Squibs and Discussions

The Uncommon Denominator: A Proposal for Consistent Reporting of Pronoun Resolution Results

Donna K. Byron*
University of Rochester

Pronoun resolution studies compute performance inconsistently and describe results incompletely. We propose a new reporting standard that improves the exposition of individual results and the possibility for readers to compare techniques across studies. We also propose an informative new performance metric, the resolution rate, for use in addition to precision and recall.

1. Introduction

To describe the merits of new pronoun resolution techniques, we often compare them with previous approaches using the performance metrics **precision** and **recall**.

Precision $P = \frac{C}{A}$ where:
C = pronouns resolved correctly
A = total pronouns attempted

Recall $R = \frac{C}{T}$ where:
C = pronouns resolved correctly
T = total pronouns in the test data

Precision computes how well a technique did what it was designed to do, and is not at issue here. Recall is intended as a more general performance measure, yet R scores are difficult to interpret due in part to varying methods of calculating T. T includes only the pronouns that were included in the study rather than all pronouns in the data set. But since different studies consider different sorts of pronouns to be in scope, R scores from different studies are difficult to compare. Also, since the pronouns in scope for a study might represent a large or small percentage of the pronouns in the corpus, R reveals little about a technique's utility for the general problem of pronoun resolution.

This paper proposes a new reporting format and a new performance measure to supplement R and P. Pronoun resolution studies differ in many respects, such as the method of calculating C and the underlying semantic assumptions, and this proposal does not address ways to make the studies themselves more consistent (for discussion of these issues, see Walker 1989, van Deemter and Kibble 1999, Mitkov 2000). Instead, we propose a reporting format that clarifies the details of a study's test data (especially those details that tend to differ between studies) and explicitly derives the numbers used to compute performance measures.

* Department of Computer Science, P. O. Box 270226, Rochester, NY 14627. E-mail: dbyron@cs.rochester.edu

2. Reporting Pronoun Resolution Performance

This squib is necessary because past reports of pronoun resolution performance have included inconsistent amounts of detail. Some provide complete details of the experimental design and results, while others (e.g., Byron and Stent 1998) fail to answer even basic questions: What pronouns did this study address? Which pronouns were resolved correctly? In order for a reader to assess performance scores, a report must describe its test data so that the reader knows exactly what the study *includes* and what it *excludes* from T. This section briefly discusses what details should be provided.

2.1 Describing the Test Data

2.1.1 Corpus Type. Each pronoun resolution evaluation is carried out over an **evaluation corpus**, for example a set of human conversations or a number of pages from a book. Details about the evaluation corpus's genre (written or spoken, news or fiction, etc.) and size (e.g., word count, number of discourse units) should be provided to help the reader understand how the corpus chosen for evaluation affected the results obtained.

2.1.2 Lexical Coverage. A report should clearly indicate which pronouns the study included, called the **coverage**, by listing each distinct pronoun type (e.g., *it*, *itself*, and *its* are shown separately). Some past reports give no coverage details at all, while others (e.g., Popescu-Belis and Robba 1997, page 97) precisely state their coverage: "/il/, /elle/, /le/, /la/, /l'/, /lui/, /ils/, /elles/." A categorical description, such as "[results are shown for] personal and possessive pronouns" (Strube 1998, page 1256) is insufficient because the author might assume that his exclusion of certain pronouns (e.g., first person pronouns) need not be mentioned since they are excluded by most other studies.

2.1.3 Exclusions. Before pronoun resolution is executed, any evaluation corpus must be brought into line with the goals of the study by marking individual pronoun tokens as included or excluded.[1] Even tokens of pronoun types covered in the study might be excluded from the evaluation. The reasons for considering tokens to be out of scope for a study are called **exclusion criteria**, and the set of pronouns remaining after all exclusions are applied to the corpus is the **evaluation set**.

