# Expectation-Driven Treatment of Difficult Referring Expressions

**Marjorie McShane**                                                        MCSHAM2@RPI.EDU

Cognitive Science Department, Rensselaer Polytechnic Institute, Troy, NY, 12180

## Abstract

Like many difficult linguistic phenomena, so-called "broad referring expressions" (BREs) – such as pronominal *this, that* and *it* – have been excluded from the purview of most natural language processing systems, being tacitly deemed unmanageably difficult. However, when building cognitively-inspired intelligent agents that are meant to have real-world utility in human-agent teams, the wholesale exclusion of difficult phenomena is neither practical nor necessary. I suggest the following strategy for incorporating the treatment of difficult language phenomena into an agent's repertoire over time. Agents are configured to automatically determine which instances of a linguistic phenomenon they can and cannot confidently treat. For the high-confidence cases, the agents carry out the language understanding (i.e., "perception") and move on to decision making and action; for the low-confidence cases, they seek clarification from their human collaborators. This paper details some strategies for resolving BREs that appear to offer high confidence solutions within the current state of the art. The analysis of BREs is distributed across language processing modules in a way inspired by principles of cognitive modeling. The data analysis and modeling strategy show that a natural language processing problem that seems impenetrable when viewed from the current mainstream perspective of supervised machine learning becomes more manageable when modeled according to human-like reasoning.

## 1. Introduction

Cognitive modeling in support of configuring sophisticated intelligent agents must include deep language understanding. In the OntoAgent research group, we define deep language understanding as the automatic generation of unambiguous, context-sensitive, ontologically grounded text meaning representations, formulated according to the Theory of Ontological Semantics (Nirenburg and Raskin 2004). Text meaning representations (TMRs) are the "perception" milestone in the perception, reasoning, action sequence of OntoAgents. For example, the TMR for sentence (1) should be as follows, where small caps indicate concepts, numerical suffixes indicate instances, and the value of evaluative modality is expressed on the abstract scale {0,1}. Coreferential categories in the example are shown by boldface, for the referring expression in question, and underlining, for its sponsor – typically, an antecedent.

(1) As for <u>math</u>, he likes **it**.

> MODALITY-1
>> TYPE            EVALUATIVE
>> VALUE           .8
>> SCOPE           MATHEMATICS-1
>> ATTRIBUTED-TO   HUMAN-1 (GENDER male)

Consider just a few details about this knowledge structure that are particularly relevant to the upcoming discussion. Sentence-initial *as for* must be recognized as a multi-word expression. The lexical sense for this expression must (a) expect the syntactic structure *As for NP, CLAUSE*; (b) include a function that establishes the coreference (if applicable) between NP and a pronominal element in CLAUSE; and (c) reanalyze the pronominal element in CLAUSE using the meaning of the NP.[1] Essentially, the input *As for math, he likes it* should be treated as a functional paraphrase of *He likes math.* Although metadata is not shown in this view of the TMR, it would include which word of input instantiates which concept and which lexical sense of each input word was selected – features that are useful for system testing and debugging.

Although we have not described *how* TMRs are generated (for that, see McShane et al., 2005, and McShane et al., In press-b), we hope that this sample TMR serves to illustrate *why* they are generated: they represent meaning using an unambiguous ontological metalanguage that (a) resolves many of the surface complexities of natural language and (b) serves as excellent input to agent reasoning. In fact, TMRs are exactly the type of knowledge structure that has been sought by the reasoning community since its inception. It is, therefore, unfortunate that for over 25 years now the natural language processing community has not sufficiently emphasized the task of extracting and representing text meaning (see Nirenburg and McShane, Forthcoming, for a historical perspective).

When working toward automating this type of natural language analysis, one recognizes that natural language offers both islands of relative simplicity and hurdles of complexity. The broad referring expressions (BREs) we focus on here represent a microcosm of that spectrum. On the simpler end are examples like (2), whose BRE (*that*) refers to the preceding proposition.[2] On the more difficult end of the spectrum are examples like (3), in which the BRE (*this*) refers to a contextual situation that cannot be identified by pointing to a span of text.

(2)  They don't trust us. **That**'s good.

(3)  {*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.

In fact, BREs are called "broad" because they can refer to practically anything: a noun phrase (NP), a full clause/proposition, a clause/proposition stripped of some features (like modality), multiple clauses/propositions, something specific that is not mentioned in the text, or something quite vague altogether. The question pursued here is: *Can we teach machines to both detect and resolve instances of BREs that are on the simpler end of the spectrum*, before turning our attention to issues of vagueness and underspecification, which theoretically and practically extend far beyond reference resolution as traditionally defined?

When talking about processing BREs, we use the term BRE to refer to any instance of *this, that* or *it* since the agent cannot know beforehand to what kind of textual or extra-textual entity a given instance of *this, that* or *it* refers.

Two of the core agent-oriented hypotheses of the research are as follows.

---

[1] Although we have not yet implemented this particular reference-oriented phrasal, we have worked extensively on phrasals in principle, as reported in McShane et al. (In press-a).

[2] Examples throughout are drawn primarily from the COCA corpus (Davies 2008-) and the Gigaword Corpus (Graff and Cieri 2003).

*Hypothesis 1*. The principles of least effort (Piantadosi et al. 2012; Zipf 1949) and reasoning by analogy (Gentner and Smith 2013) can inspire the modeling of intelligent agents. That is, the agent need not necessarily engage in maximally deep semantic and pragmatic analysis if more shallow analysis is sufficient to make some determination. The fact that some language analysis decisions appear to be "automatic" – i.e., not requiring deep analysis – is illustrated by (4):

(4) "I just lost $100 in the lottery!" "**That**'s great!"

