OntoSem and SIMPLE: Two Multi-Lingual World Views

Marjorie MCSHANE, Margalit ZABLUDOWSKI, Sergei NIRENBURG and Stephen BEALE
Institute for Language and Information Technologies (ILIT)
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250 USA
marge@umbc.edu, margalit@rcn.com, sergei@umbc.edu, sbeale@umbc.edu

Abstract

In this paper we compare programs of work that aim to develop broad coverage cross-linguistic resources for NLP: Ontological Semantics (OntoSem) and SIMPLE. The approaches taken in these projects differ in three notable respects: the use of an ontology versus a word net as the semantic substrate; the development of knowledge resources inside of as opposed to outside of a processing environment; and the development of lexicons for multiple languages based on a single core lexicon or without such a core (i.e., in parallel fashion). In large part, these differences derive from project-driven, real-world requirements and available resources – a reflection of their being practical rather than theoretical projects. However, that being said, we will suggest certain preferences regarding the content and development of NLP resources with a view toward both short- and long-term, high-level language processing goals.

1 Introduction

Ontological Semantics (OntoSem) is a multi-lingual text processing environment that takes as input unrestricted text and, using a suite of static resources and processors, automatically creates text-meaning representations (TMRs) which can then be used as the basis for any NLP application, including MT, question answering, summarization, etc. OntoSem knowledge resources are developed in coordination with each other and with the processors they serve. Some of the resources are fully language independent while others are readily parameterizable, wherein lies the cross-linguistic portability of the system. Although in this paper, we focus on OntoSem lexicons, a crucial point is that they are not built in isolation but, rather, in an integrated environment where their utility can be tested and evaluated in a variety of practical applications.

The SIMPLE project takes a different approach to achieving the dual goals of multilinguality and resource utility across applications. It aims to develop compatible (they use the term "harmonised") lexicons for 12 European languages, attempting to foresee what will be most useful for applications but without referring to any particular processors that will use the information and without building any other knowledge resources to share the burden of semantic specification. As we will show, these different points of departure lead to quite different realizations of cross-lingual lexicons for NLP.

2 Overview of SIMPLE

The SIMPLE project is developing 10K-sense "harmonised" semantic lexicons for 12 European Union languages (Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish), continuing the earlier PAROLE project, which developed 20K-sense morphological and syntactic lexicons for these languages. The lexicons are monolingual and are developed independently, with the word stock based on corpus evidence for each language. To ensure some overlap of lexical senses, certain Base Concepts of EuroWordNet must be covered in each language (462 nominal, 187 verbal and 185 adjectival Base Concepts that were culled and cleaned from EuroWordNet). This overlap will permit direct interlinking among languages; interlinking of the rest of the lexical stock is slated as future work. Pustejovsky's four Qualia (which are, essentially, properties expressing formal, agentive, constitutive and telic meanings; see Pustejovsky 1995) are used to specify certain aspects of word meaning, and a common library of 140 template types is used to guide acquisition in all languages (Lenci et al 2000a, 2000b).

Lenci et al. 2000b (p. 5) summarize the information that can be represented in a SIMPLE lexicon entry: "i) semantic type, corresponding to the template the SemU (semantic unit) instantiates; ii) domain information; iii) lexicographic gloss; iv) argument structure for predicative SemUs; v) selectional restrictions on the arguments; vi) event type, to characterise the aspectual properties of

verbal predicates; vii) link of the arguments to the syntactic subcategorization frames, as represented in the PAROLE lexicons; viii) Qualia Structure; ix) information about regular polysemous alternation in which a word sense may enter; x) cross-part-of-speech relations (e.g. *intelligent* - *intelligence*; *writer* - *to write*); xi) synonymy."

Below is the SemU for a sense of *lancet*, instantiating the template **Instrument** (from Palmer et al. 2000).

