OntoSem Methods for Processing Semantic Ellipsis

Marjorie McShane, Stephen Beale and Sergei Nirenburg

Institute for Language and Information Technologies
University of Maryland Baltimore County
{marge,sbeale,sergei}@umbc.edu

Abstract

This paper describes various types of semantic ellipsis and underspecification in natural language, and the ways in which the meaning of semantically elided elements is reconstructed in the Ontological Semantics (OntoSem) text processing environment. The description covers phenomena whose treatment in OntoSem has reached various levels of advancement: fully implemented, partially implemented, and described algorithmically outside of implementation. We present these research results at this point – prior to full implementation and extensive evaluation – for two reasons: first, new descriptive material is being reported; second, some subclasses of the phenomena in question will require a truly long-term effort whose results are best reported in installments.

1 Introduction

Syntactic ellipsis – the non-expression of syntactically obligatory elements – has been widely studied in computational (not to mention other branches of) linguistics, largely because accounting for missing syntactic elements is a crucial aspect of achieving a full parse, and parsing is required for many approaches to NLP.[1] Much less attention has been devoted to what we will call semantic ellipsis, or the non-expression of elements that, while not syntactically obligatory, are required for a full semantic interpretation of a text.[2] Naturally, semantic ellipsis is important only in truly knowledge-rich approaches to NLP, which few current non-toy systems pursue.

All definitions of ellipsis derive from a stated or implied notion of completeness. Taking, again, the example of syntactic ellipsis, this means that obligatory verbal arguments must be overt, auxiliary verbs must have complements, etc. – all of which is defined in lexico-grammatical terms. But even if a text is devoid of syntactic gaps, much remains below the surface, easily interpretable by people but not directly observable.

Typical examples of semantically underspecified elements are pronouns and indexicals (e.g., *here, now, yesterday*), whose real-world anchors must be clarified in a fully developed semantic representation (i.e., *yesterday* has a concrete meaning only if one knows when *today* is). Pronouns and indexicals, though often difficult to resolve, have one advantage over the cases to be discussed here: the *trigger* that further semantic specification need be carried out is the word itself, and the inventory of such words is well known.

By contrast, the semantically underspecified cases in the following examples are more subtle:

(1) After boosting employment the past few years, Aluminum Co. of America won't be doing any hiring this fall beyond replacing those who leave.

(2) Mitchell said he planned to work late tonight to complete the legislation.

(3) Civilians invited into the prison by the administration to help keep the peace were unable to stanch the bloodshed.

The categories of semantic ellipsis illustrated by these examples can be described as follows. (1) shows *reference resolution that relies on the reconstruction of a semantically elided category*: i.e., to understand who *those* refers to, one must understand that the implicit object of *hire* is 'employees', and that the elided head of the NP with *those* as its determiner also refers to employees (albeit a different real-world set of employees). (2) illustrates *semantic event ellipsis in configurations containing modal/aspectual + OBJECT*: i.e., the meaning

[1] Examples of NLP efforts to resolve syntactic ellipsis include Hobbs and Kehler 1997; Kehler and Shieber 1997; and Lappin 1992, among many others.

[2] Some of the types of semantic underspecification treated here are described in the literature (e.g., Pustejovsky 1995) in theoretical terms, not as heuristic algorithms. This is due, in large part, to a lack of knowledge sources for semantic reasoning in those contributions.

of *complete the legislation* is actually *complete writing the legislation*. (3) illustrates *lexical patterns with predictable event ellipsis*: e.g., *invite <person> to <location>* means 'invite someone to come/go to the location.' These examples, which illustrate the types of semantic ellipsis to be discussed below, require special treatment in our ontological semantic (OntoSem) text processing system, since its goal is to automatically produce fully specified semantic representations of unrestricted text that can then be used in a wide variety of applications.

2 A Snapshot of the OntoSem Environment

OntoSem 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 5,500 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 to "meaning procedures" (i.e., programs that carry out procedural semantics, see McShane et al. *forthcoming*) when applicable. 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, time, etc. The current English lexicon contains approximately 12K senses, including all closed-class items and the most frequent verbs, as indicated by corpus analysis.
- An onomasticon, or lexicon of proper names, which contains approximately 350,000 entries and is growing daily using automated extraction techniques.
- 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 text analyzers, which cover preprocessing, syntactic analysis, semantic analysis, and creation of TMRs.
- The TMR language, which is the metalanguage for representing text meaning.

