifting-2:** $C_b(U_{n+1}) \neq C_b(U_n)$ and $C_b(U_{n+1})$ is not the highest ranked element of $C_f(U_{n+1})$, i.e. the center was changed and is likely to change again at the next utterance.

The basic constraint proposed on center realization is given by Rule 1.

> **Rule 1** For any given utterance U_n, if any element of the $C_f(U_n)$ is realized in the utterance by a pronoun in U_{n+1} then the $C_b(U_{n+1})$ must be realized by a pronoun.

The basic constraint on center movement is given by Rule 2.[11]

> **Rule 2** Sequences of transitions are prefered in this order:
> *continuation > retaining > shifting-1 > shifting-2.*

The assumption of Brennan, Friedman and Pollard's algorithm is that these two rules are directly applied to determine the antecedent of a pronoun. Notice however, that this claim of was not made in [2] by Grosz, Joshi and Weinstein with respect to the second rule. On the contrary, [2] just argue that violations of Rule 2 would produce an increase in the inference load compared to an alternative that conforms to the rule, but are indeed quite possible in discourses. There are actually other factors that have to be accounted for on a weight basis. That is the principle of the next approach presented below.

Lappin and Leass' method[8] relies on salience measures derived from syntactic structure and a simple dynamic model of attentional state. The factors taken into account for computing

[11] Again this conforms to the extensions in [1].

salience values of an NP are its grammatical role, parallelism of grammatical roles, frequency of mention, proximity, and sentence recency. NP chains linked by coreference are treated as an equivalence class. The salience value of a class is the sum of the individual contributions of each element in it. The model of attentional state is dynamic to the point of caring for negative contributions to the salience values. For example the equivalence classes have their current salience strongly devaluated when crossing the boundary of a sentence. Pleonastic pronouns are identified and excluded of the process of pronoun resolution.

Each time a pronoun is seen, the current salience value of the possible antecedents is evaluated and they are proposed in order of their salience value to bind the pronoun. A morpho-syntactic filter is then applied to the proposed elements until one is accepted.

The algorithm is implemented as a part of a system that provides it with the syntactic representations of the parse trees (generated by a Slot Grammar parser) with information on the head-argument and head-adjunct relations.

Kennedy and Boguraev present in [7] a version of Lappin and Leass' algorithm where instead of relying on a full syntactic parsing of text, their input comes from a part of speech tagger enriched with annotations of grammatical function of the lexical items in the input text stream.[12] Each lexical item in each input sentence is marked with a linear position identification, plus morphological, lexical, grammatical and syntactic features of the item in the context it appears. So, given the text (from [7])

> "For 1995 the company set up its headquarters in Hall 11, the newest and most prestigious of CeBIT's 23 halls."

we have the following analysis that becomes the input for the resolution algorithm. The linear position index has been omitted for simplicity.

[12] This flat morpho-syntactic tagging system was developed under the Constraint Grammar project by Karlsson, Voutilainen, Heikkila and Antilla [6].

lexical item	base form	POS	morpho-syntactic features	grammatical function
For	for	PREP		@ADVL
1995	1995	NUM	CARD	@<P
the	the	DET	SG/PL CENTRAL ART	@DN>
company	company	N	SG/PL NOM	@SUBJ
set	set	V	PAST VFIN	@+FMAINV
up	up	ADV		@ADVL
its	it	PRON	GEN SG3	@GN>
headquarters	headquarters	N	NOM SG/PL	@OBJ
in	in	PREP		@ADVL @<NOM
Hall	hall	N	NOM SG	@NN>
11	11	NUM	CARD	@<P
,	,	PUNCT		
the	the	DET	SG/PL CENTRAL ART	@DN>
newest	new	A	SUP	@P-COMPL-O
and	and	CC		@CC
most	much	ADV	SUP	@AD-A>
prestigious	prestigious	A	ABS	@<P
of	of	PREP		@<NOM-OF
CeBIT's	cebit	N	GEN SG	@GN>
23	23	NUM	CARD	@QN>
halls	hall	N	NOM PL	@<P
		PUNCT		

Notice particularly how the syntactic information is poorly presented, basically giving notions of the relation of elements to a head. For instance, in the example above, the word *in* is ambiguosly tagged with two possible grammatical functions: (head of an) adverbial adjunct (@ADVL) and (head of a) postmodifier of a noun phrase (@<NOM). Even when you have a unique correct tag you may have ambiguity in the possibilities of attachment. For instance in (10) the word *behind* would be marked @<NOM, meaning that *behind the tree* is a postmodifier of a noun phrase. Only it is not said whether it modifies the door or the house. Consider also the pair of sentences in (11). Despite having different parse trees, their shallow analysis would give exactly the same sequence of tokens (ignoring the punctuation information). In general, it is hard to guess the proper structure of subordina-

tion, embedding, attachment, etc. Hence, it is remarkable how they still achieve good results with their approximations about the complex relations one needs to get to for marking coreference (e.g., commanding, embedding, co-argumentship).

(10) *The door of the house behind the tree is yellow.*

(11) *When he laughs, we know he is lying.*
If he thinks we know, he is wrong.

Due to the lack of complete configurational relation among sentence elements, there is a pre-processing task of approximating some of this constituency relations essential for either the filters or the algorithm of salience itself. This basically consists of finding all the NPs with their position range and modifier-head relations, as well as approximating contextual relations of containment, especially embedding of noun phrases into adverbial adjuncts and into other noun phrases (e.g. prepositional/clausal complements, relative clauses). This is done with the use a set of filters stated as regular expressions or phrasal grammars over an alphabet of meta-tokens like those presented in the table above. Filtering of the expletive "it" is also performed. Discourse referents resulting from the pre-processing phase contain information on grammatical function (GFUN), information on whether the noun phrase was found embedded into another noun phrase (EMBED field), and whether the noun phrase was contained in an adverbial adjunct (ADJUNCT field).[13] The details on how the noun phrases are extracted as well as the EMBED and ADJUNCT attributes are not given in [7].

3 The morpho-syntactic filter

A morphological filter is designed to block co-indexing of referential phrases that do not agree in the morphological features of gender, number and person. This accounts for our judgment that sentence (12) is acceptable but (14) is not, and that the pronoun in (13) can corefer with the subject of the sentence but not in (15) and (16).

[13] Also information concerning agreement, referential type and position in the string (given by the position of the first token of the noun phrase) is available.

(12) *John likes himself.*

(13) *Mary$_i$ thinks she $_{(i\,j)}$ is beautiful.*

(14) **John likes herself.*

(15) *Mary$_{(i)}$ thinks he $_{(*i\,j)}$ is beautiful.*

(16) *[Mary and Betty]$_{(i)}$ think she$_{(*i\,j)}$ is beautiful.*

The syntactic filter has two traditional functions. The first is to restrict to a local domain the referential expressions that can corefer with a lexical anaphor, based on syntactic and grammatical rules (see sentences (17) to (20)). The second is to block coreference of a non-reflexive pronoun with an antecedent with which it is *contra-indexed* (we also call this *disjoint reference*), also in terms of syntactic and grammatical functions (see sentences (21) to (24)).

(17) *[Dick Dastardly]$_{(i)}$ believes himself$_{(i)}$ to be the best racer.*

(18) **Dick Dastardly believes (that) himself is the best racer.*

(19) *Barney$_{(i)}$ gave Fred$_{(j)}$ a picture of himself$_{(i\,j)}$.*

(20) **Barney gave Fred a gift that wilma knew it was a picture of himself.*

(21) *[Dick Dastardly]$_{(i)}$ believes him$_{(*i\,j)}$ to be the best racer.*

(22) *[Dick Dastardly]$_{(i)}$ believes (that) he$_{(i\,j)}$ is the best racer.*

(23) *Barney$_{(i)}$ gave Fred$_{(j)}$ a picture of him(*i, * j, k).*

(24) *John$_{(i)}$ likes him$_{(*i\,j)}$.*

None of the papers tell how they implement the morphological filter. This may seem at first glance to be a simple task. However this is not the case. It is true that once the name phrases have the correct feature values, the algorithm for the morphological filter is an easy task. However, the hard task is exactly to assign the correct feature values especially of gender to the noun phrases. Consider the segments (25) and (26) below. The reader is able to understand who is *he* and *she* (*her* and *him*) in the discourses only if he knows that, e.g., the prime-minister is Margaret Thatcher and the president of U.S. is Ronald Reagan, and readers strongly rely on this factor in guiding their judgment on binding the pronouns. However in order for an automatic system to handle it, it needs world knowledge. In (27) and (28) it is even more striking the precedence of agreement over other factors working as a filter. In (27) general focusing/centering rules [10,2] make it

easy (for a reader) to interpret *she* in the second sentence as the mother, but in (28), due to agreement failure, *she* has to be interpreted as the spouse. Now notice that here it is even more difficult to get the correct gender for the spouse, because even this simple example requires an inference system to derive that *her spouse* is male.

(25) The prime-minister of the United Kingdom will meet the president of U.S.
She will tell him about her strong restrictions to his economic policies.

(26) The prime-minister of the United Kingdom will meet the president of U.S.
He will tell her about his strong restrictions to her economic policies.

(27) She asked her daughter to stop working.
She said it was time to go home.

(28) She asked her spouse to stop working.
He said it was time to go home.

For Hobbs its clear he is assuming the sentences correctly parsed and decorated with feature values wherever needed as the input for his algorithm. On the other hand, for the automatic systems of Lappin and Leass and Kennedy and Boguraev they have to rely on modules not described in the papers for the automatic morphological feature. This has to be taken into account when comparing their performance results as I will do later in this paper.

I briefly describe next the syntactic filters according to the papers, and show their weaknesses. For finding the problems and counter-examples concerning the syntactic filter I some times used terms from the presentation given by the Binding Theory to this problem (see Haegeman [3]).

