Advances in Cognitive Systems 4 (2016) 247–263 Submitted 8/2015; published 12/2016

Resolving Difficult Referring Expressions

Marjorie McShane MARGEMC34@GMAIL.COM
Petr Babkin PETR.A.BABKIN@GMAIL.COM
Cognitive Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

Some referring expressions, such as pronominal *this* and *that*, are particularly difficult to resolve automatically and, therefore, are treated minimally if at all by most reference resolution systems. Other referring expressions, such as *he, she,* and *they*, are treated by many systems but, as yet, not with sufficient accuracy. We describe a system called CROSS (**C**o**R**eference for the **O**nto**S**em2 language processing **S**ystem) that automatically selects which instances of difficult referring expressions it can treat with high precision and identifies their textual antecedents. The system uses readily computable heuristic evidence in a configuration-matching approach. The identification of textual antecedents represents an intermediate result toward full reference resolution, which requires semantic and pragmatic analysis, and which augments an intelligent agent's memory. Our evaluation shows that a language problem which seems impenetrable when viewed from the current mainstream perspective of machine learning becomes more manageable using human-inspired modeling.

1. Introduction

Reference resolution covers a spectrum of phenomena that, in terms of machine processing, range from relatively simple to extremely difficult. Among the more difficult referring expressions are so-called *broad referring expressions*, such as pronominal *this* and *that*, which can refer to either entities (1) or propositions (2) to (4).[1]

(1) "*This provision* has nothing to do with welfare reform. **It** is simply a budget-saving measure," …

(2) "*It was new equipment* and **that** is why we decided to retrieve it"…

(3) Coumadin is part of a class of pharmaceuticals known as "narrow therapeutic index" drugs. **That** means the dosages must be tightly controlled.

(4) [In the middle of a narrative about Ashley] She picked up a fork, stared at the food for a moment, then shook her head in despair. Fear had taken away her appetite. **This** can't go on, she thought angrily. Whoever he is, I won't let him do this to me. (COCA)

[1] In examples, the referring expressions to be resolved appear in boldface and their antecedents – when textually available – are in italics. Unless otherwise noted, all examples come from the English Gigaword corpus (Graff & Cieri, 2003). Invented examples and ones from the COCA corpus (Davies, 2008) are so indicated.

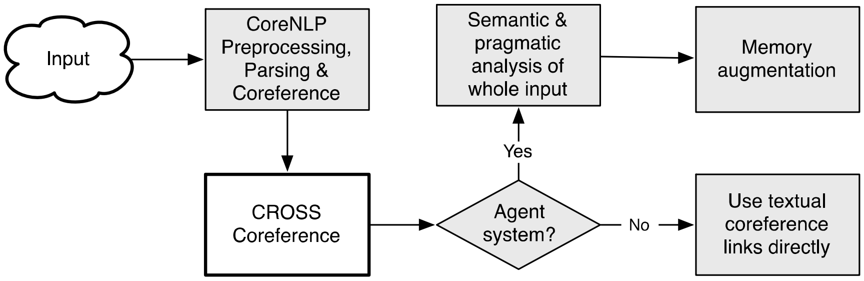
Input
CoreNLP Preprocessing, Parsing & Coreference
CROSS Coreference
Agent system?
Yes
No
Semantic & pragmatic analysis of whole input
Memory augmentation
Use textual coreference links directly

Figure 1. The configuration-based coreference resolution provided by CROSS contributes to full reference resolution, which is defined as memory augmentation for intelligent agents.

When a broad referring expression has a propositional coreferent, its meaning can be expressed in various ways. In the simplest case, it can be encoded by a text string, as in (2). Or it can be partially recoverable from a text string: in (3), *that* means "classifying a drug as 'a narrow therapeutic index drug'". Finally, it can be unavailable linguistically, requiring reasoning about the context overall: in (4), *this* means "my living, feeling, etc., the way I do", which the reader should understand from the preceding context. Because of the challenges broad referring expressions pose, most natural language processing (NLP) systems do not treat them.

In addition to untreated referring expressions, there are referring expressions that have been widely treated but have resisted high-precision results. One example is third person personal pronouns. The reason for the low precision is that resolution often requires specific world knowledge and reasoning, as illustrated by Winograd Schema examples like *The man$_i$ could not lift his son$_k$ because [**he$_i$** was so weak / **he$_k$** was so heavy]* (Levesque et al., 2012).

1.1 The CROSS System

Here we report on a system called CROSS – **C**o**R**eference for the **O**nto**S**em2 language processing **S**ystem – which identifies textual coreferents for three classes of difficult problems: broad referring expressions realized as pronouns (*this, that, it*), broad referring expressions realized as definite descriptions (e.g., *this proposal, that suggestion*), and the personal pronouns *he, him, she, her, they* and *them*. Our approach combines linguistic observations, recorded as predictive configurations, with the syntactic analysis provided by the Stanford CoreNLP tool set (version 3.4.1; Manning et al., 2014). The text level coreference links provided by CROSS can be used as a final result for knowledge-lean systems or they can serve as heuristic evidence for full reference resolution by intelligent agents. Figure 1 shows how CROSS contributes to overall language analysis in the OntoSem2 analyzer within the OntoAgent cognitive architecture (McShane & Nirenburg, 2012).

The content of CROSS – and, therefore, the genre of this system description – will be more native to linguists than to mainstream NLP developers. Linguistic phenomena are classified and then treated using knowledge-based methods. The details of each treatment strategy, including where it succeeds and fails and why, both underscore the nature of this contribution and set the stage for iterative system enhancements. In describing CROSS, we attempt to ensure reproducibility without presenting unnecessary details about the implementation. Some generalizations, operationalized as rules, apply to many aspects of reference processing and are not listed separately with each configuration. Among the more reliable rules are that coreferential

NPs must match in gender and number;[2] non-referring expressions, such as pleonastic *it* and the complementizer *that*, must be detected and excluded from reference processing; a text-span antecedent cannot be partly inside of, and partly outside of, a direct quote; and disjoint portions of quoted speech (e.g., separated by an indication of the speaker) can serve together as an antecedent. Other useful generalizations, although defeasible, are that in a nominal compound typically only the head is available as an antecedent, and given multiple candidate antecedents with the necessary feature values, the most proximate one tends to be the best choice. In contrast to these cross-configuration generalizations, some configuration-specific details must be handled to ensure reproducibility of our results; these are mentioned separately for the relevant configurations.

