**L**inguistic **I**ssues in **L**anguage Technology – **LiLT**
Volume 10, Issue 3 2015

# How to Record the Meaning of Figurative Language

**Indrek Jentson**

Published by CSLI Publications

# How to Record the Meaning of Figurative Language

Indrek Jentson, *University of Tartu*

This paper focuses on the question of what kind of data needs to be recorded about figurative language, in order to capture the essential meaning of the text and to enable us to re-create a synonymous text, based on that data. A short review of the best known systems of semantic annotation will be presented and their suitability for the task will be analyzed. Also, a method that could be used for representing the meaning of the idioms, metaphors and metonymy in the data model will be considered.

## 1 Introduction

Imagine an interlingual machine translation system. This sort of system sees a text written in some natural language as input and converts it to data that ideally contains all the necessary information about the meaning of the text in input and no longer depends on the language of the text. Based on a previously compiled dataset, the system then generates an equivalent text in another (or - why not? - the same) natural language. From a purely technological point of view, we can - in the context of the described MT system âĂŞ say that the interlingua and the dataset of meaning representation are equal, whereas there is no actual reason to insist that the format of the dataset must be human-readable. In the light of the current article I would like to point out that it is possible view the formalism of the meaning representation completely separately from the semantic analysis process. In the present case we will concentrate first and foremost on the aspect of the meaning representation itself, setting aside the question of how to extract the

*LiLT* Volume 10, Issue 3, 2015.
*How to Record the Meaning of Figurative Language.*

meaning from the text and how to generate a text based on the given representation of the meaning. A lot of work has been done during the last decade to realize an interlingua-type system like this, but the target has usually been set to processing a simple unequivocal text and handling figurative language has been left out of scope. This trend is not justified as, for example, metaphors are often unconsciously used in everyday language (Lakoff and Johnson, 2008). Now, to look at the reasons why this processing of the text containing figurative language is inconvenient, it should be noted that, in addition to a variety of problems that emerge in the course of analysis, it is not clear how to formally present ambiguous or figurative language and what could be done to improve the situation.

## 2 The existing formalisms

Below we will take a look at the systems available for present the meaning of the text written in natural language. The following list of highlighted methods is certainly not an exhaustive one; they have, however, found most recognition during the last decade.

Studying any kind of discourse, its overall structure and information conveyed by its content must be handled separately. The most general approach to discourse structure and the formalized method of annotation can be found in the international standard ISO 24617-5:2014 (discourse structure). This standard defines a framework for semantic annotation of language resources, specifies a discourse as a collection of text segments, content units and the connections between them, and provides a set of guidelines for the annotation of discourse with labeled and directed graphs. It is assumed that the content units that reflect the meaning of the text are defined as graphs, the vertices and edges of which are specified by a kind of ontology. However, the specific formalism is not determinated by the standard and it is recommended to take into account other standards published in the same annotation framework - ISO 24617-1:2012 (time and events) and ISO 24617-2:2012 (dialog acts). The group is led by Harry Bunt, under whose leadership a number of articles about the work already done have been published (Bunt et al., 2010, 2012). The plan is to draw up seven standards for that framework, the remaining four standards still being worked on.

An interesting method for representing the meaning of text is UNL - Universal Networking Language (Uchida et al., 1999, 2005) that contains all the necessary components to write down the knowledge written in natural language (Salam et al., 2012). A UNL graph is composed of universal words or concepts, their attributes and the relationships

between the universal words. In addition to graph form, there are two text format notations available for UNL. It is designed to represent discourses that are larger than a sentence. It is most commonly utilized in machine translation systems called the UNL language servers. According to the UNL language servers development strategy, the teams of native speakers of different languages create a lexicon, an encoder and a decoder for their language and thereby ensure the translation of their language to UNL and back. Ronaldo Martins has written about the preparation of language resources in connection with the UNL (Martins, 2012) whereas the thesis of Parteek Kumar (Kumar, 2012) illustrates the problem of encoding and decoding for the Punjabi language.