3.1 *Hobbs' filter*

I have extracted from Hobbs' algorithm [5] the filter for contra-indexing (recall he does not address lexical anaphora). According to him the following algorithm finds the noun phrases in a sentence to which a pronoun can not corefer.

1. Begin at the NP node immediately dominating the pronoun P.[14]
2. Go up the tree to the first NP or S encountered. Call this node X and the path used to reach it p.
   If X is an NP mark X as contra-indexed to P.
3. Mark as contra-indexed to P all NPs below node X and to the left of path p which do not have an NP or S between it and X.
   Mark as contra-indexed to P all NPs below node X and to the right of path p.
4. If node X is the highest S in the sentence STOP.
5. (otherwise) From node X go up the tree to the first NP or S encountered. Call this new node X, and call the path traversed to reach it p.
6. If X is an NP node and the path p to X passed through the N-bar node that X immediately dominates mark X contra-indexed to the pronoun P.
7. If X is an NP node mark as contra-indexed to P all NPs below X and to the right of path p.
   Otherwise (if X is an S node) mark as contra-indexed to P all NPs below node X and to the right of path p that have either an S or another NP intervening in the path between it and X.
8. Go to step 4.

Hobbs' algorithm has a subtle problem of not extending the contra-indexing process to the discourse referents previously linked by coreference. This means that it is possible to find two noun-phrases NP_1 and NP_2 that are found to be contra-indexed and indeed both be co-indexed with a third NP_3. For example, in the discourse fragment (29) *he* and *him* are contra-indexed. Hence, we know that *Paul* can not be both the antecedent of *he* and of *him*. At the moment that we find that *Paul* is the antecedent of *he*, automaticaly we discard *Paul* as the antecedent of *him*, and find it to be *John*. Hobbs would allow for that mistake, making *Paul* the antecedent of both pronouns. Here also it is somewhat contradictory the fact that in his formal algorithm this mistake come out, whereas he mentions as I said in the introduction to be assuming coreference chains to be an equivalence class. I suspect this chains may be related to the binding of relative pronouns, which is required for the syntactic filter to work properly.

---

[14] Notice that the possessive pronouns are split into a personal pronoun immediately dominated by an NP, and a possessive case marker. This NP would be the one referred by this rule. The noun phrase *his house* would have the following analysis:
$[_{NP}\ [_{Det}\ [_{NP}\ he]\ 's]\ [_{\bar{N}}\ house]]$.

(29) It was John who had eaten the cake.
Paul didn't need to make any question.
He knew him very well.

The algorithm blocks cataphora in some situations that despite unlikely for coreference are not to be thought as blocked in the sense of a filter. Step 3 would block the interpretation of *her* as *Mary* in sentence (30) below. Similarly step 7 blocks cataphoric binding by NPs below a first level of NPs and Ss. Hence, while it accepts the cataphoric reference in (31) and (32) it would block it in sentences (33) to (35). This excess in the filtering seems to have been done for pragmatic reasons. I will argue in the next section that this is actually due to the fact that the core algorithm gives cataphora a higher standing that our intuition would normally accept and Hobbs appeals to the filter to avoid large decreases in the hit ratio of the algorithm in practice.

(30) *He gave her$_{(i,j)}$ a necklace Mary$_{(i)}$ could never expect to receive from him.*

(31) *The book I gave to him$_{(i,j)}$ is still with John$_{(i)}$.*

(32) *Had he$_{(i,j)}$ been given a new heart, John$_{(i)}$ wouldn't have died.*

(33) *The book I gave to him$_{(*i,j)}$ is still John's$_{(i)}$ mother.*

(34) *I believe his$_{(i,j)}$ brother to be John's$_{(i)}$ best friend.*

(35) *Had he$_{(i,j)}$ been given a new heart, John's$_{(*i)}$ mother wouldn't be crying.*

Another observation I made at first was that Hobbs' filter didn't account for cases where the governing category (GC) of the pronoun is more ample due to ECM (exceptional case marking, using the GB terminology), when the pronoun is the subject of an infinitival clause which is the argument of a verb. Like in (36), Hobbs would not block the co-indexation of *John* and *him* (and actually it would propose *John* as the antecedent of *him*). However it's more plausible to accept that he is using another scheme for syntactic representation of the parse trees, as the one who would raise *him* to the condition of one of the objects of *believe* in the surface tree.

(36) *John$_{(i)}$ believes him$_{(*i,j)}$ to be the best.*

3.2 *Brennan, Friedman and Pollard's filter*

It's clear from both [1] and [11] that the proposal of Brennan et al. assume a syntactic filter, but they did not describe it in either paper.

3.3 *Lappin and Leass' filter*

Lappin and Leass' proposal provides a morphological filter that checks for agreement of number, gender, and person. The discourse representation and the details of the process of acquisition of feature values are not presented in [8]. The syntactic filters there are presented below. The following definitions are used:

- a phrase P is in the *argument domain* of a phrase N iff P and N are both arguments of the same head.
- P is in the *adjunct domain* of N iff N is an argument of a head H, P is the object of a preposition PREP, and PREP is an adjunct of H.
- P is in the *NP domain* of N iff N is a determiner of a noun Q and (i) P is an argument of Q, or (ii) P is the object of a preposition PREP and PREP is an adjunct of Q.
- a phrase P is *contained in* a phrase Q iff (i) P is either an argument or an adjunct of Q (*immediately contained*), or (ii) P is immediately contained in some phrase R and R is contained in Q.
- The notion of *higher argument slot* is defined by the following hierarchy of argument slots:

  Subj > Agent > Obj > (Iobj = Pobj)

where

Subj is the surface subject
Agent is the deep subject of a verb heading a passive VP
Obj is the direct object
Iobj is the indirect object
Pobj is the object of a PP complement of a verb (e.g. in 'put NP on NP')

N is a possible antecedent for a reflexive (or reciprocal) A if one of the following conditions holds:

1. A is in the argument domain of N, and N fills a higher argument slot than A.
2. A is in the adjunct domain of N.
3. A is in the NP domain of N.
4. N is an argument of a verb V, there is an NP Q in the argument or adjunct domain of N such that Q has no noun determiner, and (i) A is an argument of Q, or (ii) A is an argument of a preposition PREP and PREP is an adjunct of Q.
5. A is a determiner of a noun Q, and (i) Q is in the argument domain of N and N fills a higher argument slot than Q, or (ii) Q is in the adjunct domain of N.

It is important to note that their system has a notion of trace assumed by the algorithm. This is clear from one example for Rule 1 presented in [8] transcribed here as (37). The way this complies to rule 1 is that actually the reciprocal is bound by the trace in the object position of the verb *introduced* which refers to *people*. A problem of Rule 1, which requires a strictly higher order of the argument slots, is that since Iobj is in the same position of Pobj in the hierarchy Rule 1 would not account for (38).

(37) *Mary knows the people$_{(i)}$ who John introduced to each other$_{(i)}$.*

(38) *Mary talked to John$_{(i)}$ about himself$_{(i)}$.*

Rule 3 is actually the counterpart of Rules 1 and 2 when the head is a noun. However, I posit that the definition of NP domain is too restrictive and does not account for (39).

(39) *A book about Poirot$_{(i)}$ written by himself$_{(i)}$ would be an interesting example of embedding in an Agatha Christie plot.*

Rule 4 extends Rules 1 and 2 to a second level of depth, provided there is no "accessible SUBJECT" (in the sense of GB theory) in the embedded clause. Hence (40) is accepted but (41) is not. However, we actually need an extension to an arbitrary number of levels. For example the algorithm fails to accept (42). On the other hand it is missing a similar extension to a deeper level in Rule 3. Hence (43) and (44) are also not accepted.

(40) *They$_{(i)}$ told stories about themselves$_{(i)}$.*

(41) **They$_{(i)}$ told Mary's stories about themselves$_{(i)}$.*

(42) *They$_{(i)}$ told summaries of the stories about themselves$_{(i)}$.*

(43) *Bill's$_{(i)}$ version of the rumors about himself$_{(i)}$ is inconsistent.*

(44) *John's$_{(i)}$ mood in the picture of himself$_{(i)}$ is terrible.*

In the case where (according to GB) a verb takes as an argument a non-finite clause with overt subject, governing this NP subject through ECM seems to be accounted for here by raising the NP to the condition of an object of the main clause.

Let us turn now to the filter for non-lexical anaphora. A pronoun P is non-coreferential with a noun phrase N if and only if any of the following conditions hold:

1. P is in the argument domain of N.
2. P is in the adjunct domain of N.
3. P is an argument of a head H, N is not a pronoun, and N is contained in H.
4. P is in the NP domain of N.
5. P is a determiner of a noun Q, and N is contained in Q.

Rules 1, 2 and 4, the counterparts of rules 1, 2 and 3 for reflexives are not extended to deeper levels, and hence we have cases like (45) and (46) not blocked. Still, Rule 4 does not filter (47).

(45) *They$_{(i)}$ told stories about them$_{(*i)}$*

(46) *Bill's$_{(i)}$ version of the rumors about him$_{(*i)}$ is inconsistent.*

(47) *A book about Poirot$_{(i)}$ written by him$_{(*i)}$ would be an interesting example of embedding in an Agatha Christie plot.*

Rule 3 is probably intended to exclude non-pronominal referential expressions "c-commanded" by the pronoun from the set of candidate antecedents. However there is a clear problem in its formulation since it prevents the acceptance of genuine cases of co-indexing such as in (48) and (49). An additional requirement that P be in a higher slot than the co-argument in which N is embedded would solve the problem.

(48) *The wife of John$_{(i)}$ likes him$_{(i)}$.*

(49) *The dark side of Mary$_{(i)}$ frightens her$_{(i)}$.*

Rule 5 rules out the few positions where antecedents for possessive pronouns are prohibited. It considers in part the filtering of "i-within-i" references, where a pronoun is linked to a noun phrase that contains it. But filtering of references such as (50) are not ruled out.