Different studies apply different exclusions, and pronoun tokens that were excluded in one study might be counted as errors in another. Cataphors are a case in point. Some pronoun resolution techniques address cataphora (e.g., Lappin and Leass 1994), so the cataphors are included when calculating the performance for these techniques. Other techniques are not designed to identify cataphors, and for some of those the authors exclude cataphors from their test data (e.g., Ge, Hale, and Charniak 1998) while others include them but count the cataphors as errors (e.g., Strube and Hahn 1999). There are no standard guidelines for what exclusions are reasonable to apply, although it would be beneficial for such a standard to exist. Since performance measures are based on the number of pronouns in the evaluation set, such inconsistencies make recall scores from separate studies difficult to compare.

Because each study defines its own idiosyncratic set of exclusion criteria, it is important that performance reports clearly list which criteria were applied. Some

1 Items might be marked in the answer key or in the test corpus itself, for example, by using special part-of-speech tags. Space restrictions prevent us from discussing the additional issues of whether pleonastics and items to be resolved in the text are identified manually or automatically.

Table 1
Pleonastic constructions in English.

Extraposition	Extraposition moves a clausal subject to the predicate. Most nominal clauses can be extraposed, including participles, infinitives, relative clauses, and some prepositional clauses. Example: *It's good that you cleaned up.*
Clefts	Clefts provide contrastive stress with a dummy subject *it* and the focal NP placed after the verb. Example: *It was Pat who gave us directions.*
Idioms	Idioms often include vacuous pronouns, for example, *hit it off.*
Prop-*it*	Prop-*it* is the ascription of properties to an entity with no existential force (Quirk and Greenbaum 1973). Examples: (weather) *It is raining,* (time) *It is 5 o'clock,* and (ambient environment) *It is hot in here.*

reports provide no exclusion details at all, and even when authors do provide them, the descriptions they use are often incomplete or confusing, as in these examples:

- "7 of the pronouns were non-anaphoric and 16 exophoric" (Mitkov 1998, page 872). It is unclear what categories of pronouns this statement refers to, since exophoric pronouns are nonanaphoric.
- "Pleonastic pronouns *it* (i.e. non-anaphoric *it*) have not been included in these results" (Peral, Palomar, and Ferrández 1999, page 71). This assertion seems to incorrectly equate the categories *pleonastic* and *nonanaphoric.*
- "'It' was not counted when referring to a syntactically recoverable 'that' clause or occurring in a time or weather construction" (Hobbs 1986, page 344). These are only some of the possible pleonastic constructions. The reader is left to wonder whether all pleonastic items were excluded.

Without clear and complete exclusion details, it is impossible for future researchers to begin with the same evaluation corpus and recreate results, or for readers of the report to determine whether they think that the exclusions applied were reasonable. To aid future researchers in providing clear and complete exclusion descriptions, the terminology important for describing exclusion criteria is briefly reviewed below. Exclusion categories for nonreferential and referential items must be kept distinct.

Nonreferential items include all items lexically identical to pronouns that do not refer and that should therefore be excluded from performance statistics for pronoun resolution. In English, lexical items called **expletives** or **pleonastics** look like pronouns but are semantically vacuous. Categories of pleonastic items are defined in Table 1. Postal and Pullum (1988) describe tests to discriminate pleonastic from ordinary NPs, since the distinction is not always straightforward. In other languages, forms that sometimes function as pronouns may also be used as other parts of speech, for example, *l'* in French.

Besides pleonastic items, other tokens might be considered nonreferential by a particular study. For example, spontaneous discourse may contain pronouns in abandoned fragments that are uninterpretable to humans. In *So that'll have ok so you want all three boxcars from Dansville?* (Heeman and Allen 1995, d93-10.1,utt29), the initial false start is discarded, so the abandoned token of *that* would probably be excluded.

Referential pronoun tokens can be **anaphoric, cataphoric, exophoric,** or modified to form complete independent references (e.g., *He that plants thorns must not expect to gather roses* (Mitkov 2001). Anaphors "... point back to some previous item" (Halliday

and Hassan 1976, page 14) for their meaning. Many constituents besides pronouns can be anaphoric. Anaphors point to preceding discourse, while cataphors point to subsequent discourse. The stretch of discourse pointed to is a **sponsor**, and the pronoun and sponsor are said to **corefer** when they refer to the same, rather than to a related or inferred, entity. We reserve the term **antecedent** for coreferential base-NP sponsors. **Exophors** refer outside the discourse to entities in the discourse setting. Cornish (1986) and Mitkov (2000) note that the terms **nonreferential** and **nonanaphoric** are often conflated, as are **anaphoric** and **coreferential**, but the above definitions explain why this is incorrect.

