The Description and Processing of Multiword Expressions in OntoSem

Marjorie McShane, Sergei Nirenburg and Stephen Beale

Working Paper 07-05

March 8, 2005

Institute for Language and Information Technologies
University of Maryland Baltimore County

Abstract

The paper discusses the lexical description and runtime treatment of multiword expressions in the Ontological Semantic (OntoSem) text processing environment. We show how our formalism permits the encoding of syntactic structures and semantic entities of any degree of complexity, and how those descriptions support the automatic generation of high quality, formal text-meaning representations from unrestricted input text. We argue for the need for high quality, largely manually crafted knowledge if NLP applications are to become really sophisticated. We shift the focus from the automatic detection of MWEs (the most widespread interest of late) to the encoding of their meaning, since detection without meaning can support only shallow NLP applications.

1 Introduction

Much previous work on multiword expressions (MWEs) within computational linguistics has centered around either the automatic detection of MWEs in corpora, the treatment of MWEs in parsing, or descriptive and processing approaches to verb-particle constructions.[1] Our Ontological Semantics (OntoSem) group, by contrast, has a different interest in multiword expressions: recording the meaning of any and all types of multiword expressions such that they can be fully incorporated into the text-meaning representations (TMRs) generated by our OntoSem semantic analyzer. In this paper we describe how the syntax and semantics of multiword expressions are recorded in the OntoSem lexicon and how this information is used during parsing and TMR creation from input text. Here, as in all our work, the emphasis is on the end-to-end treatment of any and all phenomena texts can present in a practical, application-oriented system.

[1] See, e.g., the bibliography from Stanford's Multiword Expression Project http://mwe.stanford.edu/bib.html, or the the program of the ACL 2003 workshop entitled *Multiword Expressions: Analysis, Acquisition and Treatment.*

2 A Snapshot of OntoSem

OntoSem is a text processing environment that takes as input unrestricted text and carries out its tokenization, morphological analysis, syntactic analysis, and semantic analysis to yield TMRs. Text analysis relies on:

- The OntoSem language-independent **ontology**, which is represented using its own metalanguage and currently contains around 5,500 concepts, each described by an average of 16 properties ("features"), selected from the hundreds of properties defined in the ontology. The number of concepts is intentionally restricted, so that mappings from lexicons are many to one.
- An OntoSem **lexicon** for each language processed, whose entries contain (among other information) syntactic and semantic zones (linked through special variables) as well as procedural-semantic attachments that we call "meaning procedures." The semantic zone most frequently invokes ontological concepts, either directly or with modifications, but can also describe word meaning extra-ontologically, for example, in terms of parameterized values of modality, aspect, time, etc., or combinations thereof.
- A **fact repository**, which contains real-world facts represented as numbered "remembered instances" of ontological concepts (e.g., SPEECH-ACT-3186 is the 3186th instantiation of the concept SPEECH-ACT in the world model constructed during text processing as the embodiment of text meaning).
- The OntoSem text **analyzers**, which cover tokenization, morphological, syntactic and semantic analysis, and TMR creation.
- The TMR language, which is the **metalanguage** for representing text meaning, compatible with the metalanguage of the ontology and the fact repository.

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 model being built during the processing of the given text(s). Its agent is NATION-213, which is the key for the US in our fact repository. The theme of the event is the 7th instantiation of WAR-ACTIVITY in this analyzer run.

This TMR is what we call a *basic TMR*, since it reflects basic semantic dependency building, including the resolution of syntactic and semantic ambiguity. There is, however, another level of processing, during which specialized reasoners about language and the world are launched in order to further concretize TMRs. The resulting TMRs – called *extended TMRs* – show the calculated values of various modalities, aspect, time, reference resolution, speaker attitudes, etc. So, extensions to the above simple TMR would: a) include as specific a value for time as possible (i.e., the time of speech must be extracted from the text so that the past-tense verb can be interpreted), and b) attempt to link WAR-ACTIVITY-7 to the appropriate coreferential WAR-ACTIVITY in the fact repository (since the text contains ***the*** *war*, with a definite article, we know that there must be some coreferential war either in the preceding context or available as an aspect of general world knowledge, which should be stored in our fact repository). Since automated reasoning is error prone, all calculated values in extended TMRs are understood as defeasible.