Even though this context, in isolation, makes little sense, the reader has no choice but to resolve *that* with the meaning of *I just lost $100 in the lottery*, then try to figure out why someone would say this – perhaps the interlocutor is sadistic or is leading up to the Pollyannaish rejoinder, "*You could have lost $10,000!*" We as readers presumably recognize the dialog pattern [$_{Turn1}$ simple proposition] [$_{Turn2}$ *That* + copula + adjective] and automatically decide – regardless of the semantic content and pragmatic context – that *that* refers to the whole of Turn 1. For agent building, we can operationalize such expectations through the use of lexically recorded phrases and patterns (cf. recent work on construction grammar, e.g., Goldberg 2003), rather than rely on full compositional analysis to resolve all referential ambiguities.

*Hypothesis 2*. Domain-independent linguistic generalizations can help to solve difficult language analysis problems in advance of the time when agents have access to full ontological knowledge in all domains. Suggesting that agents can get by with less knowledge does not eliminate the need for developing rich knowledge bases for agents, particularly since we know that this "less knowledge-rich" (to avoid the methodologically-charged term "knowledge-lean") approach will not work in all contexts. This hypothesis essentially states that there exists a subset of automatically identifiable contexts in which full knowledge of the world is not needed to resolve the reference of BREs, and these contexts can be treated by agents in the near-term, while we are working toward developing the knowledge bases and reasoners for processing the more difficult cases.

As regards the trajectory of agent development over time, the key to configuring agents that inspire confidence in their human collaborators is making those agents capable of judging their own confidence in language analysis, reasoning and decision-making. In the near term, this will result in agents that frequently need to ask for clarification and assistance; but in the longer term, it will lead to agents that are able to perform adequately without overburdening their human collaborators.

## 1.1 Related Work

Past work contributes to the current research primarily by demonstrating that BREs remain insufficiently treated in computer systems. Most reference resolution systems (see Lee et al. 2013 for an overview of the mainstream state of the art) do not offer high-quality, if any, treatment of BREs, apart from treating (hand selected) instances of referential *it* that corefer with a NP. A knowledge-rich system was developed by Byron (2004) but covered a narrow domain. Approaches in the field of theoretical computational linguistics typically require preconditions that cannot be automatically fulfilled: e.g., Webber's (1988) is a theory of discourse deixis that requires discourse structure to be known, despite the fact that computing it automatically remains beyond the state of the art.

The main positive external contributors to this treatment of BREs have been individual linguistic observations about their usage. For example, Passonneau (1989) makes corpus-based

generalizations about the distribution of *it* versus *that*; Channon (1980) observes that *that* is often used when a set of objects has conflicting semantic features (e.g., "*I'll have a burger and a lemonade." "I'll have that too."*); and Webber (1988) notes that the local context for a demonstrative pronoun can help to disambiguate its meaning, as in *that's where, that's when,* etc.[3] These findings are in the spirit of the corpus-based, linguistic observations that represent a core contribution of the work reported here.

## 1.2 Methodology

A key methodological insight of my approach is that not all BREs are treated within a dedicated reference resolution module.[4] Instead, each BRE is treated however and whenever is most appropriate within the agent's end-to-end text analysis process. The stages of BRE treatment are shown in Row 2 of Table 1. Those in small caps will be discussed in this paper.

*Table 1.* Distributing the processing of BREs (row 2) across stages of language analysis (row 1).

| PrePro & Synt. Analysis | Basic Semantic Analysis | Extended Semantic/Pragmatic Analysis | | |
|---|---|---|---|---|
| | | Reference, Stage 1 | Reference, Stage 2 | Reference, Stage 3 |
| | a) DETECT IMPOSTOR BRES<br><br>b) DETECT DISCOURSE-FUNCTION BRES | EXPLOIT LEXICO-SYNTACTIC PATTERNS | EXPLOIT SEMANTIC PATTERNS | Use script-based reasoning[5] |

This distributed treatment of BREs must be understood within the framework of overall OntoAgent language analysis, which is comprised of: (1) preprocessing and syntactic analysis, provided by Stanford CoreNLP (Manning et al. 2014); (2) OntoAgent basic semantic analysis, which involves lexical disambiguation and the establishment of the basic semantic dependency structure (McShane, Nirenburg and Beale, In press-b); and (3) OntoAgent extended semantic/pragmatic analysis, which involves several stages of reference resolution as well as the interpretation of indirect speech acts, the detection of lexical and ontological paraphrase (McShane et al. 2008), etc.

Distributing the processing of BREs across modules not only makes practical sense, it also reflects Hypothesis 1: that agents can treat set phrases and frequent patterns first, using cognitively cheap pattern matching, and then resort to more expensive (e.g., script-supported) reasoning only when necessary.

There are five methodological differences between this work and that of current mainstream NLP, which most often utilizes supervised machine learning.

---

[3] Byron, 2004 presents a nice review of relevant past research.

[4] I argued the same for English personal pronouns (McShane and Nirenburg, 2013) and Russian personal pronouns (McShane, 2014). This global approach counters the typical mainstream-NLP decision to isolate linguistic phenomena, treating each by a separate system and assuming that prerequisites needed for one's own system will be provided externally.

[5] By "script", I mean the type of knowledge structure first discussed by Schank and Abelson (1977), and used extensively in the OntoAgent environment to support agent simulation, language analysis and agent reasoning.