Current research tends to focus only on anaphors, so nonanaphoric tokens are commonly excluded. Anaphoric pronouns with certain properties may also be excluded. Some common reasons are:

1. **Split antecedence:** The pronoun is a plural pronoun whose referent must be constructed. Example: *Pat$_i$ went to Kim$_j$'s house and they$_{i+j}$ went dancing.*
2. **Quoted speech:** Either the pronoun or its sponsor occurs in reported speech. Example: *Mr. Vinken$_i$ exclaimed, "The guy ran right in front of me$_i$."*
3. **High-order entities:** Pronouns referring to entities such as propositions and events often have sponsors that are not base NPs. Example: [*He practiced the tuba all night*]$_i$ and it$_i$ almost drove me crazy.
4. **Noncoreference:** The pronoun and its sponsor do not corefer. Example: *The Bakers$_{i+j}$ arrived next. She$_i$'s an astronaut and he$_j$'s a teacher.*
5. **Long-distance reference:** The sponsor appears outside a preset window utilized by the algorithm.

2.2 Measuring Performance

In previous studies, recall has been computed over the pronouns in scope for a study (e.g., only coreferential pronouns, only third person pronouns) rather than all referential pronouns. This makes recall rates difficult to compare because the number of pronouns in scope for different studies varies. Also, because results are stated in terms of the items that were attempted, most studies report similarly high success rates. This author has been asked, "Why is work on pronoun resolution still needed when technique X gets 93% of pronouns correct?" In fact, technique X correctly resolves 93% of singular personal pronouns that have coreferential noun phrase antecedents, which is only a fraction of the pronouns needing to be resolved. This hardly makes pronoun resolution a solved problem. But one must read the report carefully to find these details, and the fact that the question was asked demonstrates the interpretation problems that result from the performance metrics currently in use.

If the long-term goal of pronoun resolution research is to describe a process for interpreting *all* referential pronouns, there should be a performance number that indicates how a technique measures up against this goal. The metric we propose, **resolution rate**, does that by computing the percentage of referential pronouns in the evaluation corpus that were resolved correctly.

The **resolution rate** $RR = \frac{C}{T+E}$ where:

- C = number of pronouns resolved correctly
- T = all pronouns in the evaluation set
- E = all excluded referential pronouns

The denominator of *RR* includes all the pronouns that remain in the evaluation corpus after removing nonreferential items and before excluding referential tokens. Computing *RR* for a technique's performance on a variety of corpora demonstrates the technique's sensitivity to its input data. *RR* also provides a way to reward techniques that attempt to resolve more sorts of pronouns, such as cataphora or event anaphora. Obviously, *RR* applies to techniques that claim general utility but not to those designed for specific circumstances, such as the one reported in Suri, McCoy, and DeCristofaro (1999) or a technique to handle a particular phenomenon such as cataphora. *R* and *P* are still useful to show a technique's performance on the in-scope items, and they are more informative because the reader knows what percentage of the total pronouns were in scope. *R* uses the above definition of *T* as its denominator, and *P* remains unchanged.

All performance measures should be reported separately for each pronoun type covered rather than just for the test corpus as a whole. This facilitates comparing results from studies with different coverage or with test data from different genres where the mix of pronoun types might be different. It also elucidates the effect that the composition of the evaluation corpus had on the results.

3. Proposed Reporting Format: The Standard Disclosure

The **standard disclosure** includes important details, such as the coverage, performance metrics, the size and composition of the evaluation corpus, and the number of pronouns in each exclusion type, all in a user-friendly format. It includes these details in less space than would otherwise be required and spares the author from providing textual descriptions of exclusions, such as "We have only two examples of sentential or VP anaphora altogether.... Neither Hobbs algorithm nor BFP attempt to cover these examples" (Walker 1989, page 257). This leaves more space for commentary on the technique(s) being described. We describe the format as it applies to pronoun resolution studies, but it can be adapted for other categories of referring expression (e.g., descriptive NPs).