A very simple example of a TMR, reflecting the meaning of the sentence *The US won the war*, is as follows:

WIN-3
  AGENT NATION-213
  THEME WAR-ACTIVITY-7

This TMR is headed by a WIN event – in fact, it is the 3rd instantiation of the concept WIN (WIN-3) in the world "snapshot" being built during the processing of the given text(s). Its agent is NATION-213, which refers to the United States of America in our fact repository. The theme of the event is the 7th instantiation of WAR-ACTIVITY in this text. Details of this approach to text processing can be found, e.g., in Nirenburg and Raskin 2004, Beale et al 2003, Nirenburg et al 2003a,b. The ontology itself, a brief ontology tutorial, and an extensive lexicon tutorial can be viewed at http://ilit.umbc.edu.

Since OntoSem text processing attempts to do it all – meaning that any phenomenon in any language we are processing is within the purview of our approach – work on any given problem is carried out in spiral fashion: first at a rough grain size, then at a finer grain size with each iterative improvement of the system. In order both to drive and to organize work, we develop a "microtheory" for each aspect of text processing we treat: e.g., we have microtheories of mood, time, reference resolution, and many more. One of the benefits of conceiving work on a given topic in terms of a microtheory is that conceptual, algorithmic progress can occur separately from its realization in a specific application. This does not imply a disconnect between algorithms and implementations – quite the opposite: all algorithms are devised for the OntoSem environment, relying on the types of knowledge and processing it can currently provide or realistically promises to provide. Within this framework, a "big picture" of long-term work on a given topic is often clarified before all details of implementation, or complete knowledge support, become available.

In this paper we present initial results of our work on the microtheory of semantic ellipsis and underspecification, some of whose contributing phenomena can currently be well-handled in OntoSem and others of which will require long-term research and development efforts.

3 Reference Resolution that Relies on the Reconstruction of a Semantically Elided Antecedent

The reference resolution task in NLP has widely come to be understood in very narrow terms – as linking pronouns to their overt textual antecedents (a focus fueled by MUC and other similar competitions; see Sundheim

1995). However, the scope of reference-related problems is actually much broader (see, e.g., McShane and Nirenburg 2002 and McShane *forthcoming*). In this section we describe a number of cases in which reference resolution requires knowledge of semantically elided categories. That is, we are not talking simply about recovering a semantically elided category in its own right, we are talking about recovering it in order to support the correct analysis of another category in the text.

Consider the challenge of resolving the reference of *those* in example (1): *After boosting employment the past few years, Aluminum Co. of America won't be doing any hiring this fall beyond replacing those who leave*. 'Those' refers to an unspecified set of employees. The ellipsis of the head noun *employees* (or any synonym of it) is licensed by the fact that the notion of 'employees' is implicitly introduced into the discourse by the use of the word *hire* in the preceding clause (in the way described below). The real-world set of employees instantiated by the verb *hire* is not the same as the real-world set of employees referred to by the "*those*" NP. However, as this corpus-derived example shows, coreference at the level of concepts rather than instances can, in fact, license ellipsis.[3]

Most reference resolution programs rely on shallow, stochastic methods and limit potential antecedents to overt textual elements; such programs would fail to resolve this case of reference. The OntoSem reference resolution programs, by contrast, include ontological knowledge in the search space of antecedents and, accordingly, can resolve such references. To make clear how this is done, a few more words about ontological specification and TMRs are necessary.

Fillers of properties in the OntoSem ontology can be concepts, literals, numbers or ranges of numbers. A small excerpt from the ontological specification of HIRE is as follows.