1. My approach is knowledge-based, fully inspectable and extensible. It relies on introspective linguistic analysis supported by corpus analysis. Many linguistic generalizations that I have previously shown apply to other referring expressions apply to BREs as well: for example, there are parallelism and repetition effects (McShane, 2000), and many configurations are best treated as lexically recorded patterns (McShane et al., In press-a).
2. The results of the processing described here must be incorporated into TMRs, and the processing itself must be integrated with other language processing modules; this means that this task is not standalone and cannot be fully appreciated in isolation (McShane and Nirenburg 2013).
3. The agent is responsible for detecting which instances of a phenomenon it can treat and which it cannot; there is no manual pre-selection of instances that must be treated (sometimes called "markables"), which is typical for competition-style NLP development efforts (cf. the MUC co-reference resolution task (Hirshman and Chinchor 1997)).
4. The approach uses only the prerequisites the agent itself can supply – most notably, it does not utilize manually annotated corpora.
5. The evaluation metric is based on precision, not a combination of precision and recall. This seems reasonable when building agents (in contrast to standalone NLP programs) because the agent is then prepared to act with confidence in response to at least some inputs; for others, it must look to its human collaborator for clarification or corroboration.

Since this paper reports a pre-implementation study of linguistic phenomena, the evaluation suite involves manual, corpus-based analysis of component hypotheses.[6] Zooming out for a moment, the overall plan of work on reference in OntoAgent is: a) compile an inventory of hypotheses, b) test those hypotheses with humans serving as a proxy for programs, and c) implement the subset of hypotheses that proved useful (not all of the hypotheses we tested for BRE treatment were, in fact, useful). Two aspects of the work must be emphasized. First, the algorithms being tested include only the types of heuristic evidence that can be provided by currently available processors. Second, these BRE-resolution algorithms represent only a minor extension to our ongoing work on semantic analysis, multiword expressions, and reference resolution. This makes us reasonably confident in the success of the forthcoming implementation, despite the fact that implementations tend to involve some surprises.

## 1.3 Content and Organization

The paper considers, in order, the first three methods of treating BREs listed in Table 1. The paper focuses on generalizations and observations that are useful for the configuration of intelligent agents overall, using the example of OntoAgent processing but not constrained to this particular language analysis engine. The generalizations have been validated by manual corpus analysis. The goal of this paper is to show that not all BREs are impenetrably difficult to resolve, and that using agent-oriented strategies – inspired by introspective, psychologically-inspired insights – can offer realistic solutions in the near term that will support further development of automatic language understanding over time.

---

[6] Each evaluation was carried out by one undergraduate student with spot checking by me. These evaluations are sufficient to fulfill our theory-building and system-development needs; they are not claimed to represent cross-validated gold standards.

## 2. BRE Processing During Basic Semantic Analysis

Basic semantic analysis in OntoAgent is defined as lexical disambiguation and the establishment of the basic dependency structure, as shown by the TMR for sentence (1) above. During basic semantic analysis, several aspects of reference resolution should be handled (McShane 2014), including the BRE-containing configurations described below.

### 2.1  Detecting BRE Impostors

BRE impostors are non-referential – i.e., pleonastic or idiomatic – instances of *this, that,* and *it,* which should not be mistaken for referring expressions that require a sponsor. Pleonastic *it* can be treated using a lexically-recorded, pattern-based approach (cf. Boyd et al., 2005; Johnson, 2011). Similarly, idioms that contain non-referential BREs (*blow it, put a cork in it, damn it!*) can be lexically recorded and subsequently treated in the same way as other multiword expressions (McShane et al., In press-a). The main challenge in processing BRE-containing idioms expressions is the same problem posed by all multiword expressions: their potential to be polysemous, having both idiomatic and non-idiomatic senses.

We evaluated 87 BRE-containing multiword expressions using the COCA corpus (Davies 2008-). Of the examples we evaluated, some such phrases were consistently used in their idiomatic meanings, e.g., *word has it, be that as it may, wing it, to whom it may concern, that's that, that's more like it, take it easy, take his word for it,* etc. Some were reliably idiomatic only with certain punctuation: e.g., *far from it* was idiomatic when used either parenthetically (preceded by a comma or period) or emphatically (followed by an exclamation point) but not otherwise: e.g., *Once adult grouse establish their territory, they rarely stray far from it.* Other phrases were usually but not always idiomatic: e.g., *get it together* was idiomatic in 158 of 165 contexts, and *for what it's worth* was idiomatic in 139 of 143 contexts. A minority of phrases occured as readily non-idiomatically as idiomatically: e.g., *come off it* can be used (as a stable metaphor) of drugs or medications, in addition to its idiomatic sense of expressing disbelief. Finally, there were some idiosyncratically difficult cases, such as *Beat it* being used as a song title, and *any way you slice it* being used, as a pun, of meat.

Generalizing from this data, the consistently idiomatic cases represent an island of simplicity for the agent: there is no need to resolve the reference of the BRE, and the multiple words of input can be handled as a phrase, in one fell swoop. The ambiguous (potentially idiomatic or non-idiomatic) phrases, by contrast, task the agent with several aspects of text analysis: multi-word expression processing, lexical disambiguation, and – in the non-idiomatic cases – reference resolution.

### 2.2 BREs in Discourse-Oriented Multiword Expressions

Some referential BREs contribute to larger linguistic structures that are more economically treated as a whole than as the sum of their parts. For example, functionally equivalent expressions with and without BREs can be used for discourse functions like those shown in Table 2. Since OntoAgents need to be able to interpret both BRE-containing and BRE-free versions of the equivalent expressions anyway, and since the BREs in the BRE-containing versions *do not render*

*overall analysis any easier,*[7] we are configuring our agents to recognize these discourse functions using lexicalized phrasals, interpret their ontologically-grounded meanings (SEEK-CLARIFICATION, AGREE-TO, REQUEST-INFO), and avoid an unnecessary – and potentially very complex – reference resolution subtask.

