A Corpus-Based Evaluation of Centering and Pronoun Resolution

Joel R. Tetreault*
University of Rochester

In this paper we compare pronoun resolution algorithms and introduce a centering algorithm (Left-Right Centering) that adheres to the constraints and rules of centering theory and is an alternative to Brennan, Friedman, and Pollard's (1987) algorithm. We then use the Left-Right Centering algorithm to see if two psycholinguistic claims on Cf-list ranking will actually improve pronoun resolution accuracy. Our results from this investigation lead to the development of a new syntax-based ranking of the Cf-list and corpus-based evidence that contradicts the psycholinguistic claims.

1. Introduction

The aims of this paper are to compare implementations of pronoun resolution algorithms automatically on a common corpus and to see if results from psycholinguistic experiments can be used to improve pronoun resolution. Many hand-tested corpus evaluations have been done in the past (e.g., Walker 1989; Strube 1998; Mitkov 1998; Strube and Hahn 1999), but these have the drawback of being carried out on small corpora. While manual evaluations have the advantage of allowing the researcher to examine the data closely, they are problematic because they can be time consuming, generally making it difficult to process corpora that are large enough to provide reliable, broadly based statistics. With a system that can run various pronoun resolution algorithms, one can easily and quickly analyze large amounts of data and generate more reliable results. In this study, this ability to alter an algorithm slightly and test its performance is central.

We first show the attractiveness of the Left-Right Centering algorithm (henceforth LRC) (Tetreault 1999) given its incremental processing of utterances, psycholinguistic plausibility, and good performance in finding the antecedents of pronouns. The algorithm is tested against three other leading pronoun resolution algorithms: Hobbs's naive algorithm (1978), S-list (Strube 1998), and BFP (Brennan, Friedman, and Pollard 1987). Next we use the conclusions from two psycholinguistic experiments on ranking the Cf-list, the salience of discourse entities in prepended phrases (Gordon, Grosz, and Gilliom 1993) and the ordering of possessor and possessed in complex NPs (Gordon et al. 1999), to try to improve the performance of LRC.

We begin with a brief review of the four algorithms to be compared (Section 2). We then discuss the results of the corpus evaluation (Sections 3 and 4). Finally, we show that the results from two psycholinguistic experiments, thought to provide a better ordering of the Cf-list, do not improve LRC's performance when they are incorporated (Section 5).

* Department of Computer Science, Rochester, NY 14627. E-mail: tetreaul@cs.rochester.edu

2. Algorithms

2.1 Hobbs's Algorithm

Hobbs (1978) presents two algorithms: a naive one based solely on syntax, and a more complex one that includes semantics in the resolution method. The naive one (henceforth, the Hobbs algorithm) is the one analyzed here. Unlike the other three algorithms analyzed in this project, the Hobbs algorithm does not appeal to any discourse models for resolution; rather, the parse tree and grammatical rules are the only information used in pronoun resolution.

The Hobbs algorithm assumes a parse tree in which each NP node has an N type node below it as the parent of the lexical object. The algorithm is as follows:

1. Begin at the NP node immediately dominating the pronoun.
2. Walk up the tree to the first NP or S encountered. Call this node X, and call the path used to reach it p.
3. Traverse all branches below node X to the left of path p in a left-to-right, breadth-first manner. Propose as the antecedent any NP node that is encountered which has an NP or S node between it and X. If no antecedent is found, proceed to Step 4.
4. If node X is the highest S node in the sentence, traverse the surface parse trees of previous sentences 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, propose it as the antecedent. If X is not the highest S node in the sentence, continue to Step 5.
5. From node X, go up the tree to the first NP or S node encountered. Call this new node X, and call the path traversed to reach it p.
6. If X is an NP node and if the path p to X did not pass through the N node that X immediately dominates, propose X as the antecedent.
7. Traverse all branches below node X to the left of path p in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
8. If X is an S node, traverse all branches of node X to the right of path p in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent.
9. Go to Step 4.

A match is "found" when the NP in question matches the pronoun in number, gender, and person. The algorithm amounts to walking the parse tree from the pronoun in question by stepping through each NP and S on the path to the top S and running a breadth-first search on NP's children left of the path. If a referent cannot be found in the current utterance, then the breadth-first strategy is repeated on preceding utterances.

Hobbs did a hand-based evaluation of his algorithm on three different texts: a history chapter, a novel, and a news article. Four pronouns were considered: *he*, *she*, *it*, and *they*. Cases where *it* refers to a nonrecoverable entity (such as the time or weather) were not counted. The algorithm performed successfully on 88.3% of the 300 pronouns in the corpus. Accuracy increased to 91.7% with the inclusion of selectional constraints.

