The Ontological Semantic treatment of multiword expressions

Marjorie McShane, Sergei Nirenburg and Stephen Beale
Rensselaer Polytechnic Institute

Introduction

Within the fields of linguistics and natural language processing (NLP) — as in many domains of science — it is common to carve large phenomena into pieces and pursue each in isolation. Although this methodology can be supported by scientific, pragmatic and sociological rationales, we have found such splitting to be both counterintuitive and counterproductive in building the language-endowed intelligent agents we call OntoAgents. OntoAgents are multi-functional, cognitively modeled agents that are being designed to collaborate with people in dialog applications. They have simulated bodies and simulated minds, and are required to process language, learn, make decisions, and carry out simulated action, thus emulating a wide variety of human behaviors. The knowledge substrate and reasoning engines used to carry out these tasks are tightly integrated, leading to a natural preference for seeking similarities between, rather than drawing hard lines around, the different modules contributing to agent simulation and cognition.

This paper describes the unified treatment of two aspects of language analysis that have typically been treated separately by the NLP community: compositional semantics and the processing of multiword expressions (MWEs). But our point is not only that our OntoSem[1] language analysis system treats these phenomena similarly; we also define the very notion of "treatment" atypically. Whereas mainstream NLP has, over the past couple of decades, preferred to pursue the shallow analysis of both compositional semantics and MWEs,[2] we pursue deep,

1. OntoSem refers to the language processing components of OntoAgent. OntoSem is based on the theory of Ontological Semantics (Nirenburg and Raskin, 2004).

2. An example of shallow compositional semantics is semantic role labeling (e.g., Gildea and Jurafsky, 2002). An example of shallow MWE analysis is automatically detecting MWEs in support of syntactic parsing.

Lingvisticæ Investigationes 38:1 (2015), 73–110. DOI 10.1075/li.38.1.03mcs
ISSN 0378–4169 / E-ISSN 1569–9927

ontologically-grounded semantic analysis of all input. In addition, whereas mainstream NLP uses primarily knowledge-lean methods (for an overview, see Schone and Jurafsky, 2001), we use primarily knowledge-based methods. Finally, whereas mainstream NLP seeks near-term, albeit partial, solutions to natural language problems, we seek longer-term, more complete solutions.

The scope of the current analysis is the lexical recording and automatic processing of MWEs. The lexicon, along with its linked ontology, are developed in coordination with the text analyzer, so the utility of the knowledge bases can be judged by their ability to support accurate and sufficient language analysis.

The paper is organized as follows. Sections 1–3 serve as background, presenting select related work, an overview of the OntoAgent cognitive architecture, and a description of compositional semantic analysis in OntoSem. Section 4 details how this approach to compositional semantics naturally extends to MWEs; in fact, we will show that there is no clear line between compositional semantic analysis and MWE analysis when viewed from a computational-semantic perspective. Section 5 presents the results of a system evaluation that, we believe, shows the great promise of this approach. Section 6 concludes the paper with a perspective on how this work can contribute to an overdue shift in the focus of mainstream NLP, away from primarily syntactic analysis to the fundamental treatment of meaning.

1. Related work

The past decade has witnessed a boom of interest in MWEs as shown, for example, by the dozens of workshops on MWEs at NLP conferences.[3] To generalize, two threads of investigation have predominated. Most numerous are contributions reporting statistical approaches to detecting or translating MWEs. In these contributions, the meaning of MWEs is not addressed: even a correct translation achieved using statistical methods does not imply that the MWE has been understood in a way that would support intelligent agent reasoning.

The other main line of work is knowledge-oriented, and involves either classifying MWEs or recording them in lexicons. As regards classification, a cornerstone of theoretical, descriptive and computational work has been the attempt to understand to what extent idioms are fixed and to what extent they are flexible. For

3. Conferences hosting such workshops include those sponsored by ACL (The Association for Computational Linguistics), NAACL (The North American Chapter of the Association for Computational Linguistics), and EACL (The European Chapter of the Association for Computational Linguistics). Proceedings containing the many dozens of contributions are available online.

discussion, see Cacciari and Tabossi (1993), whose component articles provide particularly felicitous reviews of the literature. As regards lexicon building, this can be carried out expressly in service of NLP, or lexicons originally developed for human use can be subsequently tuned for NLP. We consider each class in turn.

Examples of MWE lexicons developed explicitly for NLP are Walenty, a valence dictionary for Polish (Przepiórkowski et al., 2014); DuELME, an electronic dictionary of multiword expressions for Dutch (Grégoire, 2010); and the MWE enhancement of the Hebrew lexicon reported in Al-Haj et al. (2013). All of these focus exclusively on the syntactic properties of MWEs in support of parsing. A contrastive study of approaches to recording the morphological properties of MWEs is provided in Savary (2008).

The other source of lexical support for NLP is resources that were initially acquired for human-oriented purposes then tuned for NLP. To our knowledge, no such resources have strong coverage of MWEs, but since they provide information about the selectional restrictions of argument-taking words, they can support analysis of some inputs that might be considered MWEs. For example, VerbNet (Kipper et al., 2008), which derives from Levin's (1993) verb classification, has been used to support syntactic analysis as well as coarse-grained semantic analysis. FrameNet (Fillmore et al., 2001), which was initially pursued as a linguistic investigation into frame semantics, has also come to be leveraged in NLP thanks to its extensive descriptions of core and non-core semantic roles, as well as its large inventory of annotated sentences.

Let us briefly compare the OntoSem approach to MWEs with these directions of work. (1) As mentioned above, most automatic MWE detection systems consider detection the end point of processing, though some developers (e.g., Sharoff, 2004 and Venkatsubramanyan and Perez-Carballo, 2004) acknowledge the need to pass off candidate MWEs to lexicographers for manual incorporation into NLP resources. The OntoSem system, by contrast, processes MWEs that have already been lexically recorded, and it pursues the goal of full semantic analysis, including disambiguation. (2) Automatic MWE detection systems often take a supply-side approach, extracting the types of entities (e.g., verb-particle constructions) that are most readily treated using the given methodology. In OntoSem, by contrast, we take a demand-side approach, considering the needs of systems first then responding to them by developing static knowledge resources and processors. (3) Although the OntoSem lexicon is being developed in conjunction with a particular language processing system (which supports evaluation of its utility), the lexicon is generic enough to be employed by other systems as well; this likens it to some of the other development efforts mentioned above (e.g., Przepiórkowski et al., 2014, and Al-Haj et al., 2013).

2. The OntoAgent environment

OntoAgent is a knowledge-based intelligent agent environment inspired by the traditional goals and motivations of artificial intelligence: attempting to achieve human-level behavior by modeling agents capable of perception, reasoning and action. OntoAgents include integrated physiological and cognitive simulations, with the latter centrally including natural language processing. A recent prototype application is Maryland Virtual Patient (Figure 1), a clinician training system in which a cohort of virtual patients can be diagnosed and treated in open-ended, interactive simulations that include a virtual tutor and additional virtual medical personnel (see, e.g., McShane et al., 2007, 2012; Nirenburg et al., 2008).

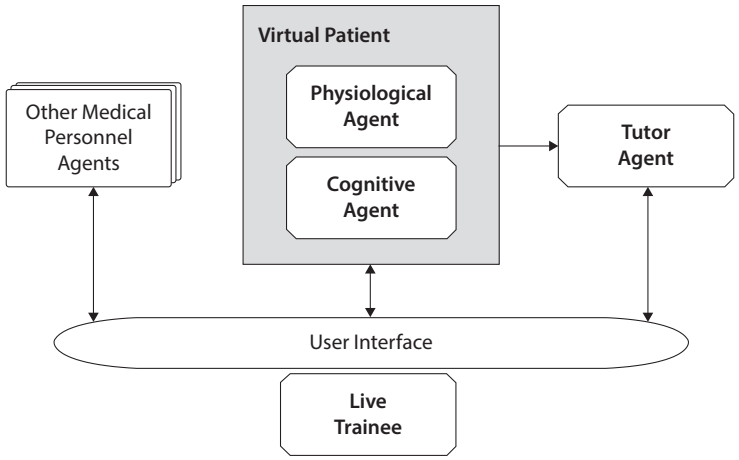

Figure 1. The Maryland Virtual Patient application.

The goal of language analysis in OntoAgent is for the agent to fully understand the meaning of textual input and use that learned information to populate its memory. Its memory then serves as input to reasoning which, in turn, leads to action.

Two tactical decisions make our work on developing deep semantic analysis capabilities both feasible and, we believe, forward-looking. First, OntoAgent is a knowledge-based artificial intelligence environment for which we manually record high-quality, machine-tractable knowledge. We do not frame manual knowledge acquisition as a response to the so-called "knowledge bottleneck" because we do not believe there is a bottleneck. The bottleneck idea emerged from the opinion that it was too costly to manually record knowledge for use in NLP. However, when the same knowledge is to be used not only for NLP but for every other aspect of agent functioning as well — physiological simulation, decision-making and simulated action — then its cost effectiveness is clear. The second key to the feasibility of deep-semantic analysis in OntoAgent is the fact that the analysis

system is embodied in agents that can evaluate their own confidence in language understanding and respond accordingly. For example, whereas high-confidence language understanding will lead directly to reasoning and action, low confidence language understanding might lead to more information gathering, such as asking the human collaborator for clarification. Using instance-level measures of confidence differs from the evaluation metrics often used by the unembodied-NLP community, which revolve around corpus-level statistics for precision and recall.

