Semantics-Based Resolution of Fragments and Underspecified Structures

Marjorie McShane, Sergei Nirenburg, Stephen Beale

Institute for Language and Information Technologies
University of Maryland Baltimore County

Department of Computer Science and Electrical Engineering
ITE 325, 1000 Hilltop Circle
Baltimore, MD 21250
marge@umbc.edu, sergei@umbc.edu, sbeale@umbc.edu

ABSTRACT. *This article presents algorithms for the interpretation of subsentential and underspecified structures in English within the theory of Ontological Semantics. The approach centers around producing a text-meaning representation of the subsentential or underspecified structure, then launching a procedural semantic routine to appropriately link its meaning to that of the previous context. The algorithms to be described are being implemented in the OntoSem text processing environment.*

KEY WORDS: *semantics, fragments, ellipsis, Ontological semantics*

1. Introduction

Under traditional approaches, the following examples will be divvied up across theories and research programs:

1. Colin Powell was appointed Secretary of State.
   a. In 2000.
   b. By George Bush.
   c. Without scandal.
2. George Bush wanted to appoint someone highly competent. Powell.
3. Who did George Bush vote for in 2004? Himself.
4. Colin Powell addressed the United Nations twice.
   a. George Bush once.
   b. Congress once.
5. Colin Powell was appointed Secretary of State.
   a. I know.
   b. I know that.

The research areas that consider the above examples within their purview are: Fragments (1, 2, 3, 4 and perhaps 5a), VP-ellipsis (5a), Gapping (4a), Question-answering (3), reference resolution of demonstratives (5b), and anaphora (3).[1] By contrast, the Ontological Semantics (OntoSem) implemented theory of natural language processing (Nirenburg and Raskin 2004) includes in its purview all phenomena of natural language. The goal of this approach is to automatically derive structured meaning from unstructured text using all available heuristics, be they from the realm of syntax, semantics, pragmatics, or world knowledge. Given the omnivorous nature of the OntoSem approach, it is not surprising that we seek out functional similarities among phenomena rather than splitting them at all possible theoretical junctures. For us, examples (1)-(5) are variations on the theme of reference resolution, with the lion's share of resolution methods carrying over among subtypes of phenomena.

Our thesis in this paper, which encapsulates the rationale of the OntoSem approach, is as follows: incorporating semantics into theoretical and practical approaches to language analysis fundamentally alters the conceptual landscape and can potentially render unnecessary those approaches that model language up to the point where semantics is expected to enter the picture and, in some unspecified way, tie up all remaining loose ends. This simplified depiction of the traditional divide and conquer approach to linguistics should not be construed negatively: it is a natural aspect of scientific investigation to circumscribe domains into a manageable scope. However, there are cases in which the assumption of too many prerequisites

[1] Contributions to each of these research programs are vast; a short list includes: fragments (Barton 1990, 1991, DATE; Merchant 2002; Morgan 1989; Stainton 2004 and forthcoming, among many others); VP Ellipsis and Gapping (Kehler 2002, Lobeck 1995); Question-answering (Ginzburg and Sag 2001); demonstratives (Byron 2004); anaphora (Safir 2004). Note that some investigators reject cross-sentential gapping, but we assume it to be a punctuation variant of typical gapping.

can lead us down a garden path, since filling those prerequisites would in large part solve the original problem. In the case of fragments and other underspecified structures, we will suggest that semantics cannot be postponed because it is the key to their use in language.

In this paper, we present algorithms for processing sentences like (1)-(5) within the OntoSem environment. These algorithms are under iterative implementation and refinement as part of ongoing work to process the full range of input found in unrestricted natural language texts. But before turning to the algorithms, let us orient our work in the broader research on related phenomena.

2. Traditional Theoretical Treatments of Fragments

The main debate about fragments revolves around the following competing hypotheses:[2]

Fragment Hypothesis 1. Fragments are full sentences with ellipsis; as such, their meaning is determined using the same syntax-semantics mapping as is carried out for full sentences. Investigators pursuing this hypothesis believe there are sufficient parallels between the syntactic form of fragments, on the one hand, and the syntactic form of corresponding elements in full sentences, on the other, to bootstrap the fragments into sentence grammar under the assumption that the fragments are complete but elliptical sentences. The implicit assumptions are (a) there is something to be gained by extending the purview of sentence-grammar theories to certain types of multi-sentence discourses, and (b) a syntactically grounded elliptical account is more explanatory and less stipulative than competing accounts (e.g., analysis within discourse grammar; see Hypothesis 2).

A recent contribution of particular rigor is Merchant 2004. Merchant argues against the direct interpretation approach because it would require independent mechanisms for such phenomena as case-marking on fragments and the possibility or impossibility of preposition-stranding within them. The drawback of the syntactic bootstrapping approach, however, is that it only works in a subset of cases, as Merchant himself points out and Stainton (forthcoming) further pursues in his response to Merchant. When syntax fails to predict the correct type or form of a fragment,[3] Merchant falls back on an underspecified semantic recovery process:

> "…Nothing in the current theory requires strict form identity of question and answer: the identity that is required is a semantic one (based on e-givenness), and hence will allow slight deviations in form provided the semantics remains constant. Thus language-particular quirks of syntax (such as the fact that there is no *wh*-form for questioning predicates directly in English) will not preclude semantically appropriate answers, even in reduced forms" (Merchant 2004: 697).

[2] See Elugardo and Stainton 2005 (Introduction) for an overview of the history of the debate in the literature.

[3] Examples, which are provided by Merchant (2002), include "What was he like?" "Hard to live with". A syntactic approach would predict that "*He was like hard to live with" should be a valid reconstruction.

As we understand it, the original justification for the bootstrapping of fragments into syntactic theory was the power of syntax to predict the form of fragments; however, if the predictive power covers only a subset of cases, and these cases cannot be systematically constrained using mechanisms within the theory, then the analysis must be evaluated as failing to achieve its original goal.

A more fundamental question regarding the bootstrapping of fragments into syntactic theory is, why do it to begin with? Since its inception, generative syntactic theory has been devoted to the level of sentence and has excluded from its purview any serious treatment of semantics, prosody, pragmatics, etc. The justification for this narrow purview is the central premise of the theory: that there is an innate language mechanism devoted exclusively to sentence-level syntax. Therefore, attempting to treat fragments, which are unarguably discourse-level utterances, using a sentence-bound theory necessitates the questions: Is the intent to expand the purview of the theory? If so, in what precise ways and how far? And how does such an expansion affect the original premise about the sentence-level mechanisms of Universal Grammar?

