The Idiom–Reference Connection

Marjorie McShane
Sergei Nirenburg
University of Maryland Baltimore County (USA)
email: marge@umbc.edu

Abstract

Idiom processing and reference resolution are two complex aspects of text processing that are commonly treated in isolation. However, closer study of the reference needs of some idioms suggests that these two phenomena will need to be treated together to support high-end NLP applications. Using evidence from Russian and English, this article describes a number of classes of idioms according to their reference needs and suggests a method of lexical encoding which, supplemented by procedural semantic routines, can adequately support the full semantic and referential interpretation of these idioms.

1 Introduction

Reference resolution and idiom processing have received much attention in natural language processing (NLP), but these phenomena are commonly treated in isolation of each other, and most treatments address only a single aspect of the respective overall problems. For example, much of the work on practical reference resolution has concentrated on establishing textual coreference relations for a subset of pronouns (e.g. Mitkov et al., 2002), and the most widely pursued aspect of idiom processing has been the automatic extraction of multi-word expressions (of which idioms are a subtype) from corpora (e.g. Baldwin and Villavicencio, 2002). Of course, some contributions in both of these subfields have ventured much wider;[1] however, we have found few practical approaches that explore the interaction of idiomaticity and reference resolution and its implications for NLP.

One might ask, why treat these phenomena together? Perhaps the best reason is to highlight the indispensability for real progress in NLP of semantic analysis that goes beyond what the most researchers are currently pursuing in practical system building. Another reason to integrate the study of reference and idioms is to address the difficulties that automatic text analyzers will encounter in detecting and processing idioms when some of their components are elided. Ellipsis, a means of expressing reference, thus, becomes an important component of this study. The approach suggested here should, we believe, alleviate some of the inherent difficulties of these complex tasks. Note that similar kinds of problems are discussed in Pulman (1993), which suggests the need for "contextual reasoning" applied to idioms, which is "the process of taking the information that can be derived linguistically from a sentence and fleshing it out with information supplied by the local context or general background knowledge" (Pulman, 1993, p. 251).

The proposed analysis delineates several categories of idioms according to their reference needs and shows how the encoding of idioms in a semantically oriented lexicon can support both basic semantic analysis and reference resolution. Although the analysis is theory- and system-neutral, the exposition follows a specific, implemented theory of natural language processing. This theory, called Ontological Semantics (Nirenburg and Raskin, 2004), colors our understanding of the nature of meaning-oriented NLP, including our treatment of reference and idioms.

Ontological Semantics seeks to achieve full semantic and pragmatic analysis of texts such that interpreted structures, rather than textual strings, serve as the input to automatic reasoners. Ontological Semantics relies on knowledge obtained through many layers of processing: preprocessing followed by morphological, syntactic, semantic and discourse analysis. The static knowledge resources, which are interconnected and all use the same metalanguage of description, are a lexicon and onomasticon for each language processed, a language-independent ontology (a knowledge base of concept types), and a language-independent fact repository (a knowledge base of concept instances). Static resources are compiled manually, using sophisticated editing environments, to ensure high quality, though we are experimenting with machine learning to speed the acquisition process. Text analysis involves the automatic

[1] See, for example, the contributions to recent workshops (e.g., ACL 2004 "Reference Resolution and its Applications" and "Multi-word Expression: Integrating Processing"; EACL 2006 "Multi-word Expressions in a Multilingual Context") and Stanford's Multi-Word Expression Project (`http://mwe.stanford.edu/`).

evaluation of semantic preferences recorded in the lexicon and ontology, as well as preferences based on stochastically trained measures of semantic distance among ontological concepts.

Within this semantically-oriented, knowledge-based environment we define reference resolution rather differently than in most NLP applications, where resolving reference is understood as linking coreferring text strings. In fact, our conceptualization of reference resolution strongly influences how we approach resolving reference in idioms and therefore must be clarified from the outset.

2 What is Reference Resolution?

We define reference resolution as the anchoring of referring expressions in the episodic memory of an intelligent text processing agent. This knowledge base of stored memories, called the fact repository, differs from the ontology in that it contains indexed *instances* of ontological concepts and their property-based interconnections. Anchoring entities in the fact repository is the culmination of semantic analysis and reference resolution.

