# Interlingual Annotation of Multilingual Text Corpora

David Farwell
Stephen Helmreich
Computing Research Laboratory
New Mexico State University
david@crl.nmsu.edu
shelmrei@crl.nmsu.edu

Bonnie Dorr
Nizar Habash
Institute for Advanced Computer Studies
University of Maryland
bonnie@umiacs.umd.edu
habash@umiacs.umd.edu

Florence Reeder
Keith Miller
Mitre Corporation
freeder@mitre.org
keith@mitre.org

Lori Levin
Teruko Mitamura
Language Technologies Institute
Carnegie Mellon University
lsl@cs.cmu.edu
teruko@cs.cmu.edu

Eduard Hovy
Information Sciences Institute
University of Southern California
hovy@isi.edu

Owen Rambow
Advaith Siddharthan
Department of Computer Science
Columbia University
rambow@cs.columbia.edu
as372@cs.columbia.edu

## Abstract

This paper describes a multi-site project to annotate six sizable bilingual parallel corpora for interlingual content. After presenting the background and objectives of the effort, we will go on to describe the data set that is being annotated, the interlingua representation language used, an interface environment that supports the annotation task and the annotation process itself. We will then present a preliminary version of our evaluation methodology and conclude with a summary of the current status of the project along with a number of issues which have arisen.

## 1 Introduction

This paper describes a multi-site NSF ITR project focusing on the annotation of six sizable bilingual parallel corpora for interlingual content with the goal of providing a significant data set for improving knowledge-based approaches to machine translation (MT) and a range of other Natural Language Processing (NLP) applications. The project participants include the Computing Research Laboratory at NMSU, the Language Technologies Institute at CMU, the Information Science Institute at USC, UMIACS at the University of Maryland, MITRE Corp. and Columbia University. In the remainder of the paper, we first present the background and objectives of the project. We then describe the data set that is being annotated, the interlingual representation language being used, an interface environment that is designed to support the annotation task and the process of annotation itself. We will then outline a preliminary version of our evaluation methodology and conclude with a summary of the current status of the project along with a set of issues that have arisen since the project began.

## 2 Project Goals and Expected Outcomes

The central goals of the project are:

- to produce a practical, commonly-shared system for representing the information conveyed by a text, or interlingua,
- to develop a methodology for accurately and consistently assigning such representations to texts across languages and across annotators,

- to annotate a sizable multilingual of parallel corpus of source language texts and translations for IL content.

This corpus is expected to serve as a basis for improving meaning-based approaches to MT and a range of other natural language technologies. The tools and annotation standards will serve to facilitate more rapid annotation of texts in the future.

## 3 Corpus

The target data set is modeled on and an extension of the DARPA MT Evaluation data set (White and O'Connell 1994) and includes data from the Linguistic Data Consortium (LDC) Multiple Translation Arabic, Part 1 (Walker et al., 2003). The data set consists of 6 bilingual parallel corpora. Each corpus is made up of 125 source language news articles along with three independently produced translations into English. However, the source news articles for each individual language corpus are different from the source articles in the other language corpora. Thus, the 6 corpora themselves are comparable to each other rather than parallel. The source languages are Japanese, Korean, Hindi, Arabic, French and Spanish. Typically, each article is between 300 and 400 words long (or the equivalent) and thus each corpus has between 150,00 and 200,000 words. Consequently, the size of the entire data set is around 1,000,000 words.

Thus, for any given corpus, the annotation effort is to assign interlingual content to a set of 4 parallel texts, 3 of which are in the same language, English, and all of which theoretically communicate the same information. The following is an example set from the Spanish corpus:

**S:** Atribuyó esto en gran parte a una política que durante muchos años tuvo un "sesgo concentrador" y representó desventajas para las clases menos favorecidas.

**T1:** He attributed this in great part to a type of politics that throughout many years possessed a "concentrated bias" and represented disadvantages for the less favored classes.

**T2:** To a large extent, he attributed that fact to a policy which had for many years had a "bias toward concentration" and represented disadvantages for the less favored classes.