The OntoAgent approach to semantics is principally and philosophically reminiscent of construction grammar (Fillmore et al., 1988). Within both approaches, it is recognized that language represents an inventory of more or less fixed patterns whose form-to-meaning correlation must often be explicitly recorded. The main difference between OntoSem work and many threads of work in construction grammar is that OntoSem is an implemented language analysis system which permits us to test our hypotheses about the type and extent of knowledge needed to support a given grain-size of semantic analysis.

3. Compositional semantic analysis in OntoSem

The OntoSem approach to natural language understanding follows the theory of Ontological Semantics (Nirenburg and Raskin, 2004). The OntoSem text analyzer takes as input unrestricted natural language text and attempts to generate a disambiguated, ontologically grounded interpretation of it, which we call a **text meaning representation (TMR)**.

Text analysis includes three main stages: (1) preprocessing and syntactic analysis, provided by Stanford CoreNLP (version 3.4.1; Manning et al., 2014); (2) basic semantic analysis, which involves lexical disambiguation and the establishment of the semantic dependency structure (McShane et al., In press); and (3) extended semantic/pragmatic analysis, which involves reference resolution, speech act analysis, and recovery from unexpected input (McShane and Nirenburg, 2013). A high-level view of the architecture is shown in Figure 2. The center module in the picture — which represents basic semantic analysis, resulting in a basic TMR — is the focus of the current discussion.

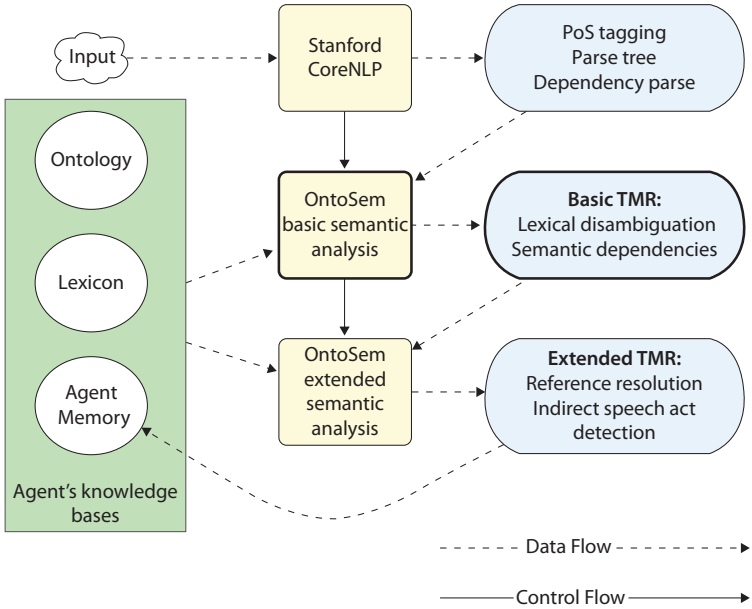

Figure 2. A high-level architecture of OntoSem text understanding. The focus of the current discussion is basic semantic analysis and the generation of basic TMRs.

As an example of basic text analysis, consider the TMR for the input *Charlie watched a baseball game*, shown in Figure 3.

```
VOLUNTARY-VISUAL-EVENT-1
    AGENT                HUMAN-1
    THEME                BASEBALL-GAME-1
    TIME                 (< find-anchor-time)        ; indicates past time
    textstring           "watched"
    from-sense           watch-v1
HUMAN-1
    AGENT-OF             VOLUNTARY-VISUAL-EVENT-1
    HAS-PERSONAL-NAME    "Charlie"
    textstring           "Charlie"
    from-sense           *personal-name*             ; "Charlie" is in the onomasticon
BASEBALL-GAME-1
    THEME-OF             VOLUNTARY-VISUAL-EVENT-1
    textstring           "baseball_game"
    from-sense           baseball_game-n1
```

Figure 3. TMR of the sentence *Charlie watched a baseball game.*

TMRs are written in the metalanguage of the ontology. Each frame is headed by a numbered instance of an ontological concept — either an OBJECT, an EVENT or a PROPERTY. Small caps distinguish ontological concepts from words of English.

Each TMR frame is described by the contextually relevant inventory of relations (including inverses) and attributes. Metadata provides traces of system processing: *textstring* indicates the word that gave rise to the frame, and *from-sense* indicates which lexical sense of the word generated the selected interpretation. For example, it is clear that *baseball game* was recorded as a phrasal in the lexicon because the textstring giving rise to the frame BASEBALL-GAME-1 is "baseball_game", and the lexical sense selected to analyze this lexeme is baseball_game-n1, the first nominal sense of baseball_game.

The ontology from which concepts are drawn is organized as a multiple-inheritance hierarchical collection of frames headed by concepts that are named using language-independent labels. It currently contains approximately 9,000 concepts, most of which belong to the general domain. OBJECTS and EVENTS are described by an average of 16 PROPERTIES, whose values can be inherited or locally specified. The facets *value*, *default*, *sem* and *relaxable-to* allow for recording more and less typical constraints on property values, as shown by the excerpt from the concept DRUG-DEALING, shown in Figure 4.

DRUG-DEALING		
IS-A	value	CRIMINAL-ACTIVITY
AGENT	default	CRIMINAL, DRUG-CARTEL
	sem	HUMAN
	relaxable-to	SOCIAL-OBJECT
THEME	default	ILLEGAL-DRUG
INSTRUMENT	default	MONEY

Figure 4. Excerpt from the ontological frame for DRUG-DEALING.

Since the OntoSem ontology is language independent, its link to any natural language must be mediated by a lexicon. Consider the two verbal senses of *address* shown in Figure 5, which use a slightly simplified formalism for readability.

```
address
    address-v1
        definition   "to talk to, give a speech to"
        example      "He addressed the audience."
        syn-struc
            subject              $var1
            v                    $var0
            directobject         $var2
        sem-struc
            SPEECH-ACT
                AGENT            ^$var1 (sem HUMAN)                    ; ^ indicates the meaning of
                BENEFICIARY      ^$var2 (sem HUMAN) (relaxable-to ANIMAL)
    address-v3
        definition   "to consider, think about"
        example      "He addressed the problem."
        syn-struc
            subject              $var1
            v                    $var0
            directobject         $var2
        sem-struc
            CONSIDER
                AGENT            ^$var1 (sem HUMAN)
                THEME            ^$var2 (sem ABSTRACT-OBJECT)
```

Figure 5. Two senses of the verb *address* in the OntoSem lexicon.

Syntactically (as shown in the syn-struc zones), both senses expect a subject and a direct object in the active diathesis, filled by the variables $var1 and $var2, respectively. However, the meanings of the direct objects are constrained differently, as shown in the respective sem-strucs. In address-v1, the meaning of the direct object (^$var2) is constrained to a HUMAN or, less commonly, an ANIMAL, whereas in address-v3 the meaning of the direct object is constrained to an ABSTRACT-OBJECT.[4] This difference in constraints permits the analyzer to disambiguate. If the direct object in an input sentence is abstract, as in *He addressed the problem*, then *address* will be analyzed as an instance of the concept CONSIDER using address-v3. By contrast, if the direct object is human, as in *He addressed the audience*, then *address* will be analyzed as SPEECH-ACT using address-v1. The semantic roles that each variable fills are explicitly indicated in the sem-struc zone as well: in both of

4. Some of these semantic constraints are actually not listed in the lexicon since they are available in the ontology: e.g., the constraints on the AGENT and BENEFICIARY of a SPEECH-ACT are recorded in the ontological concept SPEECH-ACT. Only constraints that override ontological specifications must be listed explicitly in the lexicon. However, for clarity of presentation, we make explicit in these sample lexical senses the ontological constraints that the system understands to be in effect.

the senses presented here, the meaning of $var1 (^$var1) fills the AGENT role and the meaning of $var2 (^$var2) fills the THEME role.

The examples above illustrate how lexically recorded *semantic* constraints support disambiguation, given the same syntactic structure. However, *syntactic* constraints can also support disambiguation. Consider the 4 senses of *see* shown in Figure 6. The latter two require, respectively, an imperative construction (see-v3) and a transitive construction that includes a PP headed by *to* (see-v4). These syntactic constraints, along with the associated semantic constraints, provide strong heuristic evidence for automatic disambiguation.

see-v1
  definition "to perceive visually"
  example "He saw her new car."
  syn-struc
    subject $var1
    v $var0
    directobject $var2
  sem-struc
    INVOLUNTARY-VISUAL-EVENT
      EXPERIENCER ^$var1 (sem ANIMAL)
      THEME ^$var2 (sem PHYSICAL-OBJECT)

see-v2
  definition "to consult with for advice"
  example Grandma saw her doctor."
    syn-struc
      subject $var1
      v $var0
      directobject $var2
    sem-struc
      PROFESSIONAL-CONSULTATION
        AGENT ^$var2 (sem MEDICAL-ROLE LEGAL-ROLE)
        BENEFICIARY ^$var1 (sem HUMAN)

see-v3
  definition "to refer to a portion of text"
  example "For details, see Chapter 3."
  syn-struc
    v $var0 (form imperative)
    directobject $var1
  sem-struc
    READ
      THEME ^$var1 (sem TEXT-UNIT)

```
see-v4
    definition   "to accompany someone somewhere"
    example      "He saw me to my car."
    syn-struc
        subject                 $var1
        v                       $var0
        directobject            $var2
        pp
            prep                $var3 (root to)
            obj                 $var4
    sem-struc
        ESCORT
            AGENT               ^$var1 (sem HUMAN)
            BENEFICIARY         ^$var2 (sem HUMAN)
            DESTINATION         ^$var4 (sem PLACE) (relaxable-to PHYSICAL-OBJECT)
        ^$var3 null-sem+            ; the meaning of $var3 does not contribute
                                    ; compositionally to the whole
```

Figure 6. Four verbal senses of *see* in the OntoSem lexicon.