When presented with a new text, the system must first **semantically analyze every sentence**, creating an unambiguous text meaning representation (TMR); reference is then resolved for the correct meaning of each string. The TMR contains the crucial clues for **determining which entities are referring expressions**: numbered instances of ontological concepts are referring expressions whereas properties, literal property fillers, and so on, are not. As an example, consider the following context which, although contrived, illustrates many relevant phenomena at one go.

(1) At 4:48 it became clear that the programmers couldn't finish debugging the system before the 5:00 deadline. **All hell broke loose, the boss was fit to be tied — almost strangled his project manager!**

Let us concentrate on the second sentence. In the tables below, each string or idiomatic group of strings from that sentence (top row) is associated with its corresponding semantic structure (bottom row). The concept instances set in italics must be resolved. The important thing to notice is that the system must orient around semantic structures rather than strings in order to create the correct inventory of referring expressions.

all hell	broke loose	the boss	was	fit	to be	tied		almost	strangled	his project manager
CHAOS-1		*MANAGER-1*		ANGER (RANGE 1)			*HUMAN-1*	(MODALITY-2 (TYPE EPISTEMIC) (VALUE .9) (SCOPE STRANGLE-1))	*STRANGLE-1*	*ASSISTANT-1*

Highlights of the analysis are as follows:

- Whereas *all hell* and *broke loose* could individually be referring expressions in some other context, when the are used in this idiom they together represent a single meaning, CHAOS, this instance of which is called CHAOS-1 — the first instance of the concept CHAOS encountered while processing the given text or corpus. This event, like all instances of OBJECTs and EVENTs in TMRs, requires reference resolution: it must be determined whether this is a new event to

be added to the fact repository or a reference to an event that is already recorded there. In this case, it is a new event, since the algorithm used to detect event coreference requires either (a) that there be an ample overlap of properties associated with a candidate fact repository "anchor" or (b) that the new event be referred to using a definite description (e.g., *the strike*), with the definite description triggering the search for a coreferent in the context or fact repository.

- Whereas *the boss* and *his project manager* can either be descriptors (as in *This man is a boss and that man is a project manager*) or referring expressions, here they are referring expressions and must be resolved.

- Whereas *fit* and *tied* can be referring expressions in isolation, in this idiom they are not referring expressions, nor is the idiom on the whole a referring expression: it indicates the highest value of the property ANGER.

- Although the second half of the sentence has no overt subject, *he* is the understood subject. The reference resolver must detect this missing entity and create a coreference link between it and MANAGER-1.

- *Almost* is never a referring expression: it indicates a value of less than 1 for epistemic modality scoping over the given event (here, STRANGLE-1). However, some other adverbs are referring expressions (e.g., *here, yesterday*) and must be resolved.

- STRANGLE-1, like all EVENTs, must undergo reference resolution.

Once all referring expressions have been detected, the system must **resolve them against the fact repository**. There are several possible scenarios: (a) the entity has a textual antecedent, in which case the new entity is linked to the same fact repository anchor as that antecedent; (b) the entity does not have a textual antecedent but is already a known entity (like *the earth* or *Plato*) and is linked to the existing anchor in the fact repository; (c) the entity is determined to be new and a new anchor is established for it in the fact repository. This, in a nutshell, is how reference is resolved in our semantic analysis environment, OntoSem.

Our reference resolver for English is implemented and covers all the eventualities posed by this sentence. It has not yet undergone formal evaluation. We will now describe how idioms are encoded to support this process.

The examples used for illustration are not from English, they are from Russian, a language that is not currently supported in OntoSem. The reason for using Russian examples even though the implemented system does not yet cover Russian is that Russian presents a superset of challenges for reference resolution — namely, a much wider use of ellipsis, or the null referring expression; therefore, showing that the scope of phenomena presented by Russian can be handled *a fortiori* shows that the same phenomena can be handled in English. Indeed, the OntoSem environment supports multilingual text processing, using a language-independent ontology and fact repository, and using the same types of lexicon entries regardless of the language processed (see McShane et al., 2005).