AMR - Abstract Meaning Representation - is a language that uses a non-cyclic directed graph, where nodes are concepts and edges are relationships between these concepts, for a meaning representation. This language was introduced by Irene Langkilde and Kevin Knight in connection with the natural language generator Nitrogen (Langkilde and Knight, 1998). Compiled data reflects the meaning of the source text with a certain loss which has deliberately been designed into the AMR. The authors have declared that AMR is based on the English language and oriented to the annotation of individual sentences. The best overview of the possibilities of AMR can be found in the specification of the language (Banarescu et al., 2014)[1]. In addition to machine translations (Jones et al., 2012), AMR is also used for the semantic annotation of corpora (Banarescu et al., 2013).

DRT - Discourse Representation Theory - was established in 1991 by Hans Kamp in order to address some of the problems of semantics and pragmatics, such as inter- and intrasentential anaphora, universal quantifier, tense and aspect etc (Kamp et al., 2011). In order to write down the meaning of the text, a language called the DRS is used in DRT. It has a notation of set theory, a linear notation and a specific graphical notation in the form of the so-called boxes. DRS is automatically transformable into formal logic expressions, which can then be used for calculations as e.g. proposed in (Bos, 2004). Based on this theory, software called the Boxer has been developed for the semantic processing of English text (Curran et al., 2007, Bos, 2008). DRS is used in the GMB (Groningen Meaning Bank) project for the semantic annotation of sentences (Basile et al., 2012).

In order to illustrate a representation of phraseologisms in this paper, I selected the UNL because of its flexibility and the set of special attributes, which are designed specifically for the annotation of figura-

[1] `http://www.isi.edu/~ulf/amr/help/amr-guidelines.pdf`

tive language concepts. AMR and DRS do not have special tools for figurative language. There are currently no obvious options to present multiple layers of meaning in them, but that does not mean that in these languages it is not possible to make the appropriate improvements.

## 3 Figurative language at sentence level

Works of semantic annotation of texts containing figurative language typically record figurative expressions word for word. Such an approach requires the consequent interpretation of the annotation in order to find out the conventional meaning of a given utterance.

However, if the recorded representation of the meaning gives the exact interpretation of the figurative expressions in the text, then there may be no evidence left that figurative language was used in the original text.

The problem is that in the first case the represented meaning is not necessarily unequivocal or consistent with the relationships or restrictions described in the applicable ontology. In the second case we lose the information that could be helpful in generating natural and fluent text from the recorded dataset.

For example, let us look at the sentence (1) and its annotations in GMB[2].

(1) A project aimed at providing inexpensive computers for millions of children in developing countries *has hit a snag*.

In this sentence the idiom "has hit a snag" is annotated word for word, however the real meaning of the sentence (that a project has run into an unexpected problem) remains unobtainable without additional processing (i.e. interpretation).

In order to record the meaning of figurative language in a sentence without any loss, a two-layer dataset is required. The first layer should reflect the most conventional meaning of the sentence, unambiguously and independent of the source language. This layer of data is necessary for the processing of recorded information by the computer to enable detection of contradictions and inferences.

The second layer should describe the figurative language of the original sentence. Corresponding mapping functions explain the meaning of the data links between the different layers. This information is needed for the translation process in situations where the same figurative expressions are used in the target and source languages. The mapping

[2]ID of the sentence is 74/0129 and the URL of the respective web page is `http://gmb.let.rug.nl/explorer/explore.php?part=74&doc_id=0129`

functions include metonymy (along with information on the specific type), metaphor (along with information on the metaphorical concept used) and idiom (with the information that a lexicon was used to transform the phrase).

Such interpretation of data allows us to generate sentences in the target language. Only after the sentence has been created from the first layer of data, can the outcome of sentence generation be modified according to information in the second layer. If the data in the second layer is not applicable to the target language in the sense that there is no comparable expression, the corresponding changes can be left unmade.

Let us consider with the help of an example (2)[3], how it is possible in UNL to realize a multi-layered presentation of the meaning of a sentence, using the existing options of that language.