A global rule used for disambiguation is, give preference to more specific constraints. In most cases, this rule works well: after all, when one says *I saw my doctor yesterday*, it typically refers to PROFESSIONAL-CONSULTATION — unless, of course, one adds the adjunct *at a basketball game*, in which case INVOLUNTARY-VISUAL-EVENT is the clear choice. As people, we make the latter adjustment based on the knowledge that one consults with physicians in medical buildings, not at basketball games. While such knowledge about where events typically occur is recorded in the OntoSem ontology, we are still working toward compiling a sufficient inventory of reasoning rules to exploit it. As such, we currently use the "prefer more specific constraints" rule for disambiguation despite its known limitations.

3.1 Syntactic parsing as input to semantic analysis

In the theory of Ontological Semantics, syntactic analysis is not an end in itself; instead, it provides heuristic evidence for semantic analysis. In an early implementation of Ontological Semantics, used in the Mikrokosmos machine translation system (Beale et al., 1995), we employed a homegrown syntactic analyzer that had the benefit of being specially suited to the style of lexicon described above; however, it did not have sufficient coverage of complex syntactic structures. For this reason, in about 2005 we integrated the Stanford dependency parser (Klein and Manning, 2003a,b; de Marneffe et al., 2006), which is now available within the Stanford CoreNLP Toolkit (Manning et al., 2014). As developers will surely appreciate, this integration was labor-intensive. Among the requirements was that the

output of the parser be aligned with the expectations recorded in the syn-strucs of lexical senses, even though Stanford and OntoSem use different inventories of parts of speech that reflect a different grain-size of syntactic description.[5] We recorded knowledge about these alignments as follows.

First, we automatically grouped OntoSem senses based on their syntactic dependencies. Groupings include intransitive verbs, transitive verbs, ditransitive verbs, transitive verbs with a PP argument/adjunct, etc. Then we created a generic, manually vetted, OntoSem-to-Stanford mapping for each group. As with all aspects of lexical acquisition in OntoSem, a convenient GUI supported this process.[6] Consider the example shown in Figure 7. (Note that it is a quirk of the interface that the root "time" is listed as the direct object in the screen shot: it applies to only one of the senses in the group, *spend-v2*, and is irrelevant for the mapping process we are describing.)

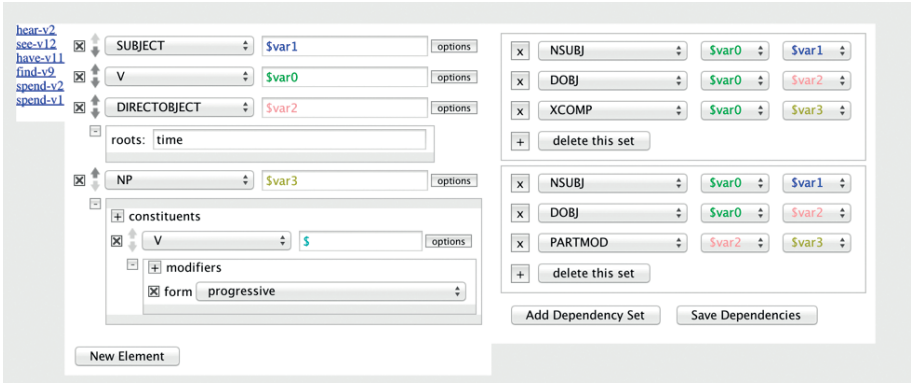

Figure 7. A GUI to support the alignment of expected syntactic dependencies in the OntoSem lexicon with the actual dependencies generated by the Stanford parser.

This syntactic class covers six verb senses (*hear-v2, see-v12*, etc.), all of which expect a subject, a direct object, and a progressive verb form as a complement, as shown by the examples below:

(1) a. *She hears him singing.*
b. *She saw him stealing the bicycle.*
c. *He has trouble doing math.*
d. *I find myself going to bed early.*

5. This process of alignment is detailed in McShane et al., In press.

6. Other interfaces in this GUI environment support basic lexical acquisition (i.e., acquisition of the linked syn-struc and sem-struc for each word sense), ontology acquisition, evaluation of TMRs, and all aspects of system testing and debugging.

e. *I spent time writing the paper.*
f. *I spent a day walking around Rome.*

The OntoSem description of this syntactic pattern is shown in the left-hand pane of Figure 7. The right-hand pane shows two Stanford dependency parses that can be generated for sentences of this structure.[7] To emphasize, what is recorded as *one* syntactic pattern in OntoSem can be treated in two different ways by the Stanford parser, *depending upon the actual words in the input sentence*. In some cases, the parser generates the dependencies *nsubj, dobj,* and *xcomp,* whereas in other cases it generates the dependencies *nsubj, dobj* and *partmod.*

No matter which dependencies the parser generates, the arguments referenced in those dependencies must be correlated with the expectations recorded in the syn-strucs of the OntoSem lexicon. This is done using numbered variables prefixed by $var.[8] For example, the *subject* of the OntoSem sense ($var1) maps to the *nsubj* of either of the Stanford parses; the *directobject* of the OntoSem sense ($var2) maps to the *dobj* of either of the Stanford parses; and the *np* (realized as a progressive verb)[9] in the OntoSem sense ($var3) maps either to Stanford's *xcomp* or *partmod*.

Since unique syn-strucs in the OntoSem lexicon number only in the hundreds, despite there being many thousands of argument-taking word senses, we were able to quickly carry out this OntoSem-to-Stanford mapping process for the whole lexicon.

Of special importance for the upcoming discussion is the fact that syntactically idiosyncratic MWEs are handled in exactly the same way as compositional argument-taking head words. For example, Figure 8 shows the OntoSem-to-Stanford alignment screen for the MWE *X {can} {not} help but Y* (curly brackets

7. Definitions of the dependencies can be found in the Stanford CoreNLP dependencies manual, available at http://nlp.stanford.edu/software/dependencies_manual.pdf. The dependencies used in Figures 7 and 8 are: nsubj (nominal subject), dobj (direct object), xcomp (open clausal complement), partmod (participial modifier), aux (auxiliary), neg (negation modifier), conj (conjunct), cc (coordination). The OntoSem syntactic descriptors that are not self-evident are infcl (the infinitival form without 'to') and verbneg (the word *not* or its contracted form, *-n't*). Naturally, the parser can return unexpected parses/mistakes. One of the benefits of this multiple-mapping strategy is that it accommodates unexpected and inconsistent parses across lexical inputs, to the extent that they are attested by the inventory of examples we test.

8. Our detection of which types of OntoSem syn-strucs can correlate with more than one Stanford output is based on testing rather than a formal methodology. For example, we know that the parser sometimes attaches PPs to the verb and sometimes to the most local NP, leading to two candidate parses for PP-inclusive inputs.

9. This syntactic description of the progressive verb form was inherited from the original implementation of the OntoSem analyzer.

indicate that multiple inflectional forms are possible). Not surprisingly, there is only one lexical sense, *help-v4,* that uses this syntactic pattern; however, no new theoretical or methodological tools were needed to prepare the system to automatically analyze this MWE.

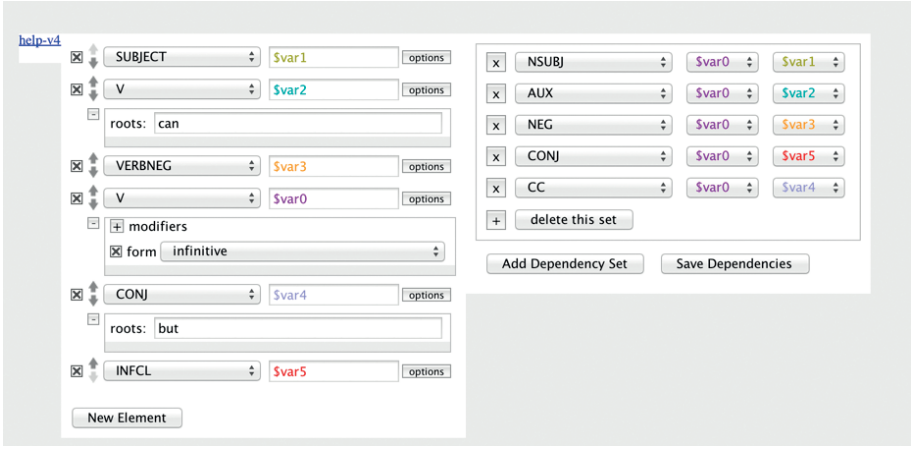

Figure 8. Aligning the expected dependencies of the OntoSem lexicon with the actual dependencies generated by the Stanford parser for the MWE *X {can} {not} help but Y*.

Once OntoSem-to-Stanford alignments have been recorded in the lexicon, the analyzer can treat compositional and MWE inputs in exactly the same way. Folded into the analysis process is the treatment of syntactic transformations, when appropriate.[10] This effectively expands the lexicon of basic diatheses to cover a large range of derived diatheses, complex sentence structures, and embedded constructions (for details, see McShane et al., In press).

It is important to emphasize that we do not consider the engineering work related to parser integration to be a part of the theory of Ontological Semantics, nor do we consider the OntoSem-to-Stanford mapping information to be a core part of the lexicon: after all, a different parser could alternatively be integrated into our environment. In most of our more theoretically-oriented work we do not even mention this level of technical detail. However, we believe that practical implementations are imperative for work in NLP, and since the syntactic treatment of MWEs has long been considered a problem for NLP systems, we hope that making manifest some aspects of how our system works will help to demystify the process.

10. For example, the passive transformation is understood to be available for non-idiomatic transitive verb senses. See below for methods of blocking transformations in idiomatic MWEs.