**T3:** He attributed this in great part to a policy that had a "centrist slant" for many years and represented disadvantages for the less-favored classes.

The annotation process involves identifying the variations between the translations and then assessing whether these differences are significant. In this case, the translations are, for the most part, the same although there are a few interesting variations.

For instance, where `this` appears as the translation of `esto` in the first and third translations, `that fact` appears in the second. The translator choice potentially represents an elaboration of the semantic content of the source expression and the question arises as to whether the annotation of the variation in expressions should be different or the same.

More striking perhaps is the variation between `concentrated bias`, `bias toward concentration` and `centrist slant` as the translation for `sesgo concentrador`. Here, the third translation offers a clear interpretation of the source text author's intent. The first two attempt to carry over the vagueness of the source expression assuming that the target text reader will be able to figure it out. But even here, the two translators appear to differ as to what the source language text author's intent actually was, the former referring to bias of a certain degree of strength and the second to a bias of a certain in a certain direction. Seemingly, then, the annotation of each of these expressions should differ.

The point here, however, is that a multilingual parallel data set of source language texts and English translations offers a unique perspective and unique problem for annotating texts for meaning.

## 4 Interlingua

Due to the complexity of an interlingual annotation as indicated by the differences described in the previous section, the representation has developed through three levels and incorporates knowledge from sources such as the Omega ontology and theta grids. Since this is an evolving standard, the three levels will presented in order as building on one another. Then the additional data components will be described.

### 4.1 Three Levels of Representation

We now describe three levels of representation, referred to as *IL0, IL1* and *IL2*. The aim is to perform the annotation process incrementally, with each level of representation incorporating additional semantic features and removing existing syntactic ones. *IL2* is intended as the interlingua, that abstracts away from (most) syntactic

idiosyncrasies of the source language. *IL0* and *IL1* are intermediate representations that are useful starting points for annotating at the next level.

#### 4.1.1 IL0

*IL0* is a deep syntactic representation. It includes part-of-speech tags for words and a parse tree that makes explicit the syntactic arguments of verbs. While the features used in this representation are purely syntactic, many disambiguation decisions, relative clause and PP attachment for example, are made at the level of semantics, rather than syntax. *IL0* is constructed by hand-correcting the output of a dependency parser (details in section 6) and is a useful starting point for semantic annotation at *IL1*, since it allows annotators to see how textual units relate syntactically when making semantic judgments.

#### 4.1.2 IL1

IL1 is an intermediate semantic representation. It associates semantic concepts with lexical units like nouns, adjectives, adverbs and verbs (details of the ontology in section 4.2). It also replaces the syntactic relations in *IL0*, like *subject* and *object*, with thematic roles, like *agent, theme* and *goal* (details in section 4.3). However, *IL1* is not an interlingua; it does not normalize over different linguistic realizations of the same semantics. In particular, it does not address how the meanings of individual lexical units combine to form the meaning of a phrase or clause. It also does not address idioms, metaphors and other non-literal uses of language. Further, *IL1* does not assign semantic features to prepositions; these continue to be encoded as syntactic features of their objects, although these might have been annotated with thematic roles such as *location* or *time*.

#### 4.1.3 IL2

*IL2* is intended to be an interlingua, a representation of meaning that is (reasonably) independent of language. *IL2* is intended to capture similarities in meaning across languages and across different lexical/syntactic realizations within a language. For example, *IL2* is expected normalize over conversives (e.g. X *bought a book from Y* vs. Y *sold a book to X)* and non-literal language usage (e.g. *X started its business* vs. *X opened its doors to customers*).

### 4.2 The Omega Ontology

In progressing from IL0 to IL1, annotators have to select semantic terms (concepts) to represent the nouns, verbs, adjectives, and adverbs present in each sentence. These terms are represented in the 110,000-node ontology Omega (Philpot et al., 2003), under construction at ISI. Omega has been built semi-automatically from a variety of sources, including Princeton's WordNet (Fellbaum, 1998), New Mexico State University's Mikrokosmos (Mahesh and Nirenburg, 1995), ISI's Upper Model (Bateman et al., 1989) and ISI's SENSUS (Knight and Luk, 1994). After the uppermost region of Omega was created by hand, these various resources' contents were incorporated and, to some extent, reconciled. After that, several million instances of people, locations, and other facts were added (Fleischman et al., 2003). The ontology, which has been used in several projects in recent years (Hovy et al., 2001), can be browsed using the DINO browser at http://blombos.isi.edu:8000/dino; this browser forms a part of the annotation environment. Omega remains under continued development and extension.