Instrument

Usem:	Lancet
BC number:	
Template_Type:	**[Instrument]**
Unification_path:	**[Concrete_entity|Artifact**$_{\text{Agentive}}$ **| Telic]**
Domain:	*Medicine*
Semantic Class:	*Instrument*
Gloss: a surgical knife with a pointed double-edged blade; used for punctures and small incisions	
Pred_Rep.:	<Nil>
Selectional Restr.:	<Nil>
Derivation:	<Nil>
Formal:	*isa* (<lancet>, <knife>: [Instrument])
Agentive:	*created_by* (<lancet>, <make>: [Creation])
Constitutive:	*made_of* (<lancet>, <metal>: [Substance]) *has_as_part* (<lancet>, <edge>: [Part])
Telic:	*used_for*(<lancet>, <cut>: [Constitutive_change]) *used_by* (<lancet>, <doctor>)
Synonymy:	<Nil>
Collocates:	*Collocates* (<SemU1>,…,<SemUn>)
Complex:	<Nil>

While the SIMPLE project is certainly producing useful resources, we would suggest that the lexical information and structure are being overly constrained by the frameworks selected, which we will comment on briefly in preparation for an extended comparison between OntoSem and SIMPLE in section 4.

EuroWordNet is being used as the anchor for semantic description in SIMPLE. However, like the original English WordNet, it is not a property-rich ontology but, rather, a hierarchical net of lexical items whose use in NLP has the same pitfalls as any non-ontological word net (e.g., lack of disambiguating power and lack of sufficient relations between entities; see Nirenburg 2004c for a discussion of the insufficiency of WordNet for NLP). In order to make up for the sparsity of information in the semantic substrate, the SIMPLE lexicons contain what would, we believe, be more efficiently recorded in a single, sufficient ontology. For example, when the lexicon acquirers for each language use the Instrument template to describe *lancet*, they must rerecord in the lexicon of each L all of the language-independent property values for this lexical item, like the values for the four Qualia (formal, agentive, constitutive, telic), the domain, the unification path, etc. This is significant redundancy and, moreover, there is no guarantee that acquirers will arrive at the same decisions, either through error, oversight or competing analyses of the phenomena in question.

Another, in our view, insufficiently explained aspect of SIMPLE is the priority given to Qualia as descriptors of lexical items. The original inventory of Qualia (from Pustejovsky 1995) consists of only four properties of the hundreds that can usefully be used to link concepts for purposes of NLP. Lenci et al. (2000b) address this issue as follows: "Although they [the four Qualia] clearly do not exhaust the semantic content of lexical items, Pustejovsky (1995) has convincingly shown that these four Qualia dimensions play a particularly prominent role in determining the linguistic behavior of word senses, as well as in the explanation of the generative mechanisms at the basis of lexical creativity. Qualia-based information can be specified for all the parts of speech, although *prima facie* it seems to be more directly suitable for the characterization of certain types of nominals". However, there is large gap between theoretical interest and practical application: in fact, because of this, the SIMPLE project has moved toward an *Extended Qualia Structure* with more fine-grained subtypes of given Qualia.[1]

In conclusion, we believe that SIMPLE is pursuing useful goals that could be pursued in even more useful ways by shifting the focus from lexicon-only work to integrated work within an environment in which ontological and lexical resources are developed together and where extant types of processors can be used to test the value of resources as they are developed.

3 Overview of OntoSem

The OntoSem approach to lexicon and ontology acquisition differs from that used in SIMPLE

[1] As an aside, we see a parallel between focusing on Qualia in lexical description and, for example, focusing on classes of verbs with respect to their alternations, as is done, for example, in Levin 1995. While the descriptions that derive from theoretically-driven research such as this can certainly be useful, when it comes to writing "well-rounded" semantic descriptions of words for large-scale systems, there is no distinction between a Quale and other properties. Similarly, the fact that a verb belongs to some group with respect to alternations is no more or less important than its other potential group membership along other parameters. See Nirenburg and Raskin 2004 for further discussion of these and related issues.

because OntoSem is an integrated text processing environment, meaning that knowledge resources are crafted hand-in-hand with each other and with processors such that responsibility for various analysis tasks can be distributed in an ideal (to the degree of our understanding) way.