3.2 All argument-taking words are construction-like

Readers might notice that, in the OntoSem approach to language processing, **the combination of syntactic expectations and semantic constraints renders every argument-taking lexical sense *construction-like***. That is, although compositional verb senses do not require that particular lexemes be used as their arguments, they do semantically constrain the set of meanings that can be used to fill case-role slots, resulting in what might be thought of as broadly specified constructions.[11] This is not a peculiar side-effect of our theory or formalism; instead, we hypothesize that this is how people think about language, and how intelligent agents configured to act like people need to learn to think about it. In short, a sufficiently fine-grained lexical specification of argument-taking words — supported by ontological knowledge about the concepts they invoke — is a long way toward being a construction, and constructions are a superclass of what are typically considered multiword expressions.

4. Lexically recording MWEs

In OntoSem, multiword expressions (MWEs) are defined as any combination of text strings or semantically-constrained syntactic categories that carry a particular meaning that we find useful to pre-record in the lexicon rather than treat compositionally at runtime. MWEs include complete phrases (*there will be hell to pay*), clauses with variable elements (HUMAN *eats his/her/their words*), verbs with particles (*think up* (a good idea)), strings that include punctuation (*nothing ventured, nothing gained*), nominal compounds (*dog bed*), and more. To put a finer point on it, we are not committed to any universal, theoretically motivated definition of which strings to treat as MWEs. Rather, there are varying degrees of compositionality which, in the context of automatic text processing, should be treated in the way that works best for the agent. Like Stock et al. (1993, p. 238), we integrate MWEs into the lexicon as "more information about particular words" rather treat them using special lists and idiosyncratic procedures.

For processing purposes, two broad classes of MWEs can be delineated: those that can be treated as lexemes, albeit with white spaces, and those that cannot. The first category, which includes entities like *vice president*, *stock market*, and *nothing ventured, nothing gained,* is rather trivial. The components must occur in the listed

11. The same is true of MWEs anchored on other parts of speech. For example, the OntoSem approach to nominal compounding — which is often considered a type of MWE — is described in McShane et al., 2014.

order, do not permit modifiers or other elements to intervene between them, and only the head of such a phrasal — usually the final word — is potentially subject to inflection. A morphological analyzer and a syntactic parser can interpret these as strings with white spaces. In the OntoSem lexicon, we record such entities as multi-part head words with an underscore indicating each white space. This approach provides simplicity and nearly perfect coverage — only *nearly* perfect because in rare cases an expletive, speaker correction or interruption might occur between the elements. (This can also, by the way, happen in the middle of regular words: *decon [ouch!] struction.*) As with all so-called unexpected input, such deviations must be handled by recovery procedures which, in our system, amount to a sequence of attempts to relax certain constraints, such as the expectation that only a blank space can intervene between components of a multiword head entry.

The rest of MWEs present a wide variety of knowledge representation and processing challenges *and opportunities*.[12] We will consider a sampling in turn, using an example-based description strategy.

4.1 Syntactic descriptions of MWEs

In the OntoSem lexicon, the syntactic description of MWEs is practically the same as for compositional argument-taking words. The main difference involves the optional use of immediate constituents (such as NP) rather than syntactic functions (such as Subject), to describe syntactic components. There are two reasons to use immediate constituents to describe MWEs. First, some MWEs are most simply described as a series of immediate constituents, with nothing to be gained by associating them with particular syntactic functions. This is the case, e.g., with *X {can} {not} help but Y*, shown in Figure 9.[13]

12. Calzolari et al., 2002 discuss the presentation of syntactic and semantic information about MWEs in multi-lingual lexicons, but the discussion is oriented toward classification and problems of MWEs rather than making specific recommendations.

13. Figure 8 showed the OntoSem-to-Stanford *linking* interface for this MWE, which involves only syntactic aspects of the MWE. Although our environment includes a basic lexicon acquisition interface (covering syntax and semantics) with a similar look and feel, we find it more useful pedagogically to present lexicon entries outside of the interface.

help-v4
definition "the phrasal: 'X cannot help but Y'. X feels he must do Y because he cannot force himself to refrain from doing Y"
example "The people in the room could not help but laugh."
syn-struc

subject	$var1		*; the subject*
v	$var2	(root can)	*; the word 'can' with any inflection*
verb-neg	$var3		*; verbal negation*
v	$var0	(form infinitive)	*; the bare infin. form of 'help'*
conj	$var4	(root but)	*; the word 'but'*
inf-cl	$var5		*; a bare-infinitive clause*

sem-struc

^$var5	
EXPERIENCER	^$var1
^$var2	null-sem+
^$var3	null-sem+
^$var4	null-sem+

meaning-procedure[14] (fix-case-role (value[15] ^$var1) (value ^$var5))

Figure 9. The lexical sense for the phrasal *X {can} {not} help but Y.*

The second reason to use immediate constituents in a syn-struc is to signal to the analyzer that canonical syntactic transformations must be blocked. For example, although the MWE *kick the bucket* is a transitive sense of the verb *kick*, the idiomatic reading does not permit the passive and middle diatheses. By convention, we block transformations by using at least one immediate constituent label in the inventory of syntactic components: e.g., rather than describing the *kick the bucket* sense of *kick* as *Subject V Directobject*, we describe it as *NP V NP*, or *Subject V NP,* or *NP V Directobject* — it only takes one immediate constituent label to block all transformations. MWEs like *X {can} {not} help but Y* are not subject to transformations for two reasons: they do not represent a canonical syntactic structure that is associated with some transformation(s), and they include immediate constituents (*conj, verb-neg, inf-cl*).

14. Meaning procedures are procedural semantic routines that compute contextual meaning. The meaning procedure *fix-case-role* is used when different verbs will require different case-roles. For example, whereas one is an EXPERIENCER of LAUGH, one is an AGENT of GIVE, even in a context like, *He could not help but give his daughter expensive birthday gifts*. (Here, the idiom does not imply that the giving was outside of his control; instead, it implies that he felt compelled to do it.) Using the fix-case-role function, the analyzer checks the case-role inventory for the contextually attested EVENT and modifies the listed case-role if necessary.

15. "Value" refers to the computed meaning of a variable.

Although the syn-strucs for MWEs tend to contain a fixed inventory and ordering of syntactic constituents, they nevertheless cover a wide variety of input sentences. Consider the variability permitted by our example, *X {can}{not} help but Y*:

i. The subject and inf-cl can be of any form or complexity: [*Even the people in the room who had come in late and really didn't know what was going on*]$_{\text{SUBJECT}}$ *couldn't help but* [*burst out laughing*]$_{\text{INF-CL}}$. The analyzer uses a small inventory of rules, in combination with the lexicon entries of the words in the input, to build up the necessary structures.
ii. The verb *can* can have various inflectional forms: *They can not <could not, could not have, couldn't have> help(ed) but laugh.*
iii. Adverbs can be freely added: *They* ***really*** *could not help but laugh* ***so loud that the neighbors heard them****.*

In sum, the syntactic descriptions of MWEs constrain the inventory and ordering (if applicable) of MWE components, but accommodate the large generative capacity of the language overall.

4.2 Semantic descriptions of MWEs

The previous example is particularly useful for showing the syntactic expressivity of the OntoSem lexicon. Semantically, however, this example is less typical since the acquirer of this lexicon entry decided to render the semantics of the whole MWE as just the meaning of the bare-infinitive clause appropriately combined with the meaning of the subject. In other words, the sentence *He could not help but laugh* — under the coarse-grained interpretation selected by the acquirer — is interpreted as the meaning of *He laughed*, as shown by the TMR in Figure 10.

```
LAUGH-1
    EXPERIENCER       HUMAN-1
    TIME              (< find-anchor-time)      ; indicates past time
HUMAN-1
    GENDER            MALE
    EXPERIENCER-OF    LAUGH-1
```

Figure 10. The TMR for the sentence *He could not help but laugh.*

The fact that no compositional semantic analysis should be applied to the elements *can, not, help* and *but* is recorded in the sem-struc using the feature "null-sem+".

One might ask, if OntoSem aims to generate fine-grained semantic analyses, why is the semantics of *cannot help but* ignored? The reason is practical: how, exactly *would* or *could* one describe the semantic nuances this collocation contributes?

One might try something like, "Irrespective of whether or not X wanted to do Y, X did Y because, given some unspecified circumstances, it would have been too difficult for X not to do Y." However, even if this were deemed a reasonable analysis, this is still little more than an English paraphrase, which is still many steps away from being a formal representation that could support useful automatic reasoning. The formal representation would be quite complex and it is not clear what goal it would serve. In short, within OntoSem, we pursue practical applications rather than flexing our knowledge representation muscles. Our primary interest is permitting agents to function in useful collaborations with people. Nuances as fine-grained as this one are tackled only when it is determined that they are necessary to satisfy the needs of a specific application.[16]

It is actually not uncommon for MWEs to contribute more pragmatic than semantic content. For example, in all of the following, the meaning of the MWE is acceptably reduced to the meaning of the main proposition, written as ^[main proposition] (recall that ^ means "the meaning of").

a. the fact that X → ^X
   The fact that you have a cold is sad → ^[[you have a cold] is sad]

b. it {turn} out that X → ^ X
   It turned out that he was right all along → ^[he was right all along]

c. it {be} just that X → ^X
   It's just that I don't want to go to school! → ^[I don't want to go to school]

d. the month of X → ^X *; same for 'city of', etc.*
   He was born in the month of January → ^[He was born in January]

These turns of phrase are not linguistically irrelevant, but their semantic contribution is too subtle to be of projected utility to our intelligent agents, which motivates a full-fledged syntactic treatment but a simplified semantic one.