We constrain the heuristic evidence used by CROSS to lexical and syntactic features, not relying on semantic analysis or deep reasoning. There are three reasons for this choice. First, it makes our results useful outside of our research paradigm, since most systems do not invoke semantic analysis. Second, high-quality deep semantic analysis is not currently achievable in open domains, leading to variable-quality results if semantics is leveraged. Third, and theoretically most important, the default supposition that semantic analysis must precede pragmatic analysis – which, naturally, includes reference processing – is a counterproductive oversimplification (McShane & Nirenburg, 2015). Even without fully understanding an input, people can often confidently establish reference relations. For example, from the dialog *"I pshacted yesterday." "That's nice."* we know that *that* refers to *I pshacted yesterday,* even though we do not know what the word *pshacted* means. Moreover, not only *can* reference be resolved pre-semantically, in some cases it must. For instance, one cannot disambiguate *stopped* in *The rain continued for hours then it stopped* without first determining what *it* refers to. The coreference between *the rain* and *it* can be hypothesized using syntactic heuristics and then confirmed by the successful full interpretation of the input as "RAIN-EVENT (PHASE END) (TIME BEFORE-SPEECH-TIME)".

1.2 Goals and Measures of Progress

The challenge of treating difficult referring expressions mirrors the biggest challenge of artificial intelligence: the need to show useful near-term progress while contributing to long-term success. Our selection of a reference-related research problem meeting these criteria reflects two tenets. First, we consider textual coreference a useful result. For knowledge-lean systems, it can serve as a final result that can be incorporated into applications. For knowledge-based systems, it is an intermediate result contributing to full reference resolution. Second, we consider it preferable to treat *some* instances of difficult referring expressions rather than no instances at all. For knowledge-lean systems, this should increase recall in applications like knowledge extraction and question answering. For agent systems, it will let the agents perform confidently in at least some situations. But selectively treating instances only makes sense if the system can independently choose what to treat. Automatic choice stands in contrast to the common NLP simplification of manually selecting so-called "markables" during corpus annotation and limiting system responsibility to them, as in the MUC-7 coreference task (Hirschman & Chinchor, 1997). This latter practice results in systems that perform much worse on open text than in stylized evaluation exercises (Mitkov, 2001).

[2] There are exceptions to this generalizations, such as *If someone*$_{SINGULAR}$ *wants to go they*$_{PLURAL}$ *should* – but we did not pursue this in the reported implementation.

In short, it is productive to view textual coreference as a task that can sometimes be carried out early and cheaply, in the spirit of the principle of least effort (Piantadosi et al., 2012). When this is the case, the resulting information can: (a) be used as an end result by knowledge-lean systems, (b) inform proposition-level semantic analysis, including lexical disambiguation, and (c) contribute to full reference resolution, defined as anchoring instances of objects and events in agent memory.

The hypothesis we pursued in this research and development effort, which springs from the corpus-based analysis of difficult referring expressions presented by McShane (2015), was that we could build a system to independently detect which instances of difficult referring expressions it could handle, treat those instances, and estimate confidence in its treatments using corpus-attested measures of precision. We will see that our evaluation results support this hypothesis.

1.3 Evaluation Preliminaries

Since evaluation results are threaded throughout the paper, some introductory comments are in order. We tasked CROSS with processing a large corpus and selecting from it those examples that it could treat using its inventory of lexico-syntactic configurations. The system treated all the examples it extracted and was judged on the correctness of its answers – i.e., precision. Recall has no place in this evaluation strategy.[3]

For development and evaluation we used different portions of the Gigaword corpus (Graff & Cieri, 2003). We compiled the gold standard against which the system would be evaluated in steps. First, two graduate students and one undergraduate student annotated the examples the system had extracted using six conventions:

[NE] If the selected entity is not actually a referring expression (e.g., pleonastic *it*), type [NE] before the example.

[ ] If there is a single perfect or near-perfect antecedent, surround it with brackets.

[Mult] If there is more than one possible antecedent, type [Mult] before the example and use multiple sets of brackets to indicate the options.

[Close] If an available text string is close to the needed antecedent but not a perfect match, type [Close] before the example and use brackets to show the best available antecedent.

[Impossible] If no text string captures the meaning of the antecedent, type [Impossible] before the example.

[Prob] If there is some other problem with the context (e.g., it is unintelligible) type [Prob] before the example.

Annotators were shown a few worked examples but given no further instructions, in contrast to earlier, well-known, annotation efforts that involved extensive guidelines that were painstakingly compiled by developers and then memorized by annotators.

[3] Calculating recall would have required annotating a corpus for all of these expressions, which would have been prohibitively expensive. The fact that precision and recall are convenient metrics for supervised learning systems does not make them necessary for all systems, particularly since corpus annotation is a well-known logjam in advancing the state of the art in automatic reference resolution.

Table 1. Some lexico-syntactic configurations, presented informally, for resolving pronominal broad referring expressions. ( ) indicates optionality, / indicates a choice, AUX indicates an auxiliary verb, NEG indicates negation, ADV indicates an adverb, boldface indicates the referring expression being resolved, and italics indicate its antecedent.

	Configuration	Example
1	ask/wonder why *CLAUSE* … **It**’s/**It** is because	If you’ve wondered why *so many 80- and 90-year-old women are named Alice*, **it**’s because … (COCA)
2	Why AUX *SMALL-CLAUSE* … **It**’s/**It** is because	“Why is *he busy*? **It**’s because of the pressure that’s being put on him,”…
3	If/when/in each case where/anytime when/whenever *CLAUSE*, (ADV) **it**’s/**it** is because	If *Myanmar seems oddly quiet*, **it** is because many are tired of struggle and just want to improve their lives.
4	Not only AUX (NEG) *this/it/NP*$_{\text{subj}}$ … **this/it**$_{\text{subj}}$	“Not only did *it* show that the emperor was very much a human being, **it** was also a grim reminder of the defeat and subserviance of their nation.”
5	*This/It/NP*$_{subj}$ (AUX) not only … **this/it**$_{\text{subj}}$	*The module* not only disables the starter, **it** shuts down the fuel injection… (COCA)
6	*This/It/NP*$_{subj}$ has/had nothing to do with … **this/it**$_{\text{subj}}$	“*This provision* has nothing to do with welfare reform. **It** is simply a budget-saving measure,” …
7	*This/It/NP*$_{subj}$ is/was not about … **this/it**$_{\mathbf{subj}}$ is about	“*This war* is not about diplomacy, he added. “**It** is about gangsterism …”
8	*This* is not (a/an) N … **it**$_{\text{subj}}$ is/**it**’s	“*This* is not the lottery. **This** is this man 's life, …”
9	*That*’s why… **That**’s why…	*That*’s why we stayed in the game and **that**’s why we won.