(2) His criticisms were right on target.

In this case, it is the conceptual metaphor ARGUMENT IS WAR and the mapping ACCURACY IS HITTING A TARGET that gives the phrase "right on target" the meaning "accurate". If we present the sentence (2) in UNL graphical form, we generally get the result shown in Figure 1.

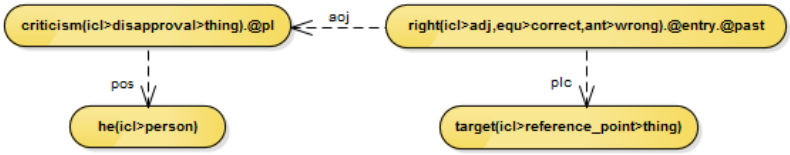


FIGURE 1 A UNL graph for sentence "His criticisms were right on target."

An experienced coder can even use the special attribute @metaphor to mark the universal words "right" and "target" in the diagram. Such a result does, however, not meet the stated purpose and leaves open the questions "what is the meaning of the phrase 'right on target' in this context" and "how should it be interpreted by the computer"?

What outcome of the semantic analysis do we expect in this case? We need to find a paraphrase that will convey the meaning of the sentence without using figurative language. How a suitable paraphrase can be constructed is a problem left beyond the scope of this paper, but at the end we expect to have two sentences with the same meaning and

[3]This example is taken from (Lakoff and Johnson, 2008)

a relationship between the concepts that refer to the different parts of some metaphorical mapping.

Taking the results of the semantic analysis of the metaphor and replacing "right on target" in the diagram with the less ambiguous concept "accurate", we get a result that is ontologically consistent and can be processed by the computer. In order to store the information about an expression used in the source text, the information about a metaphor must be added to the data using the above-mentioned substitution (see Figure 2).

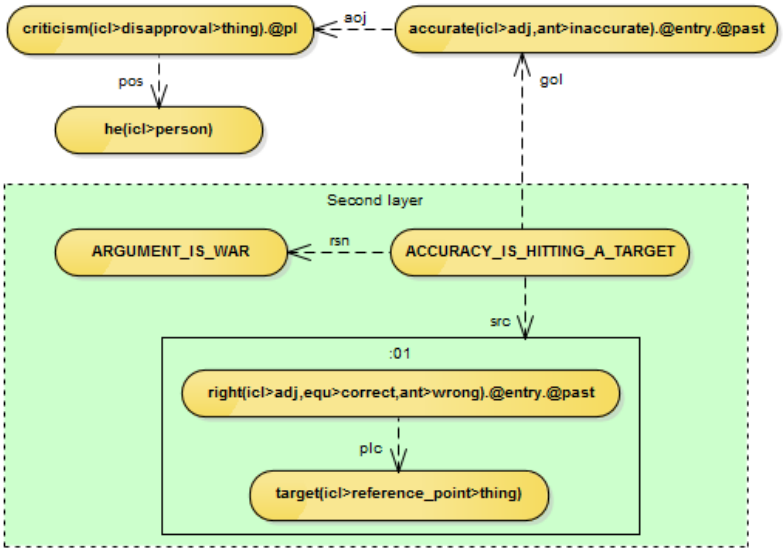


FIGURE 2 Using the second layer of data in order to retain the information about the metaphor used

The ontological explanation of the concept for the corresponding metaphor must be placed in the knowledge base used to reveal the content of the sentence. In case of UNL, such a system is the UNL Knowledgebase.

It is possible to deal with different types of phraseologisms in the same way.

Let us examine the following examples.

(3) Ann made a tongue-in-cheek remark to John.