3 Encoding Idioms to Support their Full Analysis

A cornerstone of theoretical, descriptive, computational and psycholinguistc work on idioms is the attempt to understand to what extent idioms are fixed and to what extent they are flexible (see, e.g., Cacciari and Tabossi (1993), whose component articles include extensive overviews of the literature). The competing classifications can derive from both theoretical considerations, like psycholinguistic evidence, and practical considerations, like whether an NLP system attempts to analyze only those idioms that are recorded or whether it attempts to analyze new coinages as well. The scope of the current analysis is idioms that *are* recorded as well as certain types of free modifications of them. Completely new idioms will need to be processed as "unexpected input", in a similar way as the system attempts to process metaphor and metonymy. Like Stock et al. (1993) (in Cacciari and Tabossi (1993)), we integrate idioms into the lexicon as "more information about particular words" (Stock et al., 1993, p. 238) rather than treat them using special lists and idiosyncratic procedures. In the discussion below, we look at some examples of idioms that highlight noteworthy reference resolution needs and show how our all-purpose lexical encoding mechanisms and reference resolution routines cover idiomatic input as readily as compositional input. A more detailed description of how we encode idioms and other multi-word expressions, as well as many additional examples, can be found in McShane et al. (2008).

3.1 Productive Syntactic Processes in Idioms

Each of the examples below contains an idiom in the second half, and each of those idioms shows at least one productive use of ellipsis. In the examples, the elided category, [e], and its antecedent, if syntactically available, are in boldface. Grammatical information is provided sparingly for reasons of space.[2]

(2) Nado zashchishchat' **svoix** **sotrudnikov** a ne
it-is-necessary to-defend **self's**$_{ACC.PL}$ **coworkers**$_{ACC.PL}$ and not
prinosit' **[e]** v zhertvu .
deliver$_{INFIN}$ **[e]**$_{ACC}$ as sacrifice$_{ACC.SG.FEM}$.

You should defend **your coworkers**, not sacrifice **them**.

(3) Ja ne xochu preduprezhdat' **ego**, **[e]** xochu
I don't want to-forewarn **him**$_{ACC}$ **[e]**$_{1.SG}$ want$_{1.SG.}$
zastat'**[e]** vrasplox .
to-catch**[e]**$_{3.SG.ACC.MASC}$ unawares .

I don't want to forewarn **him**, I want to catch **him** unawares.

These examples represent configurations in which ellipsis is highly promoted in non-idiomatic *and* idiomatic contexts.[3] Example (2) shows VP conjunction with the latter of two coreferential direct objects elided. Example (3) shows subject and direct object

[2] Most of the Russian examples here are from Lubensky (1995), which is a bilingual learner's dictionary of Russian idioms that provides grammatical descriptions but no special treatment of ellipsis.

[3] See McShane (2005) for discussion and extensive examples of ellipsis-promoting configurations using non-idiomatic examples. Idiomatic examples of many of the phenomena have also been found but are not presented here for reasons of space.

ellipsis in an "assertion + elaboration" strategy (see McShane (2005)), in which the topic of discourse is asserted then either restated or elaborated upon subsequently.

The above idioms are idiomatic VPs that are recorded in the OntoSem lexicon in a similar way as typical verbs, with just a few special features. Let us take the example of *v grob vgonjat'* 'to kill' (literally: to drive to the grave) as an example.

```
(vgonjat'-v1
  (def "idiom: v grob vgonjat' - to kill (drive to the grave)")
  (ex "Ja v grob vgonju tebja! I'll kill you!")
  (syn-struc
    ((subject ((root $var1) (cat n)))
     (root $var0) (cat v)
     (directobject ((root $var2) (cat n)))
     (pp ((root $var3) (cat prep) (root v)
          (obj ((root $var4) (cat n) (root grob)))))
  (sem-struc
    (KILL
        (AGENT (value ^$var1))
        (THEME (value ^$var2)))
    (^$var3 (null-sem +)) (^$var4 (null-sem +))))
```