When the annotation results were in, the coauthors manually reviewed them (with the help of the program kdiff3[4]) and selected which ones should be included in the gold standard. Often we considered more than one correct. Occasionally, we added an option not provided by the annotators. As expected, there were many differences across annotations, but most of them were inconsequential. For example, a punctuation mark could be included or excluded, a relative clause attached to an NP could be included or excluded, different annotators could select different members of a coreference chain as the antecedent, and annotators could include or exclude the label [Close]. These differences rendered moot a formal calculation of interannotator agreement.

To evaluate the system, we semi-automatically (again, with the help of kdiff3) compared the system’s answers to the gold standard, calculated precision, and carried out error analysis toward the goal of system improvement. Next we describe, in turn, the three classes of difficult referring expressions that CROSS treats: broad referring expressions realized as pronouns, broad referring expressions realized as definite descriptions, and third-person personal pronouns.

[4] Available at http://kdiff3.sourceforge.net.

2. Resolving Pronominal Broad Referring Expressions

CROSS handles instances of pronominal broad referring expressions – *this, that* and *it* – that occur in lexico-syntactically defined configurations, fill semantically constrained case role slots, or participate in syntactically simple configurations. We discuss these cases in subsections 2.1, 2.2, and 2.3, respectively.

2.1 Lexico-Syntactic Configurations for Resolution

Language does not rely exclusively on free combinations of individual words. Instead, it features multi-word expressions (*blow one's cork*), multi-part phrases (*if ... then*), semantically-paired structures (*in the first place ... in the second place*), and many types of constructions. In fact, constructions are so prevalent that they have become the cornerstone of some grammars (Fillmore et al., 1988; Goldberg, 2003).

Guided by introspection and corpus analysis, we formulated the configurations shown in Table 1, which predict the antecedents (in italics) for the referring expressions shown in boldface. These configurations leverage linguistic generalizations such as the fact that questioning the reason for something is often followed by an indication of that reason (Configuration 1), saying that something *not only* does one thing often leads to saying what else it does (Configurations 4 and 5), and a negated proposition often introduces its positive counterpart (Configurations 6 and 7). In some cases, the broad referring expression resolves to a proposition, whereas in others it resolves to a noun phrase. Although resolution to a noun phrase might seem simpler, it is still challenging because the system does not know beforehand whether the antecedent is an NP or a proposition.

Testing showed the need to incorporate some additional rules, on top of the basic configuration matching, to exclude false positives. These included:

1. For Configurations 5 to 8, if a *but* clause (double underlined in (5)) intervenes between what appears to be the two parts of the configuration (underlined), there is no match.

   (5) And for the first time in history, the Defender not only wants to introduce its own new rule for the class of boat to be raced, but also to keep *this rule* secret. **It**$_{\text{SUBJ}}$ will be disclosed to challengers at a much later stage, putting all challengers at a huge disadvantage.

2. The second half of the configuration should not be in a subordinate clause. For example, in (6) the *it's because* clause is subordinate to *I think*, so the example does not match Configuration 3. Another method of detecting this false positive example would involve recognizing that this *if* clause is paired with a different clause: *then that would be good*.

   (6) "If this could last until fall, then that would be good," he said. "I think **it**'s because it's rained, not because of the air quality."

We evaluated 27 contexts, for which 25 answers (92.6%) were correct, one (3.7%) was incorrect, as sophisticated reasoning would have been needed to select the right antecedent, and one (3.7%) was partially correct in that the system incorrectly included *then* in the antecedent for (7).

(7) If [*you have made it this far*, then] **it**'s because you have talent and the potential to do the job).

Table 2. Configurations for investigating the utility of case-role constraints.

	Configuration	**Constraint**	**Example**
1	*keyword* … **It** AUX verb$_{\text{PastParticiple}}$	*Keyword* is a typical object filler for the verb.	*The idea* is not new: **it** is being discussed by the convention on the future of Europe…
2	*keyword* … **It**$_2$ verb$_{\text{PastParticiple}}$	*Keyword* is a typical subject filler for the verb.	*The plane* hit a tree and then broke in two after **it** crashed…

The high precision of this reference resolution strategy suggests that it would be worthwhile to seek more configurations of this type. It is worth investigating whether unsupervised learning might help to identify such configurations. Supervised learning is unlikely to be feasible in the near future due to the expense of corpus annotation.

2.2 Case-Role Constraints for Resolution

If a verb imposes narrow selectional constraints on the case role that a broad referring expression fills, those constraints can guide search for the antecedent. For example, planes are typical themes of crashing (8) and ferries are typical themes of sinking (9).

(8) Several transport officials have said that flight recorders showed that the pilot's son had been at *the plane*'s controls when **it** crashed.

(9) *The ferry* was en route from Bukoba, on the western shore of the lake, to Mwanza, on the southeastern shore – both in Tanzania – when **it** sank before dawn about 30 nautical miles from Mwanza, the radio said.

As mentioned earlier, despite the fact that the antecedents in these examples are noun phrases, we still consider this broad referring expression processing because the system faces the challenge of deciding whether the antecedent in a given example is a noun phrase or a span of text. By contrast, the MUC-7 coreference task excluded instances of *it* that had text-span antecedents.

Case-role constraints belong to the realm of semantics. Although CROSS does not pursue or rely on semantic analysis, one can create list-based substitutes for semantic analysis, which is what we did for this experiment. Using a combination of corpus analysis, introspection, and consultation of resources like WordNet (Miller, 1995), we compiled a list of verbs for which either the subject or the object was narrowly constrained. We then compiled a list of typical fillers for that role; for example, the verb *abolish* often takes the objects *law, bill, agency, department, death penalty, slavery, capital punishment, draft,* and *regulation*. Our implementation used 202 verbs with an average of 50 keywords each; that average was increased by verbs like *eat, cook* and *die*, for which hundreds of food items and animals, respectively, were listed as keywords. The system sought examples matching the syntactic configurations shown in Table 2.

We evaluated 34 contexts, of which 27 (79.4%) were completely correct, five (14.7%) were incorrect, and two (5.9%) received partial credit. We awarded partial credit if the system selected the correct head but failed to include a postmodifier that all annotators selected. An example is (10), in which the correct head is in italics, whereas CROSS's selection is in brackets.

(10) "If ever there was *[a question] about the strength of our democratic institutions in the face of healthy and natural political argument*, **it** has been answered by the measured response of the American people to these extraordinary events," Clinton said.

Two of the errors involved false detection: twice CROSS failed to recognize pleonastic *it* in the phrase *it was reported*; instead, it established a coreference link between *it* and the NPs *assault* and *a clean financial report* – both of which can, in other contexts, be coreferents for *it* in the phrase *it was reported*. Two more of the errors, including (11), showed that it is not always correct to select the closest candidate antecedent that has the necessary features.