Let us further pursue the question of what could be gained by dealing with fragments within sentence grammar. Merchant (2004: 718-723) suggests that in sentences containing VP ellipsis there is an elided *do it* that permits the syntax-semantics mapping to proceed in the normal way of generative grammar (that is, if one gestured toward a chair and says *"May I?"* the structure is actually *May I [do it]?*). However, positing this extra lexical material (*do it*) offers no benefit either to syntactic theory or to the interpretative module (moreover, as Stainton, forthcoming, shows, the insertion of *do it* does not always work). From the perspective of generative grammar, this sentence needs no other elements: the auxiliary *may* licenses the ellipsis of a VP (Lobeck 1995), so from the syntactic perspective, the sentence is fine. As regards interpretation, positing a *do it* is not helpful: it leaves the semantic/pragmatic module in exactly the same situation as it would have been if the elided VP had had no content at all. Therefore, the insertion of *do it* is merely making the sentence look more like a full sentence under linguistic approaches that favor "complete" structures.

In conclusion, we must disagree with Merchant's evaluation of his analysis: that it is useful because it reduces the number of types of fragments that need to be handled by non-syntactic means. We view the relevant phenomena from the other direction: if the language mechanism needs other means of dealing with the left-over cases (what Merchant considers puzzles or matters of semantics), then it seems reasonable to assume that the same mechanism is brought to bear on all cases. What Merchant's analysis is undoubtedly useful for, regardless of one's theoretical orientation, is providing descriptive guidelines regarding licit forms of fragments. That is, the parallels between the form of fragments and the form of corresponding entities within sentences are not without use—they are simply not sufficient to support the conclusion that fragments should be subsumed under sentence grammar.

Fragment Hypothesis 2. Fragments contain exactly and only their visible elements. This analysis has been argued for most notably by Barton (1990, 1991) in terms of pragmatic theory, and by Stainton (2004, forthcoming) from the philosophy

of language standpoint. This approach attaches no a priory pre-eminence to "full" sentences in language, instead recognizing subsentential utterances as valid linguistic entities in their own right. Among the most compelling votes for this approach is the inventory of fragment types that cannot be accounted for by syntactic reconstruction followed by ellipsis.

In developing a discourse model for fragments, Barton (1990, 1991) proposes the X^{max} Generalization, which says that a grammar generates sentences not only under the initial node of S but also with initial nodes of NP, VP, AdjP, AdvP, and PP. She then uses mechanisms that are already available in generative grammar to account for many decisions about the acceptability and unacceptability of fragments. This work is appealing on philosophical, practical and even aesthetic fronts—it *feels* right to not have to posit (then move, under Merchant's account) then elide so many categories. However, Barton's discourse theory has no access to an integrated linguistic, semantic and world model to act as a conceptual substrate. As such, crucial elements of the analysis must be *assumed* rather than modeled. For example, in analyzing the discourse

D: He was paid.
P: By check?

Barton (1990: 118-119) explains that *by check* fills the instrument role of the verb *paid*, thereby functioning as an elaboration. But the process of assigning *by check* the role of *instrument* of *pay* is not specified.

Such explanatory and modeling gaps are not unique to Barton's discourse theory or to the domain of fragments. Discourse theories are known to suffer from the necessity of referring to less readily formalized notions – one reason why it has been so popular to bootstrap phenomena into the syntactic level despite the necessary losses. As Gentner et al. say of research on analogy, another cross-modular linguistic process (Cf. Section 4, discussion of sentence 4): "Complex explanatory analogies have until recently received little attention in psychology, perhaps because such analogies require fairly elaborate representations of meaning" (Gentner 1983: 166). So does the treatment of fragments.

The fragment debate is very engaging reading and, as Barton (in press) correctly points out, it has honed in on crucial data that must be accounted for by theoreticians. However, as our critique thus far has suggested, we find insufficiencies with it that we attribute to the overly narrow purviews of the theoretical frameworks levied to treat it. We do not believe that a tug-of-war of purviews will ultimately lead either to explanations of what is actually going on in a human's language processor or to sufficient computer models. Thus, we must call into question Barton's point of departure: "By specifying the exact contribution of each component [syntax, semantics, discourse, pragmatics], the development of a theory of nonsentential constituents has to examine central and controversial issues in linguistics, such as the autonomy of syntax, the interaction between grammar and pragmatics, and the nature and development of a pragmatic model" (1990 p. xii). Instead, what is needed, we suggest, is an approach that appeals to all levels of language and world knowledge and crucially involves semantic interpretation.

Naturally, fragments are not the only linguistic elements whose full analysis requires semantic and world knowledge: all use of language ultimately relies on this foundation. However, in certain domains – like fragments, ellipsis, reference resolution and what Jackendoff calls "enriched composition" (as in the famous Nunberg example of a waitress referring to a customer as *the ham sandwich*) (2002: 388-391), the insufficiency of non-semantic accounts is particularly striking. As Jackendoff says of enriched composition, it "show[s] how the understanding of sentences is a rich interaction between grammar, independent well-formedness conditions on conceptual structure, and the construal of context" (ibid, 388). He objects to dismissing such phenomena as "mere pragmatics" (ibid, 388), instead calling enriched composition a "conventionalized piece of meaning" (ibid, 389) that is both part of pragmatics and part of language. To fragments one can attribute the same description.

Working within the theory of Ontological Semantics provides the opportunity to exploit precisely the type of elaborate representation of meaning that has, until now, eluded the field – an advantage that has come at the cost of twenty years' development of the theory, along with its related knowledge resources (ontology and lexicon) and processing engines. Informed by this theory and the possibilities for deep analysis it offers, we suggest a reevaluation of past work on fragments and related issues of reference resolution.