### 4.3 The Theta Grids

Each verb in Omega is assigned one or more theta grids specifying the arguments associated with a verb and their theta roles (or thematic role). Theta roles are abstractions of deep semantic relations that generalize over verb classes. They are by far the most common approach in the field to represent predicate-argument structure. However, there are numerous variations with little agreement even on terminology (Fillmore, 1968; Stowell, 1981; Jackendoff, 1972; Levin and Rappaport-Hovav, 1998).

The theta grids used in our project were extracted from the Lexical Conceptual Structure Verb Database (LVD) (Dorr, 2001). The WordNet senses assigned to each entry in the LVD were then used to link the theta grids to the verbs in the Omega ontology. In addition to the theta roles, the theta grids specify syntactic realization information, such as Subject, Object or Prepositional Phrase, and the Obligatory/Optional nature of the argument. For example, one of the theta grids for the verb "load" is listed in Table 1 (at the end of the paper).

The set of theta roles used, although based on research in LCS-based MT (Dorr, 1993; Habash et al, 2002), has been simplified for this project. This list (see Table 2 at the end of the paper), was used in the Interlingua Annotation Experiment 2002 (Habash and Dorr).[1]

### 4.4 Incremental Annotation

As described earlier, the development and annotation of the interlingual notation is incremental in nature. This necessitates constraining the types and categories of attributes included in the annotation through the beginning phases. As was mentioned with prepositions, the annotation to include role assignment is yet to be determined, although it has been well-discussed. Other topics not addressed here, but

[1] Other contributors to this list are Dan Gildea and Karin Kipper Schuler.

considered for future work include time, aspect, location, modality, type of reference, types of speech act, causality, etc.

## 5 Annotation Tool

We have assembled a suite of tools to be used in the annotation process. Some of these tools were previously existing resources that were gathered for use in the project, and others were developed specifically with the annotation goals of this project in mind. Since we are gathering our corpora from disparate sources, we need to standardize the text before presenting it to automated procedures. For English, this involves splitting the text into sentences, but for other languages, it may involve segmentation, chunking of text, or other "text ecology" operations. The text is then processed with a dependency parser, the output of which is viewed by the researchers in TrED (Hajič, et al., 2001), a graphically-based tree editing program, written in Perl/Tk[2]. The revised deep dependency structure produced by this process is the IL0 representation for that sentence.

In order to derive IL1 from the IL0 representation, annotators use Tiamat, a tool developed specifically for this project. This tool was developed to support the annotators in their task, and enables viewing of the IL0 tree with easy reference to all of the IL resources described in section 4 (the current IL representation, the ontology, and the theta grids). This tool provides the ability to annotate text via simple point-and-click selections of words, concepts, and theta-roles. The IL0 is displayed in the top left pane, ontological concepts and their associated theta grids, if applicable, are located in the top right, and the sentence itself is located in the bottom right pane[3]. An annotator may select a lexical item (leaf node) to be annotated in the sentence view; this word is highlighted, and the relevant portion of the Omega ontology is displayed in the pane on the left. In addition, if this word has dependents, they are automatically underlined in red in the sentence view. Annotators can view all information pertinent to the process of deciding on appropriate ontological concepts in this view. Following the procedures described in section 6, selection of concepts, theta grids and roles appropriate to that lexical item can then be made in the appropriate panes.

Evaluation of the annotators' output would be daunting based solely on a visual inspection of the annotated IL1 files. Thus, a tool was also developed to compare the output and to generate the evaluation measures that are described in section 7. The reports generated by the evaluation tool allow the researchers to look at both gross-level phenomena, such as inter-annotator agreement, and at more detailed points of interest, such as lexical items on which agreement was particularly low, possibly indicating gaps or other inconsistencies in the ontology being used.