(11) *American Airlines Flight 587* twice ran into turbulence left by [a jumbo jet], including a blast of air that sent *it* careening sideways just seconds before **it** crashed…

As our experiment showed, case-role constraints can aid reference resolution even if implemented outside of a semantic analysis system using only word lists. We hypothesize that machine learning could generate a much larger, still high-quality, inventory of verb-argument pairs, thus greatly increasing the coverage of this approach. However, the strategy will clearly have the greatest impact in systems that *do* involve semantic analysis, since disambiguating events in conjunction with their case-role fillers is a cornerstone of semantic analysis.

2.3 Resolution in Syntactically Simple Contexts

Resolving broad referring expressions that refer to propositions is particularly challenging. In order to detect such instances – separating them from instances that refer to NPs – we focused on four configurations, in which the italicized verbs reflect any combination of values of tense, aspect, and mood:

- despite this/that
- because of this/that
- this/that *is* why/because
- this/that *means, leads to, causes, suggests, creates, makes*

This list by no means exhausts the inventory of collocations in which broad referring expressions tend to have a propositional antecedent, but it was sufficient to support experimentation.

For these collocations, we attempted to determine the contexts in which antecedents for these expressions could be selected with high confidence using only lexico-syntactic heuristics.[5] Inspired by our past work on resolving verb phrase ellipsis (McShane & Babkin, 2015), we hypothesized that the most readily treatable contexts would be those in which the clause containing the broad referring expression was directly preceded by what we call a *simple clause*. Informally, a simple clause contains few, if any, competing candidate antecedents. It can be realized as a full sentence (12), or as the clause preceding the broad referring expression-clause in the same sentence (13).

(12) *They live far from their homes*. **That** makes them stronger than if they formed a real community.

[5] Past research (e.g., Byron, 2004) has shown that text-span antecedents for broad referring expressions are almost always contiguous with the broad referring expression clause. However, the question remains how far back the antecedent extends – i.e., how many clauses it contains.

(13) "*Strong Serbia is not to the liking of some powers abroad*, and **that**'s why they are trying to break it up with the help of the domestic traitors," he said.

Defined in terms of the output of the CoreNLP dependency parser, simple clauses contain none of the dependencies *advcl*, *parataxis, ccomp, rcmod, complm, dep, conj* (with verbal arguments), *xcomp* (with a lexically recorded matrix verb as the governor), or *aux* (not involving a tense marker).

By contrast, when a broad referring expression is preceded by a non-simple clause, real-world reasoning is often required to resolve its reference. For example, both (14) and (15) contain clausal conjunctions in the sentence that precedes the broad referring expression. However, whereas in (14) the most complete antecedent for *that* is the preceding two sentences, in (15) the latter conjunct alone serves as the antecedent.

(14) *For maximum absorption, take your multi supplement with meals, not on an empty stomach. And make sure the meal isn't totally fat-free*. **That**'s because the fat-soluble vitamins in multis (beta-carotene/vitamin A, vitamins D and E) need a little fat to get inside you… (COCA)

(15) "Police will go pass some prostitutes on the corner and *harass some kids having a disagreement*. **It**'s because we're young." (COCA)

Constraining treatable instances to automatically detected simple clauses offers high precision but limited coverage. For this reason, we extended the notion of simple clause using three relaxation strategies that we describe in turn.

Relaxation Strategy 1. CROSS permits modalities like *I believe*, to scope over the main proposition, since they do not introduce another main verb to compete as the antecedent.

(16) "I believe *Jenny will swim faster than she ever has in Barcelona*, and **that** means she has a good chance of bringing home five medals, though the color is still to be determined"… (COCA)

However, allowing for modalities raises the question of whether the modality should be included in, or excluded from, the antecedent. For example, whereas the antecedent-clause modality is excluded from the antecedent in (16), it must be included in the antecedent in a slightly different invented example (17).

(17) "*I believe Jenny will swim faster than she ever has in Barcelona*, and **that** is why I bet big money on her."

We were unable to arrive at high-confidence, broad-coverage generalizations for treating modality when resolving broad referring expressions. This contrasts with our successful generalizations about treating modality in verb phrase ellipsis reconstruction, reported in McShane and Babkin (2015). For now, CROSS *includes* modalities in all antecedents by default, but a more sophisticated treatment remains for future work.

Relaxation Strategy 2. CROSS permits the antecedent clause to include any number of relative clauses such as *he has played against* in (18), which are included in the resolution:

(18) *Every team he has played against has targeted him* but **that** makes him a better player.

Relaxation Strategy 3. CROSS carries out one type of automatic sentence trimming – removing references to the speakers of direct speech – to create simple clauses out of certain types of non-simple clauses.[6] The trimmed material is indicated using strikethrough.

(19) ~~“~~*Energy efficiency is really the name of the game in terms of what we can do now,*~~" she said, adding that she was disappointed that Bush did not adopt a more proactive stance on global warming, despite urging on the part of Blair~~.” **That**’s why today I'm calling on the president to show real leadership,” she said, adding it was unacceptable to adopt a stance that other nations blamed for high greenhouse gas emissions, such as China and India, take steps first.

To weed out residual false positives, CROSS excludes examples in which the referring expression started a quotation but the candidate antecedent did not contain any quoted material.

We evaluated 60 contexts, of which 50 (83.3%) answers were completely correct. The rest deserved partial credit, for which we delineated two categories: five were mostly correct (8.3%) and five were somewhat correct (8.3%). To be considered mostly correct, the antecedent could contain a benign additional element, such as an adverbial clause that modifies the antecedent clause (20).

(20) [After earlier taking the men’s and women’s individual events, *they took the women’s team gold Thursday, winning a shoot-out with China 242-238*.] **That** made Kim Jo-sun a double gold medalist.

To be considered somewhat correct, the selected antecedent must include the actual antecedent but could also include elements that really should have been stripped, such as the main-clause subject and verb in a sentence whose subordinate clause is the antecedent.

(21) [The library said *its copy of the tome*, *previously thought to be 100 years old, in fact dates from between 1660 and 1675*.] **That** means it was printed not long after the original Guru Granth Sahib was compiled in 1604.

To recap, this ellipsis resolution strategy addresses a very difficult class of broad referring expressions but performs with sufficient precision to be useful in applications. We believe that we are on the right track both in operationalizing the notion *simple clause* and in relaxing its definition to cover more contexts. The relaxation strategies could, no doubt, be improved to increase both precision and coverage. The evaluation underscores that the binary metric of *correct/incorrect* is not sufficient to judge systems that address this difficult problem.

