Natural Language Engineering **16** (3): 197–243.
doi:10.1017/S1351324910000070

Interlingual annotation of parallel text corpora: a new framework for annotation and evaluation

BONNIE J. DORR[1], REBECCA J. PASSONNEAU[2], DAVID FARWELL[3], REBECCA GREEN[4], NIZAR HABASH[2], STEPHEN HELMREICH[3], EDUARD HOVY[6], LORI LEVIN[5], KEITH J. MILLER[7], TERUKO MITAMURA[5], OWEN RAMBOW[2] and ADVAITH SIDDHARTHAN[8]

[1]*Institute for Advanced Computer Studies, University of Maryland, AVW Williams Building 3153, College Park, MD 20742, USA*
e-mail: bonnie@umiacs.umd.edu
[2]*Center for Computational Learning Systems, Columbia University, 475 Riverside Drive MC 7717, New York, NY 10115, USA*
e-mails: {becky, habash, rambow}@cs.columbia.edu
[3]*Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88001, USA*
e-mails: {david, shelmrei}@crl.nmsu.edu
[4]*OCLC Online Computer Library Center, Inc., 6565 Kilgour Place, Dublin, OH 43017-3395, USA*
e-mail: greenre@oclc.org
[5]*Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213-3890, USA*
e-mails: {lsl, teruko}@cs.cmu.edu
[6]*Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA*
e-mail: hovy@isi.edu
[7]*The MITRE Corporation, 7515 Colshire Drive, Mc Lean, VA 22102-7539, USA*
e-mail: {freeder, keith}@mitre.org
[8]*Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE, Scotland, UK*
e-mail: advaith@abdn.ac.uk

(*Received 1 March 2008; revised 1 March 2010; accepted 31 March 2010*)

Abstract

This paper focuses on an important step in the creation of a system of meaning representation and the development of semantically annotated parallel corpora, for use in applications such as machine translation, question answering, text summarization, and information retrieval. The work described below constitutes the first effort of any kind to annotate multiple translations of foreign-language texts with interlingual content. Three levels of representation are introduced: deep syntactic dependencies (IL0), intermediate semantic representations (IL1), and a normalized representation that unifies conversives, nonliteral language, and paraphrase (IL2). The resulting annotated, multilingually induced, parallel corpora will be useful as an empirical basis for a wide range of research, including the development and evaluation of interlingual NLP systems and paraphrase-extraction systems as well as a host of other research and development efforts in theoretical and applied linguistics, foreign language pedagogy, translation studies, and other related disciplines.

1 Introduction

This paper focuses on an important step in the creation of a system of meaning representation and the development of semantically annotated parallel corpora, for use in applications such as machine translation, question answering, text summarization, and information retrieval. Three levels of representation are introduced: deep syntactic dependencies (IL0), intermediate semantic representations (IL1), and a normalized representation that unifies conversives, nonliteral language, and paraphrase (IL2). The resulting annotated, multilingually induced, parallel corpora will be useful as an empirical basis for a wide range of research, including the development and evaluation of many different NLP systems and research efforts.

The importance of linguistically annotated parallel corpora and multilingual annotation tools is now widely recognized (Véronis 2000), yet there are currently few cases of annotated parallel corpora, and those that exist tend to be bilingual rather than multilingually induced (Garside, Leech and McEnery 1997). Moreover, much of the previous work on linguistic annotation of corpora has focused on the annotation of sentences with syntactic information only, e.g., part-of-speech tags (Francis and Kucera 1982), syntactic trees (Marcus, Santorini and Marcinkiewicz 1994), and positionally determined argument labels (Kingsbury and Palmer 2002; Kipper, Palmer and Rambow 2002).[1] This paper focuses on the next step in the creation of a system of meaning representation and the development of semantically annotated parallel corpora, for use in applications such as machine translation, question answering, text summarization, and information retrieval.

We present an approach to semantic annotation of parallel corpora, validated in multiple English translations of texts in six languages (Arabic, French, Hindi, Japanese, Korean and Spanish), and evaluated in several ways.[2] Three levels of representation are introduced: deep syntactic dependencies (IL0), intermediate semantic representations (IL1), and a normalized representation that unifies conversives, nonliteral language, and paraphrase (IL2).

This work is the result of a collaboration across six sites on project called Interlingual Annotation of Multilingual Text Corpora (IAMTC), sponsored by the National Science Foundation. The participants included Computing Research Laboratory at New Mexico State University (NMSU), the Language Technologies Institute at Carnegie Mellon University (CMU), the Information Science Institute (ISI) at the University of Southern California, the Institute for Advanced Computer Studies at the University of Maryland (UMD), the MITRE Corporation, and Columbia University.[3] The novelty of the work comes not only from the development

[1] More recently, resources such as OntoNotes (Hovy *et al.* 2006; Pradhan *et al.* 2007) and PropBank (Kingsbury *et al.* 2003) have been used for the purpose of bootstrapping deeper semantic content from word sense information.