*Table 2*. Discourse functions that can be expressed with and without BREs.

| Discourse Function | BRE-containing | BRE-free |
|---|---|---|
| **Ask for clarification** | "I have one question." "And **that** is?" | "I have one question." "Yes?" |
| **Express agreement** | "I'll be happy to do **that**." | "Sure." |
| **Ask a question** | "The battery is charged, isn't **it**?"[8] | "Is the battery charged?" |
| **Introduce speech** | He said **this**: "I like you." | He said, "I like you." |

Of course, analysis does not end at recognizing these discourse functions, since these functions are realized by concepts whose case-roles must be filled: e.g., REQUEST-INFO requires its THEME role to be filled by the meaning of the proposition being questioned. OntoAgents have language-oriented reasoning capabilities that carry out this analysis (cf. the question-answering skills of virtual patients in the Maryland Virtual Patient application (McShane et al. 2013)). The point here is that the same methods for detecting discourse functions and filling their case roles apply *no matter which linguistic paraphrase is used* in the text, so there is no need for special BRE treatment of the Column 2 locutions.

As regards the last pattern in Table 2, we evaluated both its precision (*Are there false positives?*) and its coverage (*Can anything other than a speech act verb be used?*) using the COCA corpus (Davies 2008-). We compiled an inventory of speech-act verbs and cognitive verbs, searched for this pattern using each of those verbs, and evaluated the first 22 hits (if available) for each verb. The results fall into three groups. **Group 1** verbs had a high number of hits and the hits represented the pattern we were looking for in a high percentage of cases: *say* (21/22), *imagine* (21/22), *hear* (20/22), *think* (18/22), *believe* (17/22), *offer* (17/17) and *ask* (16/22). **Group 2** verbs had a low number of hits (between 1 and 3) but high predictive power (all hits matched the pattern): *admit, mention, agree with, disagree with, repeat, concede, declare*. **Group 3** verbs had fewer than 10 hits and low predictive power – on average, fewer than 50% of examples matched our pattern: *explain, clarify, hint at, describe, confirm, express, maintain, quote, illustrate*.

We then further examined the false positives for Group 1 and discovered two useful generalizations suggesting algorithm enhancement rules. First, the agent should recognize and prune out repetitions (and paraphrases) of the "introductory" verb, which would solve 6 of 19 errors: e.g., *Imagine this: imagine fifty spirited horses in a single team…* Second, for the verb *ask*, the question that is asked can be preceded by an explanatory introduction that intervenes between *this* and the question that serves as its postcedent: e.g., *"And I'm asking this: Judaism and Christianity. How do you balance, for example, say salvation and the Messiah?"* Both of these algorithm enhancements point to rather widespread aspects of human language use: people often

---

[7] Although it might seem that *And that is?* should point to the preceding NP and therefore represent an "easy" reference resolution task, there are actually many counterexamples involving complex NPs and subordinate structures; so our initial attempts to separate out this reference task proved counterproductive.

[8] This is called a tag question and, although it has slightly different semantic connotations than a direct question (the speaker thinks the proposition is true), the difference is inconsequential for most downstream reasoning tasks.

repeat things without intending any special semantic or pragmatic force; and even seemingly "paired" structures like this one ("ask this: *question*") are often not sequential or not paired at all. (How many rhetorical questions remain permanently unanswered?) These complexities of the language system do not invalidate the utility of the pattern-based generalizations proposed here, they simple expand the inventory of eventualities that must be covered by our expectation-based approach to language analysis.

## 3. Exploiting Lexico-Syntactic Expectations

Some lexico-syntactic configurations are highly predictive of reference relations, almost irrespective of the meaning of the text, as illustrated by our lottery example (4). We have compiled evidence (as yet unpublished) that *structural and lexico-semantic parallelism* can support the resolution of pronominal referring expressions, including but not limited to BREs. To take just one example, coreference can quite confidently be established if feature-matching pronouns occur as sequential subjects of coordinate clauses.[9]

(5) "It isn't an item of beauty," Head admits, "but **it** comes out on top of all the things we have run in the wind tunnel."

Here, we introduce 3 types of predictive lexico-syntactic configurations that supplement such structurally-oriented generalizations.

*Configuration 1: A negated proposition introduces its positive counterpart.* Certain types of negated propositions are often followed by their positive counterparts, and these negative-positive pairs narrow the search space for BRE sponsors. In fact, for the patterns illustrated in Table 3, we found few corpus examples that did not show the expected coreference relations.

*Table 3.* Lexico-syntactic patterns predicting BRE resolution.

| Pattern | Example |
|---|---|
| not only AUX (neg) this/it/$NP_{subj}$ … this/$it_{subj}$ | Not only will it be perfect, **it**'ll be never forgotten. |
| this/it/$NP_{subj}$ (AUX) not only … this/$it_{subj}$ | The module not only disables the starter, **it** shuts down the fuel injection… |
| this/it/$NP_{subj}$ has/had nothing to do with … this/$it_{subj}$ | This has nothing to do with politics. **It** has everything to do with strengthening our country. |
| this/it/$NP_{subj}$ is/was not about … this/$it_{subj}$ | Also education is not about money. **It**'s about discipline. |
| this is not (a/an) [n*] … $it_{subj}$ | This is not class warfare. **It**'s math. |

Most of the corpus examples that failed to show the expected BRE-sponsor pair did not include a BRE in the second part to begin with, so there was no reference issue: *Not only will it be perfect, you'll be so happy!*

*Configuration 2.* "Topicalization" strategies, as illustrated by sentence (1), are comprised of a topic followed by a comment. The topic can be introduced by several lexico-syntactic configurations such as, *As for X, As far as X is concerned, Regarding X,* etc. If the comment contains the BRE *it,* it often corefers with the topic as long as the following three conditions hold:

---

[9] See Lee et al. 2013 for evidence of how difficult automatic pronoun resolution has proven to be.