4. Resolving Broad Referring Expressions Realized as Definite Descriptions

Definite descriptions, such as *that decision*, can refer to propositions, making them a type of broad referring expression.

(22) The Supreme Court, however, still may decide *whether to take up Microsoft’s appeal*. **That decision** is expected as early as October.

[6] This use of trimming was inspired by our work on VP ellipsis resolution, as reported in McShane, Nirenburg and Babkin (2015).

Table 3. The antecedent for “That proposal” can occur in many lexico-syntactic configurations.

Antecedent types for *that proposal*	**Sample preceding contexts, with antecedents italicized**	***That proposal* requires resolution**
[proposed$_{ADJ}$ N]$_{NP}$	*The proposed plan* was announced last night.	*That proposal* is good.
proposal/suggestion$_{NP}$ to V$_{INF}$	*The proposal/suggestion to build a park* came up last night.	
proposal/suggestion$_{NP}$ CLAUSE	*The proposal that everybody should help* came up last night.	
propose$_{V}$ to V$_{INF}$	The committee is proposing *to build a park.*	
propose/suggest$_{V}$ NP	They proposed/suggested *the initiation of the program*.	

To avoid the unwieldy term “broad referring expressions realized as definite descriptions”, we will refer to these as this/that-NPs. Our hypothesis was that, given an instance of a this/that-NP (*this decision*), if the immediately preceding context contained a word with the same or a synonymous stem (*decision, decide, decided, settled upon*), then the latter could be used to identify the antecedent. We explored this hypothesis using a test suite of deverbal nouns: *admission, advice, argument, belief, decision, finding, increase, plan, proposal, request*, *requirement, research,* and *suggestion*. We searched for examples in which they were used in configurations like the ones shown in Table 3.

All of these configurations let the system identify the portion of text that contains the antecedent. But whereas the first three allow for a simple NP-to-NP coreference link, the last two require additional analysis. For these, the actual antecedent is not the whole verbal structure but, rather, the filler of the THEME slot of a PROPOSE event. This can be readily seen using the semantic interpretation of the sentence *The committee is proposing to build a park*, presented using the metalanguage of Ontological Semantics (Nirenburg & Raskin, 2004).

```
PROPOSE-1
    AGENT          SET-1
    THEME          BUILD-1
SET-1
    MEMBER-TYPE    HUMAN
    CARDINALITY    > 1
    AGENT-OF       PROPOSE-1
BUILD-1
    THEME          PARK-1
    THEME-OF       PROPOSE-1
```

In this meaning representation, the elements in small caps are ontological concepts, whose instances are indicated by numerical suffixes. Each frame contains a head and a set of properties, including their inverses. Returning to our example – *The committee is proposing to build a park.*

That proposal is good – the expression *that proposal* refers to the BUILD event that is the THEME-OF the PROPOSE event from the preceding context. In a semantically-oriented system, this coreference could be precisely established by coreferring *that proposal* with the semantic frame in boldface. The issue is how to approximate this coreference in a nonsemantic, string-level system like the one we are reporting. We decided on the strategy shown in (23), in which CROSS inserts the event concept in question (DECISION), as well as the case role in question (THEME), and uses brackets to show the text string whose meaning serves as the THEME.[7]

(23) The Supreme Court, however, still may decide [DECISION: THEME whether to take up Microsoft's appeal]. **That decision** is expected as early as October.

In the case of nominal antecedents, we counted as correct either the selection of the full nominal or the EVENT: THEME strategy, as shown by the different bracketed structures in (24).

(24) The meeting with ministers from Poland, Hungary, Bulgaria, the Czech Republic, Slovakia, Romania, Lithuania, Estonia and Latvia focused on [a draft EU plan [PLAN THEME: to welcome the easterners into the Union 's border-free single market]]. **That plan** is due to be approved …

Since the EVENT: THEME strategy was not used by annotators, the authors manually judged agreement between the annotators' selections and system output.

We evaluated 34 contexts. Two were problematic, apparently requiring a plural antecedent (e.g., *the 10-micron proposals*) for a singular referring expression (*the proposal*). We excluded those from the evaluation. Of the remaining 32 examples, 24 (75%) were completely correct. As shown by (25), the system *did* resolve referring expressions that included restrictive postmodification: *that decision* <u>*to refuse to back U.S.-led war in Iraq*</u> was coreferred with the previous NP *Canada's decision not to send troops to Iraq.*

(25) On a good will trip to Capitol Hill on Thursday, Martin introduced himself to top lawmakers with whom he discussed issues ranging from prescription drug importation to Canada's decision [DECISION THEME: not to send troops to Iraq]. **That decision** to refuse to back the U.S.-led war in Iraq put strain on relations between Ottawa and Washington last year.

Although this coreference might seem redundant, consider how difficult it would be to automatically recognize that *not sending troops to Iraq* is being presented as equivalent to *refusing to back the U.S.-led war in Iraq*.

We gave partial credit for four (12.5%) of the 32 responses. In these cases, the system selected the correct head but included more (26) or less (27) context than the annotators.

(26) The United States put forward [*a very practical proposal* at the last round of talks]. We want to see results to moving forward on **that proposal.**

(27) Ratner's group already filed *[a formal request in Germany] to try to bring an investigation against Rumsfeld and other current and former Bush officials for either ordering aiding or failing to prevent the torture*. German federal prosecutors rejected **that request** in April…

Four examples (12.5%) were treated incorrectly. One involved a parsing error: the string *that proposals* was not a NP; it was a complementizer followed by a plural noun. In another example,

[7] These EVENT: THEME pairs were manually listed for each of the verbs included in the experiment.

the antecedent was not available in the context. In two more examples, the system overlooked the nearest (correct) antecedent and instead selected a more distant antecedent of a fitting semantic type. Overall, though, this strategy clearly has the potential to be useful when expanded to cover a larger inventory of nouns derived from or semantically associated with verbs.

5. Resolving Difficult Personal Pronouns

Although many NLP systems treat personal pronouns, they have not achieved very high precision for third-person personal pronouns due to the need for contextual and real-world knowledge and reasoning. We have found it useful to apply the same configuration-based, confidence-oriented methodology to these pronouns as we applied to broad referring expressions. As with the latter, our goal is to achieve high precision for an automatically selected subset of instances. Instances not covered by our methods can be treated by engines that offer better coverage. As shown in Figure 1, OntoSem2 currently uses the coreference resolver in Stanford's CoreNLP toolkit (Lee et al., 2013; Manning et al., 2014) for this purpose.