3. A Snapshot of OntoSem

OntoSem (the implementation of the theory of Ontological Semantics; Nirenburg and Raskin 2004) is a text-processing environment that takes as input unrestricted raw text and carries out preprocessing, morphological analysis, syntactic analysis, and semantic analysis, with the results of semantic analysis represented as formal text-meaning representations (TMRs) that can then be used as the basis for many applications. Text analysis relies on:

- The OntoSem language-independent ontology, which is written using a metalanguage of description and currently contains around 8,500 concepts, each of which is described by an average of 16 properties.
- An OntoSem lexicon for each language processed, which contains syntactic and semantic zones (linked using variables) as well as calls for procedural semantic routines when necessary. The semantic zone most frequently refers to ontological concepts, either directly or with property-based modifications, but can also describe word meaning extra-ontologically, for example, in terms of modality, aspect and time. The current English lexicon contains approximately 30,000 senses, including most closed-class items and many of the most frequent and polysemous verbs, as targeted by corpus analysis. The base lexicon is expanded at runtime using an inventory of lexical rules. (An extensive description of the lexicon, formatted as a tutorial, can be found at http://ilit.umbc.edu.)
- An onomasticon, or lexicon of proper names, which contains approximately 350,000 entries.

- A fact repository, which contains real-world facts represented as numbered "remembered instances" of ontological concepts (e.g., SPEECH-ACT-3366 is the 3366th instantiation of the concept SPEECH-ACT in the world model constructed during the processing of some given text(s)).
- The OntoSem syntactic-semantic analyzer, which covers preprocessing, syntactic analysis, semantic analysis, and the creation of TMRs. Instead of using a large, monolithic grammar of a language, which leads to ambiguity and inefficiency, we use a special lexicalized grammar created on the fly for each input sentence (Beale, et. al. 2003). Syntactic rules are generated from the lexicon entries of each of the words in the sentence, and are supplemented by a small inventory of generalized rules. We augment this basic grammar with transformations triggered by words or features present in the input sentence. These transformations are similar in many respects to those found in XTAG (Schabes et al. 1988, XTAG-Group, n.d.).
- The TMR language, which is the metalanguage for representing text meaning.

OntoSem knowledge resources are at this time acquired primarily manually (though note that the knowledge acquirers use a variety of efficiency-enhancing tools – graphical editors, enhanced search facilities, capabilities of automatically acquiring knowledge for classes of entities on the basis of manually acquired knowledge for a single representative of the class, and the like). The ontology has been under continuous development, with varying levels of effort, for around 20 years. It took approximately 2.5 years of work by a PhD-level linguist to compile the current lexicon. (Although the OntoSem environment has always utilized an English lexicon, previous versions aimed for a coarser grain-size of description and did not reflect recent theoretical and practical advances). The onomasticon was extracted automatically from corpora and structured sources. The fact repository is populated automatically from text-meaning representations. Knowledge acquisition is largely driven by lacunae found during the processing of actual texts; it is expedited using OntoSem's DEKADE environment (see McShane et al. 2005a). We are currently working on developing a "push me pull you" knowledge acquisition strategy that incorporates machine learning (ML) of lexicon and ontology into our knowledge-rich environment: the more knowledge we learn with the help of ML, the more resources we will have to support the learning of still more knowledge. We do not consider the "knowledge bottleneck" to be anywhere near the impasse that many make it out to be: acquiring knowledge simply requires effort, no different from or more extensive than the effort currently being exerted in creating annotated corpora.

A high-level view of OntoSem text processing is shown in Figure 1.

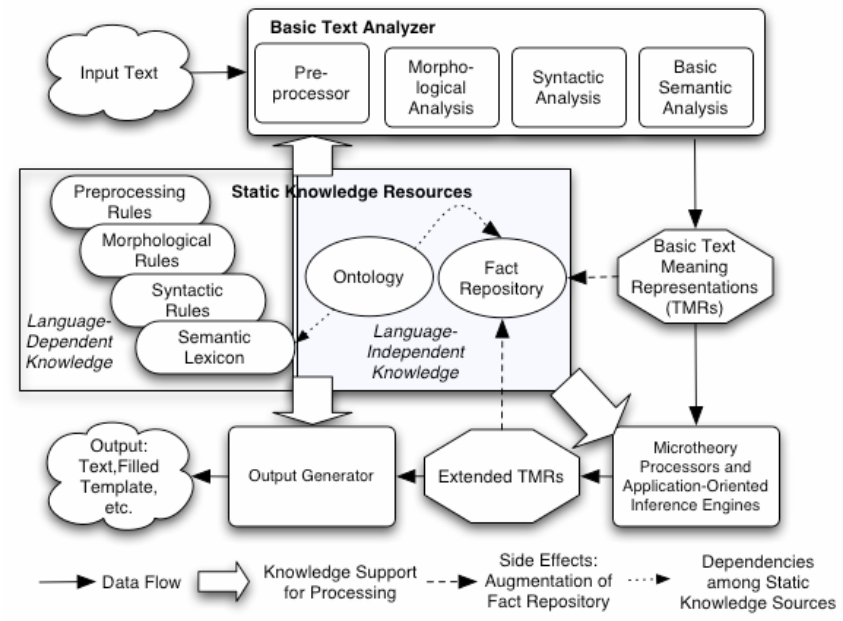

Figure 1. A High-Level View of OntoSem

TMRs represent propositions connected by discourse relations (see Nirenburg and Raskin 2004, Chapter 6 for details). Propositions are headed by instances of ontological concepts, parameterized for modality, aspect, proposition time, overall TMR time, and style. Each proposition is related to other instantiated concepts using ontologically defined relations (which include case roles and many others) and attributes. Coreference links form an additional layer of linking between instantiated concepts. OntoSem microtheories devoted to modality, aspect, time, style, reference, etc., undergo iterative extensions and improvements in response to system needs as diagnosed during the processing of actual texts. TMRs have been used as the substrate for question-answering (Beale et al. 2004), machine translation (Beale et al. 1995) and knowledge extraction, and were also used as the basis for reasoning in the question-answering system AQUA, where they supplied knowledge to enable the operation of the JTP (Fikes et al., 2003) reasoning module.

4. Processing Fragments and Underspecified Structure

In all of the examples (1)-(5), the follow-up sentence has “needs”: in order to derive its full, discourse grounded interpretation, its meaning must be linked to the meaning of the previous discourse. This does *not* suggest that the fragments lack meaning in and of themselves; quite the contrary (we recommend Robert Stainton’s body of work on fragments for a convincing discussion of the relevant issues). However, like

so many referential entities in language, full interpretation defies a sentence-bound approach.