(50) *The boss$_{(i)}$ of his$_{(*i,j)}$ son was sick.*

An important characteristic of Lappin and Leass' approach is that it confers the status of an equivalence relation to coreference, and hence contra-indexing happens made between pairs of equivalence classes instead of pairs of noun phrases. Hence, when it is faced with the example (29) where Hobbs would fail, it correctly rejects the co-indexing of *him* to *Paul* on the basis that *he* is co-indexed to *Paul* and *him* is contra-indexed to all the class to which *he* belongs.

A third filtering element presented by Lappin and Leass which is of extreme importance in English is not concerned with the possible antecedents, but with the pronoun itself: the filtering of the pleonastic occurrence of 'it' in a sentence. Having defined a class of *modal adjectives* (e.g. possible, convenient, important, etc.) and a class of *cognitive verbs* (e.g. assume, expect, etc.) pleonastic occurrences of 'it' are discarded by looking for occurrences of certain constructions. Among them:

- It is **ModalAdj** that **S**
- It is **CogVerb-ed** that **S**
- It is time to **VP**

Also included are some variants where the verbs are preceded by negation particles or modals.

3.4 *Kennedy and Boguraev's filter*

The binding domain for reflexives (and reciprocals) is determined by using grammatical function information and precedence relations. If the GFUN of the reflexive is *subject*, then the closest preceding discourse referent with a GFUN value of *subject* is identified as a (unique) possible antecedent. In the rest of the cases (when the GFUN of the reflexive is either *indirect object* or *oblique*[15]) both the closest preceding subject and the closest preceding direct object that is not separate from the anaphor by a subject are identified as possible antecedents. The rule is clearly weak, since it does not account for noun phrases in determiner position as antecedents as in (51). Also embedding is not accounted for and *John* would be taken to be the antecedent of *himself* in (52).

(51) *Once again Carlos used the example involving Mary's$_{(i)}$ picture of herself$_{(i)}$.*

(52) The *man$_{(i)}$ at* the table John was serving was hurting *himself$_{(i)}$.*

Their version of the theoretical conditions for disjoint references is as follows.

[15] As in Lappin and Leass' Kennedy and Boguraev's syntactic framework never assigns grammatical function of subject to reflexives as is the case in Binding Theory under ECM.

Condition 1: A pronoun can not corefer with a co-argument.
Condition 2: A pronoun can not corefer with a non pronominal constituent which it both commands and precedes.
Condition 3: A pronoun can not corefer with a constituent which contains it.

The above conditions are rather weak. Condition 1 corresponds exactly to Lappin and Leass' Rule 1 for disjoint reference, but no extension is made for the cases where the pronoun is contained in deeper levels of the co-argument. Also adjuncts are not accounted for (Lappin and Leass' Rule 2), and the counterpart for noun phrases (Lappin and Leass' Rule 4) is also missing. Hence it would fail filtering not only (45), (46), (47), but also (53) and (54).

(53) *John$_{(i)}$ was pictured as a criminal in the film about him$_{(*i)}$.*

(54) *John's$_{(i)}$ picture of him$_{(*i)}$ is getting tired of being considered as an odd case of disjoint reference.*

Condition 2 corresponds to the probable intended meaning of Lappin and Leass' Rule 3 (which I argued was ill formulated). However the requirement that the pronoun precede the non-pronominal constituent excludes (55) from being filtered.

(55) *Mary introduced John's$_{(i)}$ advisor to him$_{(*i)}$.*

Finally, Condition 3 corresponds to an "i-within-i" filter and is stronger than Lappin and Leass' Rule 5 which only considers possessives.

The implementation of the filter is obviously harder than for Lappin and Leass and indeed an important issue, because the conditions above can not be precisely checked without a full parse tree. The computationally viable approximations of these conditions under the flat constituency analysis of the input is presented below.

Condition 1: Implemented by finding all pronouns with grammatical function GFUN value equal to *direct object*, *indirect object*, or *oblique* which follow a discourse referent with GFUN *subject* or *direct object*, as long as no *subject* intervenes. Such pairs of discourse referent-pronoun are identified as disjoint.[16]

[16] In the original paper ([7]) the condition is stated as "Find all DISCOURSE REFERENTS with grammatical function GFUN value equal to *direct object*, *indirect object*, or *oblique* which follow a PRONOUN with GFUN *subject* or *direct object* as long as no *subject* intervenes. Such pairs of discourse referent-pronoun are identified as disjoint". I considered that the capitalized referential expressions (DISCOURSE REFERENT and PRONOUN) were unintentionally switched by the authors and I corrected it in my definition.

Condition 2: Implemented by locating for every non-adjunct and non-embedded pronoun the set of non-pronominal discourse referents in its sentence which follow it, and marking these pairs disjoint.
Condition 3: Makes use of the observation that a discourse referent contains every object to its right with a non-nil EMBED value. The algorithm identifies as disjoint a discourse referent and every pronoun which follows it and has a non-nil EMBED value, until a discourse referent with the EMBED value nil is located. Additionally coreference between a genitive pronoun and the NP it modifies is ruled out here.

I now analyze how the implementation conforms to the stated conditions. Actually it is clear that many flaws should arise due to the lack of a precise analysis of constituency relations.

The implementation of condition 1 does not account well for embedding. As a first case, take sentence (56), partially annotated with grammatical functions where relevant. The rule would filter out the co-indexing of *him* to *John* because there is no intervening subject. It could indeed be pointed out that *the fact that John did not come* is a subject and could be thought of as intervening, since it contains *John*. I consider this a good solution for the problem, but according to Kennedy and Boguraev they use exclusively the starting position of the noun phrase chunk to compare precedence. However (57) and (58) do not have the same straightforward solution. In both cases there is no subject at all between *Mary* and *her*, and hence the algorithm would block the co-indexing. The point that should be clear is that the proposal can not cope with embedding. It is hard to see how could one figure out in the general case, from the shallow information on constituency relations, that *Mary* is not a co-argument of *her* because it is in a different level of embedding.

(56) *The fact that John/Subj$_{(i)}$ did not come doesn't make him/DirObj$_{(i\,j)}$ bad.*

(57) *John gave the CD Mary/Subj$_{(i)}$ liked to her/IndObj$_{(i\,j)}$.*

(58) *John asked the kid who claimed to know Mary/DirObj$_{(i)}$ about her/Oblique$_{(i\,j)}$.*

Now, before one thinks of just making the restrictions stronger in order to avoid filtering out the above sentences, let us show that the problems happen the other way too. In sentence (59) there is an intervening subject between *John* and *him* and neverthe-

less their co-indexing has to be filtered out, because the intervening subject should not count since it is at an inner level. Some treatment of embedding could alleviate these problems, but would not completely eliminate them.

(59) *John/Subj$_{(i)}$ asked to keep a copy of the paper you/Subj submitted with him/Oblique$_{(*i\,j)}$.*

Now consider the sentence (60) under Condition 2. There is a genuine case of commanding of a non-pronominal referential expression (*the kid*) by the pronoun *he*. However since both are embedded, this is not captured by the implementation of the condition and hence not filtered out.

(60) *The fact that he$_{(*i\,j)}$ saw the kid$_{(i)}$ in the mirror doesn't make him the kid he saw.*

Again embedding is a problem for Condition 3. Consider sentence (61). The pronoun *he* has non-nil EMBED value; however it is not attached to John, but to the table, and hence co-indexing is acceptable. The inverse problem arises at (62) where it is not likely that the approach can provide such a treatment as to allow *the driver* to be co-indexed with *his*.

(61) *The table at the left of John$_{(i)}$ which he$_{(i\,j)}$ uses for reading newspaper is cherry wood.*

(62) *The driver$_{(i)}$ in his$_{(i\,j)}$ truck knew much more than the drivers in the house.*

The identification of the expletive "it" is done by searching for certain typical contexts where it happens, such as when it occurs as the subject of verbs in a class that typically includes verbs like "seem" and "appear", or when it occurs as the subject of adjectives with clausal complements (e.g., *The weather made it impossible to play tennis.*).

3.5 *General considerations about the syntactic filter*

The papers discussed here can be split into two kinds: some assuming a previous syntactic analysis, which would give a correct constituency hierarchy, allowing us to obtain correct grammatical functions, argument structure, commanding relations, etc., as is the case of [5,1]; and others assuming that the sentence is automatically analyzed by a preprocessor which in a more or less precise and correct way finds out the relations among the constituents (as in [8,7]). The latter obviously lose comparative points in evalua-

tions (even in a qualitative analysis) because their algorithms for coreference embed the finding of these constituent relations, e.g. in the case of K/B, where the filter suffers from the flat syntactic analysis, compared to e.g. Hobbs, where you assume you have already a syntactic tree. Consider the sentences below:

> The son(i) of his(*i,j) boss fell from the 20th floor.
> The picture [of John(i)] [of his(i,j) office] fell from the 20th floor.
> The picture [of the son(i) [of his(*i,j) boss]] fell from the 20th floor.
> The mystery [of John's(i) picture [of his(i,j) boss]] was in the frame!
> The fear [of John's(i) picture] [of his(i,j) boss] was evident

Attachment of PP is absolutely essential for coreference. And it seems impossible for this to be done without some semantic treatment. Hence Hobbs and Brennan commit a strong sin of assumption on this point. On the other hand, because their approaches are based are based on real systems, Lappin and Leass and Kennedy and Boguraev have this problem considered in the evaluations.

4 The core of the approaches

4.1 *Hobbs' naive approach*

Hobbs' algorithm simultaneously looks for possible antecedents and discards the contra-indexed NPs whenever they are found. Since the part related to contra-indexing was already factored out (presented in the previous section), the core of the algorithm can now be greatly simplified, assuming that in the process of search, before proposing some NP as antecedent, that NP is checked for contra-indexing. Remember that since cataphora is included, the antecedent can also occur after the pronoun in the raw text. The algorithm is presented below.