2.2 Centering Theory and BFP's Algorithm

Centering theory is part of a larger theory of discourse structure developed by Grosz and Sidner (1986). These researchers assert that discourse structure has three compo-

nents: (1) a linguistic structure, which is the structure of the sequence of utterances; (2) the intentional structure, which is a structure of discourse-relevant purposes; and (3) the attentional state, which is the state of focus. The attentional state models the discourse participants' focus of attention determined by the other two structures at any one time. Also, it has global and local components that correspond to the two levels of discourse coherence. Centering models the local component of attentional state—namely, how the speaker's choice of linguistic entities affects the inference load placed upon the hearer in discourse processing. For example, referring to an entity with a pronoun signals that the entity is more prominently in focus.

As described by Brennan, Friedman, and Pollard (1987) (henceforth, BFP) and Walker, Iida, and Cote (1994), entities called **centers** link an utterance with other utterances in the discourse segment. Each utterance within a discourse has one **backward-looking center** (Cb) and a set of **forward-looking centers** (Cf). The Cf set for an utterance U_0 is the set of discourse entities evoked by that utterance. The Cf set is ranked according to discourse salience; the most accepted ranking is by grammatical role (by subject, direct object, indirect object). The highest-ranked element of this list is called the **preferred center** (Cp). The Cb represents the most highly ranked element of the previous utterance that is found in the current utterance. Essentially, it serves as a link between utterances. Abrupt changes in discourse topic are reflected by a change of Cb between utterances. In discourses where the change of Cb is minimal, the Cp of the utterance represents a prediction of what the Cb will be in the next utterance.

Grosz, Joshi, and Weinstein (1986, 1995) proposed the following constraints of centering theory:

Constraints
For each utterance U_i, in a discourse segment D, consisting of utterances of $U_1 \ldots U_m$:

1. There is precisely one Cb.
2. Every element of the Cf-list for U_i must be realized in U_i.
3. The center, $\text{Cb}(U_i, D)$, is the highest-ranked element of $\text{Cf}(U_{i-1}, D)$ that is realized in U_i.

In addition, they proposed the following rules:

Rules
For each utterance U_i, in a discourse segment D, consisting of utterances of $U_1 \ldots U_m$:

1. If some element of $\text{Cf}(U_{i-1}, D)$ is realized as a pronoun in U_i, then so is $\text{Cb}(U_i, D)$.
2. Transition states (defined below) are ordered such that a sequence of Continues is preferred over a sequence of Retains, which are preferred over sequences of Shifts.

The relationship between the Cb and Cp of two utterances determines the coherence between the utterances. Centering theory ranks the coherence of adjacent utterances with transitions that are determined by the following criteria:

1. whether or not the Cb is the same from U_{n-1} *to* U_n;
2. whether or not this entity coincides with the Cp of U_n.

Table 1
Centering transition table.

	$Cb(U_n) = Cb(U_{n-1})$	$Cb(U_n) \neq Cb(U_{n-1})$
$Cb(U_n) = Cp(U_n)$	Continue	Smooth Shift
$Cb(U_n) = Cp(U_{n-1})$	Retain	Rough Shift

BFP and Walker, Iida, and Cote (1994) identified a finer gradation in the Shift transition, stating that Retains were preferred over Smooth Shifts, which were preferred over Rough Shifts. Table 1 shows the criteria for each transition.

Given these constraints and rules, BFP proposed the following pronoun-binding algorithm based on centering:

1. **Generate** all possible $Cb - Cf$ combinations.
2. **Filter** combinations by contraindices and centering rules.
3. **Rank** remaining combinations by transitions.

Walker (1989) compared Hobbs and BFP on three small data sets using hand evaluation. The results indicated that the two algorithms performed equivalently over a fictional domain of 100 utterances; and Hobbs outperformed BFP over domains consisting of newspaper articles (89% to 79%) and a task domain (Tasks) (51% to 49%).

2.3 The S-List Approach

The third approach (Strube 1998) discards the notions of backward- and forward-looking centers but maintains the notion of modeling the attentional state. This method, the S-list (salience list), was motivated by the BFP algorithm's problems with incrementality and computational overhead (it was also difficult to coordinate the algorithm with intrasentential resolution).

2.3.1 The S-List. The model has one structure, the **S-list**, which "describes the attentional state of the hearer at any given point in processing a discourse" (Strube 1998, page 1252). At first glance, this definition is quite similar to that of a Cf-list; however, the two differ in ranking and composition. First, the S-list can contain elements from both the current and previous utterance while the Cf-list contains elements from the previous utterance alone. Second, the S-list's elements are ranked not by grammatical role but by information status and then by surface order.