Full interpretation is at the heart of the Ontological Semantics approach to text processing. Text-meaning representations (TMRs) represent the semantics of all aspects of texts, including ontologically grounded meaning, modality, aspect, time, speaker attitudes, and all aspects of reference resolution. Reference resolution is, in fact, a good illustration of our conception of full interpretation. Within OntoSem we understand processing reference as detecting all referring expressions in a text or a corpus and associating them with their anchors in the fact repository (FR), which is a database of interlinked real-world instances of objects and events extracted from text after it has been interpreted by our analysis system. The information in the FR both supports the processing of any given text (it is a substrate of computer-tractable knowledge) and is supplemented by information from that text. Under this conception of full resolution of reference, the text string *Colin Powell* is not resolved until it is linked to its anchor in the FR, if there is one, or instantiates a new FR anchor, if none yet exists. Thus, the OntoSem engine must try to link every pronoun, relative date (*last week*), relative time (*later*), definite descriptions (*that man*), etc., not only to other co-referential elements in the given text, but to the actual anchor in the ever-growing world model. This is reference beyond co-reference.[4]

This tangent into reference resolution does not merely serve as general background, it is essential to an understanding of how OntoSem treats fragments and underspecified structures. They all require reference resolution and, as such, are subject to our battery of reference resolution methods. Consider a typical case of reference: the pronoun *he* is used to refer to a male person or animal who was referred to in the preceding context (we leave aside the case of visual cues introducing him into the discourse). The fact that this referring expression must be linked to a previously mentioned male animal is **triggered** by the lexical form of the referring expression: all pronouns are lexically defined to trigger the process of reference resolution, which in OntoSem is represented by a call to a procedural semantic routine—what we call a Meaning Procedure (for further discussion of meaning procedures, see McShane et al. 2004). Most pronouns, moreover, include some heuristics to constrain the search for a coreferent, like a given value for number or person.[5] Once the need for reference resolution has been triggered, and available lexically specified heuristics have constrained the search space, a general battery of heuristics is launched to select among the remaining candidates. Of the large inventory of heuristics pertaining to syntax, semantics and pragmatics, the

[4] Cristea et al. 2005 also posit non-surface layers of representation to be used in reference resolution. However, what they call a "semantic layer" is not a full-blown, disambiguated semantic interpretation as we define it in OntoSem; instead, their semantic layer is surface tokens supplemented by a few features and, when applicable, reference-related links among them. The goal of Cristea et al.'s research program, as we understand it, is to maximize reference resolution potential up to the point where full semantics and world knowledge are needed. We, by contrast, are building the semantic and world knowledge that supports richer left-hand sides of analysis rules.

[5] We borrow the term *coreferent* from Byron 2004.

relevant subset is automatically extracted for any given context. Ideally (but not always in practice, as yet), the combination of relevant heuristics leads to one high-confidence resolution for the referring expression. Although fragments and other underspecified structures present different triggers for the need for reference resolution, once the process has been triggered, the generalized heuristics-based reference resolution procedures are launched.

Another important aspect of reference resolution that carries over to the treatment of fragments and underspecified structures is the reliance on the TMR, the semantic representation of text meaning, as the basis for reasoning. Although syntactic heuristics have their place in our approach, most reasoning is done at the level of semantic interpretations of text rather than uninterpreted text strings (see McShane 2005 for further discussion of the treatment of ellipsis in OntoSem).

In the subsections that follow we describe the processing of each of our original five sentences. We use the first sentence as the basis for a relatively (considering space constraints) in-depth description of the production of TMRs in OntoSem and how reasoning at the level of TMR is carried out. Discussion of further sentences draws comparisons and makes extensions from that original description.

(1) Colin Powell was appointed Secretary of State. In 2000.

Our basic approach to analyzing fragments is: a) generate whatever semantic interpretation is possible from the fragment itself; b) detect the "needs" as yet unfilled in that semantic interpretation; c) attempt to fill those needs using any and all available heuristics; d) once those needs are filled, verify that the original semantic interpretation is valid, otherwise, amend it.

The OntoSem analyzer is well suited to analyzing subsentential structures because rather than rely on a large, monolithic syntactic grammar of a language, which leads to ambiguity and inefficiency, we use a special lexicalized grammar created on the fly for each input sentence (Beale, et. al. 2003). Syntactic rules are generated from the lexicon entries of each of the words in the sentence, and are supplemented by a small inventory of generalized rules. This basic grammar is augmented by transformations that are triggered by words or features present in the input sentence. Since syntactic structure is built bottom-up, there is no need for an input to be a traditional full sentence: in other words, no canonical S needs to be in the left-hand side of the rule for every syntactic parse.

When the analyzer encounters the input *In 2000,* it will search through the many lexical entries for *in* and select the one that is syntactically and semantically most appropriate – in this case, *in-prep10*:

```
(in-prep10
  (cat prep)
  (def "temporal; followed by month, year, century, etc.")
  (ex "He came in January. His change of career in 2002 surprised us.
       What happened in the fifth century B.C.?")
```

```
(syn-struc
  ((root $var1) (cat (or n v))
    (pp ((root $var0) (cat prep) (obj ((root $var2) (cat np)))))))
(sem-struc
  (^$var1 (sem EVENT) (time (value ^$var2)))
  (^$var2 (sem (or MONTH YEAR DECADE CENTURY)))))
```

OntoSem lexical entries are written in an extended LFG formalism using LISP-compatible format. This entry is read as follows.

The syntactic structure (syn-struc) indicates that the input covered by this sense of *in* should contain a constituent headed by a noun (N) or verb (V) followed by a prepositional phrase (PP). All syntactic elements are associated with variables, which permit their linking to elements in the semantic structure (sem-struc). The variable associated with the head word, here *in*, is always $var0; it does not have an explicit sem-struc linking since the whole entry is describing the meaning of $var0 in the given configuration.

The sem-struc says that the meaning of $var1 ("meaning of" is indicated by a caret (^)) is some ontological EVENT whose time is the same as the time of the meaning of $var2. Moreover, it is specified that the meaning of $var2 must represent a MONTH, YEAR, DECADE or CENTURY (this entry predicts that one cannot say, for example, **in Monday*, since *Monday* is an instance of the ontological concept DAY).