## 6 Annotation Task

To describe the annotation task, we first present the annotation process and tools used with it. Then, the annotation manuals are described. Finally, setup issues relating to negotiating multi-site annotations are discussed.

### 6.1 Annotation process

The annotation process was identical for each text. For the initial testing period, only English texts were annotated, and the process described here is for English text. The process for non-English texts will be, *mutatis mutandis*, the same.

Each sentence of the text is parsed into a dependency tree structure. For English texts, these trees were first provided by the Connexor parser at UMIACS (Tapanainen and Jarvinen, 1997), and then corrected by one of the team PIs. For the initial testing period, annotators were not permitted to alter these structures. Already at this stage, some of the lexical items are replaced by features (e.g., tense), morphological forms are replaced by features on the citation form, and certain constructions are regularized (e.g., passive) and empty arguments inserted. It is this dependency structure that is loaded into the annotation tool and which each annotator then marks up.

The annotator was instructed to annotate all nouns, verbs, adjectives, and adverbs. This involves annotating each word twice – once with a concept from Wordnet SYNSET and once with a Mikrokosmos concept, which are merged, or at least intertwined in Omega. One of the goals and results of this annotation process will be a simultaneous coding of concepts in both ontologies, facilitating a closer union between them.

In addition, users were instructed to provide a semantic case role for each dependent of a verb. In many cases this was "NONE" since adverbs and conjunctions were dependents of verbs in the dependency tree. LCS verbs were identified with Wordnet classes and the LCS case frames supplied where possible. The user, however, was often required to determine the set of roles or alter them to suit the text. In both cases, the revised or new set of case roles was noted and sent to a guru for evaluation and possible permanent inclusion. Thus the set of event concepts supplied with roles will grow through the course of the project.

[2] http://quest.ms.mff.cuni.cz/pdt/Tools/Tree_Editors/Tred/

[3] To be added to paper at later date.

### 6.2 The annotation manuals

Markup instructions are contained in three manuals: a users guide for Tiamat, a definitional guide to semantic roles, and a manual for creating a dependency structure. Together these manuals allow the annotator to (1) understand the intention behind aspects of the dependency structure; (2) how to use Tiamat to mark up texts; and (3) how to determine appropriate semantic roles and ontological concepts. In choosing a set of appropriate ontological concepts, annotators were encouraged to look at the name of the concept and its definition, the name and definition of the parent node, example sentences, lexical synonyms attached to the same node, and sub- and super-classes of the node. All these manuals are available on the IAMTC website.

### 6.3 The multi-site set up

For the initial testing phase of the project, all annotators at all sites worked on the same texts. Two texts were provided by each site as were two translations of the same source language (non-English) text. To test for the effects of coding two texts that are semantically close, since they are both translations of the same source document, the order in which the texts were annotated differed from site to site, with half the sites marking one translation first, and the other half of the sites marking the second translation first. Another variant tested was to interleave the two translations, so that two similar sentences were coded consecutively.

During the later production phase, a more complex schedule will be followed, making sure that many texts are annotated by two annotators, often from different sites, and that regularly all annotators will mark the same text. This will help ensure continued inter-coder reliability.

In the period leading up to the initial test phase, weekly conversations were held at each site by the annotators, going over the texts coded. This was followed by a weekly conference call among all the annotators. During the test phase, no discussion was permitted.

Some of the issues arose in discussing how certain constructions should be displayed and whether each word should have a separate node or whether certain words should be combined into a single node. In view of the fact that the goal was not to tag individual words, but entities and relations, in many cases words were combined into single nodes to facilitate this process. For instance, verb-particle constructions were combined into a single node. For example, in a sentence like "He threw it up", "throw" and "up" were combined into a single node "throw up" since one action is described by the combined words. Similarly, proper nouns, compound nouns and copular constructions required specialized handling. In addition, issues arose about whether annotators should change dependency trees; and in instructing the annotators on how best to determine an appropriate ontology node.