(a) *it* is not pleonastic, (b) the topic is not plural, and (c) the topic does not refer to a human. Of the 53 examples from COCA that we evaluated, 25 showed the expected coreference (e.g., (6) and (7)), 21 used pleonastic *it* (e.g., (8)), and 7 included referential *it* that did not refer to the discourse topic (e.g., (9)) – i.e., they were false positives.

(6) As for <u>plot</u>, **it**'s only there to propel the reader through the story…

(7) As for <u>Guantanamo</u>, closing **it** is complicated…

(8) As for <u>Eden</u>, **it** was discovering that lie that shattered her self image.

(9) Although Martimo can install big diesels on request (1,000 hp Cats), <u>the 56</u> [i.e., the Martimo 56 Cruising Motoryacht] is designed to run most efficiently with smaller engines. As for speed, **it** will cruise at 28 knots with a pair of 800 hp Volvo D 12s …

Ideally, pleonastic *it* will have been previously detected in the processing pipeline, meaning that 78% (25/32) of the examples that should make it to this stage of reference processing show the expected coreference relation. A common source of false positives – semantic ellipsis – is illustrated by (9). Here, *speed* refers to *the speed **of the Martimo 56 Cruising Motoryacht***, not some abstract notion of *speed*; however, detecting and resolving this ellipsis is as difficult as resolving the reference of the BRE, so this subtask cannot be brushed aside as a prerequisite that is expected to be fulfilled externally. Examples like (9) suggest the need for a strategy of cross-checking. That is, reference hypotheses due to "topicalization" should be considered preliminary, overridable by the semantic-feature analysis presented in Section 4. In (9), the narrow selectional constraints of *cruise* (a yacht can cruise; speed cannot) strongly suggest that the subject refers to a boat rather than to *speed*.

*Configuration 3*. Full sentence repetition has the pragmatic function of emphasis: *This is bad. This is bad.* Such repetitions are best treated by adding the feature EMPHASIS to the TMR of the first sentence, rather than individually analyzing each sentence and dutifully establishing cross-sentential co-reference relations but missing the point. This is a good example of why it is counterproductive to isolate individual language analysis processes: What "credit" should a system receive for coreferring these two instances of *this* when the pragmatic force of the 2nd utterance lies elsewhere?

These three configurations are just a sampling of what is likely a much larger inventory of highly predictive lexico-syntactic constructions that can be exploited for the resolution of BREs. Constructing such an inventory is, I believe, a very promising program of work that will provide tangible improvement in the language processing capabilities of agents, resulting in increases both in their verisimilitude and in their utility in applications.

## 4. Exploiting Semantic Expectations

This section describes three of many ways in which domain-independent semantic expectations can be used to help resolve BREs. Each subsection presents a powerful hypothesis which is simple to formulate and understand, but which requires refinement before it can be most usefully incorporated into language analysis by intelligent agents.

## 4.1  Selectional Constraints

If the verb selecting a BRE as its argument imposes tight selectional constraints on the case-role the BRE fills, those constraints should help to confidently identify the reference sponsor. To explore this hypothesis, we carried out a corpus study focusing on the BRE-containing configurations in Table 4.

*Table 4*. Using selectional constraints to guide BRE resolution.

| | Pattern | Constraint | Example |
|---|---|---|---|
| 1 | $keyword_1$ … $It_1$ aux $verb_{PastPart.}$ | $keyword_1$ is a typical object filler for verb | Enron says the <u>deal</u> looks favorable because **it** was *negotiated* in 1992 |
| 2 | $keyword_2$ … $It_2$ $verb_{PastPart.}$ | $keyword_2$ is a typical subject filler for verb | The <u>water</u> was drinkable because it *boiled* for several minutes. |

The first stage of work involved compiling a test list of verbs for which either the subject or the object was narrowly constrained, then compiling a list of typical fillers for that role: e.g., the verb *abolish* often takes the objects *law, bill, agency, department, death penalty, slavery, capital punishment, draft, regulation*, etc.[10] We used a total of 202 verbs with an average of nearly 60 keywords each; however it should be noted that the average was pulled up by verbs like *eat/cook* and *die*, for which hundreds of food items and animals, respectively, were listed as keywords.

We then searched the Gigaword corpus (Graff & Cieri, 2003) for the patterns in Table 4 and manually determined to what extent the simplest possible hypothesis was true: i.e., the sponsor for the BRE was *keyword*. Our analysis resulted in the identification of four necessary enhancements to this baseline rule. For graphic flow of the exposition, we first provide a set of relevant corpus examples, which we then refer to in turn while discussing the four enhancements.

(6)  Residents said they were running out of food in a city that had its electricity cut two days ago. Some wounded Iraqis bled to death, and a family was buried under the ruins of their <u>house</u> after **it** was *bombed* by a U.S. jet, Saadi said.

(7)  "Price-fixing is a <u>crime</u> whether **it** is *committed* in a local grocery store or the halls of a great auction house."

(8)  A <u>holiday</u> honoring Vid, the ancient Slavic god of healing, has become one of the most fateful days on the Serb calendar. Now known as St. Vitus Day, **it** is *celebrated* June 28. In short, BRE resolution absolutely must not be extracted from the rest of processing.