As regards the richness of semantic description for MWEs, it is the same as for compositional entities in the OntoSem lexicon. Apart from expressing meaning through a direct or modified ontological mapping, as shown in the examples of *address* and *see* above, meaning can be expressed using extra-ontological descriptors, such as values of aspect or mood. For example, the meaning of *Z {is} for X to Y,* shown in Figure 11, is "X must (Y theme Z)." This is expressed as obligative modality with a value of 1 scoping over the meaning of the main proposition.

16. For a discussion of the grain size of description in service of applications, see Nirenburg and Raskin (2004, Chapter 9). Of course, it is much easier to preserve semantic nuances in an application like machine translation, in which a transfer-based component can include idiom-to-idiom correspondences.

```
for-prep10
    definition   "phrasal: 'Z is for X to Y' — indicates who must be the agent of the event"
    example      "This problem is for the chairman to solve."
    syn-struc
        np          $var1                 ; This problem
        v           $var2 (root *be*)     ; is
        pp
            prep    $var0 (root for)      ; for
            obj     $var3                 ; the chairman
        inf-cl      $var4 (cat inf-cl)    ; to solve
    sem-struc
        MODALITY                          ; a meaning that scopes over propositions
            type    OBLIGATIVE            ; indicates necessity
            value   1                     ; the highest value on the abstract scale {0,1} of necessity
            scope   ^$var4                ; the event that is necessary: solving (the problem)
        ^$var4                            ; ^[solve]
            AGENT   ^$var3                ; ^[chairman]
            THEME   ^$var1                ; ^[problem]
        ^$var2 null-sem+                  ; 'for' is not compositionally analyzed
```

Figure 11. The lexical sense for the MWE *Z is for X to Y*. Comments following semi-colons provide a trace of how the example sentence correlates with the elements in the syn-struc and sem-struc.

Another example of a semantic interpretation of a MWE is shown in Figure 12, which is the sense for *X {pay} homage to Y*. The lexicon acquirer analyzed this idiom as mapping to the ontological concept PRAISE. In the sem-struc, ^$var2 (linked to *homage*) and ^$var3 (linked to *to*) are attributed null compositional semantics since their meaning is already taken care of by the basic dependency structure, PRAISE (AGENT…) (THEME…).

```
pay-v16
    definition   "phrasal: 'X pays homage to Y' — analyzed as 'X praises Y'"
    example      "The citizens paid homage to the queen."
    syn-struc
        np          $var1
        v           $var0
        np          $var2 (root homage)
        pp
            prep    $var3 (root to)
            obj     $var4
    sem-struc
        PRAISE
            AGENT   ^$var1
            THEME   ^$var4
        ^$var2      null-sem+
        ^$var3      null-sem+
```

Figure 12. The lexical sense for the MWE *X pays homage to Y.*

4.3 Compositional aspects of non-compositional elements

As shown by the examples above, one of the widely-used descriptors for MWE lexicon entries is "null-sem+", which indicates that the given element should not be compositionally analyzed. Its utility is clear in idioms like *{kick} the bucket,* in which there is no bucket. However, two problems can arise when attributing null compositional semantics to components of MWEs: (1) if the element is the head verb of the structure, its tense/aspect features will be lost, and (2) if the element is modified, the modifier will lose its target. Since these problems require different solutions, we discuss them separately.

Recovering features of head verbs with null compositional semantics. In most verbal senses, the verb is the head word and its features — such as tense, mood and aspect — are naturally available to the semantic analyzer. One such example is *be-v7* (shown in Figure 13), which is headed by the verb *be*.

```
be-v7
    definition   "phrasal: 'X [be-pres] [xcomp]' — indicates that X is going to happen in
                  the future, it is planned"
    example      "The president is to meet with the delegates in the lobby."
    syn-struc
        np          $var1
        v           $var0
        xcomp       $var2
    sem-struc
        ^$var2
            AGENT   ^$var1
            TIME    > find-anchor-time          ; indicates future time
```

Figure 13. Lexical sense for *NP is to Verb*.

Contrast this with situations in which the feature-carrying verb is *not* the head word of the lexical sense, as in *in-prep15* (Figure 14). Here, the verb cannot head the entry because many copular verbs are possible, such as *seem to be, appear to be, might be, could be* and so on. When the verb is too variable to be listed as the head word, a non-variable element lexically anchors the MWE — in this case, *in*.

```
in-prep15
    definition   "phrasal: 'X *be* in surgery' = X *be* the experiencer of surgery"
    example      "John was in surgery."
    syn-struc
        np                  $var1
        v                   $var2 (root *be*)       ; *be* indicates any copular verb
        pp
            prep            $var0 (root in)
            obj             $var3 (root surgery)
    sem-struc
        refsem1[17]
            PERFORM-SURGERY
                EXPERIENCER   ^$var1
        ^$var3 null-sem+
        ^$var2 null-sem+
    meaning-procedure (apply-meaning (strip-features (value ^$var2)) (value refsem1))
```

Figure 14. Lexical sense for *X *be* in surgery*.

The copular verb does not add compositional meaning to the structure, so it is attributed null semantics (^$var2 (null-sem+)). However, it does contribute features. Those features are recovered using the procedural semantic routine called

17. Refsem# is a device for marking coreference between concept instances.

apply-meaning, which is recorded in the meaning-procedure zone of the entry. It says, formalism aside: strip the features from the verb ($var2) and apply them to the overall semantic interpretation recorded in the sem-struc (PERFORM-SURGERY).

Modification of constituents with null compositional semantics. Modifiers do not always *semantically* modify the constituent upon which they are *syntactically* dependent. This phenomenon is well known from the literature on adjectives. Consider the following three examples, drawn from Raskin and Nirenburg (1998), who in turn reference Vendler (1963, 1968):

- a beautiful dancer can be a beautiful *woman* who dances or a woman (or a man) who *dances* beautifully
- a comfortable chair is a chair that *people feel* comfortable sitting in
- a fast car is a car that has the potential to *go* fast.

Similarly, in the idiomatic utterance, *He kicked the bloody bucket!* the word *bloody* does not apply to a bucket, it conveys a negative speaker attitude toward the event of this person dying.

Our algorithm for treating modifiers within MWEs is for the system to analyze the MWE as indicated in the sem-struc, then attempt to attach the meaning of unaccounted-for modifiers — which were syntactically hosted by non-compositional elements — to the meaning of the entire structure using generalized processes for meaning composition.

Let us trace this process for the input, *He kicked the bloody bucket!* The example has both literal and figurative readings: the man in question could have kicked a bucket covered in blood, or he could have died, with the speaker expressing a negative attitude toward this event. As with all disambiguation in OntoSem, this disambiguation is heuristic. If there is no coreferential category for *bucket* in the preceding context to explain the use of the definite article, then the idiomatic sense is preferred.[18] The TMR that should be generated from the idiomatic reading of this sentence is shown in Figure 15. It relies on the lexical senses shown in Figures 16 and 17.

18. Of course, there are other valid hypotheses for explaining a definite description with no coreferent: the object could be visible in the real-world context, the given NP could be always definite (*the sun*), or it could be used in a script that assumes the existence of the object (*When milking a cow, first you take* the bucket *and put it under the cow*). For more on reference resolution in OntoAgent, see McShane and Nirenburg, 2013.

```
MODALITY-1
    TYPE            EVALUATIVE
    SCOPE           DIE-1
    VALUE           .1              ; a very low value on the {0,1} scale
    ATTRIBUTED-TO   *speaker*
DIE-1
    EXPERIENCER     HUMAN-1 (GENDER MALE)
    time            < find-anchor-time
```

Figure 15. TMR for the sentence *He kicked the bloody bucket.* The negative speaker attitude is expressed by a very low value of evaluative modality.

```
kick-v2
    definition   "phrasal: 'kick the bucket' — die"
    example      "His uncle kicked the bucket."
    syn-struc
        np              $var1
        v               $var0
        directobject    $var3   (root bucket) (number sing) (contains $var2 (root the))
    sem-struc
        DIE
            EXPERIENCER   ^$var1
        ^$var2      null-sem+
        ^$var3      null-sem+
```

Figure 16. Lexical sense for *kick the bucket*. Syntactic constraints indicate that the direct object must be headed by the singular noun *bucket* and must contain the determiner *the*.

```
bloody-adj1
        definition   "related to blood"[19]
        example      "That's awfully bloody meat."
        syn-struc
            $var0       (cat adj)
            $var1       (cat n)
        sem-struc
            ^$var1
                RELATION        BLOOD
```

19. We have found no reason to split senses to account for nuances such as *covered in blood* (a bloody knife) vs. *leaking/squirting blood* (a bloody wound), vs. *otherwise involving blood* (a bloody massacre). Decision-making about whether to split or bunch senses combines rules of thumb — e.g., all basic syntactic diatheses must be covered — with judgments about the grain-size of semantic analysis required of the system for a given application.

```
bloody-adj2
    definition   "indicates speaker dissatisfaction"
    example      "He's a bloody fool!"
    syn-struc
        $var0          (cat adj)
        $var1          (cat n)
    sem-struc
        MODALITY
            TYPE       EVALUATIVE
            SCOPE      ^$var1
            value      .1
```

Figure 17. Two lexical senses of *bloody*.