In the dataset of the meaning of the sentence (3), it is possible to replace the idiom "tongue-in-cheek" with an appropriate adjective

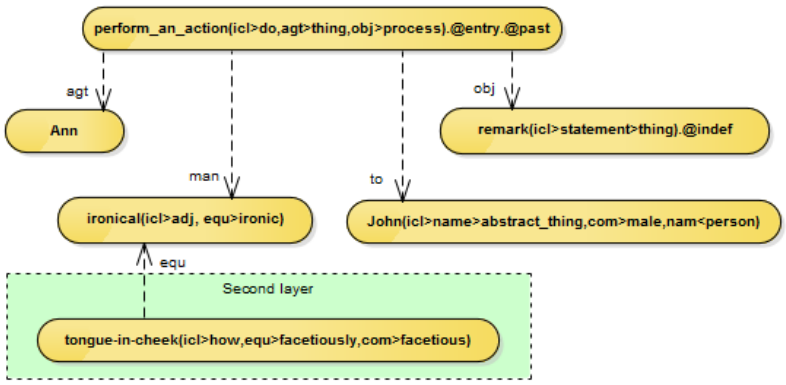


FIGURE 3 A UNL graph for sentence "Ann made a tongue-in-cheek remark to John."

"ironic" using the idiom lexicon. In the UNL, a corresponding second layer for the meaning of the sentence can be added with the relation *equ(tongue-in-cheek(icl>how, equ>facetiously, com>facetious), ironical(icl>adj, equ>ironic))*, as shown in figure 3. In this case using the relation *equ* is rather simple solution, but when the mapping function is simply a relation between two phrases there is no need for a more complicated representation.

Let us look at a representation of the sentence (4), where the metonymic concept THE PLACE FOR THE EVENT is used.

(4) Watergate changed our politics.

The word "Watergate" in this sentence represents the major event in the United States of America that has become synonymous with political scandals. Without additional context is safe to assume that here "Watergate" refers to the original event and the representation on the sentence must include concept "The Watergate scandal", as shown in figure 4.

How would the proposed method deal with the common occurrence of metaphors that involve multiple components of a sentence (5)?[4]

(5) He ran into several obstacles on the path to the presidency.

The answer to that question depends on the goal of the semantic analysis and the outcome of the interpretation. If the goal is to find an unambiguous representation of the sentence as in our case, then it is

[4]This question and sentence were posed by an anonymous reviewer.

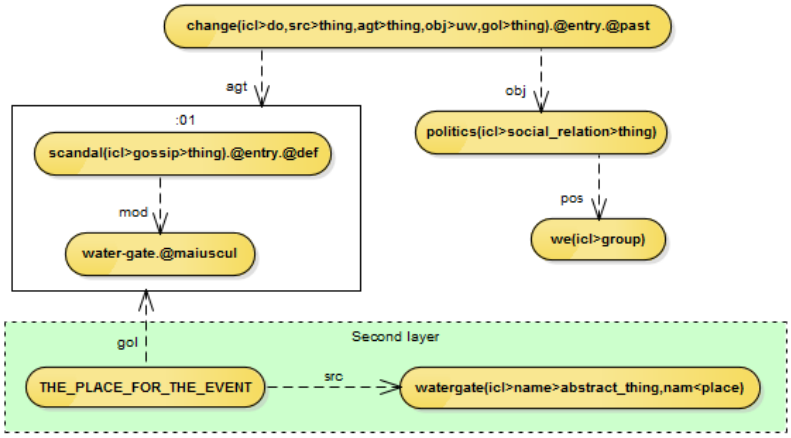


FIGURE 4 A UNL graph for sentence "Watergate changed our politics."

sufficient to replace the figurative language phrases with the substitutions derived from suitable database of metaphors. If the goal is to focus on a more general understanding, then the connections with metaphors at a sentence level can be added, but this information is only useful if it helps to analyze other sentences in a discourse. For this sentence the use of a metaphor such as LONG-TERM PURPOSEFUL ACTIVITY IS A JOURNEY is not apparent and requires certain assumptions to be made. Of course we can make a connection between the metaphor and the sentence, but based on this particular metaphor there is no obvious paraphrase to use for the sentence as a whole.

Figure 5 shows one possible solution to conveying graphically the essential meaning of the sentence (5).