1. Begin at the NP node immediately dominating the pronoun. Call it X.
2. If node X is not the highest S in the sentence, then:
   (a) 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.
   (b) If X is an NP node not contra-indexed to the pronoun, propose X as the antecedent.

(c) 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 that is encountered which is not contra-indexed to the pronoun.
(d) Traverse all branches of node X to the right of path p in a left-to-right, breadth-first manner. Propose as the antecedent any NP node encountered which is not contra-indexed to the pronoun.
(e) Go to step 2.

3. Otherwise (if node X is the highest S in the sentence) traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered it is proposed as antecedent.

Hobbs' algorithm gives priority to intra-sentential coreference. Actually it only goes beyond sentence boundaries after all intra-sentential NPs have been considered and discarded (say, due to morphological filtering). Hence in (63), where the most plausible antecedent for *he* is *Paul*, the algorithm would determine it to be *John*.

The problem is especially bad because even cataphora is preferred to an antecedent from a previous sentence. The pronoun *he* in (64) would be bound by *John* instead of by *Paul*, the correct choice. Actually cataphora is being given a higher standing that our intuition would suggest. Lappin and Leass [8] and Kennedy and Boguraev [7] strongly penalize cataphora in their salience algorithms w.r.t. other NPs (even from previous sentences). I see the restriction mentioned in the previous subsection to cataphora in Hobbs' filter as a patch to reduce the damage that this higher standing cause.

(63) Everybody respects Paul.
John thinks he is the best player in the team.

(64) Paul is very charismatic.
Whenever he decides to do something, John follows him.

The intra-sentential search consists basically of climbing the parse tree, starting at the pronoun and ending at the root of the sentence; stopping at each NP or S encountered; then making a breadth-first left-to-right search to NPs, first to the subtree to the left of the path that leads to the pronoun (real anaphora), and then to the right subtree (cataphora).

I make the following additional observations about the method. First recall that as mentioned in the previous section it does not extend the notion of contra-indexing beyond the sentence boundaries. But notice how this concept become transparent to the above version of the algorithm once it was separated from the contra-indexing part. Hence, this could be thought of as being easy to fix by maintaining a chain of coreference as long as the pronouns are resolved, and then extending the contra-indexing to whole equivalence classes instead of just the NPs found by the filter. However, consider the fragment (65). The current algorithm would allow for the co-indexing of *him* to *John* (the instance that appears in the first sentence, uttered by A). Now suppose we make a case for contra-indexing to be extended to a class (as it actually should). The algorithm for filtering would mark *him* as contra-indexed to *John* from the second sentence.[17] But then, *he* would be prohibited from being bound by the former instance of *John*. The only exit from this problem would be to assume that not all cases where a referential expression is excluded are to be considered filters in the usual sense. But then one must embed this split in the algorithm, which may be why Hobbs avoided touching this problem.

(65) *A: Why did you give John these long vacations?*
B: *I gave him the vacations because I was aware of all the trouble that he went through when Bill invaded the church and killed John's mother for vengeance.*

The second observation is that Hobbs' algorithm, despite being naive in formulation and seeming to have an arbitrary choice of ordering between NPs, actually captures three very important notions on judging coreference relations. The first is that co-arguments (plus the adjuncts of the same head) of the NP containing the pronoun are preferred as antecedents over other noun phrases embedded in this arguments/adjuncts. This leads to the choice for breadth-first search. The second is that given a set of arguments/adjuncts of the same head to be considered, there is a marked statistical preference for subjects over objects, objects over obliques, and obliques over adjuncts and this order matches in general the order in which they appear from "left to right" in the surface structure of English sentences. The third notion is a statistical preference for antecedents closer to the pronoun, and the algo-

[17] As I pointed out earlier, in order to reduce the damage caused by the high standing of cataphora given by the order of the search, it just arbitrary filters out some unlikely positions to the right of the pronoun, as it is the case here.

rithm takes this into account by starting the search at the pronoun and traversing the levels of embedding from the inside out. I emphasize that it is these perceived statistical preferences that are nicely captured by Hobbs and give his algorithm a reasonably high starting hit ratio. However I will argue in a later section that on the other hand, this strategy makes it very difficult to fix the algorithm for cases where those rigid statements on locality and depth and linearity are not respected.

The algorithm is clearly very sensitive to the particular system used to represent the parse tree, since the relative depth of two noun phrases can vary depending on the grammar system so as to influence the decisions of the algorithm. Consider the sentences (66) and (67) below. It is not clear how Hobbs would represent their parse trees, for example, by laying all the verbs at the same depth or by branching the tree at each verb occurrence. Therefore it is not clear whether the algorithm would choose the same antecedent for *it* (*phone*) in the second sentence as in the first, or whether it would choose *hat*.

(66) *[The man with the hat] picked up [the phone] [when it rang].*

(67) *[The man with the hat] didn't consider picking up [the phone] [until it rang twice].*

Finally, to show that the proposed order fails in many cases we do not need to go too far. I transcribe one of Hobbs' first examples from [5] as (68), where in order to find the antecedent of the pronoun *it*, the algorithm first proposes *536*, then *the castle*, and finally *the residence*, which is the right one. Hobbs would argue for the introduction of selectional constraints (dates and castles can not move[18]) to reject the first two alternatives.

(68) The castle in Camelot remained the residence of the king until 536 when he moved it to London.

4.2 *Brennan, Friedman and Pollard's approach based on centering*

I present below the algorithm for processing an utterance U_n. An important parameter of the centering algorithm is the rules for determining the forward-looking centers, which in [1] consider only the grammatical functions. The order of precedence is:

subject > object > 2nd object > other subcategorized functions > adjuncts.

[18] By believing that castles can indeed move and by knowing nothing about such king I resolved the pronoun in favor of the castle when reading.

1. Construct the proposed anchors[19] for U_n:
   (a) Order the referring expressions (REs) in U_n by grammatical relation.
   (b) Create the set of possible C_f lists. Each list will have one entry for each referring expression, ordered according to (a). Each entry has the referential expression and a pointer to a discourse element it (possibly) refers to. If the RE is a proper name the entry will refer to the same discourse element of a previous mention of this name. If the RE is a pronoun, the entry will be linked to one of the previously introduced discourse elements with which it agrees in its morphological features. There will be one entry in the list for each possible combination of matching pronouns and discourse elements.
   (c) Create a list of possible backward centers containing all elements of C_f (U_{n-1}) followed by NIL for the possibility of finding no C_b for U_n.
   (d) Create a list of the proposed anchors by making the cross product of the lists of the two previous steps.
2. Filter the proposed anchors:
   (a) Eliminate anchors that contain contra-indexed pairs[20] in the proposed C_f.
   (b) For any proposed anchor, find the most highly ranked element of C_f (U_{n-1}) which is also in the C_f of the proposed anchor. If this element is not the proposed C_b for the anchor, eliminate the anchor. This conforms to the definition according to which C_b (U_n) should be the most prominent element in C_f (U_{n-1}) that is realized in U_n.
   (c) Eliminate anchors where none of the entities realized as pronouns in the proposed C_f list equals the proposed C_b (this accounts for conformity to the Rule 1 of the framework).
3. Classify and Rank:
   (a) Classify each anchor in the remaining set of proposed anchors according to the kind of transition relation from U_{n-1} to U_n that it engenders.
   (b) Take the most highly ranked anchor according to the transitions it gives rise to, and make it be the C_b and C_f of U_n. If there is more than one most highly ranked anchor we have reached a case that the algorithm can not solve.

---

[19] An anchor is a pair (C_b, C_f), the backward – and forward-looking centers of an utterance.

[20] I.e., two pairs where the referential expressions are contra-indexed but that nevertheless point to the same possible discourse element representation.

I present below some problems particular to this approach (other more general problems will be covered in a later section). One is that it does not account properly for intra-sentential coreference. Indeed it has a strong preference for inter-sentential binding, just opposite to Hobbs'. It never gets the intra-sentential candidate, when there is the possibility of binding by making a smoother transition according to the centering rules. For instance it would get (69) wrong, binding *his* to *John* instead of *Paul*.

(69) *John was buying furniture yesterday.*
Paul was studying for his exams.
Nobody seemed to be available.

(70)[21] is an example of another problematic case. *His* in the third sentence can not be resolved (to bind *Ira*), because the antecedent is not in the C_f of the second sentence. The approach never works when the antecedent comes from a sentence beyond the previous one.[22] A similar case is (71).[23] Again the pronoun *it* in the third sentence refers to the meeting introduced in the first sentence, which is not considered by the algorithm. This time there is an alternative, and the antecedent will be wrongly taken to be *my office*.

(70) *I want to schedule a meeting with Ira.*
It should be at 3 p.m.
We can get together in his office.
Invite John to come, too.

(71) A: *I want to schedule a meeting with Harry, Willie and Edwina.*
B: *We can use my office.*
A: *It won't take very long.*
B: *So we can have it in the conference room.*

A third problem centers on the robustness of the criteria for ordering the forward-looking centers. Consider the pair of fragments (72) and (73). The first sentence in each contains a cleft. It seems that the cleft raises the importance of the fronted referential expression, and it is not easy to judge whether *John* or *Fred* introduces the most prominent entity. In the first case the pronoun *he* co-specifies with *Fred*, whereas in the second fragment it co-specifies with *John*. Now consider the third fragment (74), where the continuation of the discourse in the second sentence is neutral to the resolution of the pronoun. Can we guess who *he* is?