The analyzer produces the basic semantics (DIE (EXPERIENCER HUMAN)) using the MWE entry shown in Figure 16, but it needs to account for the unbound modifier, *bloody*, which has the two senses shown in Figure 17. It can use either of those meanings to modify either component of the nascent TMR: DIE or HUMAN. So, the dying can be related to blood, the human can be related to blood, the speaker can be dissatisfied with the dying or the speaker can be dissatisfied with the human. In the abstract, this decision would be difficult to make; however, the analyzer has additional evidence — it knows that this modifier is an unanchored modifier in a MWE. This knowledge suggests that preference should be given to modifier meanings that express speaker attitudes and/or scope over the entire proposition. Relying on such heuristics, even if they are defeasible, is essential in order to break through the many problems of residual ambiguity in text analysis.

4.4 To create MWEs or not to create them?

Our practical work has made it clear there is no hard line between what should and should not be recorded as a MWE. Consider the phrase *cast a spell on/over*, which was recorded in the OntoSem lexicon using the MWE sense shown in Figure 18. This sense analyzes the phrase as meaning X is the AGENT of a BEWITCH event whose THEME is Y.

```
cast-v3
    definition   "phrasal: cast a spell over/on"
    example      "She cast a spell over the mean cab driver."
    syn-struc
        subject          $var1
        v                $var0
        directobject     $var2 (root spell)
        pp
            prep         $var3 (root (or over on)) (opt +)[20]
            obj          $var4
    sem-struc
        BEWITCH
            AGENT        ^$var1
            THEME        ^$var4
        ^$var2 null-sem+
        ^$var3 null-sem+
```

Figure 18. Lexical sense for *cast a spell over*.

Recording this as a MWE has the benefit of simultaneously resolving the highly polysemous components *cast* and *spell*. However, this approach carries a consequence. Since it formally attributes null compositional semantics to *spell* (its meaning is folded into the interpretation BEWITCH), any modification of *spell* will be unbound: *X cast a **terrible/wicked/playful** spell on Y*. However, there is a solution that offers the best of both worlds. We can incorporate benign redundancy into the MWE description, as shown in the modified lexical sense in Figure 19.

```
cast-v3a
    definition   "phrasal: cast a spell over/on"
    example      "She cast a spell over the mean cab driver."
    syn-struc
        subject          $var1
        v                $var0
        directobject     $var2 (root spell)
        pp
            prep         $var3 (root (or over on)) (opt +)
            obj          $var4
    sem-struc
        BEWITCH
            AGENT        ^$var1
            THEME        ^$var4
            INSTRUMENT   ^$var2
        ^$var3 null-sem+
```

Figure 19. An improved entry for *cast a spell over*, which includes benign redundancy and directly supports inputs containing a modification of *spell*.

20. "Opt +" indicates that this prepositional phrase is an optional constituent.

This sense conveys the meaning of the MWE as X is the AGENT of a BEWITCH event whose THEME is Y and whose INSTRUMENT is a SPELL. This is redundant because the ontology already describes the event BEWITCH as having the INSTRUMENT SPELL. However, if *spell* is modified in the input, this representation provides an explicit target for that modification.

This example serves as a fitting conclusion to our overview of the lexicographic treatment of MWEs in OntoSem because it underscores the practical orientation of the enterprise. The lexicon acquirer's goal is to foresee challenges in semantic analysis and identify opportunities to prepare the system to best overcome those challenges. The test of the quality of that decision-making is to use the acquired knowledge to support automatic semantic analysis and evaluate the results. It is to that evaluation process that we now turn.

5. Evaluation

Formulating an evaluation suite to validate our approach to recording and processing MWEs was challenging for two reasons. First, there is no obvious definition for MWEs, given that argument-taking words and MWEs represent a continuum of compositionality, and, within OntoSem, we treat them all the same way. The second challenge was to evaluate the processing of MWEs without having to simultaneously evaluate every aspect of deep semantic analysis that is required to treat a typical sentence of input. The evaluation suite described below is an attempt to provide a useful metric of progress while serving the main strategy of knowledge-based system development: error analysis leading to iterative improvements in the static knowledge bases and processors.

MWE selection. For this evaluation, we defined "MWEs of interest" as those multi-component lexical senses that listed one or more non-prepositional, non-particle roots in their syn-struc zones. To take two examples from the discussion above, in *cast-v3* (*X {cast} spell on/over Y*) the direct object must be realized as the root *spell*, and in *in-prep15* (*X *be* in surgery*), the object of the preposition must be realized as the root *surgery*.

Neither the inventory of MWEs covered in the lexicon, nor the lexicon entries themselves, were modified prior to evaluation: all MWEs that were evaluated were acquired during regular lexical acquisition over the past decade.

Using this automatic MWE-detection method, 382 target MWE senses were extracted from our lexicon.

Corpus search. Since analyzing a large corpus syntactically and semantically is resource-intensive, and since only a small percentage of sentences was expected to be within the purview of the evaluation, we first used a string-based method

of detecting potentially relevant examples. We searched the Wall Street Journal corpus of 1987 for sentences in which all lexically specified roots in the MWE occurred within 6 tokens of the head word. For example, to detect candidate examples of the MWE *something {go} wrong with X* the word *something* had to be attested within 6 tokens preceding *go/went/goes*, and the words *wrong* and *with* had to be attested within 6 tokens following *go/went/goes*. This filtering yielded a corpus of 182,530 sentences, which included potential matches for 286 of our 382 target MWEs. We then selected the first 25 candidate hits per MWE, yielding a more manageable set of 2001 sentences.

Syntactic filtering. We then syntactically analyzed those 2001 sentences using Stanford CoreNLP. If the parse of a sentence did not correspond to the syntactic requirements of its target MWE — i.e., if the actual dependencies returned by Stanford did not match the expected dependencies recorded in the OntoSem lexicon — the sentence was excluded. (Recall that the initial candidate extraction method was quite imprecise — we did not expect it to return exclusively sentences containing MWEs.) This pruning resulted in 804 sentences that syntactically matched 81 of our target MWEs. We then randomly selected a maximum of 2 sentences per target MWE, resulting in a corpus that was an appropriate size for our manual evaluation procedure: 136 sentences.

Semantic filtering. These 136 sentences were semantically analyzed using the OntoSem analyzer. They were analyzed like any other inputs: the analyzer was free to select any lexical sense for each word of input, either using or not using MWE senses. To put the lexical disambiguation challenge in perspective, consider the following statistics:

- The average sentence length in the evaluation corpus was 22.3 words.
- The average number of word senses for the head word of the MWE was 23.7. This number is so high because verbs such as *take* and *make* have over 50 senses apiece due to the combination of productive meanings and light-verb usages (e.g., *take a bath, take a decision, take place*, etc.).
- The average number of word senses for each unique root in the corpus was 4.

To summarize, the analyzer was tasked to resolve all of the syntactic and semantic ambiguities in these inputs using an approximately 30,000-sense lexicon that was not tuned to any particular domain. The processing strategy showed no experiment-specific preference for the leveraging of MWEs over compositional analysis.

Since TMRs for long sentences can run to several pages of output, we used a TMR-simplification program to automatically extract the minimal TMR constituents represented by the MWE. For example, in (2)–(5) the listed TMR excerpts were sufficient to determine that the MWEs (whose key elements are in caps) were treated correctly.[21]

(2) *The company previously didn't PLACE much EMPHASIS ON the development of prescription drugs and relied heavily on its workhorse, Maalox.*

EMPHASIZE-1[22]
AGENT FOR-PROFIT-CORPORATION-1
THEME DEVELOP-1

(3) *"I'm sure nuclear power is good and safe, but it's impossible in the Soviet bloc," says Andrzej Wierusz, a nuclear-reactor designer who LOST his JOB and was briefly jailed after the martial-law crackdown of 1981.*

ASPECT-1
SCOPE WORK-ACTIVITY-1
PHASE END
WORK-ACTIVITY-1
AGENT HUMAN-1

(4) *Mr. Hodel, in a statement, PAID HOMAGE TO the accord and said it represented "a significant step in bringing together special-interest groups which often have strikingly different views about the offshore leasing program."*

PRAISE-1
AGENT HUMAN-1
THEME SETTLEMENT-1

(5) *The appeals court investigation has never been officially made public, but published accounts of the report say the investigative committee concluded that Judge Hastings allegedly PERJURED HIMSELF in the bribery trial.*

PERJURY-1
AGENT HUMAN-1

Of course, to fully appreciate the TMRs for these examples, readers would benefit from having access to complete specifications of contributing ontological

21. The full TMRs were available to evaluators and were consulted as needed.

22. The negation is taken care of by a modality frame available in the full TMR.

concepts. This, alas, was impractical for reasons of space.[23] Still, the gist of these excerpts should be accessible to most readers based on the similarity between English words and the English-like names of ontological concepts.

Manual evaluation. The evaluation was carried out by McShane with targeted double-checking and selective error attribution by Beale.

The main question posited in the evaluation was, *Did the analyzer correctly detect each MWE and compute the meaning of the non-variable portion correctly?* This was judged by evaluating whether the TMR frame(s) representing the meaning of the MWE was/were headed by the correct concept(s).[24] The decision was binary: If that head was (or those heads were) correct, then the MWE interpretation was judged correct; if not, the MWE interpretation was judged incorrect.

In examples (2)–(5), not only did the analyzer correctly detect the MWE, it also correctly disambiguated the fillers of the AGENT and THEME case-roles. For example, in (2) it selected FOR-PROFIT-CORPORATION as the analysis of the ambiguous *company* (another available meaning is "a set of humans"), and it selected DEVELOP as the analysis of *development* (other available meanings are "a novel event" and "a residential development").

However, to reiterate, we did *not* require that case-role fillers be correctly disambiguated in order to mark a MWE interpretation as correct. This is because the disambiguation of slot fillers can require much more than clause-level heuristics. For example, in (6), the MWE analysis was correct: *to look forward to* means (approximately) to *want* the state of affairs to occur, which is represented in TMR by the highest value of volitive modality scoping over what is wanted. However, the filler of one of the key slots in this TMR — the SCOPE of the modality — is probably not contextually correct.