When the analyzer searches through all the lexical senses of *in*, it selects the best one based on a combination of syntactic and semantic matching. In the case of the input *In 2000*, no matter what sense it selects, the syntactic match will be suboptimal because there is nothing explicit for the PP to modify. However, among the suboptimal choices, it will select *in-prep10* as having the best semantic match of all the candidate entries. This choice is based on the fact that, among the candidate analyses for *2000* returned by the pre-processor, one is YEAR, which perfectly matches the semantic constraints of this sense. (Space does not permit a full description of OntoSem's robust ontologically-based disambiguation; see Nirenburg and Raskin 2004 for details).

Once *in-prep10* has been selected as the best lexical match for the input, the analyzer generates what we call a Basic TMR, which includes basic semantic dependencies among the already disambiguated constituents. The Basic TMR for the input *In 2000* is as follows:

YEAR-1

textpointer	2000
ABSOLUTE-TIME	(YEAR 2000)
TIME-OF	EVENT-1

EVENT-1

textpointer	*none*
TIME	YEAR-1

Every instance of an ontological concept generated during text processing is appended with a distinct instance number. Instance numbers begin fresh for each

new corpus analyzed. YEAR-1 is instantiated from the input *2000* when it is contextually disambiguated in combination with the preposition that introduces it. Its interpretation is "absolute-time (YEAR 2000)". It is cross-referenced in the TMR as the TIME-OF EVENT-1.

EVENT-1, by contrast, was not instantiated by a text element, it was instantiated by the lexical sense that the analyzer selected as the best of the possible options. The sem-struc of that sense said that this meaning of *in* indicates the time of some EVENT; and even though that EVENT is not specified, it is still implied as a result of the compositional semantics of *in 2000*. In our parlance, the omission of the EVENT that is lexically expected by this sense of *in* **triggers** the process of reference resolution. So, in the Basic TMR, before procedural semantic routines have reasoned about the nature of the EVENT, it is referred to simply as EVENT (one of the three top bifurcations of the OntoSem ontology: OBJECT, EVENT, PROPERTY). The trace that reference resolution needs to be carried out on the EVENT is the filler *none* in the slot for textpointer. The Extended TMR, which shows the results of all applicable types of reasoning, will show the full, contextually bound interpretation of the fragment.

Within OntoSem, the basic algorithm for coreferencing events is similar to that for coreferencing objects: create an inventory of candidate coreferents then score them using a combination of heuristics. Our methods are unlike typical knowledge-lean methods for reference resolution (see, e.g., Ruslan Mitkov's extensive publications) in that they rely on not only the typical syntactic and distance heuristics, but also on ontologically grounded semantics and real-world knowledge recorded in the fact repository. For example, the implied event in the utterance *In 2000* will never be coreferred with an event whose Fact Repository entry shows it to have occurred in 1977.

In our very short input text, there is, of course, only one candidate event: the appointing of Colin Powell as Secretary of State. (We use such a short context for illustration since our point here is not to showcase OntoSem's powers of disambiguation but, rather, to suggest a new way of conceptualizing fragments.) The meaning of that sentence will be reflected in TMR as follows:

Colin Powell was appointed Secretary of State.

SOCIAL-EVENT-1
textpointer APPOINT
EFFECT HAS-SOCIAL-ROLE-1
time (< (find-anchor-time))
HAS-SOCIAL-ROLE-1

DOMAIN	HUMAN-1
RANGE	SECRETARY-OF-STATE-1
CAUSED-BY	SOCIAL-EVENT-1
HUMAN-1	
textpointer	Colin_Powell
HAS-NAME	((FIRST Colin) (LAST Powell))
DOMAIN-OF	HAS-SOCIAL-ROLE-1
FR-REFERENCE	HUMAN-FR24[6]
SECRETARY-OF-STATE-1	
textpointer	Secretary_of_State
RANGE-OF	HAS-SOCIAL-ROLE-1

When the coreference engine searches for a coreference link for the underspecified EVENT in the fragment *In 2000*, it will accept SOCIAL-EVENT-1 as the coreferent since neither this local TMR nor the FR contains any evidence to contradict this linking. The Extended TMR for the fragment, which shows the reference resolution, will be:

YEAR-1	
textpointer	2000
absolute-time	(YEAR 2000)
time-of	event-1
EVENT-1	
textpointer	*none*
time	YEAR-1
corefer	***SOCIAL-EVENT-1***

Functionally, the fragment *In 2000* supplies additional information about the preceding proposition. Many fragments have this role, including those presented as alternative continuations in example (1): *By George Bush. Without scandal.* These are processed exactly the same as *In 2000.* First, the analyzer carries out compositional semantics on the fragments themselves, relying on the lexicon to suggest preferences among the many senses of *by* and *with* based on the meaning of their complements. The use of a PP without an explicit modified element triggers the search for that modified element in the preceding context, and that search proceeds as described above.

The only outstanding issue concerns the use of results of reference resolution to verify, disambiguate or overturn the original interpretation of the fragment. Consider again the fragment *By George Bush*. The meaning of *by* cannot be fully disambiguated outside of context: *George Bush* could either be the agent of the understood event (*appointing someone Secretary of State* in our example), or he could be located next to some object (e.g., *Colin Powell wasn't sitting in the*

[6] This indicates that reference resolution at the level of the Fact Repository (database of real-world instances of ontological concepts) has been carried out. Colin Powell is the 24th instance of HUMAN persistently stored in that knowledge base.

audience, he was standing on the podium. By George Bush.) In other words, both senses of *by* are equally accepting of a HUMAN as their object, meaning that the analyzer will posit the same score for both analyses of the fragment and must wait for the EVENT co-reference to be established before settling on a preferred interpretation.[7] Once the co-reference for the EVENT has been established, the OntoSem analyzer carries out disambiguation of the fragment in the usual way.

Having described, using one extended example, how the OntoSem analyzer interprets fragments, we now comment more briefly on salient aspects of the other examples in question, focusing on the classes of phenomena they represent and the unity of approach to their analysis taken within the Ontological Semantics framework.

(2) George Bush wanted to appoint someone highly competent. Powell.