HIRE		
AGENT	sem	SOCIAL-ROLE
	default	BUSINESS-ROLE
	relaxable-to	CORPORATION
THEME	sem	SOCIAL-ROLE
LOCATION	sem	PLACE
	default	BUILDING

The fillers for ontological properties can be specified on various *facets,* including: *sem*, which indicates typical selectional restrictions; *default,* which indicates the default filler(s), if any (i.e., this is more tightly constrained than *sem*); and *relaxable-to,* which shows acceptable relaxation of typical selectional restrictions. So, whereas the most typical AGENT of hiring is somebody in a business role (children of BUSINESS-ROLE include MANAGER, CHAIRMAN, VP-CORPORATION and others) it is perfectly normal for any person in a social role to hire someone (e.g., I, as a homeowner, can hire a gardener), and even corporations can be metonymically said to hire people. As concerns the THEME of hiring, it is always a person in a social role, and no defaults or extensions to that specification are required. (Note that SOCIAL-ROLE is a child of HUMAN in the ontology.)

When a concept is instantiated in a TMR, its entire description becomes part of the TMR, and any property fillers actually provided by the text are indicated using the *value* facet. Fillers on the value facet are appended with an instance number, just like the main concept being instantiated. So an excerpt from the instantiation of HIRE (minus over a dozen properties that are not assigned specific values from the text) in the TMR for sentence (1) is as follows, with information explicit in the text shown in boldface:

HIRE-47		
AGENT	sem	SOCIAL-ROLE
	default	BUSINESS-ROLE
	relaxable-to	CORPORATION
	value	**CORPORATION-4165**
THEME	sem	SOCIAL-ROLE

In other words, the fact that certain properties of a concept are not overtly mentioned in a text does not mean that the properties themselves or information about their typical fillers is stricken from the TMR: this information is available in the TMR, just as it is available to a person when he is interpreting a text.

The OntoSem algorithm for resolving the reference of *those* can be briefly outlined as follows:

1. From the list of candidate antecedents that is generated during the processing of each sentence, exclude those with incompatible grammatical features (in this case, those in the singular).
2 Compare potential antecedents using (a) weighted heuristics of the same type as are used in most stochastic reference resolution programs, based on features such as text distance, grammatical function, etc, and (b) comparison of the semantic similarity between *those* (as suggested by the selectional restrictions imposed by its selecting verb) and each antecedent.

The two key differences between our approach and stochastic ones are that, for us, semantic comparison is a heavily weighted heuristic, and implicit properties of

[3] It is noteworthy that many elliptical phenomena permit matching at the conceptual rather than instance-based level. For example, in Russian one can say the equivalent of *They were selling stocks at a good rate so I bought,* in which case the direct object of 'bought' is elided and understood to represent some subset of the original set of stocks being sold (see McShane *forthcoming* for details).

TMR-instantiated concepts are accessible in the search space. In example (1), this means that the THEME of HIRE, which is semantically specified as SOCIAL-ROLE, is a potential source of the semantics of *those.* Since there are no other viable candidates to supply the elided semantic content, SOCIAL-ROLE will be understood as the conceptual head of the NP whose determiner is *those*.

Continuing with the example of HIRE, consider example (4) which, like all examples cited in this paper, was drawn from a news corpus.

(4) Although early placement statistics show that hiring by Wall Street has declined dramatically, students are not exactly flocking to the factory floor. For example, preliminary statistics show that hiring by investment banks has been cut in half, from 22% of graduates in 1987 to 11% this year.

The practical need for resolving the semantic ellipsis of the theme of *hire* in this passage becomes clear when one seeks to interpret the phrase *from 22% of graduates in 1987 to 11% this year*. Syntactically speaking, this phrase is difficult to parse, as it is appended to the main clause in a rather "telegraphic" way: i.e., it is doubtful that most parsers have a rule to specifically target this sentence structure (ours does not). Interpreting this phrase relies primarily on semantics, i.e., an understanding that the graduates are coreferential with the semantically elided object of *hire.*

In OntoSem, difficult cases of parsing are handled using what we call "recovery" procedures. If a perfect parse cannot be arrived at in the initial run of the parser – where the most typical syntactic dependency structures are sought – the parser can invoke several levels of recovery rules, as needed (see Beale et al. 2003 for details). Among these recovery rules is the option to apply the semantics of a constituent to the nascent TMR without recourse to its syntactic function. This type of recovery reflects our general desire to leverage semantic knowledge more and rely on syntax less.