(6) *We LOOK FORWARD TO the result.*

MODALITY-1		
SCOPE	ANY-NUMBER-1	
VALUE	1	*; on the abstract scale {0,1}*
ATTRIBUTED-TO	SET-1	
TYPE	VOLITIVE	

This TMR states that what was looked forward to was some number (one sense of *result*), whereas what is probably being looked forward to is some state of affairs

23. OntoSem static knowledge resources are not currently distributable.

24. In some cases, the semantic description of a lexical item can involve more than one head. For example, in (3), both the frames for ASPECT and for WORK-ACTIVITY needed to be correct, since *lose one's job* centrally involves both of these elements of meaning.

(another sense of *result*). But in all fairness, the analyzer *could* be correct since the sentence *could* be uttered in a math class by students waiting for their resident genius to solve a problem. Lacking extra-clausal heuristic evidence, the analyzer arrived at comparable scores for both analyses and randomly selected between them.[25] Extra-sentential reasoning is far afield from our current goal of evaluating the utility of our approach to recording and processing MWEs.

Examples (7) and (8) offer further insights into why we did not include case-role disambiguation into this evaluation. All four salient case-role fillers in these examples of the MWE *X {pose} problem for Y* were analyzed incorrectly, even though analysis of the MWE was correct.

(7) *The changing image did however POSE a PROBLEM for the West.*

(8) *But John McGinty, an analyst with First Boston Corp., said he believed dissolution of the venture won't POSE any PROBLEM for Deere.*

Two of the errors — involving the analyses of *the West* and *Deere* — were due to the mishandling of proper names, which can be remedied by additional onomasticon-related knowledge engineering. One error — the analysis of *the changing image* — could not be avoided using the sentence-level context provided by our examples: i.e., *image* can be a pictorial representation or an abstract conceptualization. And the final error — the analysis of *dissolution of the venture* — results from a failure to simultaneously recognize the metaphorical usage of *dissolve* and select the correct sense of the polysemous noun *venture*. These examples underscore the fact that even a small corpus of naturally occurring text can offer a broad spectrum of challenges presented by natural language.

The results. To reiterate the evaluation setup, for each target MWE, the system semantically analyzed a maximum of two examples that were selected as candidates by virtue of lexico-syntactic correlation with a target MWE sense. The syntactically-oriented selection/filtering process included no semantic heuristics, so when it came time for semantic analysis, the analyzer could choose either the MWE reading or any compositional reading available for the input strings.

The evaluation suite included 136 examples, 98 of which were judged to be correct according to the evaluation standards described above: i.e., the head(s) of the excerpt TMR frame(s) were correct. This resulted in a precision of 72%. However, the numbers themselves are relatively unimportant. What *is* important is what the evaluation taught us about MWE processing and the ways in which we

25. Randomly selecting among same-scoring semantic analyses is only one of many possible system configurations. The analyzer could also be configured to return all candidate analyses that score within a threshold of the highest score.

can improve it over time through enhancements to the static knowledge resources and processors.

Lessons learned. We divide the results of the evaluation study into four naturally occurring categories.

Category 1. For the MWEs listed in Table 2, both of the evaluated examples actually contained the MWE and the analyzer correctly selected that interpretation.

Table 2. MWEs for which all examples were analyzed correctly.[26]

X {welcome} Y as Z	it *be* thought that Y	X {think} so
X {succeed} Y as Z	X {spend} time with Y	X {spend} time Y-ing
X {serve} as Y	X {return} verdict	X {pose} problem for Y
X {set} fire to Y	X {perjure} myself <yourself, etc.>	X {perform} examination on Y
X {perceive} Y as Z	X {pay} homage to Y	X {pay} tax
X {pay} attention to Y	X {lose} job	X {look} forward to Y
X would like to Y	it is/was hoped that X	X {make} effort
X {make} sure that Y	X {make} it clear that Y	X {make} request
it {make} sense	X {fill} vacancy	X {fall} short in Y
X {fall} asleep	it {is} expected that X	X {provide} service to Y
X {did} so	X {find} Y guilty	X {decline} to comment
X {decline} comment	it {costs} X to Y	X {classify} Y as Z
it *be* believed that	X {conduct} business	X {turn} a profit
X {presents} a problem	it *be* assumed that	X {answers} {question}
X {let} Y know Z	X {take} shape	X {take} issue with Y
X {plead} guilty	X {plead} innocent	X {submit} resignation
X {reach} decision	X {give} signal	X {give} speech <lecture>
X {give} advice	X {have} no idea <clue>	X {take} trip
X {have} trouble Ying		

Category 2. For the following MWEs, both of the evaluated examples were analyzed incorrectly — the examples used the MWE sense but the analyzer selected a compositional (non-MWE) interpretation: *sun {set}; X {make} noise <sound>; X {put} pressure on Y; X {work} together to Y; X {attract} attention; X {take} place; X {take} part in*. The source of these errors was not immediately apparent, and

26. We refer to MWEs using a shorthand in which { } indicates any inflectional form of the word; ***be*** indicates any copular verb; X, Z, Y indicate arguments; and *word <word1, word2>* indicates options. Modifiers that are not a central part of the MWE — e.g., 'a/the' in *X {return} verdict* are not included in these shorthand representations.

tracking them would have required end-to-end system debugging. As a reminder, the analyzer had to select from an average of 23.7 word senses for each MWE head, each having their own inventories of expected syntactic and semantic constraints, which "competed" to be used in the analysis of each input.

Category 3. For the following MWEs, both of the evaluated examples actually contained the MWE but, whereas the analyzer correctly analyzed one of the two inputs using the MWE, it incorrectly analyzed the other as not containing the MWE: *X {go} unnoticed <undetected>; X {reserve} the right to Y; X {take} note of Y; X {take} the form of Y*. Although this might suggest that the MWE analysis is always correct and should be given a very large scoring bonus during analysis, this is not always the case, as will become clear presently.

Category 4. For several MWEs, the automatic analysis of one or more of the examples was incorrect and the nature of the mistakes, which was readily apparent upon investigation, deserves comment. There were two main sources of errors: insufficient lexical specification of the MWE, and polysemy (including literal vs. metaphorical use) of the MWE.

Lexical specification issues. In some cases, a certain type of constituent must *not* appear in a MWE — an eventuality that we had not anticipated during lexical acquisition. For example, the MWE *X {can} tell (that) Y* idiomatically means UNDERSTAND, as in *I can tell that you're tired.* However, if the indirect object of *tell* is overt, the MWE reading is excluded. So, whereas the analyzer correctly analyzed (9) as containing this MWE, it misanalysed (10), which uses a compositional meaning of *tell*.

(9) *American Legal Systems, a New York legal support firm, installed a complex computer system to handle the thousands of pages of exhibits and objections — as well as a buzzer system and tote board so the judge COULD TELL who was speaking.*

(10) *He added: "All I CAN TELL you is it doesn't apply to the president."*

Similarly, the lexicon contains the MWE *X {go} and Y,* which can be semantically reduced (with minimal stylistic losses) to *X Y:* e.g., *He went and drank my soda → He drank my soda.* However, adding a particle or PP to {go} excludes the MWE reading, as shown by examples (11) and (12), which were analyzed incorrectly.

(11) *"I had to GO out AND work."*

(12) *"My parents always lived below their income level, but there was an assumption when we were growing up that we would GO to college AND GO into a profession."*

In fact, the idiom the lexicon acquirer had in mind should be more narrowly specified not only with respect to the inadmissibility of PPs and particles, but also with respect to tense and aspect, since it seems that only simple present and future tenses are permissible — a hypothesis that requires corpus vetting: *He went and drank my soda. / If I leave the room, he'll go and drink my soda. / ?He is going and drinking my soda.*

Another noteworthy example involves the MWE *X {take} a seat*, which the lexicon describes as SIT (PHASE BEGIN). If one adds the modifier *back* to *seat*, a different MWE is intended: *X takes a back seat to Y* means 'X defers to Y'. Since the lexicon did not include this second MWE, the analyzer mistakenly used the 'take a seat' analysis in a sentence containing *take a back seat*.

Another source of errors derived from the lexicon acquirer's overreaching in providing synonyms for components of MWEs — a problem reminiscent of pitfalls of thesaurus-based query expansion in applications such as question answering and knowledge extraction. For example, the *MWE X {place} emphasis <focus> on* — meaning EMPHASIZE — was listed as permitting the synonyms "value, importance, significance, worth", none of which reliably conveys the meaning EMPHASIZE in this configuration. So, whereas the analyzer correctly analyzed a sentence containing *place the **emphasis** on*, it incorrectly analyzed example (13), which uses *place a **value** on*.

(13) *HMO America said the letter of intent it signed with Mount Sinai doesn't PLACE a VALUE ON the transaction.*

A similar error occurred due to the erroneous expansion of the MWE *X {concede} goal* such that it included the synonym *X {concede} **point***. The latter led the analyzer to impose a sports-oriented reading on (14):

(14) *Mr. Poehl CONCEDES the POINT, saying the Bundesbank is "in a difficult phase of central bank policy."*

For the analyzer to correctly analyze this sentence, the lexicon must include the MWE *X {concede} point*, which should map to an ACKNOWLEDGE event whose THEME is the meaning of the coreferent of the referring expression *the point*.

Yet another type of suboptimal lexical acquisition decision involved underspecification of necessary MWE components. For example, the MWE *X {give} time to Y* was recorded as meaning VOLUNTEER. However, in order for this meaning to obtain, (a) the modifier of *time* must be a possessive pronoun ("X {give} **his <her, etc.>** time to Y") and (b) X must refer to a human. These constraints did not hold in (15) and (16), which the analyzer misanalyzed.