This lexical sense is headed by the verb, *vgonjat' 'drive'*. The syntactic zone (syn-struc) says that the verb takes a subject, direct object and prepositional phrase with no unusual syntactic constraints, meaning that the structure is open to the same sorts of variability — like different verbal tenses and aspects, free word order, syntactic transformations and ellipsis — as is typical of non-idiomatic Russian. The only special syntactic feature is that the roots of the lexical components of the prepositional phrase are explicitly listed: *v* (into) and *grob* (grave). The semantic zone (sem-struc) records the semantic interpretation: it is headed by a KILL event whose AGENT and THEME are productively analyzed as the meaning of the subject and direct object, respectively. The meanings of *v* (into) and *grob* (grave), which are — under this analysis — non-compositional, are attributed null semantics.

Two aspects of semantic interpretation require comment. First, in most contexts this idiom is not used to threaten actual killing; however, the same can be said for the lexeme *kill* used in the threat *I'll kill you!*; this aspect of interpretation is clearly extra-lexical. Second, although it is likely that a person who did not know this idiom would be able to interpret its meaning using the meanings of the component elements, most NLP systems would struggle. Once we decide to record a phrase as idiomatic to ease processing, the level of transparency of the components becomes unimportant.

Analysis of a clause that uses *vgonjat' v grob* 'kill' will generate three referring expressions that must be resolved: the AGENT of the killing, the THEME of the killing (we will not quibble here about which case role to choose for the person killed), and the act of killing. These referring expressions might be realized, for example, as HUMAN-23, HUMAN-24 and KILL-4 in a given text meaning representation. Once the system has arrived at these analyses, reference resolution proceeds as it would for any referring expressions, whether or not they were part of an idiom: textual coreferents — recorded as semantic entities in TMR — are sought and, whether or not they are found, the referring expression is anchored in the fact repository. If we look at what is

special about processing the reference in idioms, then, there are only two aspects: (1) ensuring that productive syntactic processes are permitted only if applicable, and (2) ensuring that the correct inventory of referring expressions — understood as semantic structures — is generated.

Let us compare this treatment of idioms to the one proposed by Villavicencio et al. (2004). They treat the potential variability of idioms using the notion of semantic decomposition. If an idiom can be paraphrased in a syntactically parallel way, it is decomposable (*spill the beans → reveal a secret*), even though non-standard meanings need to be assigned to each component. The fundamental differences between their approach and ours relate to semantic encoding and reference resolution. For Villavicencio et al., the semantics of idioms is conveyed by paraphrases with other linguistic elements (*spill → reveal, beans → secret*). For us, semantics is formulated using the ontologically grounded metalanguage of OntoSem. As regards the initial syntactic parse, both approaches seem to offer the same coverage of syntactic variability, and resources could be shared with seemingly minimal work devoted to format conversion.

3.2 Essentially Frozen Idioms

We have just shown how syntactic processes — specifically, various types of ellipsis — can apply to idioms in a language, and how the lexical encoding of such idioms allows for syntactic variability. Other idioms, by contrast, are syntactically frozen. Such idioms are commonly treated as strings with spaces, but this only works if absolutely no modifiers or other entities (e.g., "ahem") can intervene. If intervening material is possible, it is preferable to encode the idiom using separate syntactic constituents. However, if one records the components individually, the analysis system must understand that diathesis transformations, ellipsis, pronominalization, etc., are not applicable. In OntoSem we label frozen syntactic constituents using immediate constituents, like NP, rather than grammatical function labels, like subject. Since transformations apply only to grammatical functions, they become automatically inapplicable if immediate constituents are used. However, since all constituents are still listed individually, intervening material and free modification are permitted in the usual way, as in *He kicked the bloody bucket!*

Of course, treating free modifications of non-compositional parts of an idiom or other multi-word expression (MWE) is not trivial, as described in some depth in McShane et al. (2008). To summarize that discussion, our basic approach to treating modifiers within MWEs is to analyze the MWE as indicated in the sem-struc, then attempt to attach the meaning of "orphan" modifiers to the meaning of the entire structure using generalized processes for meaning composition. In the case of *He kicked the bloody bucket*, the basic meaning will be rendered in the text meaning representation as (DIE-1 (EXPERIENCER HUMAN-1)). The modifier *bloody* has two senses in our lexicon, semantically described as (RELATION BLOOD) and (EMPHASIS .7). We have a rule that prefers the stylistic interpretation in the case of non-compositional idioms. So the final text meaning representation will be (DIE-1 (EXPERIENCER HUMAN-1) (EMPHASIS .7)). ((emphasis .7) indicates a high value for the property EMPHASIS on the abstract scale $\{0,1\}$.)