In this case the following replacements were made:

- the idiom **ran into something** was replaced with the phrase **experienced something unexpectedly** based on a dictionary look-up (Heacock, 2003);
- **obstacles** was replaced with **difficulties** based on the metaphor DIFFICULTIES ARE IMPEDIMENTS;
- **the path to the presidency** was replaced with **the process of becoming a president** based on the metaphor THE MEANS OF ACHIEVING PURPOSES ARE ROUTES.

It is interesting to observe that each replacement can be done individually without altering the meaning of the sentence.

As before, the goal was that in the final dataset there must exist

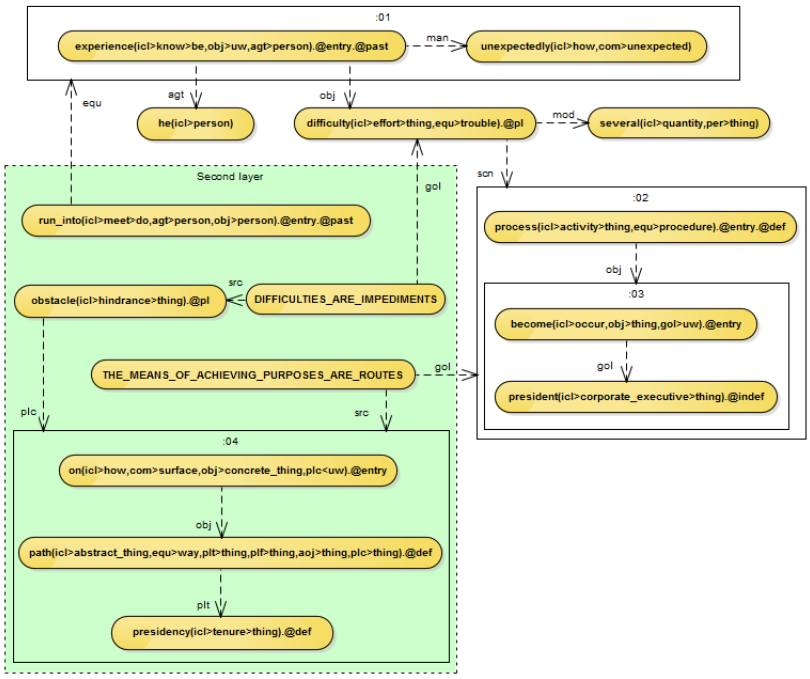


FIGURE 5 A UNL graph for sentence "He ran into several obstacles on the path to the presidency."

an unambiguous machine-processable concept, as well as an equivalent concept for the literal meaning. The specific realization of the two-layered representation depends on the language and the system used.

## 4 A metaphorical concept in a knowledgebase

In order for the computers to be able to adequately analyze figurative language on sentence level, there must exist relevant language resources and ontologies. If, for example, the knowledge about a metaphor needs to be recorded, then the required information would be the substitutions within the metaphorical concept, which refer to the corresponding concepts in the source and target domains, allowing us to find a suitable machine-processable concept for the sentence. A conceptual approach to the placement of a metaphor into a knowledge base is shown in class diagram 6.

Sometimes, it might happen that in the applicable ontology there exists a concept with a specific meaning for a certain figuration. When

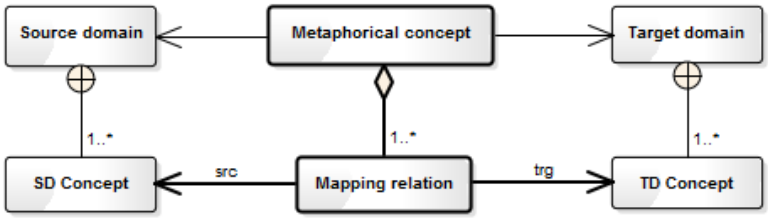


FIGURE 6 Class diagram: A metaphorical concept in a knowledge base

we use such concept in order to represent a meaning of a figuration in a discourse, then it may seem that there is no immediate cause for binding the concept-figuration pair with the relevant metaphorical concept. However, the usability of such representation of a figuration in comparisons or estimates remains unclear, and the result therefore would be undesirable. Consequently, it is important not to include the concepts corresponding to the figurative speech phrases in such ontology, which is originally meant to contain the concepts involved in the calculations.