In order to highlight the added value of CROSS, we chose to evaluate only third-person, non-reflexive pronouns: *he, him, she, her, they,* and *them.* We excluded first- and second-person pronouns (*I, me, you, we* and *us*), as well as the reflexive pronouns (e.g., *himself*), because they are too easy and their inclusion would have artificially inflated the system's score. We excluded *it* for the opposite reason: it cannot be confidently treated without prior detection of pleonastic and idiomatic instances. The configurations we tested are described informally in Table 4 and are illustrated by examples (28) to (33), in turn.

(28) The Warwickshire all-rounder Roger Twose has been named in the New Zealand squad to tour India beginning in October. Now *he* has taken the decision to make his life in New Zealand and **he** goes with our blessing and best wishes.

(29) Established Zulu actors were used in the dubbing process, which took more than a month, *he* said. **He** said senior politicians, among them PWV provincial premier Tokyo Sexwale, and leading movie personalities had been invited to the gala…

(30) Rabin also accused Iran of controlling the Islamic fundamentalist group Hezbollah, which has been blamed for several terrorist attacks. But *he* said **he** believed the weapons flow through Syria had slowed in recent months.

(31) President Bill *Clinton* warned Saturday that **he** would veto any attempt by Republicans to scrap plans to put 100,000 additional police on US streets in line with his prized crime-fighting package.

(32) The *survivors* of the family live under one roof. **They** live frugally on rice and beans distributed by the church.

(33) In addition, some 170 US soldiers will go to Saudi Arabia to take two Patriot missile *batteries* out of storage and transfer **them** to Kuwait, the Pentagon said.

We tasked the system only with identifying the nominal head of the antecedent, not the entire noun phrase, which might include a determiner, adjectives or relative clauses. Enhancing the resolution to full NP selection should be straightforward.

Table 4. Configurations for resolving the personal pronouns *he, him, she, her, they,* and *them*. Feature matching means having the same value for person, number and gender.[8] In all cases, the pronoun and its antecedent are at the same level of quotation. "Sequential" implies that there are no other categories of the given type intervening. Examples (30) to (35) illustrate the configurations. Evaluation outcomes are correct, partial credit, and incorrect, presented as percentages. The last column shows the number of examples evaluated.[9]

	Configuration Description	Example	Correct	Partial	Incorrect	# Exs.
1	Sequential, string-matching subjects of coordinated clauses	(30)	100	0	0	20
2	Sequential feature-matching subjects of speech-act verbs[10]	(31)	90	0	10	20
3	Sequential string-matching subjects in a main clause + subordinate structure[11]	(32)	100	0	0	32
4	Sequential feature-matching subjects in a main clause + subordinate structure	(33)	85	0	15	20
5	Sequential feature-matching subjects of identical verbs	(34)	90	5	5	20
6	Direct objects of sequential coordinate clauses	(35)	85	0	15	20

Most mistakes reflected the need for real-world reasoning, such as (36), which matched – but was incorrectly treated by – Configuration 4, "Sequential feature-matching subjects in a main clause + subordinate structure."

(36) [Tears] of joy and grief poured from the *two teams* as **they** lined up for the medal ceremony.

One example, (37), was genuinely ambiguous: some annotators thought the antecedent was *they,* whereas others thought that it was *delegations*. CROSS selected *they* and was marked correct.

(37) The US search for evidence of the al-Qaeda terror network in Somalia has come up with nothing, the president of Djibouti, Ismael Omar Guelleh, said on Sunday. *They* have sent delegations and **they** have looked at the whole coast.

One partial-credit case involved quantifier interpretation in the noun phrase *1,500 kilos of high-quality explosives*. Rather than select the logical head, *explosives*, CROSS selected *kilos.* When

[8] "Gender" here refers to the English distinction between *he* (masculine animal), *she* (feminine animal), and *it* (non-human animal or inanimate).

[9] Our initial corpus sampling resulted in a sufficient number of hits for only one of our configurations: *sequential string-matching subjects in a main + subordinate structure*. Therefore, we ran a larger corpus to increase hits for the rest. For this second run, we forewent formal annotation of the extracted examples and, instead, two graduate students simply checked CROSS's results.

[10] We used 17 speech-act verbs: *say, admit, declare, explain, mention, express, clarify, state, announce, remark, note, add, reply, respond, repeat, explain,* and *confirm*.

[11] Formally, they are both subjects of a ccomp or advcl dependency. Definitions of the dependencies can be found at http://nlp.stanford.edu/software/dependencies_manual.pdf in the Stanford CoreNLP dependencies manual.

we enhance CROSS to select the full NP antecedent, we expect problems like these to resolve. The next section discusses how this module can be used in conjunction with a broader-coverage reference resolution engine such as the one available in the CoreNLP tool set.

6. CROSS within the Bigger Picture

Let us reiterate some noteworthy features of CROSS. It treats broad referring expressions, which are not treated by most systems. The system bypasses the complexity and expense of manually annotating all instances of a given type of referring expression by focusing efforts on the automatically selected subset of instances that it knows how to treat. CROSS works in fully automatic mode, requiring no manual corpus massaging; this means that the reported precision should be achievable for any new corpus of the same genre (although we would not expect our configurations to perform as well on highly elliptical or informal texts). The resolution strategies are psychologically motivated and reflect the hypothesis that some aspects of reference resolution can be captured without relying on deep semantic and pragmatic reasoning. And the resolution strategies are inspectable, which not only permits developers to introduce iterative improvements, but also permits an intelligent agent to explain its language-processing decisions to its human collaborators.

CROSS is intended to be used as a high-confidence supplement to a broader-coverage coreference engine. As shown in Figure 1, OntoSem2 incorporates the reference resolver of the CoreNLP toolset. Both of these work presemantically and resolve a subset of referring expressions. Since they both treat the personal pronouns *he, him, she, her, they,* and *them*, we decided to compare their resolution accuracy for the inventory of examples included in the evaluation reported in Section 5.[12] Let us linger for a moment on the outcome and implications of this comparison. Whereas both systems did pretty well for configurations 1 to 4, CROSS substantially outperformed CoreNLP for configurations 5 and 6, scoring about 30 percent higher for the first and 60 percent higher for the second. This means that CROSS's results for these configurations should certainly be preferred. However, it raises the question of how accurately CoreNLP resolves third person personal pronouns overall, and how to choose between resolutions in cases of disagreement.

Lee at al. (2013) report an extensive evaluation of the CoreNLP reference resolver but it does not provide exactly what we need: the precision of third person pronoun coreference treated in isolation. We can, however, roughly estimate that precision from the statistics for overall system precision presented in their Table 8. To understand the numbers, one must understand the system architecture. The CoreNLP reference resolver is designed as ten sieves called in order of decreasing accuracy. The sieve that handles third-person pronouns is the last, least accurate, one.