3.1 Explanation of the Format

Table 2 is a sample disclosure for a fictional study comparing a new technique, Technique Beta, with an existing baseline Technique Alpha, on the same English evaluation corpus. Footnotes in this example are provided to assist in explaining the format and would not be included in an actual disclosure. Italicized row and column headings indicate parts of the disclosure that will vary depending on the study being reported (they need not be italicized in an actual disclosure), while items not in italics are invariant portions of the format.

The header to the disclosure lists the evaluation corpus used as well as its genre and size. In the table proper, a data column is provided for each lexical type covered by the study; all types that are not addressed in this study are summarized in the "Out of Scope" column. Because pronouns are a closed word class in English, pronoun types are best described by showing the different lexical forms. Some flexibility is allowed; for example, one might wish to collapse the categories for "*He/She*" or "*Him-/Herself.*" In other languages, or for other forms of referring expressions such as descriptive noun phrases, column headings would instead be category labels.

The first data row, "A: Raw Word Count," contains the count of all tokens of that lexical form in the evaluation corpus. The next section details nonreferential exclusions, resulting in subtotal row "B: Sum Nonreferential." More details could be provided in this section at the researcher's discretion; for example, different categories

Table 2
Sample standard disclosure for a fictional study.

Evaluation corpus name: Peanut dialogues (Babar et al. 1994)
Genre: Two-party problem-solving dialogues
Size: 15 dialogues, 937 turns, 31 minutes total speaking time

Pronoun Lexical Types:[a]	*Her*	*She*	*Herself*	*He*	*Him*	*His*	*Himself*	*It*	*Its*	*Itself*	Out of Scope[b]	Total
A: Raw Word Count	22	25	3	89	44	7	14	94	12	1	186	497
Nonreferential Exclusions[c]												
Pleonastic	0	0	0	0	0	0	0	6	0	0	2	8
Abandoned Utterance	0	0	0	1	0	1	0	0	0	0	2	4
B: Sum Nonreferential	0	0	0	1	0	1	0	6	0	0	4	12
C: Total Referential (A−B)	22	25	3	88	44	6	14	88	12	1	182	485
Referential Exclusions[d]												
Plural	0	0	0	0	0	0	0	0	0	0	120	120
Demonstrative	0	0	0	0	0	0	0	0	0	0	36	36
1st/2nd Person	0	0	0	0	0	0	0	0	0	0	24	24
Reported Speech	0	0	0	1	0	0	0	0	0	0	2	3
Event Anaphora	0	0	0	0	0	0	0	15	0	0	0	15
D: Sum Ref Exclusions	0	0	0	1	0	0	0	15	0	0	182	198
E: Evaluation Set (C−D)	22	25	3	87	44	6	14	73	12	1	0[e]	287
Results												
Technique Alpha												
F:#*Correct: Ante (Inter)*	7/7	16/17	0/3	35/45	20/21	2/3	0/14	30/41	2/3	0/1	0	112 (82%)
F:#*Correct: Ante (Intra)*	15/15	7/8	0/0	35/42	20/23	3/3	0/0	24/32	9/9	0/0	0	113 (86%)
Errors: Cataphora	0	0	0	7/7	0	0	0	3/3	0	0	0	10
Errors: Long Distance	0	2/2	0	4/4	0	0	0	4/4	0	0	0	10
G:#*Correct: Refs*	21	22	0	67	38	5	0	52	11	0	0	216 (75%)
Errors: Chaining	1	0	0	0	1	0	0	0	0	0	0	2
Resolution Rate (G/C)	100%	88%	0%	76%	86%	83%	0%	59%	92%	0%	0%	45%
New Technique Beta												
H:#*Correct: Ante (Inter)*	5/7	17/17	3/3	45/45	15/21	2/3	13/14	34/41	3/3	1/1	0	138 (90%)
H:#*Correct: Ante (Intra)*	15/15	7/8	0/0	31/42	24/31	3/3	0/0	27/32	6/9	0/0	0	113 (85%)
Errors: Cataphora	0	0	0	7/7	0	0	0	1/3	0	0	0	8
Errors: Long Distance	0	2	0	4	0	0	0	4	0	0	0	10
I:# Correct: Refs	20	23	3	76	38	5	13	61	8	1	0	248 (86%)
Errors: Chaining	0	0	0	1	2	0	0	0	0	0	0	3
Resolution Rate[f] **(I/C)**	90%	92%	100%	86%	86%	83%	93%	69%	67%	100%	0%	51%