---

21 From Sidner [10].
22 This is actually more restrictive than recovering a "global focus".
23 Also from Sidner [10].

(72) *John knew it was Fred who had eaten the green hobgoblins. He was very hungry and couldn't resist to the temptation.*

(73) *John knew it was Fred who had eaten the green hobgoblins. But he decided not to tell anybody.*

(74) *John knew it was Fred who had eaten the green hobgoblins. He decided to take a bath ...*

Finally, let us observe this interesting example from Sidner [10]. The discourse fragment (75) apparently breaks one of the fundamental rules of centering, since two forward locking centers of the first sentence are realized in the second sentence, but the one realized by a pronoun is not the most prominent. However notice that something special is going on there. The more prominent candidate, *Alfred and Zohar*, is indeed realized by an special case of non-pronominal anaphoric reference. This could suggest some adaptations in the formulation of centering.

(75) *After playing baseball, Alfred and Zohar had ice cream cones. The boys thought they tasted really good.*

4.3 *Lappin and Leass' salience approach*

Among all the approaches considered for this survey, Lappin and Leass [8] are the first to consider seriously the fact that there are many factors affecting anaphora resolution, and that despite the fact that there is some statistically observed order of preference among these factors in determining the antecedent of a pronoun (that is basically the idea behind the initial success of Hobbs' algorithm), it is clear that for an arbitrary occurrence of a pronoun in a sentence we can not tell in advance which factor should dominate the search for an antecedent, without analyzing the specific circumstance of occurrence of the pronoun. Their approach combines a number of known factors, including syntactic (constituency relations from the parse tree), grammatical (function of the constituents w.r.t. the head), lexical (agreement features), and pragmatic and discourse related (parallelism), ranking the NPs by the points conferred to them according to their conformance to the rules for each factor. Rather than to individual expressions, weights are attributed to coreference chains, which are seen as equivalence classes of the referential expressions linked. Newly introduced referential expressions

(including pronouns before being resolved) are considered singleton classes. Once pronouns are linked to an equivalence class by coreference, the weight is recalculated for the resulting class. The weighting is dynamic and takes into account that old (not recently used) classes of referential expressions must have their weights reduced.

I describe now Lappin and Leass' process of analysis of a new sentence. Before starting a new sentence, what we have is a set of equivalence classes of discourse referents[24] that reflects the coreference chains of pronouns resolved from previous sentences. Each of these classes has a salience value. Before starting the analysis of the new sentence, these values are 'aged', i.e. they are degraded by a factor of 2 (hence classes that are not 'refreshed' by new pronominal references for a long time end up being removed from the list).

The initial steps in the processing of the sentence comprise finding the new NPs, filtering the pleonastics and applying the syntactic filters as presented in the previous section.[25]

Then an initial estimate is made of the weight of the new discourse referents introduced in the sentence. This estimate is the sum of the weights of all factors applying to the NP. These factors are presented in the table below, with their weight contributions and a brief explanation of their applications. Lappin and Leass refer to them as non-local because their contributions to the NPs final weight are independent of the pronoun being resolved. We will see later that when a given pronoun is to be resolved, the final weight values local to the resolution of this particular pronoun are evaluated considering then the so called local factors.

---

[24] Notice that two discourse referents would be in the same class if the NPs which introduced than are in the same coreferentiality chain. They does not necessarily refer to the same entity. Recall my usage of coreference is the usage Sidner gives to co-specification in [10].

[25] By now, we would have contra-indexed pairs still restricted to a local intra-sentential domain only, since no pronoun of the current sentence has yet been bound to extend the contra-indexation to its equivalence class. I assume on what follows that as far as pronouns are being bound the restrictions are extended accordingly.

Factor	Weight	Description
Sentence recency	100	Conferred to all NPs in the current sentence.[26]
Subject emphasis	80	Added to NPs that appear in subject position.
Existential emphasis	70	Added to predicate nominal NPs in existential constructions, e.g. *papers* in *There are four papers to read.*
Accusative emphasis	50	Conferred to NPs in direct object position.
Indirect object and oblique complement emphasis	40	Added to NPs head of an indirect object and to NPs head of a PP object.
Head noun emphasis	80	Conferred to an NP which is not embedded in another NP.
Non-adverbial emphasis	50	Added to an NP if it is not contained (embedded) in an adverbial PP demarcated by a separator (e.g. adverbial PPs dislocated to the beginning of the sentence, separated by a comma from the rest).

Finally we come to the resolution of the pronouns in the sentence. This is accomplished by taking them in order of appearance (from left to right in the surface string). For a reflexive the binding is done to the most salient referent in its binding domain. Hence in both (76) and (77) below, it would take *John* to be the antecedent of *himself*. While in (76) there seems to exist a preference for this interpretation by readers, that is certainly not the case for (77), where *David* would be prefered as the antecedent.

(76) John gave David a picture of himself.

(77) John showed David to himself (in the mirror).

To resolve the non-reflexive pronouns, which is the most interesting case, we take the updated list of equivalence classes of discourse referents with their current estimated non-local weights[27] and calculate their final weight w.r.t. the particular pronoun being solved. The local factors Lappin and Leass describe for that purpose are listed below.

[26] Notice that actually this is conferred to every NP when they appear. Hence this can be seen as a initial value for the discourse referent. The use of this is clearly as an adjusting parameter for the algorithm.

[27] Notice that the new discourse referents introduced in the sentence are considered themselves singleton classes. Each time a pronoun is resolved, its class is coalesced to the class it was bound by, and the salience value of the new class is revaluated.

- Parallelism: a small reward of 35 points is applied to the NPs which fill the same grammatical role as the pronoun being resolved.
- Cataphora: a strong penalization of -175 points is applied to NPs that appear to the right of the pronoun being resolved.

In case of a tie between two or more candidate antecedents, the closer is chosen (where close is measured in the surface string, with no preference for left or right direction).

I will show in chapter 6 that there are many important factors which are not considered either by this or the other approaches described here. Hence even if we make an optimal choice of weights at the training phase it is expected that many examples will be solved incorrectly. Indeed, the statistical measures of success with an optimal setting of weights will give us an upper bound of the possibilities allowed by the set of factors considered.

Subordination relations are not taken into account in general (exceptions are the emphasis to relative clauses and to adverbial clauses demarcated by a separator). Hence, in (78)[28] *it* would be bound to *the controller* when the correct antecedent would be *the indicator*.

(78) *This green indicator is lit when the controller is on.*
It shows that the DC power supply voltages are at the correct levels.

The salience approach seems to be hindered by not considering other kinds of coreference, both non-pronominal and with demonstratives. In the example below, Lappin and Leass report the error of the method in assigning *them* to *the users* and suggest that other mechanisms could be used to reject this. However, notice that if the anaphoric reference of *these objects* to *a user profile and a system distribution directory entry* was realized, then these two referential expressions would form an equivalence class which would contain the correct antecedent, *these objects*, and would be the most salient, assuring the success in the resolution of *them*. Because this approach makes strong use of frequency of reference, it is particularly sensitive to the correct account of other forms of anaphoric relations.

(79) *The users you enroll may not necessarily be new to the system and may already have a user profile and a system distribution directory entry.*
&vfc. checks for the existence of these objects and only creates them as necessary.

[28] Taken from [8].

Notice also in (79) that there is much induction going on due to the parallelism between the two coordinated verbal phrases in the second sentence. Lappin and Leass clearly give a very weak account of parallelism. I return to this point in section 6.

I finally point out that there are two aspects to be considered in the approach. The general framework of salience weights, which I consider very promising, and the particular complete proposal with the choice of factors and weights presented in [8]. With respect to this latter aspect the approach is not so outstanding, since the factors are basically the same syntactic factors used by the other systems. What could be the differential factors, as parallelism are not properly accounted for.

4.4 *Kennedy and Boguraev's approach*

Basically Kennedy and Boguraev take the salience approach of the previous subsection, make some conceptual improvements, and apply it by replacing relations normally obtained from the complete constituent structure with approximate relations empirically obtained from their flat morpho-syntactic structures.

Part of their good results despite the poor syntactic analysis (and hence higher susceptibility to errors) is due to improvements in the salience formulation by including new factors and changing some weights w.r.t. Lappin and Leass. The new factors are the *context emphasis* and *possessive emphasis*. The sensitivity to context is evaluated through the use of a text-segmentation algorithm that determines topically coherent segments of text. A discourse referent is assigned the points if it has been introduced in the current context. Possessive emphasis checks whether the discourse referent is in determiner position of an NP (e.g. as *John* in *John's house*). The *indirect object and oblique complement emphasis* is split into two differently weighed factors, the second being reduced to 30. I present below the complete set of factors with their weight.

Factor	Weight
Sentence recency(SENT-S)	100
Context emphasis(CNTX-S)	50
Subject emphasis(SUBJ-S)	80
Existential emphasis(EXST-S)	70
Possessive emphasis(POSS-S)	65
Accusative emphasis(ACC-S)	50
Indirect object emphasis(DAT-S)	40
Oblique complement emphasis(OBLQ-S)	30
Head noun emphasis(HEAD-S)	80
Non-adverbial emphasis(ARG-S)	50

To determine SUBJ-S, POSS-S, ACC-S, DAT-S, the algorithm just checks whether the GFUN value of the discourse referent is subject, possessive, direct object or indirect object. Similarly for HEAD-S and ARG-S the attributes EMBED and ADJUNCT respectively are verified. If the NP was the complement of a preposition then OBLQ-S applies.

Among the local factors, the formulation of the parallelism reward is changed. Instead of rewarding discourse referents with the same grammatical function as the pronoun, as Lappin and Leass do, Kennedy and Boguraev propose to reward a discourse referent when the pair consisting of its grammatical function and that of the pronoun is the same as the pair from a previously identified anaphor-antecedent pair. Here it is not clear which previously identified pairs are considered. It is intuitively reasonable that these pairs should be "aged" and removed from consideration after a while (e.g. when segment changes are detected), otherwise in long discourses, after a while almost all new pairs would be rewarded.