(15) *Even a 90-day delay before Congress grants immunity to key witnesses probably won't GIVE the independent counsel enough TIME to complete major aspects of his investigation.*

(16) *The move to postpone the vote will GIVE the companies TIME to try to settle their differences over how to resolve the government antitrust concerns that have snagged their merger plans and caused Hughes this week to call off the deal.*

Polysemy of MWEs. MWEs, like most lexical items, can be polysemous, with the multiple meanings either being rather stable — and presumably recorded in a speaker's mental lexicon — or being generated on the fly, as by creative metaphorical extensions. For example, the MWEs *X {take} a bath* and *X {take} a look at Y* are each recorded with a single literal sense in our lexicon, even though the former can refer to sustaining a financial loss, as in (17), and the latter can refer to thinking about abstract objects or ideas.

(17) *In this instance, Morgan TOOK a BATH in Eurodollar floating-rate notes, particularly perpetual floating-rate notes.*

Both of the examples of *take a bath* and *take a look at* in our evaluation corpus involved the non-literal usage and, accordingly, were misanalyzed.

Yet another multiply ambiguous use of the verb *take* is found in the MWE *it takes X to Y*. Our recorded sense means "X is required for Y", as in *It takes money to live comfortably.* Corpus example (18), however, reflects a different meaning entirely.

(18) *Today IT will "TAKE appropriate steps to present a proposal directly to (Pennsylvania Enterprises') shareholders and take such other action as we determine to be appropriate."*

A similarly polysemous MWE is *X {let} Y go*, which was lexically recorded with the single meaning LIBERATE (as in *She let the bird go*.) However, the words *let* and *go* can be used in close proximity in a whole range of syntactic and semantic contexts, including those shown in (19) and (20).

(19) *If the Germans LET things GO as they are, then a further weakening of the dollar could move them dangerously toward zero growth.*

(20) *David Doniger, a lawyer for the Natural Resources Defense Council, said, "To me it is outrageous to think that because people can use sunscreen that the government should LET the ozone layer GO to hell."*

Example (19) shows a different MWE sense of *X {let} Y go*, whereas (20) shows a compositional usage of *let* plus the idiom *go to hell*.

We believe that the evaluation exercise validated the feasibility of our approach to lexically recording MWEs and then leveraging that knowledge during automatic semantic analysis. Close analysis of the results of the evaluation has led to constructive suggestions about avenues of future improvement of the lexicon. These improvements will address well-known challenges in automatic natural language understanding (polysemy of semantic heads, polysemy of case-role fillers, and metaphorical usages) that are applicable both to MWEs and to compositional elements alike.

6. Conclusions

This paper has presented the OntoSem approach to recording and processing MWEs within the context of a comprehensive, semantically-oriented text understanding system. We have shown that there is no reason to accord MWEs special treatment: they can be lexically recorded and computationally analyzed using the same methods that are used at the level of individual words.

Whether MWEs are recognized as a separate category or not, the core text processing challenge remains the same: residual ambiguity. The good news about residual ambiguity is that it can often be overcome by tightening lexical and ontological descriptions. The bad news about residual ambiguity is, possibly surprisingly, exactly the same as the good news: it needs to be treated by tightening lexical and ontological descriptions. This means high-quality knowledge engineering. And for high-quality knowledge engineering to take place at a reasonable scale, the field needs to expand its definition of what is valuable in NLP-oriented endeavors. As long as value is attributed almost exclusively to machine learning methods, and as long as knowledge acquisition, in the guise of corpus annotation, is placed primarily in the hands of only lightly trained annotators (which, naturally, imposes constraints on the level of sophistication of such annotations), the field will have no choice but to continue to focus on language problems that can be treated using surfacy methods. This method-driven restriction of the problem space will certainly not last forever: the ceiling of results is not sufficient even for run-of-the-mill NLP applications, let alone language-endowed intelligent agents. We hope that this paper will contribute to compiling a critical mass of evidence for the feasibility of semantically-oriented NLP that will give rise to a paradigm shift that is long overdue.

Acknowledgments

This research was supported in part by Grant N00014-09-1-1029 from the U.S. Office of Naval Research. All opinions and findings expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research. Our thanks to reviewers of *Lingvisticæ Investigationes* for their helpful comments on a draft of this paper.

References

Al-Haj, H., Itai, A., & Wintner, S. (2014). Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2), 130–170. DOI: 10.1093/ijl/ect036

Beale, S., Nirenburg, S., & Mahesh, K. (1995). Semantic analysis in the Mikrokosmos machine translation project. In *Proceedings of the 2nd Symposium on Natural Language Processing*, Bangkok, Thailand.

Cacciari, C., & Tabossi, P. (1993). *Idioms: Processing, structure and interpretation*. Lawrence Erlbaum and Associates, Inc.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC 2002) (pp. 1934–1940), Las Palmas, Canary Islands.

de Marneffe, M., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Fillmore, C., Kay, P., & O'Connor, C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, 64, 501–538. DOI: 10.2307/414531

Fillmore, C., Wooters, C., & Baker, C. (2001). Building a large lexical databank which provides deep semantics. In B. Tsou & O. Kwong (Eds.), *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 2001.

Gildea, G., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3), 245–288. DOI: 10.1162/089120102760275983

Grégoire, N. (2010). DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44, 23–39. DOI: 10.1007/s10579-009-9094-z

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42(1), 21–40. Springer Netherland. DOI: 10.1007/s10579-007-9048-2

Klein, D., & Manning, C. D. (2003a). Fast exact inference with a factored model for natural language parsing. In *Proceedings of Advances in Neural Information Processing Systems 15 (NIPS 2002)*, (pp. 3–10). Cambridge, MA: MIT Press.

Klein, D., & Manning, C. D. (2003b). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* - vol. 1.

Levin, B. (1993). *English verb classes and alternations*. University of Chicago Press.

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

McShane, M., Fantry, G., Beale, S., Nirenburg, S., & Jarrell, B. (2007). Disease interaction in cognitive simulations for medical training. In *Proceedings of MODSIM 2007* (pp. 74–80). Virginia Beach, Sept. 11-13, 2007.

McShane, M., Beale, S., Nirenburg, S., Jarrell, B., & Fantry, G. (2012). Inconsistency as diagnostic tool in a society of intelligent agents. *Artificial Intelligence in Medicine (AIIM)*, 55(3), 137–148. DOI: 10.1016/j.artmed.2012.04.005

McShane, M., & Nirenburg, S. (2013). Use of ontology, lexicon and fact repository for reference resolution in Ontological Semantics. In A. Oltramari, P. Vossen, L. Qin, & E. Hovy (Eds.), *New trends of research in ontologies and lexical resources: Ideas, projects, systems* (pp. 157–185). Springer. DOI: 10.1007/978-3-642-31782-8_9

McShane, M., Nirenburg, S., & Beale, S. (In press). Language understanding with Ontological Semantics. *Advances in Cognitive Systems*.

McShane, M., Beale, S., & Babkin, P. (2014). Nominal compound interpretation by intelligent agents. *Linguistic Issues in Language Technology (LiLT)*, vol. 10. 1–36. July.

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

Nirenburg, S., McShane, M., & Beale, S. (2008). A simulated physiological/cognitive "double agent". In J. Beal, P. Bello, N. Cassimatis, M. Coen, & P. Winston (Eds.), *Papers from the AAAI Fall Symposium, Naturally Inspired Cognitive Architectures*. Washington, D.C., Nov. 7–9. AAAI technical report FS-08-06, Menlo Park, CA: AAAI Press.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., & Woliński, M. (2014). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*. Association for Computational Linguistics and Dublin City University, 83–91.

Raskin, V., & Nirenburg, S. (1998). An applied Ontological Semantic microtheory of adjective meaning for natural language processing. *Machine Translation*, 13(2-3), 135–227. DOI: 10.1023/A:1008039100776

Savary, A. (2008). Computational inflection of multi-word units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1, 1–53.

Schone, P., & Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*, Pittsburgh, PA.

Sharoff, S. (2004). What is at stake: A case study of Russian expressions starting with a preposition. In *Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing* (pp. 17–23). July. DOI: 10.3115/1613186.1613189

Stock, O., Slack, J., & Ortony, A. (1993). Building castles in the air: Some computational and theoretical issues in idiom comprehension. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure and interpretation* (pp. 229–248). Lawrence Erlbaum and Associates, Inc.

Vendler, Z. (1963). The grammar of goodness. *The Philosophical Review*, 72(4), 446–465. DOI: 10.2307/2183030

Vendler, Z. (1968). *Adjectives and nominalization*. The Hague: Mouton.

Venkatsubramanyan, S., & Perez-Carballo, J. (2004). Multiword expression filtering for building knowledge. In *Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing* (pp. 40–47). DOI: 10.3115/1613186.1613192

Abstract

The Ontological Semantic Treatment of Multiword Expressions

This paper describes, and presents a formal evaluation of, the Ontological Semantic approach to automatically analyzing multiword expressions. It shows how multiword expressions can be lexically recorded and processed in the same way as compositional argument-taking words. It suggests that the component modeling strategies are psychologically plausible and hold promise for supporting the development of sophisticated, language-endowed intelligent agents.

Keywords: multiword expressions, phraseology, idioms, computational semantics

Authors' address:

Marjorie McShane, Sergei Nirenburg, Stephen Beale
Cognitive Science Department, Carnegie 108
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY, 12180
USA

{margemc34, zavedomo, stephenbeale42}@gmail.com

Received 28/07/2014