OntoSem takes as input unrestricted raw text and carries out preprocessing, morphological analysis, syntactic analysis and semantic analysis, with the results of semantic analysis represented as formal text-meaning representations (TMRs) that can then be used as the basis for a wide variety of NLP applications. Text analysis relies on:

- The OntoSem language-independent **ontology**, which is written using a metalanguage of description and currently contains around 5,500 concepts, each of which is described by an average of 16 properties. In all, the ontology contains hundreds of properties (which cover the same territory as the Qualia plus much more). Fillers for properties can be other ontological concepts or literals.
- An OntoSem **lexicon** for each language processed, which contains syntactic and semantic zones (linked using variables) as well as calls to "meaning procedures" (i.e., programs that carry out procedural semantics, see McShane et al. 2004a) when applicable. The semantic zone most frequently refers to ontological concepts, either directly or with property-based modifications, but can also describe word meaning extra-ontologically, for example, in terms of modality, aspect, time, etc. The current English lexicon contains approximately 12K senses, including all closed-class items and the most frequent verbs, as indicated by corpus analysis. This English lexicon took less than 1 person year to build and can (as described below) be ported to other languages.
- An **onomasticon**, or lexicon of proper names, which contains approximately 350,000 entries and is growing daily using semi-automated extraction techniques.
- A **fact repository**, which contains real-world facts represented as numbered "remembered instances" of ontological concepts (e.g., SPEECH-ACT-3366 is the 3366th instantiation of the concept SPEECH-ACT in the world model constructed during the given run of the analyzer).
- The OntoSem **text analyzers**, which cover preprocessing, syntactic analysis, semantic analysis, and creation of TMRs. They are largely parameterizable and thus can be ported to other languages.
- The **TMR language**, which is the metalanguage for representing text meaning. A very simple example of a TMR (simple because most of the sentences we process are much longer), which reflects the meaning of the sentence *He asked the UN to authorize the war,* is as follows:

REQUEST-ACTION-69
AGENT	HUMAN-72
THEME	ACCEPT-70
BENEFICIARY	ORGANIZATION-71
SOURCE-ROOT-WORD	ask
TIME	(< (FIND-ANCHOR-TIME))

ACCEPT-70
THEME	WAR-73
THEME-OF	REQUEST-ACTION-69
SOURCE-ROOT-WORD	authorize

ORGANIZATION-71
HAS-NAME	UNITED-NATIONS
BENEFICIARY-OF	REQUEST-ACTION-69
SOURCE-ROOT-WORD	UN

HUMAN-72
HAS-NAME	COLIN POWELL
AGENT-OF	REQUEST-ACTION-69
SOURCE-ROOT-WORD	he ; *ref. resolution done*

WAR-73
THEME-OF	ACCEPT-70
SOURCE-ROOT-WORD	war

Details of this approach to text processing can be found, e.g., in Nirenburg et al. 2004a,b. The ontology itself, a brief ontology tutorial, and an extensive lexicon tutorial can be viewed at http://ilit.umbc.edu.

OntoSem has been used with languages including English, Spanish, Chinese, Arabic and Persian, to varying degrees of lexical coverage (e.g., earlier, less fine-grained English and Spanish lexicons contained 40K entries and were used for MT in the Mikrokosmos project). What makes OntoSem amenable to efficient cross-linguistic usage is that many of the resources are either fully language independent (the ontology, the fact repository, the TMR metalanguage) or parameterizable in well understood ways. Here we focus on exploiting cross-linguistic similarity for lexical acquisition, but a similar analysis could be applied to the OntoSem analyzers.

3. 1 OntoSem Lexicons

A basic verbal lexicon entry in OntoSem looks as follows (in presentation format):

watch
  watch-v1
    synonyms "observe"
    anno
      definition "to observe, look at"
      example "He's watching the competition."

```
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').

Apart from mapping directly to an ontological concept, there are many other – and more complex – ways to express meaning in OntoSem. For example, one can map to an ontological concept with modified property values: e.g.,

- **Zionist** is described as a POLITICAL-ROLE that is the AGENT-OF a SUPPORT event whose THEME is Israel.
- **asphalt (v.)** is described as a COVER event whose INSTRUMENT is ASPHALT.
- **recall (v.** as in *they recalled the high chairs*) is described as a RETURN-OBJECT event that is CAUSED-BY a FOR-PROFIT-CORPORATION and whose THEME is ARTIFACT, INGESTIBLE or MATERIAL.

There are also a number of fully or partially non-ontological ways of describing meaning, like the use of parametric values of mood or aspect. For example, the auxiliary *might* as in *He might come over* is described using the modality 'epistemic', which deals with the truth value of a statement:

```
syn-struc
       subject          $var1    cat n
       v                $var0    cat v
       inf-cl           $var2    cat v
sem-struc
   ^$var2
       epistemic    .5
       agent        ^$var1
meaning-procedure
    fix-case-role (value ^$var1) (value ^$var2)[2]
```

Another set of extra-ontological semantic descriptors is used for time expressions, as shown by the example of *yesterday* below.

```
syn-struc
   root      $var1          cat v
   mods      root $var0     cat adv
             type           pre-verb-post-clause
sem-struc
  ^$var1
       time
         combine-time
              (find-anchor-time) (day 1) before
```

As already shown in the examples of *might* and *yesterday,* calls to procedural semantic routines (which may or may not be listed in the meaning-procedure zone of the lexicon entry) are used widely in OntoSem lexical description. This reflects the fact that many aspects of meaning cannot be statically described but, rather, must be computed. An advantage of developing lexical resources within a processing environment is being able to assign responsibility for portions of semantic composition to resources best suited for them.

In addition to the means of lexical expression described above, OntoSem lexicon entries can include entities of any degree of complexity, including phrasals of any profile, as reported in McShane et al. 2004b.

3.2 Porting OntoSem Lexicon Entries Across Languages

As is clear from the examples above, OntoSem provides significant expressive power semantically (not to mention syntactically, which we do not pursue here). Expressive means include mapping to the ontology (which itself is rich in property-value descriptors), mapping to the ontology with lexical supplementation of properties, or referring to extra-ontological microtheories like those that treat time, reference resolution, comparison, ellipsis resolution, modality, aspect, etc. What must be emphasized, however, is how language neutral – and therefore portable across languages – the semantic descriptions are. Whereas it is typical to assume that lexicons are language-specific whereas ontologies are language-independent, most aspects of OntoSem sem-strucs are language-independent, apart from the linking of specific variables to their counterparts in the syn-struc.

[2] This meaning procedure reassigns a case-role if the listed AGENT case-role is inappropriate considering the meaning of $var1 and/or $var2: e.g., in *the truck might come,* truck is a THEME of a MOTION-EVENT, not an AGENT, and in *I might get sick,* I am an EXPERIENCER of a DISEASE event, not an AGENT of it.

Stated differently, if we consider sem-strucs – no matter what lexicon they originate from – to be building blocks of the representation of *word meaning* (as opposed to concept meaning, as is done in the ontology), then the job of writing a lexicon for L2 based on the lexicon for L1 is in large part limited to a) providing an L2 translation for the head word(s), b) making any necessary syn-struc adjustments and c) checking/modifying the linking among variables in the syn- and sem-strucs. This conception of cross-linguistic lexicon development derives in large part from the Principle of Practical Effability (Nirenburg and Raskin 2004), which states that what can be expressed in one language can *somehow* be expressed in all other languages, be it by a word, a phrase, etc.

Apart from this theoretical justification for conceptualizing the sem-strucs as building blocks for lexical representation, there are two practical rationales: supporting consistency of meaning representation across languages and using acquirer time most efficiently in large-scale lexical acquisition.

As regards consistency, the potential for paraphrase must be considered when building multi-lingual resources. For instance, 'weapons of mass destruction' can be described as the union of CHEMICAL-WEAPON and BIOLOGICAL-WEAPON, or it can be described as WEAPON with the ability to KILL > 10,000 HUMANs (the actual number recorded will be treated by the analyzer in a fuzzy fashion; however, it would be less than ideal for a lexicon for L2 to record 10,000 while a lexicon for L3 recorded 25,000). While both representations are valid, it is desirable to use the same one in all languages covered. In addition, the decision of how to describe a notion – whether by ontologizing it, describing it using extra-ontological means, describing it using an existing concept with additional properties and values defined – is often a judgment call. It would not be desirable for the acquirer of German to map the word *Schimmel* 'white horse' to the concept HORSE with the lexical restriction COLOR: WHITE, while the acquirer of some other language that also has a word for 'white horse' introduced an ontological concept specifically for this entity. Again, while both representations are valid and, in this case, semantically equivalent, the general tendency should be to strive toward uniformity where possible.