The elements of the S-list are separated into three information sets—**hearer-old discourse entities** (OLD), **mediated discourse entities** (MED), and **hearer-new discourse entities** (NEW)—all of which are based on Prince's (1981) familiarity scale. The three sets are further subdivided: OLD consists of **evoked** and **unused** entities; MED consists of **inferrables**, **containing inferrables**, and **anchored brand-new discourse intrasentential** entities; NEW consists solely of **brand-new** entities.

What sorts of NPs fall into these categories? Pronouns and other referring expressions, as well as previously mentioned proper names, are *evoked*. *Unused* entities are proper names. *Inferrables* are entities that are linked to some other entity in the hearer's knowledge, but indirectly. *Anchored brand-new* discourse entities have as their anchor an entity that is OLD.

The three sets are ordered by their information status. OLD entities are preferred over MED entities, which are preferred over NEW entities. Within each set, the ordering is by utterance and position in utterance. Basically, an entity of utterance x is preferred over an entity of utterance y if utterance x follows utterance y. If the entities are in the same utterance, they are ranked by position in the sentence: an entity close to the beginning of the sentence is preferred over one that is farther away.

2.3.2 Algorithm. The resolution algorithm presented here comes from Strube (1998) and personal communication with Michael Strube.

For each utterance ($U_1 \ldots U_N$): for each entity within U_i:

1. If U_i is a pronoun, then find a referent by looking through the S-list left to right for one that matches in gender, number, person, and binding constraints. Mark entity as EVOKED.[1]
2. If U_i is preceded by an indefinite article, then mark U_i as BRAND-NEW.
3. If U_i is not preceded by a determiner, then mark U_i as UNUSED.
4. Else mark U_i as ANCHORED BRAND-NEW.
5. Insert U_i into the S-list given the ranking described above.
6. Upon completion of U_i remove all entities from the S-list that were not realized in U_i.

In short, the S-list method continually inserts new entities into the S-list in their proper positions and "cleanses" the list after each utterance to purge entities that are unlikely to be used again in the discourse. Pronoun resolution is a simple lookup in the S-list.

Strube did perform a hand test of the S-list algorithm and the BFP algorithm on three short stories by Hemingway and three articles from the *New York Times*. BFP, with intrasentential centering added, successfully resolved 438 pronouns out of 576 (76%). The S-list approach performed much better (85%).

2.4 Left-Right Centering Algorithm

Left-Right Centering (Tetreault 1999) is an algorithm built upon centering theory's constraints and rules as detailed in Grosz, Joshi, and Weinstein (1995). The creation of the LRC algorithm is motivated by BFP's limitation as a cognitive model in that it makes no provision for incremental resolution of pronouns (Kehler 1997). Psycholinguistic research supports the claim that listeners process utterances one word at a time. Therefore, when a listener hears a pronoun, he or she will try to resolve it immediately; if new information appears that makes the original choice incorrect (such as a violation of binding constraints), the listener will go back and find a correct antecedent.

Responding to the lack of incremental processing in the BFP model, we have constructed an incremental resolution algorithm that adheres to centering constraints. It works by first searching for an antecedent in the current utterance;[2] if one is not found, then the previous Cf-lists (starting with the previous utterance) are searched

1 In the original S-list formulation, pronouns are not the only entities that can be marked as EVOKED; nominal anaphora and previously mentioned proper names (to name just two) can also be EVOKED (Strube and Hahn 1999). In our implementation, pronouns are the only entities that can fall in this category.

2 In this project, a sentence is considered an utterance.

left to right for an antecedent:

1. **Preprocessing**—from previous utterance: $Cb(U_{n-1})$ and $Cf(U_{n-1})$ are available.
2. **Process utterance**—parse and extract incrementally from U_n all references to discourse entities. For each pronoun do:
   (a) Search for an antecedent intrasententially in Cf-partial(U_n)[3] that meet feature and binding constraints.
   If one is found, proceed to the next pronoun within utterance. Else go to (b).
   (b) Search for an antecedent intersententially in $Cf(U_{n-1})$ that meets feature and binding constraints.
3. **Create Cf**—create Cf-list of U_n by ranking discourse entities of U_n according to grammatical function. Our implementation used a left-to-right breadth-first walk of the parse tree to approximate sorting by grammatical function.

It should be noted that while BFP makes use of Rule 2 of centering theory, LRC does not since Rule 2's role in pronoun resolution is not yet known (see Kehler [1997] for a critique of its use by BFP).