This example shows reference specification, or cataphora. Typically, reference resolution involves concretizing the meaning of a pronoun or other underspecified entity by linking it to a preceding referential expression in the context. Here, by contrast, we have a referential expression that offers the concretization of a more generalized expression in the context. As with all matters of reference resolution, we conceptualize reference specification in terms of *triggers* for the need for reference resolution and *combined heuristics* to carry it out.

In this case, the trigger is the NP functioning as a fragment. The OntoSem analyzer has a special syntactic rule—one of its few non-lexically grounded rules—that requires bare sentential NPs be referentially linked to the discourse context. Without this rule, the bottom-up syntactic analyzer would analyze the fragment as an NP, compute its semantics, and be perfectly satisfied to stop processing the sentence at that point. Only NP fragments require such a rule because in all other cases the need for reference resolution is lexically triggered: all PPs, adjectives and adverbs are lexically described as needing a modified element, so the lack of one in a corresponding fragment will trigger reference resolution; similarly, any missing arguments in verbal fragments will be interpreted as triggers for reference resolution. The actual process of reference specification is the same as for traditional reference resolution: the meaning of the target element (here, *Powell*) is compared to the meaning of candidate coreferents (here, *George Bush* and *someone highly competent*), and the highest scoring option – based on weighted heuristics – is selected. Here, *someone highly competent* is selected because its features are compatible with those of the target, *Powell*: HUMAN (GENDER: MALE) (number: singular). *George Bush* is excluded because entities with different surnames cannot corefer.

(3) Who did George Bush vote for in 2004? Himself.

[7] The control structure is outside the scope of this paper. Suffice it to say that the process of postponing ambiguity resolution in semantic interpretation is similar to that used in syntactic parsing.

Question-answer contexts have often been treated as a special topic (see, e.g., Ginzburg and Sag 2001). Despite clear practical reasons for this—most notably, the need for effective question-answering systems in the near term—there is no convincing theoretical evidence that answer fragments need to be distinguished from other fragments, or further, that fragments in general need to be distinguished from more complete utterances, or from incomplete utterances of different kinds (the thrust of our argumentation). There are at least two justifications for treating answer fragments like any other utterance: first, the economy of effort achieved by reusing resources and approaches across phenomena; second, the frequency with which question-answer contexts defy canonical expectations. A recent search of a Wall Street Journal corpus showed that, with surprising frequency, questions are not followed by their answers: they are followed by other questions or discourses that in a roundabout way provide an answer. Therefore, rather than create a specialized approach to Q/A contexts, we launch our generalized methods on them.

Let us assume that we are processing a text containing the sentences in (3).[8] The TMR for the question will be headed by a concept instance of REQUEST-INFO, which is instantiated due to the question mark. The THEME of REQUEST-INFO (i.e., the information being sought) is an unspecified HUMAN who is the THEME of both the ELECT event and REQUEST-INFO. The analyzer knows to instantiate a HUMAN as the filler because the ontological specification of ELECT constrains the THEME to a HUMAN. Thus the TMR for this sentence is:

REQUEST-INFO-1		
	THEME	ELECT-1.THEME
	textpointer	*question-mark*
ELECT-1		
	AGENT	HUMAN-3
	TIME	TIME-2
	THEME	(HUMAN-4 (THEME-OF (REQUEST-INFO-1)))
	textpointer	vote_for
HUMAN-3		
	textpointer	George_Bush
	AGENT-OF	ELECT-1
TIME-2		
	ABSOLUTE-TIME	(YEAR 2004)
	TIME-OF	ELECT-1
	textpointer	2004

The first step in analyzing the subsequent fragment, *himself*, is to look up this word in the lexicon, where it – like all pronouns – is described syntactically as a pronoun and semantically using a combination of a basic ontological mapping (ANIMAL, not AGENT), inherent features (GENDER: MALE, NUMBER: SINGULAR), and a call to a meaning procedure that seeks out its necessary coreferential category. The procedural attachment for *himself* includes a weighted set of heuristics of a primarily

[8] That is, we are not building a Q/A system that will seek to answer the first sentence by outputting the second.

syntactic nature, since reflexive pronouns can typically be resolved almost exclusively using the syntax of their local sentence. However, when those heuristics are inapplicable, as in the case of fragments, semantics can take over. That is, syntactic heuristics are relaxable, which is a crucial means of handling not only certain classes of "expected" utterances, as in our example, but also all kinds of "unexpected" input, as one finds frequently in genres like email and colloquial spoken language.

Step by step, the resolution of *himself* goes as follows. The basic, lexically encoded semantic description of *himself* is ANIMAL whose case role is not AGENT. A search of candidate coreferents will exclude George Bush (since this entity is assigned the case-role AGENT) and will return a high score for HUMAN-4 (the semantic analysis of *Who*), which is the THEME of the ELECT event. So, *Who* and *Himself* form a chain of coreference. Once this chain is established, the reference resolution engine is rerun for *himself*, which still has an outstanding coreference "need" (it has not been linked to any real-world entity). This time, it will find a local AGENT to corefer with – namely, HUMAN-3, which is George Bush. This example shows that the detection of reference "needs" and their resolution can occur in cycles until all needs have been satisfied.

(4) Colin Powell addressed the United Nations twice.
- **a. George Bush once.**
- **b. Congress once.**

In this pair of examples, the lexical trigger for the reconstruction of an EVENT is the use of an adverb (*once*) without an event to modify (that is, the system does not need a special rule to say that a sentence composed of an NP followed by an adverb is missing a necessary verb). An underspecified EVENT is inserted into the TMR for the fragment, and its property "textpointer *none*" triggers the need for reference resolution, as described for sentence (1). Prior to reference resolution, the system cannot guess the case-role of George Bush or Congress, so these constituents are listed as fillers of the generic CASE-ROLE.[9] The Basic TMR for the fragment *George Bush once*, therefore, is as follows:

```
HUMAN-4
        textpointer       George_Bush
        CASE-ROLE-OF      EVENT-4
EVENT-4
        textpointer       *none*
        CASE-ROLE         HUMAN-4
        CARDINALITY       1                ; from textpointer 'once'
```

The next step is to seek the specification of the EVENT in the preceding context. Note that, in this case, we are not seeking a coreference of an event *instance* we are seeking coreference of an event *type*. This need not be stipulated because a general

[9] For example, in the input *John got sick twice. Mary only once*, John and Mary are EXPERIENCERs (not AGENTs) of ANIMAL-DISEASE, as defined in the ontology.

rule in the analyzer blocks coreference of events with incompatible case-role fillers. When incompatibility arises, it relaxes the nature of the coreference to type-coreference rather than instance coreference. In our example, type-coreference will be established with the SPEECH-ACT (instantiated from *addressed*) in the TMR for the preceding sentence. The AGENT of that SPEECH-ACT is the HUMAN, Colin Powell, and the BENEFICIARY is the SOCIAL-ORGANIZATION, United Nations.