(9)  In the worst atrocity, some 5,000 men, women and children were slaughtered in the <u>border town of Halabja</u> in March 1988 when **it** was *bombed* and shelled with cyanide gas.

(10)  Although Imayev was talking about fighting and encirclements, <u>Kuraly, a village of about 5,000 people</u>, could not have appeared more peaceful. Like many Chechen villages, **it** was *bombed* by Russian airplanes during the fighting that started in December

---

[10] Similar information is available in the OntoAgent ontology and lexicon but, since we wanted broader lexical coverage of specific concepts, we used a different list-building strategy for this experiment.

(11)   The 20-year-old started playing cricket at the <u>Soweto Cricket Club</u> soon after **it** was *built* 10 years ago.

*Enhancement 1*. Syntactic analysis is needed to avoid false positive keyword analyses. For example, if keywords are part of a compound, they must be the final (head) element of that compound. For example, if a context contains "*... the restaurant garage … It was bombed*", then *garage,* not *restaurant* is a candidate sponsor for *it.*

*Enhancement 2*. Candidate sponsors must be ranked according to recency, with the most recent being favored. E.g., although in (6) both a *city* and a *house* can be bombed, the sponsor for *it* is the more proximate *house.*

*Enhancement 3.* Establishing chains of coreference is essential, both for overall processing and for evaluating the results of BRE resolution. E.g., if the system selects *crime* as a sponsor in the (7), this resolution should be considered correct since *crime* is in a coreference chain with the semantically more informative *price-fixing.* It is noteworthy that it is easier to teach a system to recognize generic correspondences like "commit ~ crime" (7) and "celebrate ~ holiday" (8) than more specific ones like "commit ~ price-fixing" or "celebrate ~ St. Vitus Day" (the system's onomasticon might well lack *St. Vitus Day*, and its preprocessor may or may not productively recognize proper nouns headed by *Day* as holidays).

*Enhancement 4*. During preprocessing, several types of complex-entity analysis need to be carried out to support BRE-resolution, such as "TYPE *of* INSTANCE" constructions (9), appositives (10), and proper nouns with meaningful head words (11).

   Outstanding problems involve the usual suspects, such as vagueness (in (12), is *it* the library or the palace?) and indirect referring expressions, such as metonymy (13).

(12) After the meeting, Kinkel and Mubarak inaugurated a public library in a renovated palace overlooking the Nile. **It** was *built* with a German grant of 5.5 million marks (dlrs 3.9 million)

(13) <u>The stolen van Gogh</u>, he said, has special value because **it** was *painted* in the last six weeks of the artist's life.

   To sum up, selectional constraints can provide strong heuristic guidance for resolving some BRE arguments. The knowledge for this can be provided either by word lists (as in the corpus analysis carried out here) or by a lexicon/ontology pair, as used for OntoAgent processing (McShane et al., In press-b). The identified necessary enhancements to the baseline approach of "resolve the BRE to the preceding keyword" should be well within the capabilities of many language processing environments, including OntoAgent.

## 4.2  Predicate Nominal Meaning

When a BRE is used as the subject of a predicate nominal construction – *BRE copula NP* – it would seem that the meaning of the NP should reliably guide the search for the BRE's sponsor. In some cases, this works well: in (14) *that* refers to a *year,* and 1971 is a year; and in (15) *that* refers to a *place*, and *the prison* is a *place.*

(14) Back to <u>1971</u> for a moment. **That** was the year Texas Stadium opened …

(15) <u>The prison</u> became for me the symbol of Soviet system. **That** was the place where …

However, the actual scope of eventualities for BREs reference in predicate nominal constructions is quite broad. When studying an inventory of automatically extracted examples, we found it useful to classify the eventualities into five categories.

*Type 1*. The particular phrase almost never has a textual antecedent and is best treated as idiomatic: e.g., *This is [personal name]* (*This is Marvin.*); *This is [name of broadcast]* ("This is Showbiz Tonight on HLN news and views.").

*Type 2*. The NP is vague and therefore does not usefully constrain the search for its sponsor. This was, unfortunately, the most common outcome in our search of the pattern *This|that [be] the [N\*]* in the COCA corpus (the most common head nouns returned were *way, problem, thing, reason, difference, case, point, subject, reason, reality, goal, theory, message, conclusion*). In order to best leverage our overall strategy of relying on high-confidence heuristics, we compiled a list of predicate nominals that *do* have reasonable sponsor-constraining potential and limited our analysis to those configurations. This corpus-attested list includes: *car, church, city, country, day, guy, location, man, person, place, plane, road, school, street, time, town, year, woman*. The evaluation numbers below reflect our analysis of 466 examples in which the predicate nominal was headed by one of these nouns.

*Type 3*. The context preceding the predicate-nominal construction contains exactly one candidate sponsor that either (a) matches the head noun of the NP, (b) is its synonym, or (c) is its hyponym or hypernym. E.g., in (15) *prison* is a hyponym of *place*. 14% of evaluated contexts were of this type.

*Type 4*. The context contains exactly one candidate sponsor whose identification requires the type of processing available from state-of-the-art preprocessors, such as proper name recognition (*I love **America**. It is the place where I was born*) and the identification of numbers and dates (e.g., 1971 in (14) is a year). 41% of evaluated contexts were of this type.