A new local factor is added that rewards locality. A candidate antecedent is rewarded when it is in the same subordinate context as the pronoun. In my opinion this is a very important difference from Lappin and Leass. It negates the effects of the head-noun and non-adverbial emphasis that relatively penalizes an embedded noun phrase (since it is not emphasized) for cases when it is in the same embedded context of the pronoun. The assumption is that the prominence of a candidate should be determined with respect to the anaphor. The reward just puts the candidate at the level it would have if it were not in the embedded context.

The surprising result in the analysis of this version is that basically all the information needed may be obtained from the input, which comes with grammatical functions. And the results reported by the authors of [6] of this tagging process are on the order of 97% for recall and 95% for precision. Indeed the precision in noun phrase detection might be a concern. However, they also deal with this in [6] with good results. Hence, the relatively good results we will see next, in section 5 are not so surprising, and much of the differences may be due to the filter which is not so good.

5 Some quantitative results

In this section I show some published quantitative results comparing the performances of the algorithms described.

In [5] Hobbs reports the results of the analysis of 100 occurrences of third person pronouns (personals and possessives) from three different kinds of text.[29] The overall results are 88.3% correctly resolved pronouns. However, from the 300 occurrences, in only 132 is there a real conflict in determining the antecedent, and on these the hit ratio is 72%. By using selectional constraints the overall results rise to 91.7%, and to 81.8% over the subset of 132 conflicts. The pleonastic occurrences of 'it' were not counted.

Walker [11] compares Hobbs with the centering-based algorithm, and her results for Wheels, Newsweek and a collection of task dialogues are as follows. Notice how dialogues are harder to cope with.

Source	Brennan et al.	Hobbs
Wheels	88	90
Newsweek	89	79
Dialogues	51	49

Lappin and Leass report two sets of results: the first was obtained from a training corpus used to tune the salience weights, and the other from a blind test. Surprisingly the results are similar. All the material came from computer manuals (560 occurrences of third person pronouns, including reflexives, for the training, and 360 occurrences for the blind test). However, the conditions of the blind test make it slightly opaque to the interpretation of the re-

[29] William Watson's Early Civilization in China, pp. 21-69; the first chapter of Arthur Hailey's novel Wheels; and the July 7, 1975 edition of Newsweek, pp. 13-19, beginning with the article "A Ford in High Gear".

sults.[30] One of this conditions is that the random selected sentences are each preceded by only one previous sentence. Another is that they filter the set of sentences so as to keep only sentences with at least two elements in the candidate list, where the actual antecedent appears in the list. Also they do not include pleonastic pronouns and reflexives.

As for the characteristic of the input, despite the fact that they use the slot grammar system to parse the sentences, they edit them slightly, by making lexical substitutions, to overcome parser inaccuracies.

The table below includes in the last column the results of the application of Hobbs' algorithm to the blind test corpus. The algorithm was implemented in their underlying framework, and the filter was factored out and replaced by their own filter. Hence the differences should reflect the core strategies.

	TRAINING CORPUS		BLIND TEST		
	Occurrences	Hits	Occurrences	Hits	Hobbs' hits
Intra-sentential	471	86%	290	89%	81%
Inter-sentential	89	81%	70	74%	87%
Total	560	85%	360	86%	82%

Kennedy and Boguraev [7] achieve an overall accuracy of 75.5% in the analysis of 27 texts containing 306 occurrences of third person pronouns (including lexical) from a random selection of genders, including press releases, product announcements, news stories, magazine articles, WWW pages, etc. They claim that only a small number of errors can be attributed to the absence of configurational information, and that most of the differences compared to Lappin and Leass' results are due to the different kinds of texts (computer manuals are expected to be much more well behaved than the variety of sources Kennedy and Boguraev used, e.g., containing quoted passages in-line), and to gender mismatch at the input (35% of errors).

[30] The reasons for this masking seems to be partially the compatibility with other approaches they wanted to compare, e.g. Hobbs', which, for instance, does not deal with pleonastics and reflexives.

6 Conformance to the notion of algorithm and analysis of the quantitative results

I start this section by recalling one version of the standard definition of algorithm. I then argue that there are different degrees of conformance to this definition which lay in a continuous and imprecise line. I would like to call 'algorithmicity' this characteristic of adherence to the concept of the algorithm. It measures how confident we should be that the solution proposed is really a computational solution. The proposals given in this paper should have their quantitative results and their claims for partial success evaluated taking into account their adherence to it.

An algorithm is a finite, unambiguous definition of a finite set of operations. Each operation should be mechanically executable, and given any input, the process of deciding the order in which to perform the operations must be effective. Moreover the number of operations executed must be finite for any given input. Based on this definition we can see that Hobbs' presentation of his method for resolving pronouns is indeed algorithmic. Given a parse tree (or a sequence of parse trees for the discourse), it is easy to see how to compute pronoun resolution through the steps he gives.

The first definition of the three conditions for disjoint reference given by Kennedy and Boguraev is not algorithmic, e.g., given the kind of input they assume, to compute co-argumentship is not an obvious task at all. Indeed it is probably impossible to formulate a general algorithm that gives the correct answer. Then they present a set of approximation rules for computing the conditions taking into account the kind of input they have. Here we have a point to discuss. One can argue that the formulation they gave is not an algorithmic definition in the traditional sense. However, it is easy to see that given any input we can effectively compute whether the conditions are true or not for the input. And this is indeed the way we frequently work: a specification in a higher level of abstraction is regarded as algorithmic if anyone can see that an algorithm in the traditional sense can be derived from it. Hence I consider Kennedy and Boguraev's set of implementation rules as an algorithm.

The connection of algorithmicity and the analysis of quantitative results is basically that the results can be given credit as long as we can see they were obtained through the use of an algorithm. And what is frequently the case is that authors do not make this process easy, due to their not-so-well-grounded underlying assumptions or due to underspecified conditions, as observed by

Walker in [11]. While it is perfectly natural for assumptions to be made, say, about the existence of other components in the system to pre- or co-process the input, when not grounded in reality or not fully explained, they can blur ones judgment of the results, especially when the analysis is comparative with other methods that make different assumptions.

So let us focus on Hobbs' assumption that we have as input the parse trees with the correct attachments of prepositional phrases. Consider sentences (80) and (81). He argues that in (80), *of his truck* is an argument of the head *driver*, while *in his truck* is an adjunct of it, with the (partially) bracketed parse trees given by (82) and (83). His algorithm assumes their input trees come with this precision in the level of attachment in order to be able to say that *he* is contra-indexed with *driver* in (80) but not in (81).

(80) *Mr. Smith saw a driver of his truck.*

(81) *Mr. Smith saw a driver in his truck.*

(82) $[_{NP}\ [_{Det}\ a]\ [_{\bar{N}}\ driver\ [_{PP}\ of\ he's\ truck]]]$

(83) $[_{NP}\ [_{Det}\ a]\ [_{\bar{N}}\ driver]\ [_{PP}\ in\ he's\ truck]]$

I now argue that this is too strong an assumption and make a case that pronouns sometimes need to be resolved before deciding the construction of the parse tree, and may even help this construction. Consider the the sentence (84). (85) and (86) are the bracketings corresponding to the two possible prepositional attachments, enriched by the coreferentiality possibilities that they engender for the pronoun *his*. Now consider the three fragments where the sentence appears in context. I make the following claims about the interpretation of the pronoun *his* in the second sentence in each case. For (87) *his* co-specifies with *the man* and the correct attachment is the one given by (86). For (88) the attachment is the same but *his* refers to *John*. Finally for (89) *his* refers to *John* and the attachment is given by (85). I claim that the attachment can not be decided until the pronoun is resolved. On the contrary it seems to be the case that the context induces the choice for the antecedent of the pronoun and then the parse tree.[31] Hence we are forced to conclude that the apparent algorithmicity of the solution of the problem is a little reduced by this too strong assumption.

[31] One could argue the possibility that the context induces the parse tree and then the pronoun is resolved. However, a careful analysis of the examples indicates that this can not be the case. The parse tree can not be decided without regard for the discourse element expressed by the pronoun.

(84) *The son of the man carrying the umbrella about whom his girlfriend was talking is Peter.*

(85) *[The son$_{(i)}$ [of the man$_{(j)}$ carrying the umbrella [about whom his$_{(*i,*j,k)}$ girlfriend was talking]]] is Peter.*

(86) *[The son$_{(i)}$ [of the man$_{(j)}$ carrying the umbrella] [about whom his$_{(*i,j,k)}$ girlfriend was talking]] is Peter.*

(87) *The girlfriend of the man carrying the umbrella was talking about one of his sons.*
The son of the man carrying the umbrella about whom HIS girlfriend was talking is Peter.
(HIS = the man)

(88) *John was listening his girlfriend talking about one of the sons of the man carrying the umbrella.*
The son of the man carrying the umbrella about whom HIS girlfriend was talking is Peter.
(HIS = John)

(89) *John was listening his girlfriend talking about the man carrying the umbrella.*
The son of the man carrying the umbrella about whom HIS girlfriend was talking is Peter.
(HIS = John)

Brennan, Friedman and Pollard on the other hand are assuming that all noun phrases are available with their grammatical functions. If at first it seems that they do not need as strong assumptions on the input as Hobbs does, a closer look will show otherwise. They assume the existence of a module that provides contra-indexing information, and this module is the one which most need the precise structural relations among constituent noun phrases (e.g. commandment, embedding, etc.). Actually by assuming the filter in their hand-simulated test they start with an advantage compared to Hobbs'. How significant this handicap is hard to determine. I suspect it to be indeed small.

Since Kennedy and Boguraev assume neither parse trees as input nor the existance of hidden filters, their results should be seen as disadvantageous, compared to the other approaches (Recall from the previous section that Lappin and Leass edit their test sentences to avoid incorrect parses).