As concerns acquirer time, composing sem-strucs is, by far, the most time- and effort-intensive aspect of writing OntoSem lexicon entries. This derives from the wealth of expressive means; the fact that microtheories of time, reference, etc., are naturally built during lexicon development (recall that our environment is fully integrated with processors); and the fact that ontology development occurs hand-in-hand with lexicon development. Therefore, work on the first lexicon entry that describes a word sense – regardless of the language of origin – takes much more time than editing a word sense for a new language. Moreover, although in the worst case some editing of entries is necessary for L2, L3, etc., in most cases no such editing is needed. Although one might hypothesize this state of affairs based on cross-linguistic principles, we have tested it in the lexicon-porting experiment described below.

3.3 An English to Polish Lexicon Porting Experiment

For the experiment, a bilingual English/Polish computational linguist took the English OntoSem lexicon as a seed and experimented with various porting methods into Polish.

The primary insight was that while manually porting individual lexical senses is quite straightforward and *will* save time over acquisition from scratch, porting lexicons wholesale is rather more complex. That is, manually providing translations for the senses in L1 is a conceptually relatively simple task, complicated only by the need for the occasional remapping of variables, editing of syntactic structures, omission of given senses due to language lacunae (e.g., a phrasal encoded in L1 might not occur in L2 in a fixed form), etc. However, if one attempts either to (semi-)automate the acquisition process and/or use L1 as a seed lexicon for more "creative" acquisition of L2, the space of options becomes quite broad and must be constrained programmatically in order to actually benefit from the reuse of semantic descriptions.

For example, if a well-trained acquirer of L2 is using L1 as a seed, questions that arise include: Should the base lexicon be left as is (considering that it is known to have incomplete coverage) or should one attempt to improve its quality and coverage while building L2? Should L2 acquisition be driven by correspondences in head words or simply by the content of sem-struc zones (e.g., all English senses of *table* will be in one head entry, and typically will be acquired at once; should all senses of all L2 translations of *table* be handled at once during L2 acquisition or should the L2 acquirer wait until he comes upon sem-strucs that represent the given other meanings of the L2 words)? To what extent should the regular acquisition process – including ontology supplementation – be carried out on L2? The answers to all of these, and more such, questions

depend entirely upon available resources and should be informed by (a) experiments to determine what works best for a given acquirer, and (b) the goals of a given project.

As regards automation, the experiment found that automatically mapping L2 words to L1 OntoSem entries works very well (at well over 90%) when the machine-tractable L1-L2 resource used to support this process has one sense of the given word in the given part of speech and the OntoSem lexicon also has one sense for the given part of speech. The extraction and matching of such senses represents a well-defined, extremely time-efficient task, especially for specialized terminology that tends to have only one sense in any language. When the mapping between senses in the L1-L2 lexicon and the OntoSem lexicon is more than one to one, manual linking of senses (which do not always correspond among the languages) has proved necessary, with the potential benefits of a time-saving interface becoming immediately clear.[3]

4 *Pudding* in SIMPLE and OntoSem

Now we return to the comparison between SIMPLE and OntoSem. We use the example of *pudding*, which is cited in numerous documents related to SIMPLE. The Qualia (in italics) and their values (in boldface) for this word are: *formal* – **substance**; *constitutive* – **ingredients**; *telic* – **eat**; *agentive* – **make**. The stated rationale for encoding these qualia values in SIMPLE lexicon entries is that they are needed to understand the semantics of the sentences like the following (from Lenci et al. 2000b):