An excerpt from the core of the TMR for the second sentence in (4) will look as follows (with COMMERCIAL-BANK-8 representing the string *investment banks*):

HIRE-50

AGENT	sem	SOCIAL-ROLE
	default	BUSINESS-ROLE
	relaxable-to	CORPORATION
	value	**COMMERCIAL-BANK-8**
THEME	sem	SOCIAL-ROLE

And the TMR for the syntactically unattached component, *from 22% of graduates in 1987 to 11% this year*, will look as follows (*from* and *to* have lexical senses that indicate the start-value and end-value of a range when their complement is a number):

START-VALUE		
DOMAIN	SOCIAL-ROLE-977	
	AGENT-OF	GRADUATE-COLLEGE
RANGE	.22	
YEAR	1987	
END-VALUE		
DOMAIN	SOCIAL-ROLE-978	
	AGENT-OF	GRADUATE-COLLEGE
RANGE	.11	
YEAR	*find-anchor-year* ; a call to a procedural semantics program	

In short, the head of the core TMR expects a SOCIAL-ROLE as the THEME of HIRE, and the domain of the syntactically unattached segment of the sentence is namely a SOCIAL-ROLE. The direct, and correct, hypothesis is to link the unattached TMR fragment namely to the filler of the THEME of HIRE, which is exactly what our semantic analyzer does.

Cases in which semantically elided elements are crucial to the interpretation of other sentence elements are not rare. Another example taken from the same domain of hiring (we remain in this domain only for simplicity of exposition) is shown in (5).

(5) For one thing, in 20 states and the District of Columbia, it's illegal to discriminate in hiring or promotions on the basis of marital status.

In order to interpret the connection of *marital status* to the rest of the proposition, one must corefer the HUMAN in the DOMAIN of the concept MARITAL-STATUS to the implicit THEME of HIRE and PROMOTE.

LEGALITY-ATTRIBUTE-4

DOMAIN	sem	SOCIAL-EVENT
	value	**DISCRIMINATE-23**
RANGE	sem	YES, NO
	value	**NO**

DISCRIMINATE-23

AGENT	sem	HUMAN
	relaxable-to	CORPORATION
		ORGANIZATION
THEME	sem	MENTAL-OBJECT
	value	**HIRE-65**
		PROMOTE-53
BENEFICIARY	sem	HUMAN
	relaxable-to	CORPORATION
		ORGANIZATION
CAUSED-BY	sem	EVENT
	VALUE	**MARITAL-STATUS-1**

MARITAL-STATUS-1
DOMAIN HUMAN
RANGE SINGLE, MARRIED, WIDOWED, DIVORCED

As these examples show, there is a concrete, corpus-attested need to resolve many instances of semantic ellipsis, namely, the need to use implicit information as the antecedent for coreferring categories.

4 Semantic Event Ellipsis in Configurations Containing a Modal/Aspectual + OBJECT

In English and many other languages, modals and aspectuals can take nominal complements. Those complements can, semantically, be of two types: OBJECTs and EVENTs. If the syntactic object semantically represents an EVENT, then there is no semantic ellipsis, as in *The delegates began the **conversation** at noon*, whose simplified TMR is as follows:

SPEECH-ACT-35333
PHASE begin
AGENT DELEGATE-2223
TIME 12.00

In other words, since *conversation* is mapped to the event SPEECH-ACT, it naturally has a PHASE and an AGENT and a TIME and there is no semantic ellipsis. Examples of this type are frequent in texts, as shown by examples (5)-(7):

(5) Dataproducts has since started a restructuring that started the still-raging bidding wars

(6) Nomura started a credit-card venture with American Express Co.

(7) The spokesman said Maxicare hopes to complete the reorganization by early 1990

If the syntactic object semantically represents an OBJECT, then the semantics of the implied verb must be recovered. For OntoSem text processing, two subtypes of such cases are important: those in which the object refers to an institution, program, etc., and the elided verb predictably means "initiate, found", and those in which the object refers to something else and the verbal semantics must be inferred based on the meaning of the overt categories. Examples of the first subtype include the following:

(8) She'll be the first one to leave it and start a fourth party.

(9) Brazil started an ethanol program about 15 years ago.