*Type 5*. The context is tricky in some way – not necessarily too difficult to be automatically resolved using our "light linguistic" methods but requiring some type of additional rule and/or introducing some degree of uncertainty. For example, the sponsor might not be grammatically identical to what is expected (*English* vs. *England* in (16)); there might be more than one candidate sponsor (*this country* vs. *Czarist Russia* in (17)); there might be an elliptical or inexact coreference (the sponsor in (18) is actually *when someone is over 70*); the head word might be used non-literally ('road' in (19)); and so on, for the many more types of complexities that natural language exhibits.

(16) But **English** soccer has a reputation it still can't shake off, no matter how hard it tries. This is the country that exported soccer violence back in the 1970s and '80s

(17) "My grandparents came to **this country** crammed into tight ship quarters from Czarist Russia because they believed this was the country where their votes would be counted"…

(18) "Every year it is the same cast of characters, the Czechs, Russians, Finns and Swedes," Hitchcock said. "But it depends on the big-time players and if the big-time players are engaged then that is the country that wins."

(19) Asked whether there's a risk of another Great Depression if Congress doesn't approve a $700 billion bailout package, Palin said, "Unfortunately, **that** is the road that America may find itself on."

In our corpus evaluation, we defined "tricky" quite broadly, to include many contexts, like (16), that are actually straightforward enough to resolve as long as the system knows what to look for:

e.g., the name of a citizen of a country can stand in for a country. Using this broadly encompassing definition, 42% of the corpus examples we evaluated were in some way "tricky". However, the important point for our precision-oriented agents is that they will not attempt to resolve cases in which they cannot detect a single, high-confidence sponsor. So "tricky" contexts will certainly impose a hit on recall, but not necessarily on the precision of resolving BREs in predicate nominal configurations.

Note, however, that even if the system correctly points to the sponsor in these contexts, this does not fully resolve the meaning of the construction. For example, if *that* is resolved to 1971 in (14), the agent still has to combine a semantic interpretation of *1971 was the year* with a semantic interpretation of *Texas stadium opened.* We can facilitate the agent's doing this by creating phrasal lexical senses for configurations like YEAR *[be] the year (when/that) X;* LOCATION *[be] the place (where) Y;* HUMAN *[be] the person (who/that) Z*; and so on. The semantic interpretation of those senses will explicitly assemble the semantic pieces, yielding X (TIME YEAR-1); Y (PLACE LOCATION-1); Z (AGENT HUMAN-1). This same 2-part algorithm – i.e., point to the sponsor then assemble the semantic pieces using a phrasal lexical entry – can also be used for copular constructions such as *This is who/what/where/when/how….*

### 4.3 Bad Things Should Stop

One of many domain-independent generalizations is that people want bad things to stop. So, given an utterance like *This must stop!* one expects *this* to be something bad. The corresponding hypothesis we explored was: If we compile a list of "bad" events/states, and if the context immediately preceding *This must stop!* includes a word on that list, then that word should be the sponsor for the BRE.

In order to test this hypothesis, we needed a list of "bad events", which in the current lingo are called negative sentiment terms. Although we found an automatically compiled list of this type (Liu et al. 2005), it included many words that were either not events or were not sufficiently negative to wield strong predictive power (*gibe, flirt*). So, to support our experimentation, we manually compiled a list of over 400 negative-sentiment events, using introspection combined with manual inspection of both WordNet and Liu et al.'s list. Then we tested our hypothesis against this list using the configurations shown in Table 5, which simply state that some "bad" event or state is located in the clause or sentence preceding a statement that some bad thing must stop or is intolerable.

*Table 5.* Configurations expressing bad things that should stop or that are intolerable.

| | Sent./clause 1 contains | Sent./clause 2 | Example |
|---|---|---|---|
| 1 | bad event/state_i | BRE_i must/has to stop/end/be stopped/etc. | "This war is in no way acceptable to us. **It** must stop immediately"… |
| 2 | bad event/state_i | BRE_i is/was unacceptable/ intolerable/etc. | "…1,200 people were detained and packed in here, in building 19-6. **This** is unacceptable in a member country of the Council of Europe"… |

Our corpus analysis suggested that the patterns in Table 5 should be useful overall for predicting BRE sponsors, but only with certain enhancements.

*Enhancement 1*. Negative-sentiment terms should guide BRE resolution only *after* more confident reference resolution strategies have fired. For example, lexico-syntactic parallelism

13

seems to have stronger predictive power than a negative sentiment term, so *it* in (20) should be resolved to *this incident* by leveraging the "SUBJ copula ADJ" repetition "*this incident is unacceptable… it is unacceptable*".

(20) "<u>This incident</u> is unacceptable to the national authority and to the Palestinian people and free world. **It** is unacceptable at all levels."

*Enhancement 2.* Syntactic analysis is needed to avoid false positives both when matching the pattern and when matching the events in our list. For example, *It must stop to refuel* – as might be said of a vehicle – does not match our pattern: it has the extra complement *to refuel*. Similarly, *rebels* can serve as a BRE sponsor only when used as a verb (*He rebels every day and this must stop*), not as a plural noun.

*Enhancement 3.* In many cases, multiple entities are referred to by a BRE, which requires dynamic list concatenation – something that can be needed for plural referring expressions like *they* as well. E.g., in (21), all of the underlined negative events must be concatenated into the sponsor for *it*.

(21) "The stories we are hearing of the <u>harassment of political opponents</u>, <u>detentions without trial</u>, <u>torture</u> and <u>the denial of medical attention</u> are reminiscent of our experiences at the hands of apartheid police. **It** must stop now"…

In addition to investigating "it must stop/it is intolerable" contexts that contained a readily identifiable "bad event", we investigated "it must stop/it is intolerable" contexts that lacked such an event. We found that BRE resolution could be driven by another semantic generalization: *Any event that must be stopped must currently be going on.* Fortunately, there are at least three linguistic clues that an event is in progress: a) use of a verb in the progressive aspect, b) an adverbial expressing duration, and c) a verb expressing an increase or decrease in something. For instance, in invented example (22), the progressive aspect (*has been playing*) and the time adverbial (*for two hours straight*) suggest that *playing his recorder* is the sponsor of the BRE, despite the fact that recorder playing can be quite nice if done well and within reason.