The point of this, by necessity, brief description of the OntoSem environment is that our goal is to automatically produce fine-grained semantic representations of texts which draw upon lexical and ontological resources as well as reasoning about language and the world. These representations can be, and in fact have been, used as the basis for many applications. For example, our TMRs were used as the basis for reasoning in the question-answering system AQUA, where it supplied knowledge to enable the operation of the JTP (Fikes et al., 2003) reasoning module. Details of this approach to text processing can be found, e.g., in Nirenburg and Raskin forthcoming and Nirenburg et al. 2003. The ontology itself, a brief ontology tutorial, and an extensive lexicon tutorial can be viewed at http://ilit.umbc.edu. See Nirenburg et al. 2004 for a description of our evaluation methodology and the current level of OntoSem TMR output.

3 Basics of OntoSem Lexical Specification

The lexical description and processing of MWEs in OntoSem differs very little from the description and processing of single tokens: a few special expressive means for syntax are necessary (e.g., means of listing punctuation within an MWE), but no special expressive means for semantics are needed. This is natural since complex semantics is encoded all the time in TMRs; encoding it for MWEs simply means doing it in the lexicon explicitly rather than preparing the system to do it compositionally, as is done for non-MWE input. So MWEs are handled as a regular part of our work, not a special topic. In this way OntoSem differs from most NLP systems (cf. Sag et al.'s representative title "Multiword Expressions: A Pain in the Neck for NLP").

A simple OntoSem lexicon entry for a transitive verb (in presentation format) is as follows:

```
watch
  watch-v1
    synonyms "observe"
    anno
       definition  "to observe, look at"
       example "He's watching the demolition team."
    syn-struc
       subject          $var1    cat n
       v                $var0    cat v
       directobject     $var2    cat n
    sem-struc
       VOLUNTARY-VISUAL-EVENT
         agent          ^$var1
         theme          ^$var2
```

The syntactic structure (syn-struc) says that this is a transitive sense of *watch* and the semantic structure (sem-struc) says that a VOLUNTARY-VISUAL-EVENT – which is a concept in our ontology – must be instantiated in the TMR. The variables are used for linking, so, for example, the syntactic subject is linked to the meaning of the AGENT of the VOLUNTARY-VISUAL-EVENT (^ is read 'the meaning of').

When writing syntactic structures for single token head words, like garden variety transitive, bitransitive or intransitive verbs, we refer to syntactic functions like subject, direct object, indirect object, xcomp (a verbal complement headed by an infinitive), comp (a clausal complement headed by an optional 'that'), etc., as in the example above. By contrast, when writing syntactic structures for MWEs we often refer to immediate constituents, like NP, N, Adj, Prep, Conj, etc. This is necessary because of the often idiosyncratic makeup of MWEs. We can specify any sequence of elements, including punctuation and optional elements, without the superfluous (for purposes of our syntactic/semantic analyzer) requirement of creating from them canonical "well formed" syntactic structures. The drawback of this is that the ordering of these constituents must be assumed to be fixed by the analyzer.[2] This is in contrast to the case where functional categories, such as subject and directobject, are used: for these, generalized rules of ordering can be applied.[3]

Although in Section 4 we will illustrate OntoSem's approach to MWEs in some detail, let us begin with one example of a moderately complex syntactic structure for a MWE. The phrasal *X cannot help but Y* is recorded with all of its syntactically optional constituents as a verbal sense of *help.* This syn-struc covers input such as *He could not help but laugh, He could not help but pity her, He cannot help but think to himself that this is all wrong,* etc.

```
syn-struc
   subject     $var1
   v           $var2    root can
   verb-neg    $var3                ; indicates 'not'
   v           $var0    form infinitive  ; bare verb form
   conj        $var4    root but
   v           $var5    form infinitive
   pp
      $var6    root to  opt +
      obj      $var7
   comp        $var8    opt +
```

This syn-struc exemplifies the possibility of specifying any combination of syntactic constituents in an OntoSem lexicon entry. With this as a baseline, we now turn to what we consider the most important aspect of the description of MWEs: their meanings.

4 An Extended Example

To further describe the OntoSem approach to MWEs we will use selected examples from the more than 40 senses of the verb *go* in the OntoSem lexicon of English. The selected senses represent phrasal entities that can be disambiguated by the OntoSem syntactic-semantic analyzer based on their detailed syntactic and semantic descriptions. As an ordering principle for the description of MWEs in OntoSem, we delineate three types of entries based upon the means by which given senses of MWEs are recorded so as to optimize parsing and semantic disambiguation.