(10) Quebecor started the Philadelphia Journal.

The OntoSem lexicon contains a number of lexical senses of *start, finish,* etc. that cover such cases: e.g., one sense specifies the THEME to be an ORGANIZATION, and heads the semantic description with the concept FOUND-ORGANIZATION; another specifies the THEME to be a MENTAL-OBJECT and heads the semantic description with INVENT (as in 'He started a new semantic theory'). This type of semantic ellipsis is discussed more fully in Section 5.

The second subtype requires procedural semantic analysis to recover the meaning of the implied event. Examples of such contexts include the following:

(11) Mitchell said he planned to work late tonight to complete the legislation [elided WRITE].

(12) He conscripted 700,000 slaves to finish the Great Wall [elided BUILD].

(13) Most service businesses can complete their books within three weeks after a period has ended [elided BOOKKEEPING].

(14) Next Inc.... has finished the first version of its operating-system software [elided DESIGN-SOFTWARE].

(15) Manufacturers Hanover this week started a new series of ads that push "Power Savings" [elided BROADCAST].

The OntoSem lexical sense that covers these contexts includes a procedural attachment called *seek-event-specification*, which attempts to dynamically recover the meaning of the semantically elided events. That is, it seeks concepts for which the meaning of the subject and direct object provided in the text are most specifically constrained. For example, in (11), the program will seek an EVENT for which the default AGENT is SENATOR (Mitchell was a senator at the time[4]) and the default THEME is BILL-LEGISLATIVE; and in (12), the program

[4] We can expect that earlier in the text he was referred to using a more complete appellation which either overtly described him as a senator or provided sufficient information for our reference-resolution program to link him to his fact-repository entry, where his SOCIAL-ROLE of SENATOR is listed. Reference resolution using fact-repository information has been implemented but not widely tested yet. The problem of identifying him as the same person that has just been elected Chairman of Disney is outside of the purview of this paper.

will seek an EVENT for which AGENT is SLAVE and the default THEME is WALL (the basic ontological mapping of *Great Wall*, though a number of properties are defined in its fact repository entry, like LOCATION: China, LENGTH: 5000 km). If more than one match is found, all options are retained in the TMR for possible later disambiguation based on further context. If no matches are found using the *default* facet, matches using the *sem* facet are sought. In the worst case (the maximal level of semantic relaxation), the only thing the semantic analyzer can say about the elided EVENT is that there is, indeed, an unspecified EVENT that has the text-specified AGENT and THEME.

Two points must be emphasized: a) the OntoSem lexicon records our expectations that dynamic semantic-ellipsis resolution will be necessary in certain types of contexts, which can be specified based on reference to ontological types of OBJECTS; and b) the resolution of semantic event ellipsis is supported by the property-defined relationships between ontological OBJECTs and EVENTs.

5 Lexical Patterns with Predictable Event Ellipsis

Ontological semantics has practical aims, which means, among other things, that extending the lexicon to include complex entities and thus bypass the need for their runtime compositional semantic treatment is a valid methodological option. A good case in point is the lexicalization of common cases of semantic ellipsis. Like any lexicalization, this does not offer full coverage; however, like all lexicalization, it does provide concrete information about concrete phenomena that can be immediately exploited. Here we present just a few examples of the lexicalized treatment of semantic ellipsis as an illustration of our omnivorous approach to improving the overall quality of text processing.

The verb *invite*, when followed by a prepositional phrase or adverb indicating location (or destination) directly or metonymically, actually means 'invite to come/go to that place'; the verb of motion is semantically elided. Examples include (16)-(19):

(16) Civilians *invited into the prison* by the administration to help keep the peace were unable to stanch the bloodshed.

(17) "If they *invited* us *back* tomorrow to govern the mainland, frankly we would hesitate," Vice Foreign Minister John H. Chang told a U.S. governor's delegation.

(18) All 13 OPEC oil ministers were *invited to the meeting*.

(19) He often is one of a handful of top aides *invited into the Oval Office* for the informal sessions at which President Bush likes to make sensitive foreign-policy decisions.