[12] This comparison was coarse grained because of the complexities in assigning partial credit, and because the systems generate different outputs. Whereas CoreNLP selects full coreference chains, CROSS only seeks coreference pairs, and whereas CoreNLP selects full antecedents, CROSS currently only selects the heads of antecedents. Therefore, this comparison provides only for ballpark generalizations.

Table 5. Summary of evaluation results for all modules of CROSS. The rightmost columns show the percentage of resolutions that were correct, that received partial credit, and that were incorrect.

Expression Type	Configuration Type	Correct	Partial Credit	Incorrect
Pronominal *this, that, it*	Lexico-syntactic configurations	92.6	3.7	3.7
	Case-role constraints	79.4	5.9	14.7
	Simple contexts	83.3	most credit: 8.3 some credit: 8.3	0
this/that-NPs	Lexico-syntactic configurations	75	12.5	12.5
Difficult personal pronouns: *he, she, they, him, her, them*	Sequential subjects, coord.	100	0	0
	Sequential subjects, speech acts	90	0	10
	Sequential string-match., subord.	100	0	0
	Sequential non-string-match., subord.	85	0	15
	Sequential subjects, identical verbs	90	5	5
	Sequential coord., direct objects	85	0	15

Lee et al.'s Table 8 shows that the system's precision decreased with the addition of each sieve, with the absolute scores for precision overall differing across evaluation metrics: MUC: 60.9, B^3: 73.3, and BLANC: 79.3. Clearly, CROSS's precision is higher for the pronouns it treats, so its coreference links should be preferred over those of CoreNLP when there are discrepancies. CoreNLP developers are aware that third-person pronouns are their weak link, having traced 28.7 percent of overall system errors to pronoun handling. They write, "Implementing a richer model of pronominal anaphora using syntactic and discourse information is an important next step" (Lee et al., 2013).

Saying that the highest-scoring coreference vote wins has different implications for different settings. For knowledge-lean applications, the winning resolution should be used as the answer. By contrast, for a knowledge-rich agent system, all presemantic reference votes should be tentative until semantic analysis can verify or overturn them – a process carried out in the module called "Semantic and pragmatic analysis of whole input" in Figure 1. For example, both CoreNLP and CROSS will fail to correctly resolve *he* in the example *My father talked to the surgeon and then he operated*, but the selectional constraints on SURGERY recorded in the ontology will permit the OntoSem2 semantic analyzer to override the surface reference votes on semantic grounds. Furthermore, many instances of referring expressions – such as referential verbs and pronominal broad referring expressions not caught by CROSS's configurations – will not be attempted presemantically and are addressed through combined semantic and reference analysis.

Table 5 summarizes the evaluation results for all of the modules of CROSS. Across categories (not individual examples), average performance was 88 percent correct, 7.6 incorrect, and 4.4 percent partially correct. The fact that CROSS can resolve some examples of difficult referring expressions with nearly full confidence ("nearly" because evaluation suites can fail to cover the full range of natural language phenomena) validates the utility of this knowledge-based approach. However, although we are quite satisfied with these numbers, we should point out what they do *not* show. CROSS has low recall, covering a small percentage of instances of each type of referring expression it treats. We do not know exactly how low because determining that would require a prohibitively expensive corpus annotation effort. The expense derives from the need to invent far more sophisticated corpus annotation guidelines and methods than are typically in use, hire well-trained annotators, permit those annotators to work slowly and carefully, and invent both theoretical and practical methods for dealing with ambiguity and vagueness in the use of referring expressions. However, even though recall is currently low, this configuration-based methodology is extensible – in fact, we have barely scratched the surface of its potential utility. What remains is more knowledge engineering. Although this has been devalued over the past two decades in the excitement over statistical methods, we believe it must return if we hope to develop human-level cognitive systems.

7. Comparisons with Others

Work on reference has long been marked by a divide between conceptual contributions and system building. Descriptive and theoretical linguists have posited analyses that can be quite impressive and satisfying for human consumers but remain difficult to implement because necessary prerequisites cannot be automatically fulfilled. For example, Webber's (1988) theory of discourse deixis requires discourse structure to be known, but computing it automatically and with high confidence remains beyond the state of the art. The Centering Theory approach to pronominal coreference (Grosz et al., 1995) captures many intuitions about the distribution of referring expressions, but is difficult to operationalize due to problems in defining key concepts (see Poesio et al., 2004 for discussion). And the reference challenges presented in McShane (2009) arguably must be addressed in comprehensive, long-term programs of work, such as that of Ontological Semantics.

Developers of coreference resolution systems have responded in three ways to the prerequisite problem. The first is to constrain the domain and explicitly supply all prerequisites as Byron (2004) did in treating broad referring expressions. This approach has two-pronged utility: on the one hand, it advances our understanding of how to operationalize the treatment of difficult linguistic phenomena, and on the other hand, it can support applications in the given domain or other domains similarly modeled. A second response to the prerequisite problem is to manually provide prerequisites via corpus annotation. For example, as noted earlier, past reference resolution competitions (see, e.g., Hirschman & Chinchor, 1997) have provided competitors with annotated training and evaluation corpora: the annotations indicate which referring expressions must be resolved (the "markables") and provide manually vetted syntactic features. The benefit of this approach is that it optimizes the head-to-head comparison of supervised learning methods (for reviews of learning methods brought to bear for reference resolution, see Zheng et al., 2011, and Lee et al., 2013). The shortcoming is that systems thus trained have little utility in real-world applications, whose text inputs will normally not be annotated (for insightful discussions, see Mitkov, 2001, and Stoyanov et al., 2009). A third response to the prerequisite problem is the one taken here: requiring systems to work in open domains and answer for all prerequisites, but not requiring them to treat every instance of every linguistic phenomenon. In this, CROSS shows a strategic similarity to Baldwin's CogNIAC system (1997), which treats only high-confidence instances of referring expressions. However, these systems' actual rule sets are quite different since CROSS, unlike CogNIAC, does not address phenomena that are adequately handled by other available technologies: e.g., the CoreNLP reference resolver (Lee et al., 2013) does quite well on reflexive pronouns, first and second person pronouns, and referring expressions for which there is only one feature-matching candidate in the window of coreference.