Such meaning composition is not specific to multi-word expressions: our semantic analyzer carries out the same process in all cases when meaning must be recovered

from an incomplete parse. The latter may be due to insufficient coverage of the syntactic parser, lexical lacunae that confound the parser, or unexpected (ungrammatical, highly elliptical, etc.) input.

Returning to our main point about how to encode essentially frozen idioms, encoding their components as separate entities provides the best of both worlds: frozen components, fixed word order, and the possibility of intervening strings that typically act as modifiers. One Russian idiom that fits this description is shown below.

(4) Ishchi-svishchi vetra v pole .
Look-for-whistle-for$_{IMPER}$ wind in field .
'You'll never find him/her/it/etc.'

```
(ishchi-svishchi'-v1
   (def "idiom: ishchi-svishchi vetra v pole
                'you will never find him/her/it/etc.'")
   (syn-struc
      ((root $var0) (cat v) (form imperative)
       (np ((root $var1) (cat np) (root vetra)))
       (pp ((root $var2) (cat prep) (root v)
       (np ((root $var3) (cat np) (root pole)))))
   (sem-struc
      (modality
        ((type potential)
         (value 0)
         (attributed-to (sem HUMAN))
         (scope (value refsem1))))
      (refsem1
         (FIND
             (AGENT (sem human))
             (THEME (sem all))
             (time (> (find-anchor-time))))
      (^$var1 (null-sem +))
      (^$var2 (null-sem +))
      (^$var3 (null-sem )))
   (meaning-procedure
      (seek-specification
         ((value find.modality.attributed-to)
          (resolve-1st-sing)))
      (seek-specification
         ((value find.agent) (resolve-2nd-sing)))
      (seek-specification
         ((value find.theme) (resolve-3rd))))
```

The syntactic description should be self-evident based on the examples and description above, but the semantic structure requires commentary.

The variables $var1, $var2 and $var3 are attributed null semantics because they do not contribute to compositional meaning — that is, this idiom ("look for whistle for wind in the field") is completely semantically opaque.

The sem-struc is headed by a modality statement: *it is impossible* introduces modality of the type 'potential' with a value of 0. This modality is attributed, by default, to

the speaker. It scopes over a proposition headed by FIND, and the latter is ontologically defined as taking an AGENT and a THEME case role.

The semantic representation includes four referring expressions that must be resolved: (1) the speaker, to whom the modality is attributed; (2) the FIND event itself, which will be a new anchor in the fact repository; (3) the AGENT of finding, which is the interlocutor; and (4) the THEME of finding, which must be contextually computed.

The OntoSem analyzer would resolve the reference of the instance of FIND in the usual way; this requires no further comment. What does require further comment, however, is the way in which we guide the analyzer's efforts to resolve the underspecified instances of HUMAN, HUMAN and ALL that represent the speaker, the interlocutor and the object of the FIND event, respectively. We provide this guidance in the *meaning-procedures* zone of the lexicon entry, which contains calls to procedural semantic routines that are launched at run time. For example, we need to know who the speaker is so that the modality can be attributed to the correct real-world person. This is done using the "seek-specification" meaning procedure. The first argument of this procedure is what we are seeking the specification of (i.e., to whom the modality is attributed), and the second argument is the function that will let us determine this — i.e., "resolve-1st-sing", which is, incidentally, the same routine used to seek the referent of the pronoun *I*. The latter meaning procedure includes ordered routines testing for many cases including:

- the pronoun *I* being used in a context in which another pronoun *I* (which itself should have been resolved earlier) can serve as an antecedent: **I** like chocolate ice cream and always choose it if **I** have the option.