The lexicon sense that covers this use of *invite* in (16) and (18) is as follows, in presentation format (the lexicon sense that covers (17) has an adverb of location/destination instead of a PP):

```
invite-v2
  def  "+ pp of destination, implies 'invite to come'"
  ex "She invited him to Paris"

  syn-struc
    subject          root $var1      cat n
    v                root $var0
    directobject     root $var2      cat n
    pp-adjunct       root $var3      cat prep
                     root (or to onto into on)
                     obj     root $var4   cat n

  sem-struc
    INVITE
        AGENT        value ^$var1
        THEME        MOTION-EVENT
                        DESTINATION  value ^$var4
                        AGENT  value ^$var2
     ^$var3  null-sem +
```

The syntactic structure (syn-struc) says that this sense of like requires a subject, direct object and PP, and that the PP must be headed by the word *to, onto, into* or *on.* The semantic structure (sem-struc) is headed by an INVITE event, whose AGENT is the subject of the clause (note the linked variables) and whose theme is a MOTION-EVENT. The AGENT and DESTINATION of the MOTION-EVENT are the meanings of the direct object and prepositional object, respectively, of the input clause. (We gloss over formal aspects of the entry that are tangential to the current discussion.) Note that there is no verb of motion in the input text: MOTION-EVENT is lexically specified since it is a predictable semantically elided aspect of meaning in the given configuration.

Another lexical item for which we can predict a particular type of semantic ellipsis is *forget*. When the direct object of *forget* semantically represents a PHYSICAL-OBJECT, there is an elided TAKE event, as shown in (20).

(20) "This is the slowest day I've seen this year," said Peter Canelo, a market strategist at Bear Stearns Cos. "I've only had one call all day from a real investor and he just forgot his umbrella."

Thus, the OntoSem lexicon has a special sense for *forget* + PHYSICAL-OBJECT that is selected by the semantic analyzer in contexts like (20).

Obviously, a lexicon that anticipates instances of semantic ellipsis must be quite detailed and, as a result, relatively expensive to build. The OntoSem lexicon falls into both of these categories. However, expensive does not mean prohibitive, and we believe that the ultimate utility of such a knowledge resource will fully justify its compilation. The rate of acquisition for open-class words and phrases in OntoSem depends primarily on the type of entity being acquired, be it argument-taking or not. A conservative estimate for lexical acquisition for OntoSem, based on a recent acquisition drive, is as follows:

- acquisition of argument-taking word and phrase senses: 6 words/hr * 6 hrs./day * 5 days/week * 50 weeks/yr = 9,000 senses/year
- acquisition of non-argument-taking word and phrase senses (about 5 times as fast): 9000 * 5 = 45,000 senses/year

According to these estimates, and considering that many more words are non-argument-taking than are argument-taking, we might realistically expect to increase the size of the lexicon by around 100,000 senses per year if given 3 full-time acquirers supported by one full-time ontology developer. In short, large volumes of high-quality knowledge can be achieved in real time.

6 Evaluation

In response to the current evaluation standards in NLP (which are more suited to and informative for stochastic-based systems than knowledge-based ones), we have recently developed a novel evaluation methodology that assigns scores as well as blame for errors to various aspects of the TMRs generated during OntoSem text processing. While percentage scores for correct vs. incorrect results can provide a general evaluation of system function, it is blame assignment that drives development. Blame assignment is determined by processing each sentence multiple times: first without manual intervention, then with the correction of preprocessor errors, then with the correction of syntax errors. The rationale behind these loops of correction and reevaluation is that "low level" mistakes like preprocessor errors or lack of coverage of some syntactic construction require different development action than more weighty (from our point of view) errors in semantic interpretation that might result from gaps in knowledge, insufficient reasoning engines, etc.

The first experiment with our new evaluation regime produced the following results (reported on in detail in Nirenburg et al, 2004): the analyzer was shown to carry out word sense disambiguation at over 90% and semantic dependency determination at 87% on the basis of correct syntactic analysis and on sentences of an average length of over 25 words with 1.33 unknown words on average per input sentence. Outstanding errors in semantic analysis were due, in most cases, to non-literal use of language (which is one of our topics of ongoing investigation). Although this first formal experiment was limited to WSD and semantic dependencies, testing of other modules – like those for reference resolution and ellipsis – will soon be added to the formal evaluation regime. At this stage, evaluation work is slow, but we are well into the development of an evaluation and correction environment that promises to significantly speed up both evaluation and system enhancement.