The preference for searching intrasententially before intersententially is motivated by the fact that large sentences are not broken up into clauses as Kameyama (1998) proposes. By looking through the Cf-partial, clause-by-clause centering is roughly approximated. In addition, the antecedents of reflexive pronouns are found by searching Cf-partial right to left because their referents are usually found in the minimal S.

There are two important points to be made about centering and pronoun resolution. First, centering is not a pronoun resolution method; the fact that pronouns can be resolved is simply a side effect of the constraints and rules. Second, ranking by grammatical role is very naive. In a perfect world, the Cf-list would consist of entities ranked by a combination of syntax and semantics. In our study, ranking is based solely on syntax.

3. Evaluation of Algorithms

3.1 Data

All four algorithms were compared on two domains taken from the Penn Treebank annotated corpus (Marcus, Santorini, and Marcinkiewicz 1993). The first domain consists of 3,900 utterances (1,694 unquoted pronouns) in *New York Times* articles provided by Ge, Hale, and Charniak (1998), who annotated the corpus with coreference information. The corpus consists of 195 different newspaper articles. Sentences are fully bracketed and have labels that indicate part of speech and number. Pronouns and their antecedent entities are all marked with the same tag to facilitate coreference verification. In addition, the subject NP of each S subconstituent is marked.

The second domain consists of 553 utterances (511 unquoted pronouns) in three fictional texts taken from the Penn Treebank corpus, which we annotated in the same manner as Ge, Hale, and Charniak's corpus. The second domain differs from the first in that the sentences are generally shorter and less complex, and contain more *he*s and *she*s.

3 Cf-partial is a list of all processed discourse entities in U_n.

3.2 Method

The evaluation (Byron and Tetreault 1999) consisted of two steps: (1) parsing Penn Treebank utterances and (2) running the four algorithms. The parsing stage involved extracting discourse entities from the Penn Treebank utterances. Since we were solely concerned with pronouns having NP antecedents, we extracted only NPs. For each NP we generated a "filecard" that stored its syntactic information. This information included agreement properties, syntactic type, parent nodes, depth in tree, position in utterance, presence or absence of a determiner, gender, coreference tag, utterance number, whether it was quoted, commanding verb, whether it was part of a title, whether it was reflexive, whether it was part of a possessive NP, whether it was in a prepended phrase, and whether it was part of a conjoined sentence. The entities were listed in each utterance in order of mention except in the case of conjoined NPs. Conjoined entities such as *John and Mary* were realized as three entities: the singular entities *John* and *Mary* and the plural *John and Mary*. The plural entity was placed ahead of the singular ones in the Cf-list, on the basis of research by Gordon et al. (1999).

Conjoined utterances were broken up into their subutterances. For example, the utterance *United Illuminating is based in New Haven, Conn., and Northeast is based in Hartford, Conn.* was replaced by the two utterances *United Illuminating is based in New Haven, Conn.* and *Northeast is based in Hartford, Conn.* This strategy was inspired by Kameyama's (1998) methods for dealing with complex sentences; it improves the accuracy of each algorithm by 1% to 2%.

The second stage involved running each algorithm on the parsed forms of the Penn Treebank utterances. For all algorithms, we used the same guidelines as Strube and Hahn (1999): no world knowledge was assumed, only agreement criteria (gender, number) and binding constraints were applied. Unlike Strube and Hahn, we did not make use of sortal constraints. The number of each NP could be extracted from the Penn Treebank annotations, but gender had to be hand-coded. A database of all NPs was tagged with their gender (masculine, feminine, neuter). NPs such as *president* or *banker* were marked as androgynous since it is possible to refer to them with a gendered pronoun. Entities within quotes were removed from the evaluation since the S-list algorithm and BFP do not allow resolution of quoted text.

We depart from Walker's (1989) and Strube and Hahn's (1999) evaluations by not defining any discourse segments. Walker defines a discourse segment as a paragraph (unless the first sentence of the paragraph has a pronoun in subject position or unless it has a pronoun with no antecedent among the preceding NPs that match syntactic features). Instead, we divide our corpora only by discourses (newspaper article or story). Once a new discourse is encountered, the history list for each algorithm (be it the Cf-list or S-list) is cleared. Using discourse segments should increase the efficiency of all algorithms since it constrains the search space significantly.

Unlike Walker (1989), we do not account for false positives or error chains; instead, we use a "location"-based evaluation procedure. Error chains occur when a pronoun P_{i_2} refers to a pronoun P_{i_1} that was resolved incorrectly to entity E_k (where P_{i_2} and P_{i_1} evoke the same entity E_i). So P_{i_2} would corefer incorrectly with E_k. In our evaluation, a coreference is deemed correct if it corefers with an NP that has the same coreference tag. So in the above situation, P_{i_2} would be deemed correct since it was matched to an expression that should realize the correct entity.