Notes on the format:
[a]Pronouns shown as column headings are those included in this (fictional) study. Other studies would have different column headings depending on their coverage or the language of the evaluation corpus.
[b]Plurals, demonstratives, 1st/2nd person, reported speech, and event anaphora in this example.
[c]Categories in this section differ in different languages. For example, the French *le* is both a pronoun and a determiner, so a study using a French corpus would have an exclusion category for determiners.
[d]These are the exclusions applied in our fictional study. For any particular study, the categories listed here may differ from these.
[e]All pronouns in the "Out of Scope" category have been explicitly listed, resulting in 0 "Out of Scope" pronouns remaining in the evaluation set.
[f]The numerator of *RR* is either correct referents or correct antecedents, depending on the researcher's goals.

of pleonastics could be listed separately. Identifying all the nonreferential tokens is time-consuming, but need only be performed once for each evaluation corpus. The next row, "C: Total Referential," is simply $A - B$ and is used as the denominator of RR.

The next section lists referential pronouns excluded from the test set. All the exclusions applied in the study must be itemized. Categories of pronouns that are clumped together in the "Out of Scope" column, such as demonstratives and plurals in this example, are listed individually in this section. Row "D: Sum Ref Exclusions" shows

the total tokens excluded, and the next row, "E: Evaluation Set," is $C - D$, the resulting count of pronouns that are in scope. Notice that because the table starts with raw word counts and works forward to the evaluation set, the researcher must explicitly account for each excluded token.

The final section shows the performance of the technique(s) under study. For systems that compute referents for the test pronouns, it is recommended that the correct antecedents (Ante) and correct referents (Ref) be shown separately to clarify the effect of chaining errors. We also recommend calculating performance separately for intersentential (Inter) and intrasentential (Intra) sponsors, since techniques tend to vary across this dimension. Separating the resolution details in this manner is informative; however, it is optional. The table could instead show only one number for the total correct resolutions per type of pronoun, although that would be less useful to the reader. Recall would be included for techniques that do not resolve every item attempted.

Error analysis is optional, but in light of the fact that pronouns that are excluded in one study often cause errors in another, it is highly recommended that error details be shown for classes of pronouns that are commonly excluded. Other categories of errors could be detailed as well if particular error categories are of interest in the study. The resolution rate is shown last, calculated as the number of correct resolutions divided by the number of referential pronouns in row C. If a technique reports high *RR* with this format, it is easy to tell whether its performance results from doing a few things well or from doing a mediocre job at everything.

To summarize, the important features of this format are:

1. The pronoun types included in the study are readily apparent.
2. Categories and itemized counts of excluded tokens are clearly shown.
3. *RR* can be calculated because the referential exclusions are enumerated.

3.2 The Benefits of the Standard Disclosure

By combining details of the evaluation corpus's construction with performance statistics, the standard disclosure displays many important details in one place, making them easy for readers to find. Some authors in the past have stated their performance statistics separately for each pronoun type, while others stated only one overall performance number per technique. Because a particular technique's performance can vary widely across pronoun types (for example, Hobbs's algorithm resolved 93% of instances of *he* but only 77% of instances of *it*; Hobbs 1986), reporting performance per pronoun type should become standard practice. Also, different studies choose different combinations of pronouns to investigate, and without detailed performance numbers one cannot know how the two techniques compare on the pronouns they have in common. Although the only sure way to compare two techniques is in a head-to-head test on the same corpus, results stated in the standard disclosure format leave the reader better able to judge, for example, if a technique might be appropriate to his corpus.