Category 1: Lexical Entities in the Syn-Struc Delimit the MWE

As shown above in the example *X cannot help but Y,* specific lexical items can be indicated in a syn-struc: in that example, $var2 must have the root *can,* and $var4 must have the root *but.* This is, in fact, a typical way of creating a phrasal entry in OntoSem: one constrains the input that will match a given lexical sense by associating strings with syntactic constituents. Another such example is *go unpunished,* as in *The crime went unpunished.*

```
go-v25
 anno
   definition  "phrasal: go unpunished"
   example "The crime went unpunished."
 syn-struc
    subject        $var1
    v              $var0
    adj            $var2     root unpunished
 sem-struc
    PUNISH
      theme        ^$var1
      epistemic    0
```

[2] For English, fixing the ordering of components tends not to be a big problem, but for languages with freer word order, like Russian, either special rules will be needed or the different possible orderings will need to be listed in separate lexical senses.

[3] Using syntactic function labels also supports porting of "typical" lexicon entries among languages, since properties of subject, direct object, etc., should be dealt with globally by the syntactic analyzer. MWEs, by contrast, tend to be idiosyncratic for each language, so the decreased portability due to the use of immediate constituent labels represents a relatively minor loss in efficiency.

The sem-struc says that there is no PUNISH event whose theme is the meaning of $var1, since epistemic modality with a value of 0 (which is OntoSem's way of encoding negation) is attributed to the event. Compare this with the phrasal *go free*, as in *The prisoner went free 2 years before his term was out*. The lexical sense covering this MWE would have the same syn-struc as above apart from the root of the adj being listed as *free*. Its meaning, however, would be quite different: the sem-struc would be headed by an IMPRISON event with [phase: end], and the theme of that event would be linked to the syntactic subject.

Consider one final example of a MWE with *go* for which listed lexical elements in the syn-struc delimit the application of the MWE: *whatever X says goes*, as in *Whatever the boss says goes.*

```
go-v31
  anno
    definition  "phrasal: whatever x says goes; x has full
                 authority"
    example "Whatever the boss says goes."
  syn-struc
      np              $var1    root whatever
      subject         $var2
      v               $var3    root say
      v               $var0
  sem-struc
      AUTHORITY-ATTRIBUTE
          domain      ^$var2
          range       1
      ^$var1          null-sem +
      ^$var3          null-sem +
```

In this syn-struc, we indicate that $var2 is the subject, rather than just an NP, to ensure checking of inflectional features between the subject and the verb. However, there would be no benefit to our processors gained by assigning any specific grammatical function to *whatever* since (a) there are no general rules of English that should be applied to it and (b) the MWE actually does have fixed element order, so the syn-struc might as well be interpreted as having fixed word order.

Category 2: Syntactic structure + semantic constraints on a variable delimits use

A typical method for ensuring disambiguation during OntoSem text processing is to constrain the semantic expectations of verbal arguments. For example, the phrasal verb *go down* can have different meanings depending on the theme of the event. For example, if a WATER-VEHICLE goes down, it means that it is the THEME of a SINK event, whereas if an AIR-VEHICLE goes down, it is the THEME of a FALL-AND-HIT event. Thus, the combination of syntactic structure (subject + *go* + *down*) and the semantic class of the subject (WATER-VEHICLE or AIR-VEHICLE) determines which meaning is intended. This is recorded in the lexicon as follows, using SINK as the example:

```
go-v12
  anno
      definition  "phrasal 'go down', of ships; to sink"
       example "The ship went down in the storm."
  syn-struc
      subject     $var1
       v          $var0
       adv        $var2    root down
  sem-struc
     SINK
        theme    ^$var1   sem WATER-VEHICLE
     ^$var2      null-sem +
```

This representation says that the syntactic structure contains a subject headed by $var1, plus the verb *go* in any form (indicated by $var0 – the head word in the entry), plus the adverb *down*, which is also associated with a variable for linking purposes. (If other adjuncts are present in the input sentence, their meaning is computed productively.) The semantic structure is headed by the concept SINK, whose theme is the meaning of the NP indicated by $var1. The semantic restriction on this theme is that it must be a WATER-VEHICLE. The lexical sense for the FALL-AND-HIT meaning would look the same except that the semantic constraint on the theme of FALL-AND-HIT would be AIR-VEHICLE.