3.3 Algorithm Modifications

The BFP algorithm had to be modified slightly to compensate for underspecifications in its intrasentential resolution. We follow the same method as Strube and Hahn (1999);

that is, we first try to resolve pronouns intersententially using the BFP algorithm. If there are pronouns left unresolved, we search for an antecedent left to right in the same utterance. Strube and Hahn use Kameyama's (1998) specifications for complex sentences to break up utterances into smaller components. We keep the utterances whole (with the exception of splitting conjoined utterances).

As an aside, the BFP algorithm can be modified (Walker 1989) so that intrasentential antecedents are given a higher preference. To quote Walker, the alteration (suggested by Carter [1987]) involves selecting intrasentential candidates "only in the cases where no discourse center has been established or the discourse center has been rejected for syntactic or selectional reasons" (page 258). Walker applied the modification and was able to boost BFP's accuracy to 93% correct over the fiction corpus, 84% on *Newsweek* articles, and 64% on Tasks (up from 90%, 79%, and 49%, respectively). BFP with Carter's modification may seem quite similar to LRC except for two points. First, LRC seeks antecedents intrasententially regardless of the status of the discourse center. Second, LRC does not use Rule 2 in constraining possible antecedents intersententially, while BFP does so.

Because the S-list approach incorporates both semantics and syntax in its familiarity ranking scheme, a shallow version that uses only syntax is implemented in this study. This means that inferrables are not represented and entities rementioned as NPs may be underrepresented in the ranking.

Both the BFP and S-list algorithms were modified so that they have the ability to look back through all past Cf/S-lists. This puts the two algorithms on equal footing with the Hobbs and LRC algorithms, which allow one to look back as far as possible within the discourse.

Hobbs (1978) makes use of selectional constraints to help refine the search space for neutral pronouns such as *it*. We do not use selectional constraints in this syntax-only study.

3.4 Results

Two naive algorithms were created to serve as a baseline for results. The first, "most recent," keeps a history list of all entities seen within the discourse unit. The most recent entity that matches in gender, number, and binding constraints is selected as the antecedent for the pronoun. This method correctly resolves 60% of pronouns in both domains.

A slightly more complex baseline involves using the LRC algorithm but randomizing all Cf-lists considered. So, in the intrasentential component, the ranking of the entities in Cf-partial is random. Previous Cf-lists are also randomized after being processed. This method actually does well (69%) compared with the "intelligent" algorithms, in part because of its preference for intrasentential entities.

Tables 2 and 3 include results for the different algorithms over the two domains. "Success rate" as defined by Mitkov (2000) is the number of successfully resolved pronouns divided by the total number of pronouns. Two variations of LRC are included as further baselines. LRCsurf ranks its Cf-list by surface order only. LRC ranks the Cf-list by grammatical function. LRC-F is the best instantiation of LRC and involves moving entities in a prepended phrase to the back of the Cf-list (which is still ranked by grammatical function). LRC-P ranks its entities the same way as LRC-F except that it then moves all pronouns to the head of the Cf-list (maintaining original order). This algorithm was meant to be a hybrid of the S-list and LRC algorithms with the hope that performance would be increased by giving weight to pronouns since they would be more likely to continue the backward-looking center.

Table 2
Pronoun resolution algorithms for *New York Times* articles.

Algorithm	Right	Success Rate	% Right Intra	% Right Inter
BFP	1004	59.4	75.1	48.0
Random Cf	1175	69.4	70.2	66.7
S-list	1211	71.7	74.1	67.5
LRCsurf	1266	74.7	72.0	81.6
LRC	1268	74.9	72.0	82.0
Hobbs	1298	76.8	74.2	82.0
LRC-F	1362	80.4	77.7	87.3
LRC-P	1362	80.4	77.7	87.3

Table 3
Pronoun resolution algorithms for fictional texts.

Algorithm	Right	Success Rate	% Right Intra	% Right Inter
BFP	241	46.4	81.8	43.8
S-list	337	66.1	84.4	56.5
Random Cf	367	71.1	84.3	62.5
LRCsurf	372	72.1	84.3	64.2
LRC	372	72.1	84.3	64.2
LRC-P	378	74.0	84.3	66.2
Hobbs	414	80.1	85.8	75.2
LRC-F	420	81.1	86.0	76.2

4. Discussion

For this study, we use McNemar's test to test whether the difference in performance of two algorithms is significant. We adopt the standard statistical convention of $p \leq 0.05$ for determining whether the relative performance is indeed significant.