Once the reference for the implied event in the fragment has been concretized, the CASE-ROLE of the George Bush (4a) or Congress (4b) can be further constrained. Since both HUMANs and SOCIAL-ORGANIZATIONs can be AGENTs or BENEFICIARYs, there is no simple way to choose the appropriate interpretation in each case. We must apply to a comparison of the semantic similarity between each element of the fragment and each element of the TMR containing the event that has been established as the type-coreferent. This matching process is carried out using OntoSearch, a stochastically trained engine that computes the ontological similarity between entities based on the traversal of variously weighted ontological paths (Onyshkevych 1997). In our example, it is trivial to detect a closer correlation between the two humans, and between the two organizations and use that to infer that George Bush, like Colin Powell, will be an AGENT of SPEECH-ACT, and Congress, like United Nations, will be a BENEFICIARY. However, the OntoSearch engine has been used to make far more complex analogical inferences as applied to disambiguation. We hypothesize (and expect to be able to show in the near term) that applying it to fragments, as well as more traditional instances of the elliptical process called Gapping, will actually simplify the treatment of phenomena that have represented significant barriers in non-semantic environments.

We find support for our semantics-based approach to modeling this type of analogy in the work on analogy by Dedre Gentner and colleagues. Their psycholinguistic approach, like ours, bases the description of analogical reasoning on "internal descriptions, as opposed to, for instance, lexical items" (Yan, Forbus and Gentner 2003). Among the preliminary assumptions of their Structure Mapping theory are tenets that are quite compatible with those adopted in OntoSem: "domains and situations are psychologically viewed as systems of objects, object-attributes and relations between objects"; knowledge is represented as "propositional networks of nodes and predicates"; there is a distinction between attributes and relations; both first- and second-order predicates are needed; and the representations are intended to "reflect the way people construe a situation." (Gentner 1983: 156-157). Thus, although we do not wish to overstate the comparison between psychological and computer modeling, it is encouraging that our approach is compatible with, and can potentially benefit from the insights of, research into the how and why of human reasoning.

(5) Colin Powell was appointed Secretary of State.
a. I know.
b. I know that.

This final example is intended to further emphasize the functionally affinity between processing fragments and processing other instances of reference. Under typical

syntactic approaches, both (5a) and (5b) are complete sentences: in (5b) all arguments of *know* are explicitly accounted for, and in (5a) the complement is elided according to the well-known licensing strategy permitted by auxiliaries and modals (Lobeck 1995). However, a full semantic representation of either sentence requires resolution of the content of the complement. In both cases, this need is lexically triggered: in (5a), by the ellipsis of the expected complement of *know*; in (5b) by the use of a demonstrative pronoun, which always requires reference resolution. In neither case are any heuristics provided about the complement, since *that* is a completely unspecified referring expression – essentially, a placeholder. In short, although the methods for triggering reference resolution differ in (5a) and (5b), once the process is launched, it is identical for both cases: seek an EVENT in the preceding context that is semantically compatible with the selectional restrictions of the complement of *know*. All of the reference resolution methods for this case are precisely the same as for the case of fragments: candidate coreferents are compared based on their semantic compatibility with the selecting verb, text distance, and any and all other heuristics from our general inventory that can be brought to bear in the given context.[10]

5. Discussion

In the paper we have argued for an approach to language analysis and language processing that centrally includes semantics and finds unnecessary the splitting of phenomena along theoretically imposed (and largely artificial, we would suggest) lines. Our analysis relies on a theory, a suite of knowledge resources, and a text processing engine that have been under development for two decades. The approach is built upon the belief that both theories and text-processing environments must answer for their own prerequisites. In the case of most systems and theories that assume external prerequisites, those prerequisites are not filled from the outset because they constitute the hardest part of the work; but once the sophistication to achieve the prerequisites has been achieved, it is entirely possible that the original solution will be supplanted.

---

[10] Space does not permit a sufficient description of our multifaceted approach to resolving demonstrative pronouns, but let us mention that our lexicon includes many configurations – defined syntactically and semantically – that strongly suggest one or another coreferent for the demonstrative. For example, to cover input like *Then he said this: "I will never give up."*, in which the coreferent of *this* is predictably the following quoted material, we have a phrasal lexicon entry that expects the syntactic configuration Subject Verb Direct-Object [colon] Quoted-Material, and the semantic constraints that the Subject is a HUMAN, the Verb is a SPEECH-ACT and the Direct-Object is *this/that*. Clearly, one can never exhaustively list all such cases, and generalized heuristics must be relied on in many contexts (like our example (5b)). However, exploiting an inventory of several dozen frequent corpus-attested configurations takes a significant step toward achieving coverage and confidence in the automated analysis of this very difficult aspect of reference resolution (for a nice overview of the history of work on demonstratives and the inherent difficulties in treating them, see Byron 2004).

Naturally, space constraints do not permit discussion of all relevant issues. By limiting the contexts of sample sentences, we elected to forgo showcasing OntoSem's capacity to select from among competing candidate coreferents using an extended preceding context—a capability that, while always under refinement, has been positively evaluated. We also did not discuss syntactic constraints on fragments, which has been of significant interest in the "syntax vs. pragmatics" debate. This is primarily because our current applications (we are a strongly application-oriented group) focus on analysis rather than generation; moreover, in pursuing analysis of open text, we are treating ill-formed ("unexpected") input as well as grammatical input, meaning that well-formedness is of relatively less importance for us. When we turn to generation, all syntactic constraints that have been delineated in the literature will be incorporated into our arsenal of expressive means.

The interpretation of fragments is no more a solved problem than is any other difficult aspect of automated text processing: disambiguation, metaphor, metonymy, ellipsis, implication, language-related reasoning… We approach it in the context of all of these other phenomena, acknowledging that semantics and world models are necessary *in general* and therefore should be leveraged in this domain as well.