Providing details on the exclusion criteria applied to the evaluation corpus provides a sanity check so that the reader understands how the initial corpus was pared down to become the evaluation data set. If the table shows that an unexpectedly high percentage of pronouns were excluded from testing, the reader might wonder whether the results obtained are reliable or if, on the other hand, the researcher might have overzealously tailored the evaluation set to the capabilities of the algorithm being tested. Because many past studies either did not discuss their exclusion categories at

all, described their exclusions with confusing descriptions of the sort listed in Section 2.1.3, or did not state the number of pronouns excluded, the reader must be guarded in interpreting the stated results. The tabular format suggested here does not guarantee consistent application of the exclusion categories across studies, but it does represent an improvement over current practices. Preparing exclusion data might at first seem like an extra burden. However, it must only be collected once per evaluation corpus, and much of this information is already collected during the corpus annotation process. As we demonstrated above, many authors already discuss exclusions in the body of a paper. We believe that the increased clarity that the standard disclosure format offers to the reader outweighs any small outlay of time required to prepare it.

Finally, this format allows the researcher to compute RR for general-purpose algorithms, giving the community a more realistic view of how an algorithm performs. While in the past the reader knew that a particular technique correctly resolved 93% of some subset of pronouns, he had no clear idea what that 93% represented because the process used to derive its denominator was so unclear.

4. Summary

The reporting format we propose has numerous benefits. Important details of a pronoun resolution study are in one place and easy for readers to find. The information is organized to clearly state details that may differ from one study to another so that future researchers do not need to reimplement a technique simply to remove these differences. Its tabular format consumes less space for this additional information, freeing up room in the body of a paper for analysis and discussion of the techniques under investigation. By tabulating the number of referential pronouns that are excluded, the format clarifies the composition of the test data set and enables the calculation of the resolution rate (RR), which is a more accurate general measure of performance. RR makes a nice addition to the performance metrics currently in use that state performance in terms of the in-scope pronouns. While it does not solve many of the difficulties involved in comparing techniques from different studies, this format does offer an incremental improvement over current practices.

Acknowledgments

This material is based on work supported by ONR Grant N00014-95-1-1088 and DARPA Grant F30602-98-2-0133. The author thanks James Allen, Nate Blaylock, Jason Eisner, Lucian Galescu, Brandon Sanders, Amanda Stent, and the anonymous reviewers for helpful comments on ideas developed here.

References

Byron, Donna and Amanda Stent. 1998. A preliminary model of centering in dialog. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 1475–1477.

Cornish, Francis. 1986. *Anaphoric Relations in English and French*. Croom Helm.

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.

Halliday, M. A. K. and Ruqaiya Hassan. 1976. *Cohesion in English*. Longman.

Heeman, Peter A. and James Allen. 1995. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium.

Hobbs, Jerry. 1986. Resolving pronoun reference. In Barbara J. Grosz, Karen Sparck Jones, and Bonnie Lynn Webber, editors, *Readings in Natural Language Processing*. Morgan Kaufmann, pages 339–352.

Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics, 20(4)*:535–561.

Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, 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 Anaphora Resolution Conference (DAARC2000)*, pages 96–107.

Mitkov, Ruslan. 2001. *Anaphora Resolution.* Longman.

Peral, Jesús, Manuel Palomar, and Antonio Ferrández. 1999. Coreference-oriented interlingual slot structure and machine translation. In *Proceedings of the Workshop on Coreference and Its Applications (ACL'99)*, pages 69–76.

Popescu-Belis, Andrei and Isabelle Robba. 1997. Cooperation between pronoun and reference resolution for unrestricted texts. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 94–99.

Postal, Paul M. and Geoffrey K. Pullum. 1988. Expletive noun phrases in subcategorized positions. *Linguistic Inquiry*, 19:635–670.

Quirk, Randolph and Sidney Greenbaum. 1973. *A University Grammar of English.* Longman.

Strube, Michael. 1998. Never look back: An alternative to centering. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 1251–1257.

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.

van Deemter, Kees and Rodger Kibble. 1999. What is coreference, and what should coreference annotation be? In *Proceedings of the Workshop on Coreference and Its Applications (ACL'99)*, pages 90–96.

Walker, Marilyn A. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)*, pages 251–261.