- the pronoun *I* being used within a quotation, and that quotation being the THEME of a SPEECH-ACT of which the coreferent of *I* is the AGENT: **I/Mary** said, "But **I** don't *want* strawberry ice cream!"

- the pronoun *I* being used outside of a quotation and the writer of the text being available in metadata: `<title>Understanding Your Finances</title> <author>Mary Smith</author>` ...**I** believe that the only way to understand your finances is to consult a financial advisor.

In short, using the combination of the information in the *sem-struc* and *meaning-procedures* zones we arm the analyzer with the types of the information a person would use to both understand the idiom and to resolve all implied references. (For a more detailed description of meaning procedures in OntoSem, see McShane et al. (2004).)

3.3 Subjectless Constructions

We conclude our example-based discussion with one category of phenomena in which idiom processing is actually much simpler than the processing of structurally similar compositional language since it permits preemptive disambiguation. The disambiguation in question regards subjects, which in Russian can be overt, elided or completely missing. Completely missing (uninsertable) subjects occur in the following constructions:

- In the *indefinite personal construction* a 3rd person plural verb form is used without a subject to indicate an unspecified person or people. It is used in contexts like the Russian equivalent of *They say it will rain today.*
- In the *non-agentive impersonal construction* a 3rd person singular verb is used without a subject to show that the event is non-agentive. It is used in contexts like the Russian equivalent of *He's attracted to girls like that*, whose structure-preserving paraphrase would be "[some unnamed force] attracts him to girls like that."

The difficulty in processing *productive* subjectless sentences is determining whether the verb *has* a specific subject that has been elided and must be recovered, or does *not* have a specific subject, in which case the generalized personal or non-agentive interpretation should be used. However, when it comes to *idioms* that employ these constructions, the syntax can be encoded to explicitly block a subject, and the semantics can explicitly indicate the interpretation added by the missing subject.

An idiom that employs the indefinite personal construction is shown in (5), along with the lexical sense of *bit'* 'hit' that records it.

(5) Lezhachego ne b'jut . [L-45]
Lying-down-person$_{ACC.SG.MASC.}$ not beat$_{3.PL.PRES.}$.
You don't/shouldn't kick a man/person/guy when he's down.

```
(bit'-v10
  (def "phrasal: Lezhachego ne b'jut - you shouldn't do
       something bad to someone who is in a bad position already")
  (ex "You don't/shouldn't kick a guy when he's down")

(syn-struc
  ((np ((root $var1) (cat n) (root lezhachij)
        (case acc) (gender masc) (number sing)))
   (verb-neg ((root $var2) (cat verb-neg)))
   (root $var0) (cat v) (tense present) (person third) (number pl))))

(sem-struc
   (modality                    ; ``should''
      (type obligative)
      (scope (value refsem1))
      (value 1)
      (attributed-to *speaker*))
   (refsem1
      (modality                 ; ``not''
         (type epistemic)
         (scope (value refsem2))
         (value 0)
         (attributed-to *speaker*)))
   (refsem2
      (ABUSE
         (AGENT (value refsem3))
```

```
            (THEME (value refsem4))))
      (refsem3
         (set
            (member-type human)
            (cardinality 1)
            (complete yes)))
      (refsem4 (HUMAN (EXPERIENCER-OF MISFORTUNE)))
      (^$var1 (null-sem +)) (^$var2 (null-sem +))
   (output-syntax (cl)))
```

The syn-struc should be clear based on previous examples; the only new element is *verb-neg*, which indicates a negating particle.

The sem-struc looks more complex than it actually is because many of the slot fillers require reified structures, each of which must be pointed to using numbered variables called refsems.The sem-struc is headed by obligative modality, which scopes over an epistemic modality, which scopes over an ABUSE event. The obligative modality has the value 1 (absolute obligation), whereas the epistemic modality has the value 0 (negation). Put plainly, "it is necessary not to abuse". The AGENT of the ABUSE event is the set of all people, described just as we describe the word *everyone*. The THEME of the ABUSE event is a HUMAN who is the EXPERIENCER-OF a MISFORTUNE. One might ask, why not record this idiom as a fully fixed entity with white spaces in between, rather than as a multi-part syntactic structure? For the same reason as discussed earlier: there is an outside chance of modification, so the component elements must be kept separate.