## 7 Evaluation

The evaluation criteria and metrics continue to evolve. Several courses for evaluating the annotations and resulting structures exist. In the first evaluation path, the annotations are measured according to inter-annotator agreement. For this purpose, data is collected reflecting the annotations selected, the Omega nodes selected and the theta roles described. Then, inter-coder agreement is measured in a profile that includes a Kappa measure (Carletta, 1993) and a Wood standard similarity (Habash and Dorr, 2002) for annotations, Omega selection and theta role selection. For expected agreement in the Kappa statistic, P(E) is defined as 1/(N+1) where N is the number of choices at a given data point. In the case of Omega nodes, this means the number of matched Omega nodes (by string match) plus one for the possibility of the annotator traversing up or down the hierarchy. Multiple measures (including Wood standard similarity) were used because it is important to have a mechanism for evaluating inter-coder consistency in the use of the IL representation language which does not depend on the assumption that there is a single correct annotation of a given text. The measurement tool used here has been described (Habash and Dorr, 2002).

Second, the accuracy of the annotation is measured. Here accuracy is defined as correspondence to a predefined baseline. In the initial development phase, all sites annotated the same texts and many of the variations were discussed at that time, permitting the development of a baseline annotation. While not a useful long-term strategy, this produced a consensus baseline for the purpose of measuring the annotators' task and the solidity of the annotation standard.

Finally, an attempt to evaluate the resulting representation has also been devised and is currently being implemented. Since the ultimate goal is to generate an IL representation which is useful for MT (among other NLP tasks), we measure the ability to generate accurate surface texts from the IL representation as annotated. At this stage, we are using an available generator, Halogen (Knight and Langkilde, 2000). A tool to convert the representation to meet Halogen requirements is being built. Following the conversion, surface forms will be generated and then compared with the originals through a variety of standard MT metrics (ISLE, 2003). This will serve to determine if the elements of the representation language are sufficiently well-defined and if they serve as a basis for inferring interpretations from semantic representations or (target) semantic representations from interpretations.

# 8 Conclusion

## 8.1 Accomplishments so far

In a short amount of time, we have identified languages and collected corpora with translations appropriate to those languages. We have selected representation elements, from parser outputs to ontologies, and have developed an understanding of how their component elements fit together. A core markup vocabulary (e.g., entity-types, event-types and participant relations) was selected. An initial version of the annotator's toolkit (Tiamat) has been developed and has gone through alpha testing. The multi-layered approach to annotation decided upon reduces the burden on the annotators for any given text as annotations build upon one another. In addition to developing individual tools, an infrastructure now exists for carrying out a multi-site annotation project.

In the coming months we will be fleshing out the current procedures for evaluating the accuracy of an annotation and measuring inter-coder consistency. From this, a multi-site evaluation will result with results to be reported. Regression testing, from the intermediate stages and representations will be able to be carried out. Finally, a growing corpus of annotated texts will become available.

## 8.2 Issues that have arisen

In addition to the issues discussed throughout the paper, a few others have not yet been identified. From a content standpoint, looking at IL systems for time and location should utilize work in personal name, temporal and spatial annotation (e.g., Ferro et al., 2001). Also, an ideal IL representation would also account for causality, co-reference, aspectual content, modality, speech acts, etc. At the same time, while incorporating these items, vagueness and redundancy must be eliminated from the annotation language. Many inter-event relations would need to be captured such as entity reference, time reference, place reference, causal relationships, associative relationships, etc. Finally, to incorporate these, cross-sentence phenomena remain a challenge.

From an MT perspective, issues include evaluating the consistency in the use of an annotation language given that any source text can result in multiple, different, legitimate translations (see Farwell and Helmreich, 2003) for discussion of evaluation in this light. Along these lines, the problem of annotating texts for interpretation without including in the annotations inferences from the source text.

## 8.3 Concluding remarks

This is a radically different annotation project from those that have focused on morphology, syntax or even certain types of semantic content (e.g., for word sense disambiguation competitions). It is most similar to PropBank and Framebank annotation efforts and should dovetail with those efforts. We are attacking a more abstract level of mark up (interpretation), we are attacking the assignment of well-defined meaning representation language to concrete texts, and we are addressing the issues of a community-wide consistent and accurate annotation of meaning.