There are at least two more uses of *go down* that must be covered: first, contexts in which the subject is neither a WATER-VEHICLE nor an AIR-VEHICLE, in which case *down* will be a preposition followed by a complement and the meaning will be compositionally analyzed (*He went down the stairs*); second, non-idiomatic use of *go down* if the subject is a WATER-VEHICLE or AIR-VEHICLE, as in *When the plane went down my ears popped.* This latter case cannot be prepared for lexically and the ambiguity must be resolved using script-based reasoning, which we are just beginning to

support in OntoSem. Currently, the analyzer will prefer MWE lexical senses over compositional readings (although semantic constraints can override this preference).

Entries in which the syntactic structure plus semantic constraints on a variable delimit the use of the given phrasal sense show many variations. For example, the semantic constraints can include more than one option, as in the following example:

```
go-v15
  anno
    definition  "phrasal 'go into', of documents: describe,
                 treat in detail"
    example     "That message goes into the reasons
                 for the bailout."
  syn-struc
    subject     $var1
    v           $var0
    pp
      prep        $var2   root  into
      obj         $var3
sem-struc
  ABOUT-AS-TOPIC
    domain    ^$var1   sem BROADCAST, PRINTED-MEDIA
    range     ^$var3
```

The sem-struc is headed by the ontological RELATION called ABOUT-AS-TOPIC, whose domain (which is linked to the syntactic subject) is BROADCAST/PRINTED-MEDIA and whose range (i.e., what the former is about) is the meaning of the object of *into*, referred to as $var3. By contrast, if the same phrasal, *go into*, has a human subject (*Their boss went into the reasons why they were fired*), then it has a different semantic representation altogether, which is not a relation but rather an event: DESCRIBE, whose AGENT is the syntactic subject and whose THEME is the syntactic direct object.

Whereas in some cases semantic restrictions on dependent elements are specified in a given lexical sense, often they need not be because the ontology itself provides sufficient information to ensure disambiguation. For example, the MWE *go off* should be interpreted as EXPLODE only if the subject is an EXPLOSIVE-DEVICE *(The bomb went off*).

```
go-v19
  anno
    definition  "phrasal 'go off', of explosives; explode"
    example "The bomb went off at 5 a.m."
  syn-struc
    subject     $var1
    v           $var0
    prep        $var2     root off
  sem-struc
    EXPLODE
      instrument   ^$var1
    ^$var2         null-sem +
```

In the lexical entry for this sense, however, there is no need to specify that ^$var1 is semantically limited to explosives because the ontological concept EXPLODE has its INSTRUMENT restricted to explosive devices. Therefore, this sense of *go off* will only be selected if the ontologically listed restrictions on ^$var1 match the expectations encoded for the INSTRUMENT of EXPLODE in the ontology.

The same richness of semantic description is available for MWEs as for all other entities in the OntoSem lexicon. For example, in lieu of expressing meaning through a direct or modified ontological mappings, meaning can be expressed by extra-ontological means, e.g., by values of mood, aspect or time. For example, *go out*, when used of cigarettes or fires, means to stop burning. This is represented by the concept BURN, whose theme is linked to the subject of *go out* and whose phase is indicated as "end": i.e., there is a BURN event and it is over.

```
go-v24
  anno
    definition  "phrasal 'go out'; stop burning"
    example "The fire went out before midnight"
  syn-struc
    subject     $var1
    v           $var0
    prep        $var2  root out
sem-struc
    BURN
       theme      ^$var1
       phase      end
    ^$var2     null-sem +
```

As above, this sense will be chosen only if the theme of BURN indicates a typically burnable thing, as recorded in the ontology.

Category 3: Strictly Fixed Phrases

Phrases that do not permit intra-phrase modification, like *Secretary of State* and *stock market*, can be listed as head words in the OntoSem lexicon, conjoined by an underscore for formal reasons.

```
stock_market-n1
  anno
    definition  "phrasal: a market organized for the buy-
                 ing and selling of stocks and bonds"
  syn-struc
    root    $var0
  sem-struc
    STOCK-EXCHANGE-MARKET
```

Such phrases, by virtue of their being included in the lexicon, will be identified and treated as a single word by the morphological analyzer. Both intra-phrase and end-of-the entity inflections are handled; so our morphological preprocessor will analyzer *attorneys general* as *attorney_general* plus a feature identifying it as plural, and the same will be done for *stock markets* based on the headword *stock_market*. On the other hand, if a phrase can have intra-phrase modifiers, then this approach is not possible, and one of the first two categories presented above will need to be used. For example, the prototypical phrasal *kick the bucket* cannot be treated as a single word because of the possibility of adding a modifier, such as in *He kicked the proverbial bucket*.