First, we consider LRC in relation to the classical algorithms: Hobbs, BFP, and S-list. We found a significant difference in the performance of all four algorithms (e.g., LRC and S-list: $p \leq 0.00479$), though Hobbs and LRC performed the closest in terms of getting the same pronouns right. These two algorithms perform similarly for two reasons. First, both search for referents intrasententially and then intersententially. In the *New York Times* corpus, over 71% of all pronouns have intrasentential referents, so clearly an algorithm that favors the current utterance will perform better. Second, both search their respective data structures in a salience-first manner. Intersententially, both examine previous utterances in the same manner: breadth-first based on syntax. Intrasententially, Hobbs does slightly better since it first favors antecedents close to the pronoun before searching the rest of the tree. LRC favors entities near the head of the sentence under the assumption that they are more salient. These algorithms' similarities in intra- and intersentential evaluation are reflected in the similarities in their percentage correct for the respective categories.

Although S-list performed worse than LRC over the *New York Times* corpus, it did fare better over the fictional texts. This is due to the high density of pronouns in these texts, which S-list would rank higher in its salience list since they are hearer-old. It should be restated that a shallow version (syntax only) of the S-list algorithm is implemented here.

The standing of the BFP algorithm should not be surprising given past studies. For example, Strube (1998) found that the S-list algorithm performed at 91% correct on three *New York Times* articles, while the best version of BFP performed at 81%. This 10% difference is reflected in the present evaluation as well. The main drawback for BFP was its preference for intersentential resolution. Also, BFP, as formally defined, does not have an intrasentential processing mechanism. For the purposes of the project, the LRC intrasentential technique was used to resolve pronouns that could not be resolved by the BFP (intersentential) algorithm. It is unclear whether this is the optimal intrasentential algorithm for BFP.

LRC-F is much better than LRC alone considering its improvement of over 5% in the newspaper article domain and over 7% in the fictional domain. This increase is discussed in the following section. The hybrid algorithm (LRC-P) has the same accuracy rate as LRC-F, though each gets 5 instances right that the other does not.

5. Examining Psycholinguistic Claims of Centering

Having established LRC as a fair model of centering given its performance and incremental processing of utterances, we can use it to test empirically whether psycholinguistic claims about the ordering of the Cf-list are reflected in an increase in accuracy in resolving pronouns. The reasoning behind the following corpus tests is that if the predictions made by psycholinguistic experiments fail to increase performance or even lower performance, then this suggests that the claims may not be useful. As Suri, McCoy, and DeCristofaro (1999, page 180) point out: "the corpus analysis reveals how language is actually used in practice, rather than depending on a small set of discourses presented to the human subjects."

In this section, we use our corpus evaluation to provide counterevidence to the claims made about using genitives and prepended phrases to rank the Cf-list, and we propose a new Cf-list ranking based on these results.

5.1 Moving Prepended Phrases

Gordon, Grosz, and Gilliom (1993) carried out five self-paced reading time experiments that provided evidence for the major tenets of centering theory: that the backward-looking center (Cb) should be realized as a pronoun and that the grammatical subject of an utterance is most likely to be the Cb if possible. Their final experiment showed that surface position also plays a role in ranking the Cf-list. They observed that entities in surface-initial nonsubject positions in the previous sentence had about the same repeated-name penalty as an entity that had been the noninitial subject of the previous sentence. These results can be interpreted to mean that entities in subject position and in prepended phrases (nonsubject surface-initial positions) are equally likely to be the Cb.

So the claim we wished to test was whether sentence-initial and subject position can serve equally well as the Cb. To evaluate this claim, we changed our parser to find the subject of the utterance. By tagging the subject, we know what entities constitute the prepended phrase (since they precede the subject). We developed two different methods of locating the subject. The first simply takes the first NP that is the subject of the first S constituent. It is possible that this S constituent is not the top-level S structure and may even be embedded in a prepended phrase. This method is called LRC-F since it takes the first subject NP found. The second method (LRC-S) selects the NP that is the subject of the top-level S structure. If one cannot be found, then the system defaults to the first method. The result of both tagging methods is that all NPs preceding the chosen subject are marked as being in a prepended phrase.

Table 4
Prepended phrase movement experiments over *New York Times* articles.