7 Closing Thoughts

The type of work presented in this paper might be termed a practical, progressive long-term effort.

The work is **practical** because it is being carried out within a working system that: (a) uses non-toy, real text-oriented knowledge resources – lexical and ontological – that are being built not in the hope that someday some system might be able to use them, but because they are useful right now in the system under construction; (b) has processors that cover all levels of text analysis, from preprocessing raw input text to creating semantic text-meaning representations of it; (c) has been and continues to be used in applications as diverse as machine translation, information extraction, summarization and question answering. In short, the work that we carry out on any given aspect of text processing answers a *need* encountered in real applications, and does so in a concrete, implemented and further implementable way.

The work is **progressive** in the sense that the loop of algorithm development and integration in each new version of the working system is continuous. We find it important, tactically, to view issues in natural language from a broad perspective first, with development of practical "microtheories" for their treatment progressing as need demands and resources permit. What we try *not* to do is artificially exclude from our purview those aspects of phenomena that are not easily treated at the present state of the art. Instead, we include such aspects in our algorithms to the degree possible and make sure that they are modified as soon as an advance is made in resource acquisition or algorithm fusion (e.g., incorporating stochastic methods if and when knowledge-based ones fail to produce a single, unambiguous semantic representation, as in the case of weighted heuristics for reference resolution).

The work is **long term** because we know that high-quality text processing cannot be achieved in the short term. If a phenomenon exists in a language we are processing, it is, by definition, within our purview. Our ul-

timate aim: an intelligent agent able to communicate no less fluently than you or I and in possession of human-level background knowledge about the world and language. Of course, this goal will not be realized in our lifetimes, unless adequate resources are allocated to this task and its subtasks. However, a solid foundation that in principle can accommodate any and all later needs of language processing is what we are attempting to develop while at the same time developing working applications.

References

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

Jerry R. Hobbs and Andrew Kehler. 1997. A theory of parallelism and the case of VP ellipsis. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics,* Madrid, Spain.

Andrew Kehler and Stuart Shieber. 1997. Anaphoric dependencies in ellipsis. *Computational Linguistics*, 23(3): 457-466.

Shalom Lappin. 1992. The syntactic basis of ellipsis resolution. *Proceedings of the Stuttgart Ellipsis Workshop*, 1-47.

Marjorie McShane. Forthcoming. *A Theory of Ellipsis.* Oxford University Press.

Marjorie McShane, Stephen Beale and Sergei Nirenburg. Forthcoming. Some meaning procedures of Ontological Semantics. *Proceedings of LREC 2004*, Lisbon, Portugal.

Marjorie McShane and Sergei Nirenburg. 2002. Reference and ellipsis in Ontological Semantics. *Memoranda in Computer and Cognitive Science,* MCCS-02-329. The Computing Research Laboratory, New Mexico State University.

Sergei Nirenburg, Marjorie McShane and Stephen Beale. 2003a. Enhancing recall in information extraction through Ontological Semantics. *Proceedings of the Workshop on Ontologies and Information Extraction*, Bucharest, Romania, August 2003.

Sergei Nirenburg, Marjorie McShane and Stephen Beale. 2003b. Operative strategies in Ontological Semantics. *Proceedings of HLT-NAACL-03 Workshop on Text Meaning*, Edmonton, Alberta, Canada, June 2003.

Sergei Nirenburg and Victor Raskin. 2004 (forthcoming). *Ontological Semantics,* the MIT Press, Cambridge, Mass.

Sergei Nirenburg, Stephen Beale and Marjorie McShane. 2004. Evaluating the Performance of the OntoSem Semantic Analyzer. Submitted to *ACL-04.*

James Pustejovsky. *The Generative Lexicon.* The MIT Press, Cambridge, Mass.

Roger Schank and Robert Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures.* L. Erlbaum Associates, New York.

Beth Sundheim. 1995. The MUC coreference task definition v. 3.0. *Proceedings of the 6th Message Understanding Conference*.