5 Processing MWEs

As mentioned above, processing MWEs in OntoSem is no different from processing any other lexical items. Analysis of all input is driven by syntactic and semantic expectations encoded in lexicon entries. In a non-MWE example like *watch* (Section 3), the analyzer parses the sentence drawing upon the expectations that the verb *watch* will be preceded by a subject and followed by a direct object (with possible diatheses, like passive, encoded in rules within the analyzer). In addition, the semantic constraints in the ontological frame for VOLUNTARY-VISUAL-EVENT set up expectation that the AGENT will be an ANIMAL and the THEME will be a PHYSICAL-OBJECT, PHYSICAL-EVENT or SOCIAL-EVENT. If any of these constraints are not met by the input, another lexical sense of *watch* that better meets the constraints is selected. The same processing is applied to MWEs, the only difference being that, in most cases, MWEs set up more constrained expectations based on, for example, specific lexical items being listed directly in the syn-struc. For a more in-depth description of the functioning of the OntoSem syntactic-semantic analyzer, see, e.g., Beale 2003 and Nirenburg 2003.

6 Final Thoughts

In response to the anticipated – and natural – question of how one can realistically scale up a system for which such extensive manual acquisition work is needed, we present the following for consideration.

First, *without* such knowledge, high-level NLP will not be achieved. In fact, even most practitioners in stochastic methods want and need more and deeper knowledge as heuristic support. The well-studied task of automatically extracting MWEs – while important for parsing – will not get us any closer to being able to represent their meaning. And the need for rich, formal semantic representations is great: for example, all work in reasoning requires such representations but the field is at a loss regarding where to obtain them.

Second, the ILIT team is in the process of building just such a lexicon and we see that it *can* be done in finite time. To speak in concrete numbers, it took less than 1 person year to create an English OntoSem lexicon of over 12K senses that includes, among other things, the entire closed class, the most polysemous, difficult verbs, and many MWEs. Since this preliminary work included developing approaches to the treatment of many types of phenomena (microtheories of time, reference resolution, the incorporation of lexically-specified calls to procedural semantic programs, etc.), the fact that work will significantly speed up in the future goes without saying. Moreover, importing lexicons and ontologies of terminology, as for the medical domain (a work in progress at ILIT) has much potential to increase lexicon and ontology size with little extra effort expended. That being said, however, the most difficult lexicographical and ontological work still lies in the basic, everyday vocabulary (e.g., how to describe the concept *love* in a non-iconic way).

Third, the core lexicon can be efficiently ported to other languages, as described in McShane et al. 2004. This is because the OntoSem ontology and fact repository are language independent, and the lexicon and processors are parametrizable in well understood ways.

Finally, the ubiquitous economics-oriented judgment about the impracticality of the knowl-

edge-based approach has been accepted in the field without questioning for the last fifteen years or so. But this argument is a double-edged sword. Over the same period, a large amount of resources has been expended on methods that claim no need for manual acquisition. In the process, two things became clear: a) the field relinquished the goal of attaining representations of text meaning and concentrated on lower-level tasks that fit the method of choice better; and b) the amount of manual knowledge acquisition to support these methods has proved to be anything but negligible – only instead of acquiring lexicons and other knowledge bases the acquisition centered on annotated corpora. We believe that it will be beneficial for the long-term progress in the field to reassess the utility and indispensability of recorded static knowledge resources.

References

Beale, Stephen, Sergei Nirenburg and Marjorie McShane. (2003). Just-in-Time grammar. Proceedings 2003 International Multiconference in Computer Science and Computer Engineering. Las Vegas, Nevada

Fikes, R., J. Jenkins , and G. Frank (2003). JTP: A System Architecture and Component Library for Hybrid Reasoning. Proceedings of the Seventh World Multiconference on Systemics, Cybernetics, and Informatics. Orlando, Florida, USA.

McShane, Marjorie, Sergei Nirenburg, Stephen Beale, Margalit Zabludowski. (2004). OntoSem and SIMPLE: Two Multi-Lingual World Views. Submitted to the ACL workshop on Text Meaning and Interpretation.

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

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

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

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. (2002). Multiword Expressions: A Pain in the Neck for NLP, In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, Mexico, pp. 1-15.