Let's look at a sentence 6, where the noun '*behavior*' is linked to the adjective '*unpalatable*'.

(6) He found her behavior unpalatable.

In this example, we have a metaphor that refers to our perception, specifically to the sense of taste. If there would be a concept for the phrase '*unpalatable behavior*' in an ontology, then it would probably bear the knowledge that we are dealing with some kind of behavior, but this concept is not included into some scale of values that can be used in the calculations. However, when we want to correct the described situation and we introduce the metaphorical concept PALATABLE IS PLEASANT, then it is possible to link the concept '*unpalatable*' in the domain '*flavors*' to the concept '*unpleasant*' in the domain '*subjective assessments*', in the latter concepts are placed on the better-worse scale. In the context of the noun '*behavior*', the concept '*unpleasant*' is much more usable in the comparisons than the concept '*unpalatable*'.

## 5 Conclusion

In the current article we have studied one possible method to annotate figurative language according in discourse in a way that makes the result machine-processable and retains the initial reference to the figurative language used in the text. If we want to apply this theory in practice, a number of preparatory steps must be made. In order for the

computers to be able to link the expressions of figurative language to the corresponding machine-processable concepts in the process of semantic analysis of a natural language, a lexicon of the phraseologisms is required for each language. Clarification of the meaning of metaphors in a semantic analysis requires, on one hand the presence of the appropriate domains and the concepts in the ontologies, and on the other hand the formal description of the identified metaphorical concepts in the knowledge bases that use these ontologies.

## References

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Linguistic Annotation Workshop*. Citeseer.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2014. Abstract Meaning Representation (AMR) 1.2 Specification.

Basile, Valerio, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *LREC*, vol. 12, pages 3196–3200.

Bos, Johan. 2004. Computational semantics in discourse: Underspecification, resolution, and inference. *Journal of Logic, Language and Information* 13(2):139–157.

Bos, Johan. 2008. Wide-coverage semantic analysis with Boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 277–286. Association for Computational Linguistics.

Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Bunt, Harry, Rashmi Prasad, and Aravind Joshi. 2012. First steps towards an ISO standard for annotating discourse relations. In *Joint ISA-7 Workshop on Interoperable Semantic Annotation SRSL-3 Workshop on Semantic Representation for Spoken Language I2MRT Workshop on Multimodal Resources and Tools*, page 80.

Curran, James R, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 33–36. Association for Computational Linguistics.

Heacock, Paul. 2003. *Cambridge Dictionary of American Idioms*. Cambridge University Press.

Jones, Bevan, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, and Kevin Knight. 2012. Semantics-Based Machine Translation with Hyperedge Replacement Grammars. In *COLING*, pages 1359–1376.

Kamp, Hans, Josef Van Genabith, and Uwe Reyle. 2011. Discourse representation theory. In *Handbook of philosophical logic*, pages 125–394. Springer.

Kumar, Parteek. 2012. *UNL Based Machine Translation System for Punjabi Language*. Ph.D. thesis, Thapar University.

Lakoff, George and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Langkilde, Irene and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 1, pages 704–710. Association for Computational Linguistics.

Martins, Ronaldo. 2012. Le Petit Prince in UNL. In *LREC*, pages 3201–3204.

Nunberg, Geoffrey, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language* pages 491–538.

Salam, Khan Md Anwarus, Hiroshi Uchida, and Tetsuro Nishino. 2012. Multilingual Universal Word Explanation Generation from UNL Ontology. In *24th International Conference on Computational Linguistics*.

Uchida, Hiroshi, Meiying Zhu, and Tarcisio Della Senta. 1999. *A gift for a millenium*. Tokyo: IAS/UNU.

Uchida, Hiroshi, Meiying Zhu, and Tarcisio Della Senta. 2005. *The Universal Networking Language, 2nd ed*. UNDL Foundation.