Our research results have both theoretical and practical implications. From a purely theoretical standpoint, the algorithms we describe build upon the well-known contributions cited above while offering enhancements due to the availability of the tools for semantic analysis supplied by the theory of Ontological Semantics. Thus, we argue that our approach is fruitful not only because it is implementable, but also because it more closely models actual language processing than other competing analyses – or so we hypothesize. From a practical standpoint, the work is implementable using resources and processors that currently exist and are growing in real time. Since we have not completed implementation of the microtheory of fragments, full evaluation has not been carried out. (For earlier evaluation efforts and a broad perspective on the rationale behind and evaluation of our system, see Nirenburg, Beale and McShane 2004 and McShane et al. 2005b). The OntoSem knowledge bases are available to the research community and collaboration with other research teams is welcome.

12. Bibliography/References

Barton, Ellen, *Nonsentential Constituents: A Theory of Grammatical Structure and Pragmatic Interpretation*, Amsterdam and Philadelphia, John Benjamins, 1990.

Barton, Ellen, "Nonsentential constituents and theories of phrase structure," *Views on Phrase Structure*, ed. Katherine Leffel and Denis Bouchard, Dordrecht, Kluwer, 1991, p. 193-214.

Barton, Ellen, "Competing approaches to small utterances: An ellipsis analysis vs. a nonsentential analysis," *The Syntax of Nonsententials*, ed. Ljiljana Progovac, Kate Paesani, Eugenia Casielles and Ellen Barton, Amsterdam and Philadelphia, John Benjamins, in press.

Beale, Stephen, Sergei Nirenburg and Marjorie McShane, "Just-in-time grammar," *Proceedings of the 2003 International Multiconference in Computer Science and Computer Engineering*. Las Vegas, Nevada, 2003.

Beale, Stephen, Benoit Lavoie, Marjorie McShane, Sergei Nirenburg and Tanya Korelsky, "Question answering using Ontological Semantics," *Proceedings of ACL-2004 Workshop on Text Meaning and Interpretation*, Barcelona, Spain, 2004.

Beale, Stephen, Sergei Nirenburg and Kavi Mahesh, "Semantic analysis in the Mikrokosmos machine translation project," *Proceedings of the 2nd Symposium on Natural Language Processing*, Bangkok, Thailand, 1995.

Byron, Donna K., Resolving Pronominal Reference to Abstract Entities**,** University of Rochester Computer Science Department Technical Report #815, January 2004.

Cristea, Dan, Postolache, Oana-Diana, "How to deal with wicked anaphora," ed. António Branco, Tony McEnery and Ruslan Mitkov, *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, Current Issues in Linguistic Theory, Benjamin Publishing Books, 2005, p. 17-46.

Elugardo, Reinaldo and Robert Stainton, Introduction, Reinaldo Elugardo and Robert Stainton (eds.) *Ellipsis and Non-Sentential Speech*, Dordrecht: Kluwer, 2005, p. 1-26.

Fikes, Richard, Jessica Jenkins and Gleb Frank, "JTP: A system architecture and component library for hybrid reasoning," *Proceedings of the Seventh World Multiconference on Systemics, Cybernetics, and Informatics*, Orlando, Florida, USA, 2003.

Gentner, Dedre, "Structure-Mapping: A theoretical framework for analogy," *Cognitive Science* 7: 155-170, 1983.

Ginzburg, Jonathan and Ivan A. Sag, *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*, CSLI Publications, 2001.

Jackendoff, Ray, *Foundations of Language: Brain, Meaning, Grammar, Evolution*, Oxford University Press, 2002.

Kehler, Andrew, *Coherence, Reference, and the Theory of Grammar*, CSLI Publications, 2002.

Lobeck, Anne, *Ellipsis: Functional Heads, Licensing, and Identification*, Oxford University Press, 1995.

McShane, Marjorie, *A Theory of Ellipsis*, Oxford University Press, 2005.

McShane, Marjorie, Stephen Beale and Sergei Nirenburg, "Some meaning procedures of Ontological Semantics," *Proceedings of LREC '04*, 2004.

McShane, Marjorie, Sergei Nirenburg, Stephen Beale and Thomas O'Hara. Semantically rich human-aided machine annotation. Proceedings the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, ACL-05, Ann Arbor, June 2005, p. 68-75 (2005a).

McShane, Marjorie, Sergei Nirenburg and Stephen Beale, "Text-meaning representations as repositories of structured knowledge," *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, Barcelona, 9-10 December 2005 (2005b).

Merchant, Jason, "Fragments and ellipsis," *Linguistics and Philosophy* 27(6):661-738, 2004.

Morgan, Jerry, "Sentence fragments revisited," *CLS Parasession on Language in Context*, Chicago, Chicago Linguistics Society, 1989, p. 228-241.

Nirenburg, Sergei and Victor Raskin, *Ontological Semantics*, the MIT Press, 2004.

Nirenburg, Sergei, Stephen Beale and Marjorie McShane, "Evaluating the Performance of the OntoSem Semantic Analyzer," Proceedings of the ACL Workshop on Text Meaning Representation, 2004.

Onyshkevych, Boyan, An Ontological-Semantic Framework for Text Analysis, Unpublished Ph.D. thesis, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA, 1997.

Safir, Kenneth J. *The Syntax of Anaphora*, Oxford University Press, 2004.

Schabes, Yves, Anne Abeillé and Aravind K. Joshi, "Parsing strategies with 'lexicalized' grammars: Applications to tree-adjoining grammars," *Proceedings of COLING-88*, Budapest, Hungary, 1988.

Stainton, Robert, "In defense of non-sentential assertion," ed. Z. Szabo, *Semantics vs. Pragmatics*, Oxford: Oxford University Press, 2004, p. 383-457.

Stainton, Robert. "Neither fragments nor ellipsis," ed. L. Progovac et al., *The Syntax of Nonsententials*, Philadelphia: John Benjamins, forthcoming.

Yan, J., Forbus, K., and Gentner, D., "A theory of rerepresentation in analogical matching," *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society*, 2003.

XTAG-Group. A lexicalized tree adjoining grammar for English. www.cis.upenn.edu/~xtag.