By providing an essential, and heretofore non-existent, data set for training and evaluating knowledge-based language processing systems, the resultant annotated multilingual corpus of translations is expected to lead to significant research and development opportunities for Machine Translation and a host of other Natural Language Processing technologies including Question-Answering (paraphrase and entailment relations) and Information Extraction. Not only will this lead to improved translation and language technologies but, just as importantly, it will increase our understanding of human cognitive processing.

# References


Bateman, J.A., Kasper, R.T., Moore, J.D., and Whitney, R.A. 1989. A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Unpublished research report, USC/Information Sciences Institute, Marina del Rey, CA.

Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2), 249-254

Conceptual Structures and Documentation, UMCP. http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

Dorr, Bonnie J. 2001. LCS Verb Database, Online Software Database of Lexical

Dorr, Bonnie J., 1993. *Machine Translation: A View from the Lexicon*, MIT Press, Cambridge, MA.

Farwell, David, and Steve Helmreich. 2003. Pragmatics-based Translation and MT Evaluation. In Proceedings of Towards Systematizing MT Evaluation. MT-Summit Workshop, New Orleans, LA.

Fellbaum, C. (ed.). 1998. WordNet: An On-line Lexical Database and Some of its Applications. MIT Press, Cambridge, MA.

Ferro, Lisa, Inderjeet Mani, Beth Sundheim and George Wilson. 2001. TIDES Temporal Annotation Guidelines. Version 1.0.2 MITRE Technical Report, MTR 01W0000041

Fillmore, Charles. 1968. The Case for Case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1--88. Holt, Rinehart, and Winston.

Fleischman, M., A. Echihabi, and E.H. Hovy. 2003. Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. Proceedings of the ACL Conference. Sapporo, Japan.

Habash, Nizar and Bonnie Dorr. 2002. Interlingua Annotation Experiment Results. AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA.

Habash, Nizar, Bonnie J. Dorr, and David Traum, 2002. "Efficient Language Independent Generation from Lexical Conceptual Structures," *Machine Translation*, 17:4.

Hajič, Jan; Vidová-Hladká, Barbora; Pajas, Petr. 2001: The Prague Dependency Treebank: Annotation Structure and Support. In Proceeding of the IRCS Workshop on Linguistic Databases, pp. . Univesity of Pennsylvania, Philadelphia, USA, pp. 105-114.

Hovy, E.H., Philpot, A., Ambite, J.L., Arens, Y., Klavans, J., Bourne, W., and Saroz, D. 2001. Data Acquisition and Integration in the DGRC's Energy Data Collection Project, in Proceedings of the NSF's dg.o 2001. Los Angeles, CA.

Jackendoff, Ray. 1972. Grammatical Relations and Functional Structure. *Semantic Interpretation in Generative Grammar*. The MIT Press, Cambridge, MA.

Knight, K., and I. Langkilde. 2000. Preserving Ambiguities in Generation via Automata Intersection. American Association for Artificial Intelligence conference (AAAI).

Knight, K, and Luk, S.K. 1994. Building a Large-Scale Knowledge Base for Machine Translation. Proceedings of AAAI. Seattle, WA.

Levin, B. & Rappaport-Hovav, M. 1998. From Lexical Semantics to Argument Realization. Borer, H. (ed.) *Handbook of Morphosyntax and Argument Structure*. Dordrecht: Kluwer Academic Publishers.

Mahesh, K., and Nirenberg, S. 1995. A Situated Ontology for Practical NLP, in Proceedings on the Workshop on Basic Ontological Issues in Knowledge Sharing at IJCAI-95. Montreal, Canada.

Philpot, A., M. Fleischman, E.H. Hovy. 2003. Semi-Automatic Construction of a General Purpose Ontology. Proceedings of the International Lisp Conference. New York, NY. Invited.

Stowell, T. 1981. Origins of Phrase Structure. *PhD thesis*, MIT, Cambridge, MA.