	Prepended Movement		Standard Sort	
Algorithm	**Norm**	**Pre**	**Norm**	**Pre**
LRC-F	76.21%	80.40%	79.63%	75.50%
LRC-S	75.97%	80.08%	78.81%	74.85%

Eight different corpus trials were carried out involving the two different parsing algorithms (LRC-F and LRC-S) and two different ordering modifications: (1) ranking the Cf-list after processing and (2) modifying the order of entities before processing the utterance. The standard Cf-list consists of ranking entities by grammatical role and surface order. As a result, prepended phrases would still be ranked ahead of the main subject. The modified Cf-list consists of ranking the main clause by grammatical role and placing all entities in the prepended phrase after all entities from the main clause. The second method involves reordering the utterance before processing. This technique was motivated mostly by the order we selected for pronoun resolution: an antecedent is first searched for in the Cf-partial, then in the past Cf-lists, and finally in the entities of the same utterance not in the Cf-partial. Pronouns in prepended phrases frequently refer to the subject of the same utterance as well as to entities in the previous utterance. Moving the prepended entities after the main clause entities before evaluation achieves the same result as looking in the main clause before the intersentential search.

Table 4 contains the results of the trials over the *New York Times* domain. "Prepended movement" refers to ranking the Cf-list with prepended entities moved to the end of the main clause; "Standard sort" refers to maintaining the order of the Cf-list. "Norm" means that prepended entities were not moved before the utterance was processed. "Pre" means that the entities were placed behind the main clause.

All statistics (within the respective algorithms) were deemed significant relative to each other using McNemar's test. However, it should be noted that between the best performers for LRC-F and LRC-S (movement of prepended phrases before and after Cf-list, column 2), the difference in performance is insignificant ($p \leq 0.624$). This indicates that the two algorithms fare the same. The conclusion is that if an algorithm prefers the subject and marks entities in prepended phrases as less salient, it will resolve pronouns better.

5.2 Ranking Complex NPs

The second claim we wished to test involved ranking possessor and possessed entities realized in complex NPs. Walker and Prince (1996) developed the complex NP assumption that "In English, when an NP evokes multiple discourse entities, such as a subject NP with a possessive pronoun, we assume that the Cf ordering is from left to right within the higher NP" (page 8). So the Cf-list for the utterance *Her mother knows Queen Elizabeth* would be {*her*, *mother*, *Elizabeth*}. Walker and Prince note that the theory is just a hypothesis but motivate its plausibility with a complex example.

However, a series of psycholinguistic experiments carried out by Gordon et al. (1999) refute Walker and Prince's claim that the entities are ordered left to right. Gordon et al. found that subjects had faster reading rates for small discourses in which a pronoun referred to the possessed entity rather than the possessor entity.

Table 5
Results of evaluating pronoun resolution algorithms for *New York Times* articles.

Algorithm	WP	+Gen	+Pos
LRC-F	80.04%	79.99%	78.41%
LRC-S	80.40%	80.34%	78.83%
LRC	76.15%	76.03%	74.47%

Table 6
Results of evaluating pronoun resolution algorithms for Fiction texts.

Algorithm	WP	+Gen	+Pos
LRC-F	79.73%	79.58%	79.42%
LRC-S	81.08%	81.08%	80.88%
LRC	80.31%	80.15%	79.81%

Hobbs (1978) also assumes Gordon et al.'s interpretation in his pronoun algorithm. He assumes that possessor entities are nested deeper in the parse tree, so when the algorithm does a breadth-first search of the tree, it considers the possessed NP to be the most prominent.

To see which claim is correct, we altered the Cf-list ranking to put possessed entities before possessor entities. The original LRC ordered them left to right as Walker and Prince (WP) suggest. Tables 5 and 6 include results for both domains. "+gen" indicates that only complex NPs containing genitive pronouns were reversed; "+pos" indicates that all possessive NPs were reversed, matching Gordon et al.'s study. The results indicate for both domains that Walker and Prince's theory works better, though marginally (for all domains and algorithms, significance levels between WP and +gen are under 0.05). For the *New York Times* domain, the difference in the actual number correct between LRC-S with WP and LRC-S with +pos is 1,362 to 1,337 or 25 pronouns, which is substantial ($p \leq$ 1.4e-06) over a corpus of 1,691 pronouns. Likewise, for the fictional texts, 1 extra pronoun is resolved incorrectly when using Gordon et al.'s method.

Looking at the difference in what each algorithm gets right and wrong, it seems that type of referring expression and mention count play a role in which entity should be selected from the complex NP. If an entity has been mentioned previously or is realized as a pronoun, it is more likely to be the referent of a following pronoun. This would lend support to Strube and Hahn's S-list and functional centering theories (Strube and Hahn 1996), which maintain that type of referring expression and previous mention influence the salience of each entity with the S-list or Cf-list.