(23) <u>That kid has been playing his recorder for two hours straight</u>. **This** has to stop!

Of course, multiple sources of evidence pointing to the same BRE resolution increases the agent's confidence of its BRE resolution decision.

The generalization that bad things should stop is only one of many domain-independent generalizations with the potential to guide the agent in resolving the reference of BREs. As with the other classes of phenomena presented here, our main goal is to expand the inventory of contexts in which agents can reliable resolve BREs during natural language analysis.

## 5. Discussion

This paper has described a microtheory of BRE detection and resolution that contributes to the theory of Ontological Semantic text processing implemented in the OntoAgent environment. This microtheory is formulated as a series of hypotheses realized as linguistic patterns which help to detect and resolve instances of BREs. The microtheory is amenable to implementation because all heuristic evidence relies exclusively on the output of available processors. In fact, these

algorithms represent only a modest enhancement to our approaches to semantic analysis, reference resolution, and multiword expressions.

During this exploratory stage of building this microtheory of BREs, the component hypotheses were evaluated by manual corpus analysis. This served as a proxy for machine processing. We expect the automatic analyzer to perform at a somewhat lower level of precision than our manual proxy due to potential errors by upstream processing modules.

This microtheory does not treat all instances of BREs. Instead, it treats only those that seem to be resolvable with relatively high confidence. This selectivity, as argued earlier, is realistic in applications in which agents are tasked with evaluating their own confidence in language understanding ("perception") before moving forward to reasoning and action. Future work involves extending the inventory of contexts that this microtheory can treat with high confidence.

## Acknowledgements

## References

Boyd, A., Gegg-Harrison, W. & Byron, D. (2005). Identifying non-referential *it*: A machine learning approach incorporating linguistically motivated patterns. *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, 40–47. Association for Computational Linguistics.

Byron, D. (2004). *Resolving pronominal reference to abstract entities*. Doctoral dissertation, University of Rochester. Available as Technical Report #815, University of Rochester Computer Science Department.

Channon, R. (1980). Anaphoric that: A friend in need. In Kreiman, J. Ojeda, A. (Eds.), *Papers from the Parasession on Pronouns and Anaphora*, 98-109. Chicago Linguistic Society.

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

Gentner, D. & Smith, L. A. (2013). Analogical learning and reasoning. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology*, 668–681. New York, NY: Oxford University Press.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *TRENDS in Cognitive Sciences* 7(5).

Graff, D. & Cieri, C. (2003). English Gigaword LDC2003T05. Web Download. 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.

Johnson, B. (2011). *A rule-based system for identifying pleonastic 'it'*. Master's Thesis, Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County.

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(4), 885–916.

Liu, B., Hu, M. & Cheng, J. (2005). Opinion Observer: Analyzing and comparing opinions on the web. *Proceedings of the 14th International World Wide Web Conference* (WWW:2005), May 10–14, 2005, Chiba, Japan. Their list of negative sentiment words can be found at https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/negative-words.txt

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

McShane, M. (2000). Hierarchies of parallelism in elliptical Polish structures. *Journal of Slavic Linguistics*, 8, 83–117.

McShane, M. (2014). A multi-faceted approach to reference resolution in English and Russian. In V. P. Selegei (Chief Editor), *Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference "Dialogue"*, 13, 391–409.

McShane, M. & Babkin, P. (2015). Automatic ellipsis resolution: Recovering covert information from text. *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (AAAI-15).

McShane, M. & Nirenburg, S. (2013). Use of ontology, lexicon and fact repository for reference resolution in Ontological Semantics. In Oltramari, A., Vossen, P., Qin, L., Hovy, E. (Eds.), *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, 157–185. Springer.

McShane, M., Nirenburg, S. & Beale, S. (2005). An NLP lexicon as a largely language independent resource. *Machine Translation* 19(2): 139–173.

McShane, M., Nirenburg, S. & Beale, S. (2008). Two kinds of paraphrase in modeling embodied cognitive agents. *Proceedings of the Workshop on Biologically Inspired Cognitive Architectures*, AAAI 2008 Fall Symposium, Washington, D.C., Nov. 7-9.

McShane, M., Nirenburg, S., & Beale, S. (In Press-a). The Ontological Semantic treatment of multiword expressions. *Lingvisticæ Investigationes* 38(1).

McShane, M., Nirenburg, S. & Beale, S. (In Press-b). Language understanding with Ontological Semantics. *Advances in Cognitive Architectures*.

McShane, M., Nirenburg, S., & Jarrell, B. (2013). Modeling decision-making biases. *Biologically-Inspired Cognitive Architectures (BICA) Journal*, 3, 39–50.

Nirenburg, S. & McShane, M. (Forthcoming). Natural language processing. *The Oxford Handbook of Cognitive Science*, edited by Susan Chipman.

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

Passonneau, R. J. (1989). Getting at discourse referents. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (ACL '89), 51–59.

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

Schank, R. C. & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into*

*human knowledge structures*. Hillsdale, NJ: L. Erlbaum.

Webber, B. L. (1988). Discourse deixis: reference to discourse segments. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics* (ACL '88), 113–122.

Zipf, G.K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, Mass.: Addison-Wesley Press, 1949.