Tapanainen, P. and T Jarvinen. 1997. A non-projective dependency parser. In the 5th Conference on Applied Natural Language Processing / Association for Computational Linguistics, Washington, DC.

White, J., and T. O'Connell. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of the 1994 Conference, Association for Machine Translation in the Americas

Walker, Kevin, Moussa Bamba, David Miller, Xiaoyi Ma, Chris Cieri, and George Doddington 2003. Multiple-Translation Arabic Corpus, Part 1. Linguistic Data Consortium (LDC) catalog number LDC2003T18 and ISBN 1-58563-276-7.

| Role | Description | Grid | Syntax | Type |
|---|---|---|---|---|
| Agent | The entity that does the action | Agent: *load*<br>Theme with possessed | SUBJ | OBLIGATORY |
| Theme | The entity that is worked on | Agent: *load*<br>Theme with possessed | OBJ | OBLIGATORY |
| Possessed | The entity controlled or owned | Agent: *load*<br>Theme with possessed | PP | OPTIONAL |

Table 1 : A theta grid for the verb "load"

| Role and Definition | Examples |
|---|---|
| Agent: Agents have the features of volition, sentience, causation and independent exist | • Henry *pushed/broke* the vase. |
| Instrument: An instrument should have causation but no volition. Its sentience and existence are not relevant. | • The Hammer *broke* the vase.<br>• She *hit him* with a baseball bat |
| Experiencer: An experiencer has no causation but is sentient and exists independently. Typically an experiencer is the subject of verbs like feel, hear, see, sense, smell, notice, detect, etc. | • John *heard* the vase shatter.<br>• John *shivered*. |
| Theme: The theme is typically causally affected or experiences a movement and/or change in state. The theme can appear as the information in verbs like acquire, learn, memorize, read, study, etc. It can also be a thing, event or state (clausal complement). | • John *went* to school.<br>• John *broke* the vase.<br>• John *memorized* his lines.<br>• She *buttered* the bread with margarine. |
| Perceived: Refers to a perceived entity that isn't required by the verb but further characterizes the situation. The perceived is neither causally affected nor causative. It doesn't experience a movement or change in state. Its volition and sentience are irrelevant. Its existence is independent of an experiencer. | • He *saw* the play.<br>• He *looked* into the room.<br>• The cat's fur *feels* good to John.<br>• She *imagined* the movie to be loud. |
| Predicate: Indicates new modifying information about other thematic roles. | • We *considered* him a fool.<br>• She *acted* happy. |
| Source: Indicates where/when the theme started in its motion, or what its original state was, or where its original (possibly abstract) location/time was. | • John left the house. |
| Goal: Indicates where the theme ends up in its motion, or what its final state is, or where/when its final (possibly abstract) location/time is. It also can indicate the thing/event resulting from the verb's occurrence (the result). | • John *ran* home.<br>• John *ran* to the store.<br>• John *gave* a book to Mary.<br>• John *gave* Mary a book. |
| Location: Indicates static locations---as opposed to a source or goal, i.e., the (possibly abstract) location of the theme or event. | • He *lived* in France.<br>• The water *fills* the box.<br>• This cabin *sleeps* five people |
| Time Indicates time. | • John *sleeps* for five hours.<br>• Mary *ate* during the meeting. |
| Beneficiary: Indicates the thing that receives the benefit/result of the event/state. | • John *baked* the cake for Mary.<br>• John *baked* Mary a cake.<br>• An accident *happened* to him. |
| Purpose: Indicates the purpose/reason behind an event/state | • He *studied* for the exam.<br>• He *searched* for rabbits. |
| Possessed: Indicates the possessed entity in verbs such as own, have, possess, fit, buy, and carry. | • John *has* five bucks.<br>• He *loaded* the cart with hay.<br>• He *bought* it for five dollars |
| Proposition: Indicates the secondary event/state | • He *wanted* to study for the exam. |
| Modifier: Indicates a property of a thing such as color, taste, size, etc. | • The red book sitting on the table is old. |
| Null Indicates no thematic contribution. Typical examples are impersonal *it* and *there*. | • It was raining all morning in Miami. |

TABLE 2: List of Theta Roles