Another point where Kennedy and Boguraev are in disadvantage w.r.t. the others concerns the morphological filter, since their test does not assume that feature values are correctly as-

signed at the inputs. Instead they get it from the execution of the tagger, and report that a great amount of errors are due to gender misevaluation. I made the case in an earlier section that gender is not easy to determine correctly in all cases.

Lack of information from the paper is still a point that comes to join the previously mentioned factors in blurring the evaluation of the results. Hobbs assumes the existence of a pre-processing phase in which "syntactically recoverable material" is recovered. Due to the vagueness of the paper we are allowed to think that the following is true: during the hand-simulation, everything that he considers recoverable and that helps pronoun resolution is recovered. This "cognitive" process falls short of being algorithmic.[32] It seems that co-indexation with empty elements (e.g. traces) is also assumed for both Hobbs and Brennan, Friedman and Pollard (Lappin and Leass seem to have real modules that attempt to do the same).

Brennan, Friedman and Pollard mention a hierarchy among arguments and adjuncts of a verb to create the forward looking center list. Nothing is said about referential expressions in other positions, e.g., possessors, adjuncts to an NP, or how arguments and adverbial adjuncts of embedded clauses are considered in the hierarchy. Also there is no clue as to how a second pronoun in the sentence is solved when this does not influence the transition alternatives. Consider (90) and (91). The capitalized occurrences of *they* can be bound either to *the green hobgoblins* as in (90) or to *Canny and Ball* as in (91). We can not tell what their algorithm would choose.[33]

(90) *Canny and Ball were starving when they saw the green hobgoblins. They knew THEY were nice to eat.*

(91) *Canny and Ball were starving when they saw the green hobgoblins. They knew THEY should eat some more conventional flesh but they couldn't resist to the temptation.*

Maybe the key point which separates the first two approaches presented from the last two, is that they are hand-simulated, which means that one only needs to decide things on the basis of the occurrence in the tested corpus. One thing somewhat related that is worth noting in Lappin and Leass is the preoccupation with separating the training corpus (used to tune the salience factors) from the blind test corpus.

---

[32] In the sense that human beings know how to execute the process on real inputs, but not necessarily how to give a general computational procedure that do the same.

[33] Indeed there is a third possibility of collecting all entities together.

7 Coverage

In Section 4 I covered problems specific to the proposals presented there. In this section I focus on general problems that apply to all four proposals. I try to present the examples in such a way that, without going into details of how each approach would solve them, it becomes clear that each fails to account for the problem.

Let us start with the account of plurals. Consider (92) and (93). In the first, *they* refer to the four cartoon characters, whereas in the second, it refers only to Fred and Wilma. It is impossible to get that without fully interpreting the subordinate clauses in the second sentence. Hence, either the approach gets (92) right or (93), but not both. To account for plural pronouns is much harder than singular ones, without semantics. Antecedents can be virtually any subset of entities introduced in the discourse, and understanding the situation is generally required to get the subsets that can be used as candidate antecedents.

(92) *Fred and Wilma had just finished an argument when the Rubbles arrived.*
After Barney gave a phone call they had dinner while talking to each other.

(93) *Fred and Wilma had just finished an argument when the Rubbles arrived.*
After Barney and Betty left they went to sleep still angry with each other.

In (94),[34] in order for the approaches to get *it* as being the face, it is necessary for all the approaches that they reject *the mud pack* as the antecedent. In the literature this is often described as a need for selection restrictions. The verb *feel* would not be allowed to take as argument a referential expression that denotes something unable to feel, like a (mud) pack. I point out some major issues here that show the topic is not easy as people may make it appear to put it as an assumption of their proposals. First is the simple characterization of the entities selected (e.g. how to get that a face, which is not an animate entity, can be selected by *feel* but not a pack). The second is that selection is not a binary feature, but a question of plausibility. Whether castles can be moved or not is used to resolve coreference in a sentence presented earlier from Hobbs as (68) that I repeat here as (95). Well, castles can indeed move, but it is much more plausible that the abstract entity "the residence of the king" is

[34] From Sidner [10].

moved instead. Third, selection is not static. As a discourse goes on, selection constraints change, and tables start being able to fly, as well as, why not, years to move. Still knowledge of the world or context may be involved as in (95) again, where if one knows something about Camelot and the king who once lived there, he/she is less likely to get into wrong interpretations (as I did myself).

(94) *Take the mud pack off your face. Notice how soft it feels.*

(95) *The castle in Camelot remained the residence of the king until 536 when he moved it to London.*

Some deeper account is needed to resolve the pair (96) and (97) due to Terry Winograd.[35] Indeed both city councils and women can either fear violence or advocate revolution. A much more complicated inference process than selection is required to reject *the city council* as the antecedent of *they* in (97), say, on the basis of "unlikelihood" in the context. Similarly consider the pair (98) and (99),[36] where in the first *it* means the plain, and in the second *eroson* is the antecedent. Naive selectional restriction alone is not enough to rule out *contour farming* as an antecedent considering the acceptability of (100).

(96) *The city council refused to give the women a permit because they feared violence.*

(97) *The city council refused to give the women a permit because they advocated revolution.*

(98) *The plain was reduced by erosion to its present level.*

(99) *Contour farming has reduced erosion to its present level.*

(100) *The present level contour farming was reduced to is a shame, considering that farmers should know it is the solution for erosion.*

Indeed, the knowledge required in the inference process is not only a general knowledge of language (like the meaning of verbs, etc.). Contextual knowledge is important. In (101) the friends of Alin and Lucian are able to interpret *he* as Alin, because they know he is neurotic-ally worried with his preliminary Ph.D. examinations. On the other hand *he* means *Lucian* in the second sentence to people who knows about his characteristic way of initiating telephone conversations expressed in the fragment.

---

35 Taken from Sidner [10].
36 From Hobbs [5].

(101) *Lucian called Alin yesterday.*
As usual he was studying for his WPE-I.

(102) *Lucian called Alin yesterday.*
As usual he introduced the conversation by saying "What are you doing ?".

Up to now, it may be argued that all the examples in this section conform to a general assumption that antecedents are proposed and considered in some order, and that a given antecedent is only considered after all those who precede it in the hierarchy are discarded. The next two fragments are intended to show this is not the case. In (103) *the chocolate* is interpreted as being the brown bag, while in (103) *the chocolate* is the antecedent. Under the usual ordering that all the proposals here would assume, i.e., that the chocolate precedes the bag (and the bar in the second example) in consideration,[37] what is it that causes *the chocolate* to be rejected as antecedent in the first case, but not in the second case ? My answer is that both candidates are considered and weighted and the winner is selected to be the antecedent. This argument against the linear ordering affects Hobbs' proposal and Brennan, Friedman and Pollard's.[38]

(103) A: *What were you eating?*
B: *I was eating a chocolate I got from the brown bag standing here a minute ago. Did you see it?*

(104) A: *What were you eating?*
B: *I was eating a chocolate I got from the bar a minute ago. Did you see it?*

One issue certainly all the authors of the proposals agree on, is that something else is needed to account for a *global focus* or the popping of the focus in what is called the *task environments*. This refers to entities that the discourse is about globally, or during a segment, etc. The dialogue (105)[39] exemplifies the concept. None of the algorithms would get *it* to be *the bolts*.

---

[37] I hope the reader is convinced that the problem is not about a wrong choice of the order.

[38] The original proposal of centering [2] would give a escape to the latter for they assume the existence of a partial, not necessarily linear order. Whether this would be enough to support a proposal on a similar basis as Brennan, Friedman and Pollard's, with this aspect reconsidered, it is hard to tell.

[39] From Sidner [10].

(105) A: *One of the bolts is stuck and I'm trying to use both the pliers and the wrench to get it unstuck.*
B: *Don't use the pliers. Show me what you are doing. Show me the 1/2" combination wrench.*
A: *OK.*
B: *Show me the 1/2" box wrench.*
A: *I already got it loosened.*

Parallelism is also a difficult factors to account for, as it appears in many levels of language. Roughly it consists of the presentation of two or more utterances that have similarities to a point of inducing the hearer to fill in some elided elements so as to maintain or increase that similarity. In (106)[40] parallelism explains why readers get *it* to be the wild rose instead of the green Whitierleaf, even if the second would be considered more highly ranked than the first. None of the approaches accounts properly for that. The condition of parallelism used by Kennedy and Boguraev is not even triggered here, since there is no other anaphor-antecedent pair to compare. As for Lappin and Leass, the parallelism emphasis would be used, but then, without counting weights, consider the counterpart fragment (107). Here *it* refers clearly to the gun, not to the house. Either Lappin and Leass would get the first or the second pronoun wrong. The difficult question concerns when parallelism is triggered and when it is not. In this example the word *too* is crucial to make the parallel interpretation to be used in the first but not the second fragment. Parallelism goes far beyond just matching grammatical rules. Linking words like *also* and *too*, coordination of clauses, the semantic relation between the verbs used (e.g. repetition of the verb or use of verbs with opposite meanings), etc. must all be considered.

(106) *The green Whitierleaf is most commonly found near the wild rose. The wild violet is found near it too.*

(107) *A gun was found on the house. The bullets were found near it.*

In the introduction I mentioned that the algorithms could correctly find the antecedents in sentences exhibiting phenomena like specific/non-specific reference and prototype versus copy (see (7) and (8)), without actually taking into account the semantic distinctions. Now consider Eugene Charniak's (108).[41] In the last sentence,

[40] From Sidner [10].
[41] Taken from Hobbs [5].

it refers to the kite; not to the one mentioned in the third sentence, but actually to the generic one introduced in the second sentence. None of the approaches have conditions to recognize that.

(108) *Jack invited Janet to his birthday party.*
Janet wondered if Jack would like a kite.
But Bill said Jack already had a kite.
Jack would make her take it back.