Example (6) shows an idiomatic example of the second type of obligatorily subjectless sentence: the non-agentive impersonal construction.

(6) Kakim vetrom vas
what$_{INSTR.SG.MASC.}$ wind$_{INSTR.SG.MASC.}$ you$_{ACC.PL/POLITE}$
zaneslo sjuda ?
brought$_{3.SG.NEUT.PFV}$ here$_{DIRECTIONAL}$?
What brings you here?/What are you doing here?

This idiom will be recorded under the headword *zanesti* 'bring'. The core meaning of the idiom — COME — heads the sem-struc. There are two variables in this multiword expression: the direct object, mapped to the AGENT of COME, and the spatial adverbial, mapped to the DESTINATION of COME. These are productively analyzed at run-time. The meaning of 'what wind' is, of course, attributed null semantics.

To summarize this section: recording obligatorily subjectless idioms not only provides for their semantic interpretation, it also removes ambiguity in analysis, since the "elided subject" reading is explicitly blocked.

4 Final Thoughts

This paper has presented an analysis of phenomena that extends past what any given system currently uses or requires. However, the utility of this analysis reaches well beyond the traditional goals of descriptive and theoretical linguistics. Ideally, system building in NLP should centrally involve the objective of incrementally overcoming

successively more difficult challenges and thus lead to more sophisticated systems in the future. Looking forward to the next stage can help us to develop methodological, architectural and knowledge infrastructures to facilitate progress toward future goals. The OntoSem environment does not currently work on Russian, though it has been applied, at least partially, to several languages apart from English in the past — including such different languages as Turkish, Spanish, Korean and Georgian. The reason for exploring the idiom-reference connection in Russian was to judge how well our approach, which is implemented for and works well in English, holds up cross-linguistically. Having worked the examples presented in this paper and many others, we are convinced that when the time comes, a Russian OntoSem will be configurable without the need to expand the theory and methodology that support our treatment of idioms, ellipsis and reference overall.

A reasonable question would be, why not evaluate the approach on English, since an English system already exists? The reason is purely practical: it is far more difficult and expensive to run evaluations of knowledge-based systems that treat complex phenomena than it is to run evaluations of systems that treat less complex phenomena. That being said, we are just completing a new version of our DEKADE knowledge acquisition and evaluation environment which will make it much easier than before to evaluate the results of text analysis. We expect regular evaluations to become part of our development work in the near future.

References

Baldwin, T. and A. Villavicencio (2002). A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*, pp. 98–104.

Cacciari, C. and P. Tabossi (1993). *Idioms: Processing, Structure and Interpretation*. Lawrence Erlbaum and Associates, Inc.

Lubensky, S. (1995). *Russian-English Dictionary of Idioms*. Random House.

McShane, M. (2005). *A Theory of Ellipsis*. Oxford University Press.

McShane, M., S. Beale, and S. Nirenburg (2004). Some meaning procedures of Ontological Semantics. In *Proceedings of LREC-2004*.

McShane, M., S. Nirenburg, and S. Beale (2005). An NLP lexicon as a largely language independent resource. *Machine Translation 19*(2), 139–173.

McShane, M., S. Nirenburg, and S. Beale (2008). Achieving adequacy of description of multiword entities in semantically-oriented computational lexicons. Submitted.

Mitkov, R., R. Evans, and C. Orasan (2002). A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of CICLing-2000*.

Nirenburg, S. and V. Raskin (2004). *Ontological Semantics*. MIT Press.

Pulman, S. (1993). The recognition and interpretation of idioms. In C. Cacciari and P. Tabossi (Eds.), *Idioms: Processing, Structure and Interpretation*, pp. 249–270. Lawrence Erlbaum and Associates, Inc.

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

Villavicencio, A., A. Copestake, B. Waldron, and F. Lambeau (2004). The lexical encoding of MWEs. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating processing*.