a) John refused the pudding (= refused to eat: telic);
b) That's an easy pudding (= easy to make: agentive);
c) There is pudding on the floor (= substance: formal);
d) The pudding came out well (= has been made well: agentive);
e) That was a nice bread pudding (= made of/ingredient: constitutive)

We would suggest, as before, that the lexicon is not the best place for this information and, further, that this information is incomplete. For comparison, we present our approach to describing and processing *pudding* in the OntoSem environment. Since OntoSem uses a full ontology (not a word net), the ontological specification of the concept PUDDING contains much of the needed information for processing all the above sentences containing *pudding*. Moreover, since the OntoSem ontology, lexicons and processors are developed together, their known mutual contributions drive resource acquisition. Obviously, one cannot expect the same approaches to be used in a lexicon-only project like SIMPLE. However, a non-trivial question, considering the expense of manual resource acquisition, is to what extent should we be developing resources separately from processors that can use them, especially when the nature of processors crucially affects what is needed of knowledge resources?

Below is a subset (for reasons of space) of the properties and values for the concept PUDDING in the OntoSem ontology; the first 4 are locally specified while the others are inherited.

PUDDING

IS-A	
DESSERT	
HAS-OBJECT-AS-PART	MILK, SUGAR, EGG
FATTINESS	> .6
THICKNESS	> .8
Inherited from DESSERT	
BITTERNESS	0
SALTINESS	0
SWEETNESS	> .7
SPICINESS	0
Inherited from PREPARED-FOOD	
THEME-OF	PREPARE-FOOD, BUY
PRODUCT-TYPE-OF	FOOD-SERVICE-ORGANIZATION
Inherited from FOOD	
THEME-OF	INGEST
Inherited from ARTIFACT	
CREATION-RELATION	HUMAN
COST	> 0
(Inheritance continues, from INANIMATE, PHYSICAL-OBJECT, OBJECT, ALL.)	

Since all of the necessary information about PUDDING is encoded in the ontology, the OntoSem lexicon entry for *pudding* need only contain a direct link to the concept.

The analysis of sentences (a)-(e) in OntoSem is carried out as follows. For (a), there is a lexical sense of *refuse* that expects an OBJECT (not an EVENT, as in the main sense) as its direct object. This sense *expects* the semantic ellipsis of a verb and, as such, is supplemented with a meaning procedure called 'seek-specification', which searches for the elided event. There are two sources it searches: previous TMRs, for a recent semantically viable event, and the ontology itself, for an EVENT (or EVENTs) whose default AGENT is HUMAN and default THEME is PUDDING. This

[3] Some automation of the mapping between L1-L2 multi-sense words is possible as demonstrated by Pianta, et al. 2002, but the results still require intensive manual work by an acquirer.

search procedure in some cases returns more than one candidate event to reconstruct the semantic ellipsis. While this is not always ideal, it does reflect precisely the type of lexical ambiguity that can be resolved only by contextual clues. For example, the sentence *John refused the pudding* could be used in a supermarket context to describe a situation where John refused to take/accept a free box of pudding that was being pushed upon him by a promoter. The desire to be able to treat this second reading of the sentence is the reason for treating constraints in OntoSem abductively. As far as one can tell, the constraints in SIMPLE are rigid: "telic = eat" for *pudding* is a hard constraint. In fact, the example *John refused the pudding* is representative of a much broader class of phenomena known as semantic ellipsis, the treatment of which must be carried out by procedural semantic routines (see McShane et al. 2004a for details).

Example (b) is another case that OntoSem handles through lexical and procedural semantics working in tandem. The NP *easy pudding* is actually a construction {a value on the scale DIFFICULTY + ARTIFACT} that is known to involve semantic ellipsis. Thus, we prepare for it in the OntoSem lexicon by associating this construction with the seek-specification meaning procedure, described above, which handles with equal efficacy *easy pudding* (PREPARE-FOOD), *easy song* (PERFORM-MUSIC), etc.

Example (c) is handled trivially based on the fact that PUDDING is a PHYSICAL-OBJECT and, like all PHYSICAL-OBJECTs, is ontologically defined for LOCATION.