8. Final Thoughts

As shown in Figure 1, CROSS is a module of OntoSem2, which is a new implementation of the theory of Ontological Semantics. OntoSem2 differs from its predecessor, OntoSem (McShane et al., 2016), by analyzing input incrementally rather than as full sentences. This will ultimately permit human-like behaviors, such as beginning to act before an utterance is completed. The new system continues the OntoSem tradition of integrating language understanding into a larger-scale cognitive system. For example, agents must decide how much effort to devote to understanding each input and how to behave in each situation, based on factors like their confidence in language understanding, their interpretation of their own and others' plans and goals, the risk of making a mistake, and the reversibility of actions (McShane & Nirenburg, 2015). In this task-oriented context, the utility of CROSS's selective approach to treating referring expressions should be clear. The system will attempt to resolve all referring expressions as part of the overall semantic and pragmatic interpretation of the input. If it succeeds with high confidence, it will use its newly gained knowledge to support reasoning about action. If not, it will choose among its other options, such as deferring decision making or asking a human collaborator for clarification.

Although the coreference configurations presented here are domain independent, it can also be useful to formulate more narrowly specified configurations for specific domains or applications, thereby guaranteeing the correct interpretation of frequent or critical inputs. Such configurations have the same cognitive status, and serve the same agent-building function, as multi-word expressions in the lexicon: in both cases, a ready-made answer can be selected without compositional analysis. Consider an example from Maryland Virtual Patient, a clinician training prototype system developed within the OntoAgent environment (Nirenburg et al., 2008). In this application, an intelligent agent plays the role of a virtual patient that is diagnosed and treated by a human clinician in training. Ideally, the agent will fully understand all of the human's text inputs, but, at a minimum, it must understand those aspects that directly contribute to its decision making and action, such as requests for information and action.

For example, consider the input *You need to take it twice daily with food.* In a medical context, human readers would immediately understand that *it* refers to a medication and *take* means ingest. However, this analysis is not so easy for an intelligent agent whose computational lexicon includes dozens of productive and phrasal meanings of *take,* and whose coreference resolution system can identify many candidate antecedents for a broad referring expressions like *it.* To prepare the agent to correctly interpret this frequent-for-the-application input, we can create the configuration *you* [*should, must, need to,* etc.] *take it* [*daily, twice a day,* etc.] [*with food, without food, on an empty stomach*], and we can specify that *it* corefers with the most recently mentioned medication. Although this configuration has very narrow coverage if judged against all of the uses of *it* in a typical corpus, it has very high precision for this application and can be recorded in a methodologically-motivated way, no differently from the broader-coverage configurations evaluated above. Moreover, in terms of cognitive modeling, it seems entirely justified to capture such quasi-idiomatic locutions since having to take medicine on a fixed schedule presumably holds a privileged status in a person's mental model of the medical domain.

We will end on a methodological note. The experience of corpus annotation for this project confirmed our longstanding belief that not all natural language processing systems should be subject to the "annotate then automatically evaluate" methodology that has become all but required due to influence from the statistical learning paradigm. Our system evaluation could have been equally well served – and we could have saved dozens of person hours – by having people simply check whether CROSS's answers were correct. It is also possible that broad referring expressions would have been addressed by the language-processing community sooner, and by more developers, if not for the field-wide habit of waiting for the appearance of annotated corpora to render linguistic phenomena actionable. We believe that the goals of artificial intelligence and cognitive systems will be better served by inventing new strategies to fulfill needs, rather than exclusively finding outlets for available strategies.

Acknowledgements

This research was supported in part by Grants N00014-09-1-1029 and N00014-16-1-2118 from the U.S. Office of Naval Research. Any opinions or findings expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

References

Baldwin, B. (1997). CogNIAC: High precision coreference with limited knowledge and linguistic resources. In R. Mitkov & B. Boguraev (Eds.), *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts* (pp. 38–45). Stroudsburg, PA: Association for Computational Linguistics.

Byron, D. (2004). *Resolving pronominal reference to abstract entities*. Doctoral dissertation, Computer Science Department, University of Rochester, Rochester, NY.

Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Fillmore, C., Kay, P., & O'Connor, C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, *64,* 501–38.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7,* 219–224.

Graff, D., & Cieri, C. (2003). English Gigaword LDC2003T05. Web Download at https://catalog.ldc.upenn.edu/LDC2003T05. Philadelphia: Linguistic Data Consortium.

Hirschman, L., & Chinchor, N. (1997). MUC-7 coreference task definition. Version 3. Available at http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M. & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, *39,* 885–916.

Levesque, H., Davis, E., & Morgenstern, L. (2012). The Winograd Schema Challenge. *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning* (pp. 552–561). Palo Alto, CA: AAAI Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the Fifty-Second Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Stroudsburg, PA: Association for Computational Linguistics.

McShane, M. (2009). Reference resolution challenges for an intelligent agent: The need for knowledge. *IEEE Intelligent Systems*, *24*, 47-58.

McShane, M. (2015). Expectation-driven treatment of difficult referring expressions. *Proceedings of the Third Annual Conference on Advances in Cognitive Systems*. Atlanta, GA.

McShane, M., & Babkin, P. (2015). Automatic ellipsis resolution: Recovering covert information from text (pp. 572–578). *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. Palo Alto: AAAI Press.

McShane, M., & Nirenburg, S. (2012). A knowledge representation language for natural language processing, simulation and reasoning. *International Journal of Semantic Computing*, *6*, 3-23.

McShane, M., & Nirenburg, S. (2015). The interplay of language processing, reasoning and decision-making in cognitive computing. *Proceedings of the Twentieth International Conference on Applications of Natural Language to Information Systems*. Passau, Germany.

McShane, M., Nirenburg, S., & Beale, S. (2016). Language understanding with Ontological Semantics. *Advances in Cognitive Systems*, *4*, 35–55.

McShane, M., Nirenburg, S., & Babkin, P. (2015). Sentence trimming in service of verb phrase ellipsis resolution. *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science*. Torino, Italy.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41.

Mitkov, R. (2001). Outstanding issues in anaphora resolution. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing*, 110-125. Berlin: Springer-Verlag.

Nirenburg, S., McShane, M., & Beale, S. (2008). A simulated physiological/cognitive "double agent". *Papers from the AAAI Fall Symposium: Naturally Inspired Cognitive Architectures*. (AAAI Technical Report FS-08-06). Menlo Park, CA: AAAI Press.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. Cambridge, MA: MIT Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122*, 280–291.

Stoyanov, V., Gilbert, N., Cardie, C., & Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. *Proceedings of the Joint Conference of the Forty-Seventh Annual Meeting of the ACL and the Fourth International Joint Conference on Natural Language Processing of the AFNLP.* Stroudsburg, PA: Association for Computational Linguistics.

Webber, B. L. (1988). Discourse deixis: Reference to discourse segments. *Proceedings of the Twenty-Sixth Annual Meeting of the Association for Computational Linguistics* (pp. 113–122). Stroudsburg, PA: The Association for Computational Linguistics.

Zheng, J., Chapman, W. W., Crowley, R. S., & Savova, G. K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, *44*, 1113–1122.