6. Conclusions

In this paper we first presented a new pronoun resolution algorithm, Left-Right Centering, which adheres to the constraints of centering theory and was inspired by the need to remedy a lack of incremental processing in Brennan, Friedman, and Pollard's (1987) method. Second, we compared LRC's performance with that of three other leading pronoun resolution algorithms, each one restricted to using only syntactic information. This comparison is significant in its own right because these algorithms have not been previously compared, in computer-encoded form, on a common corpus. Coding

all the algorithms allows one to quickly test them on a large corpus and eliminates human error. Third, we tried to improve LRC's performance by incorporating theories on Cf-list construction derived from psycholinguistic experiments. Our corpus-based evaluation showed that prepended phrases should not be ranked prominently in the Cf-list as Gordon, Grosz, and Gilliom (1993) suggest. Our results also showed that Walker and Prince's (1996) complex NP assumption performs marginally better than the opposite theory based on experimental results. We believe that corpus-based analyses such as this one not only increase performance in resolution algorithms but also can aid in validating the results of psycholinguistic studies, which are usually based on small sequences of utterances.

7. Future Work

The next step is to research ways of breaking up complex utterances and applying centering to these utterances. An overlooked area of research, the incorporation of quoted phrases into centering and pronoun resolution, should be explored. Research into how transitions and the backward-looking center can be used in a pronoun resolution algorithm should also be carried out. Strube and Hahn (1996) developed a heuristic of ranking transition pairs by cost to evaluate different Cf-ranking schemes. Perhaps this heuristic could be used to constrain the search for antecedents.

It should be noted that all the algorithms analyzed in this paper are syntax based (or modified to be syntax based). Incorporating semantic information such as sortal constraints would be the next logical development for the system. We believe that purely syntax-based resolution algorithms probably have an upper bound of performance in the mid 80s and that developing an algorithm that achieves 90% or better accuracy over several domains requires semantic knowledge. In short, the results presented here suggest that purely syntactic methods cannot be pushed much farther, and the upper limit reached can serve as a baseline for approaches that combine syntax and semantics.

There are several other psycholinguistic experiments that can be verified using our computational corpus-based approach. The effects of parallelism and other complex NPs such as plurals still need to be investigated computationally.

Acknowledgments

I am grateful to Barbara Grosz for aiding me in the development of the LRC algorithm and for discussing centering theory issues. I am also grateful to Donna Byron, who was responsible for much brainstorming, cross-checking of results, and coding of the Hobbs algorithm. I am thankful as well for Jenny Rogers's work in annotating the fictional texts from the Penn Treebank in the same style used by Ge, Hale, and Charniak (1998). Special thanks go to Michael Strube, James Allen, Lenhart Schubert, and Mark Core for advice and brainstorming. I would also like to thank Eugene Charniak and Niyu Ge for the annotated, parsed Penn Treebank corpus, which proved invaluable.

Partial support for the research reported in this paper was provided by the National Science Foundation under Grants IRI-90-09018, IRI-94-04756, and CDA-94-01024 to Harvard University and by DARPA Research Grant F30602-98-2-0133 to the University of Rochester.

References

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

Byron, Donna K. and Joel R. Tetreault. 1999. A flexible architecture for reference resolution. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–232.

Carter, David M. 1987. *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester, UK.

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

Gordon, Peter C., Barbara J. Grosz, and Laura Gilliom. 1993. Pronouns, names and the centering of attention in discourse. *Cognitive Science*, 17(3):311–348.

Gordon, Peter C., Randall Hendrick, Kerry Ledoux, and Chin Lung Yang. 1999. Processing of reference and the structure of language: An analysis of complex noun phrases. *Language and Cognitive Processes*, 14(4):353–379.

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

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

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

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

Kameyama, Megumi. 1998. Intrasentential centering: A case study. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, 89–112. Oxford University Press.

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

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

Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 869–875.

Mitkov, Ruslan. 2000. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, 96–107.

Prince, Ellen F. 1981. Towards a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, 223–255. Academic Press.

Strube, Michael. 1998. Never look back: An alternative to centering. In *Proceedings of the 36th Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, volume 2, pages 1251–1257.

Strube, Michael and Udo Hahn. 1996. Functional centering. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 270–277.

Strube, Michael and Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Suri, Linda Z., Kathleen F. McCoy, and Jonathan D. DeCristofaro. 1999. A methodology for extending focusing frameworks. *Computational Linguistics*, 25(2):173–194.

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

Walker, Marilyn A. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 251–261.

Walker, Marilyn A., Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–233.

Walker, Marilyn A. and Ellen F. Prince. 1996. A bilateral approach to givenness: A hearer-status algorithm and a centering algorithm. In T. Fretheim and J. K. Gundel, editors, *Reference and Referent Accessibility*, John Benjamins, Amsterdam, pages 291–306.