Example (d) is analyzed using the information that PUDDING is a PREPARED-FOOD and, as such, is the THEME-OF PREPARE-FOOD, which in turn is a child of CREATE-ARTIFACT. The lexicalized phrasal {ARTIFACT + come out + a value of evaluative modality} is mapped to CREATE-ARTIFACT, with the THEME being the given ARTIFACT and the evaluative modality being concretized based on the evaluative value of the lexical item (e.g., 'well', as in 'the pudding came out well' is mapped to 'evaluative .7'). This phrasal, of course, works for any ARTIFACT and any value of evaluative modality, so lexicalizing it once is a real savings in time and effort.

Example (e) has two possible treatments in OntoSem: on the one hand, the lexical item 'bread pudding' could (and, ultimately, should – though it is not in the OntoSem lexicon at the moment) be listed as a phrasal in the lexicon, described as PUDDING: HAS-OBJECT-AS-PART BREAD. However, if it is not listed, it is treated by our productive rules for treating noun-noun compounds. One of the N-N compound rules is that the pattern MATERIAL + N is analyzed as N:HAS-OBJECT-AS-PART:MATERIAL.

5 Conclusions

Although space does not permit us to fully describe the resources, programs and resulting TMRs for sentences (a)-(e), this snapshot of their processing underscores the point that developing resources within an environment where they are tightly coupled with processing has clear advantages over developing resources in the abstract. Of course, in the absence of a full environment, projects like SIMPLE make sense. However, the challenges of resource development outside of an environment are keenly felt by developers: as Calzolari (1999:42) reports: "A dichotomy at stake here is the one between generality of a LR [lexical resource] vs. usefulness for applications. In principle, only when we know the actual specific use we intend to do [*sic*] of a LR can we build the 'very best' LR for that use, but this has proved to be too expensive and not realistic. In practice, however, there exists a large core of information that can be shared by many applicative uses, and this leads to the concept of "generic" LR, which is at the basis for the EAGLES initiative and of the PAROLE/SIMPLE projects, to be then enhanced and tuned with other means". The only aspect of this statement that we would dispute is the unrealistic nature of building resources for particular systems. If a system, like OntoSem, creates text-meaning representations that can be used equally effectively for many applications, then there is no reason why they cannot be built specifically for the given environment. In other words, when the result of semantic analysis is a metalanguage-formulated TMR, programs of any profile can exploit this representation. Stated differently, there need not be a direct link between end applications and the input text elements or their lexical representations.

References

Lenci, Alessandro, Federica Busa, Nilda Ruimy, Elisabetta Gola, Monica Monachini, Nicoletta Calzolari, Antonio Zampolli et al. 2000a. SIMPLE Work Package 2, Linguistic Specifications, Deliverable D2.1, March 2000.

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, Antonio Zampolli. 2000b. SIMPLE: A General

Framework for the Development of Multilingual Lexicons. *Proceedings of LREC 2000*.

Levin, Beth. 1995. *English Verb Classes and Alternations*. Chicago: University of Chicago Press.

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

McShane, Marjorie, Sergei Nirenburg and Stephen Beale. 2004b (ms.). The description and processing of multi-word expressions in OntoSem. Available at http://ilit.umbc.edu/RecentPubl.htm.

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

Nirenburg, Sergei, Stephen Beale and Marjorie McShane. 2004b (forthcoming). Evaluating the performance of the OntoSem semantic analyzer. *ACL 2004 Workshop on Text Meaning and Interpretation*..

Nirenburg, Sergei, McShane, Marjorie, Stephen Beale. Forthcoming. 2004c (forthcoming). The rationale for building resources expressly for NLP. *Proceedings of LREC 2004*, Lisbon, Portugal.

Palmer, Martha, Ralph Grishman Nicoletta Calzolari, Antonio Zampolli. 2000. Standardizing multilingual lexicons. Paper presented at the workshop on Web-Based Language Documentation and Description 12-15 December 2000, Philadelphia, USA.

Pederson, Bolette Sandford and Britt Keson. SIMPLE - Semantic information for multifunctional plurilingual lexica: Some examples of Danish concrete nouns. *SIGLEX99: Standardizing Lexical Resources Workshop, ACL99.*

Pianta, Emanuele, Luisa Bentivogli and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January 21-25, 2002.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cam*,* 28(1):11-21.