An interesting aspect of pronoun resolution concerns semantic neutrality. Sidner discusses (109) arguing that despite the ambiguity as to whether *it* refers to the whole situation or to the bear, this choice is semantically neutral. Consider further the fragment (110). It is not clear whether *it* means the (sound of the) stereo, the party, or the whole situation. On the other hand, it seems that hearers would not be confused with the discourse, and would not consciously perceive or report some ambiguity in it. However it does not seem to be the case that semantic neutrality exists among the choices. Rather, it seems that the binding of the pronoun remains "loose" in the mind of the hearers.

(109) *One of the black bears got loose in the park the other night.*
It frightened all the campers and generally caused panic.

(110) *Mike turned on his stereo late in the night yesterday and started a party.*
We couldn't stand with it and called the police.

Finally there is no account for the odd cases where the pronoun is used referentially, but with no explicit antecedent in the discourse as in (6).

8 Modifiability

I will not expand much on this topic, but I want to point out that in the analysis of a proposal for solving a problem, it is very important to know about the chances it has of surviving in the long term. I actually believe that more important than its current success concerning the comparative results is the perspective for improvement in the long term. Rather than algorithms that give 80% results but seem to have an upper bound of 85%, I prefer algorithms which give 70% but such that one can see as possible to improve them to achieve the 100%.

Hobbs' algorithm is one that is considered to be stuck. And the main reason is that one can only consider an antecedent in the discourse after all others that precede it in the search are discarded. Hence, one can not talk about different degrees of plausibility, or weighting candidates according to different criteria. For example, it is true that we can use selectional constraints on verbs to help find the antecedent, but for Hobbs, either we consider that castles can move or that castles can not move. Either babies can be boiled or they can not be boiled.[42] The reality is that babies can be boiled and castles can move, given the context, and provided there is nothing else more likely to be subjected to these actions. Also, how could parallelism be accounted for? How could local or global focus be accounted for?

The approach by Brennan, Friedman and Pollard is different in two aspects. First, we have to consider that the main concern of the proposal is to use centering. And centering does not constrain the rules for ordering the referential expressions in the C_f. Hence more elaborate methods can be achieved for that task as suggested in [2].[43] Also it can be improved by ordering the candidate antecedents for intra-sentential coreference, in cases where the decision is to be made among candidates in the same sentence.[44] Nothing can be done, however, against the strong preference for candidates from the preceding sentence. The second aspect is that if we consider centering by itself, not limited to the particular use they make of it in [1], it is clear that centering is an important factor for deciding coreference. To this respect recall that Grosz, Weinstein and Joshi propose the framework to deal with the local attentional component in discourse. On one side they are aware of the importance of the global component. On the other they are aware that these are not the only factors to be considered in coreference resolution.

Among the four, I consider the most promising framework to be the one by Lappin and Leass, and this is because it is the only one that can account for the many factors that influence pronoun resolution without an a priori constraining order. For instance we could think of inserting in it a factor that takes into account centering, as well as global focus, with proper weights.

---

42 This is a reference to the famous sentence *If the baby does not thrive on raw milk, boil it*[4].

43 Recall also the discussion about cleft in Section 4, and consider the use of partial order among candidates with some additional mechanism for decision.

44 This is indeed suggested by Walker in [11].

Regarding Kennedy and Boguraev's approach, I consider this to be an engineering version of Lappin and Leass' to answer current needs for robust, domain independent subsystems for coreference resolution with some acceptable rate of error. I do not believe that a flat morpho-syntactic analysis can be used alone to derive the kind of constituency relations needed in discourse analysis with the required precision.

9 Conclusion

I presented four algorithms for pronoun resolution, analyzing them individually, and some times comparatively. This may give the impression that these are very distinctive approaches. In the section on coverage however, it begins to appear that this is not the case. Indeed all of the approaches are very similar in the sense that they mostly consider syntactic aspects of coreference, mainly the relative position w.r.t. the pronoun and the grammatical function. Indeed, although they have each some peculiarities in their heuristics used to account for the different factors, e.g., by favoring intra-sentential to inter-sentential coreference, or by considering embedding or not, etc., the results do not substantially change. It seems that they just play with the accidental statistical distribution of the cases without really coping with the missing factors that would make a difference.

To better make the point, let us consider one detail reported by Lappin and Leass. Among the 360 pronouns of their blind test, when they deactivate the parallelism reward, they have two instances of pronouns whose resolutions correctly disagree with the standard test, and four instances that incorrectly disagree with it. Had the relation been something like 1 correct disagreement to 100 incorrect disagreements, I would say that the factor was correctly accounted for, but 2 to 4 clearly indicates that they are just touching an accidental statistical consequence of parallelism. The real factors involved in parallelism are not (properly) accounted for.

Although I believe in the syntactic factors as they are considered in the approaches, other non-syntactic factors have to be included. Once they are accounted for, we may end up concluding that the differences among the general heuristics adopted by the proposals are irrelevant. These factors, as I showed in section 6, should involve semantic preferences, plausibility of the choices, and detection of contexts that induce parallel interpreta-

tive behavior. Binary static accounts (e.g. residences move, castles never move) would probably be a big step ahead in improving the statistical results (as actually demonstrated by Hobbs).[45] I suspect that a more appropriate and detailed account of parallelism, which could be done to a great extent by only considering lexical semantics and structural relations, would also be a big step. But in order to account for the phenomenon of pronominal coreference as a whole, there is clear need for dealing with general world knowledge, discourse acquired knowledge, capacity of local logical inference, among other higher order cognitive aspects of language. Indeed, examples like (92) and (93) seem to require processing capability comparable to full semantic interpretation of the discourse.

As a general framework, I consider Lappin and Leass' to be the best (among those studied here),[46] because it is the only one which is ready by conception to support some of these improvements,[47] offering a mechanism for controlled bias according to different factors, as opposed to the other two approaches that structurally embed fixed choices (e.g. intra-versus inter-sentential dominance). Even plausibility (as opposed to binary judgments) could be weighted. I also recall that the approach does not imply an a priori order among candidates.

As complete proposals we can summarize the approaches as follows. Hobbs' naive approach is stuck by conception. Its purpose was mainly to show how a naive fixed strategy of ordering of antecedents considering only relative structural position could indeed bring good statistical results. Centering framework gives some constraints for pronoun resolution and some rules of preference among transitions. The rules of preference are an important factor to be accounted for in a weight basis. Brennan, Friedman and Pollard, however use them as a decision procedure, which actually constitutes the major limitation of their approach. Lappin and Leass have a promising framework (which for instance could account for centering transition preferences). But the kind of factors accounted for in the paper [8] are indeed similar to the others. Kennedy and Boguraev's approach, as I mentioned before, is mostly a robust, all-purposes engineering version of Lappin and Leass' with surprisingly good results.

---

45 However this indeed seems to be playing with statistics.
46 Included here Kennedy and Boguraev's engineering style of implementation.
47 Although it is not clear to which extent and how properly such extensions could be accounted for in the framework.

I posited that claims for computational approaches must be clear regarding their adherence to algorithmic feasibility. In particular, assumptions are a critical issue. Although they are obviously important for isolating the particular phenomenon to be covered, one has to be very clear how grounded such assumptions are, keeping in mind that between cognitive abilities of human beings and algorithms there is a big distance. In this sense special care must be taken for hand-validated proposals. Using the approaches analyzed here, I discussed the interpretation of published results, suggesting that the information conveyed by numbers presented in papers is limited if computational groundedness and implications of assumptions and under-specifications are not fully understood.

The proposals for filtering modules were extensively covered and two conclusions can be drawn. One is that despite being a topic studied for long time, it seems that there is still no adequate complete account for filtering that makes possible the construction of computational systems that makes allways correct decisions. Indeed, different formulations on which the different computational approaches are based make them fail in different cases. The other observation is that the judgment on locality for binding is not uniform in the sense that if for one side there are some relative positions of the antecedent where binding is strongly accepted or rejected, there are other positions where the judgments from natives are not so strong. This could either be attributed to competence/performance reasons, or it could suggest that locality constraints are not hard constraints, but maybe something to be account by a negative penalization factor with strong weight.

Acknowledgements

I would like to thank Bonnie Webber for the valuable discussions and sugestions throughout all the development of this paper; Robin Clark, for the discussions on filtering formulations and judgement of disjoint reference; Ellen Dickey, who dedicated to the painful task of revising this text of a non-native English writer; and Connie Parkes and Jason Baldridge for judging the acceptability of the many sentences I produced while trying to disclose the filters.

References

[1] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In Proceedings of 25th Annual Meeting of the Association for Computational Linguistics, pages 155-162, Stanford, CA, 1987.

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

[3] Liliane Haegeman. Introduction to Government and Binding Theory. Blackwell, Cambridge, USA, 1994.

[4] Graeme Hirst. Anaphora in natural language understanding: A survey. In Lecture Notes in Computer Science, 119. Springer-Verlag, 1981.

[5] Jerry R. Hobbs. Resolving pronoun references. Lingua, 44(2/3):311-338, 1978.

[6] Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Antilla. Constraint grammar: A language-independent system for parsing free text. Mouton de Gruyter, Berlin and New York, 1995.

[7] Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In Proceedings of 16th International Conference on Computational Linguistics (COLING '96),
Copenhagen, Denmark, 1996.

[8] Shalom Lappin and Herber J. Leass. An algorithm for pronominal anaphora. Computational Linguistics, 20(4):535-561, 1994.

[9] Randolph Quirk and Sidney Greenbaum. A Concise Grammar of Contemporary English. Harcourt Brace Jovanovich, San Diego, 1973.

[10] Candace L. Sidner. Focusing the comprehension of definite anaphora. In Michael Brady and Robert C. Berwick, editors, Computational Models of Discourse, pages 267-330. MIT Press, Cambridge, MA, 1983.

[11] Marilyn Walker. Evaluating discourse processing algorithms. In Proceedings of 27th Annual Meeting of the Association for Computational Linguistics, pages 251-261, Vancouver, Canada, 1989.