[2] Our focus was on the annotation of the English translations of each source-language text where divergences are already readily apparent and interannotator agreement is more easily measured; although source-language annotation was also conducted, the results are not the focus of this study, but will be examined in a future.

[3] Different aspects of this work have been published previously (Dorr *et al.* 2004; Farwell *et al.* 2004; Mitamura *et al.* 2004; Reeder *et al.* 2004; Rambow *et al.* 2006).

of a (language-independent) multilevel interlingual representation, but also from improved methodologies for designing and evaluating such representations. More specifically, this project has the following objectives:

- Development of a framework for interlingual analysis based on a careful study of parallel English text corpora, translated from six different languages.
- Annotation of these parallel corpora using two tiers of an agreed-upon three-tiered interlingual representation.
- Development of semantic-annotation tools for use broadly by the parallel-text processing community (a tree editor, annotation interface, etc.). These tools enable effective and relatively problem-free annotation at six different sites and subsequent merging of the results.
- Design of new metrics and undertaking of various evaluations of the interlingual representations, ascertaining the degree of annotator agreement and providing a means for choosing the granularity of meaning representation that is appropriate for a given task.

The impact of this research stems from the depth of the annotation and the evaluation metrics for the annotation task. These enable research on both parallel-text processing methods and the modeling of language-independent meaning. To date, such research has been impossible, since corpora have been annotated at a relatively shallow (semantics-free) level, forcing NLP researchers to choose between shallow approaches and hand-crafted approaches, each having its own set of problems. We view our research as paving the way toward solutions to representational problems that would otherwise seriously hamper or invalidate later larger annotation efforts, especially if they are monolingually motivated.

Other linguistic annotation projects include Semeval data (Moore 1994), FrameNet (Baker, Fillmore and Lowe 1998) PropBank and VerbNet (Kingsbury and Palmer 2002; Kipper *et al.* 2002), and OntoNotes (Hovy *et al.* 2006; Pradhan *et al.* 2007). The corpora resulting from these efforts have enabled the use of machine learning tools (including stochastic methods) that have proven much better than hand-written rules at accounting for the wide variety of idiosyncratic constructions and expressions found in natural languages. However, machine learning approaches have in the past been restricted to fairly superficial phenomena.

The work described below constitutes the first effort of any kind to annotate multiple translations of foreign-language texts with interlingual content.[4] The resulting annotated, multilingually induced, parallel corpora will be useful as an empirical basis for a wide range of research, including the development and evaluation of interlingual NLP systems and paraphrase extraction systems as well as a host of other research and development efforts in theoretical and applied linguistics, foreign language pedagogy, translation studies, and other related disciplines. For example, current automatic approaches to paraphrase extraction (Barzilay and Lee 2003;

[4] The broader impact of this research lies in the multilingually induced parallel resources it provides, and in the annotation procedures and agreement evaluation metrics it has developed. Downloadable versions of our results are available at http://aitc.aitcnet.org/nsf/iamtc/results.html.

Pang, Knight and Marcu 2003; Dolan, Quirk and Brockett 2004; Bannard and Callison-Burch 2005; Callison-Burch, Koehn and Osborne 2006; Madnani *et al.* 2007) rely on bilingual parallel corpora. Annotation of corpora with appropriate interlingual information allows one to take paraphrase extraction to the next level, where interlingual content – rather than another language – becomes the pivot for monolingual paraphrastic pairs.

The next section describes the corpora we have annotated. Section 4 defines the different levels of interlingua and the representation language we use to annotate our corpora. Section 5 presents an interface environment that is designed to support the annotation task and describes the process of annotation itself. Section 6 describes our approach to evaluation: the data partitioning and agreement coefficients, the use of a new distance metric called MASI (Measuring Agreement for Set-valued Items), the scope of the annotation task, and the results of our evaluation. Finally, we look in more detail at various combinations of annotators to identify subsets that would achieve reliability above 0.70, and individuals who are relatively more consistent in comparison with other annotators. We conclude with the current status of the project and future directions.

2 Related work

There are very few multilingual annotation efforts that have produced readily usable and publicly available results on the scale of the current project. The EDR Corpus (Miyoshi *et al.* 1996) was obtained by collecting a large number of example sentences and analyzing them on morphological, syntactic, and semantic levels. The Japanese Corpus contains approximately 200,000 sentences, and the English Corpus contains approximately 120,000 sentences. However, the corpus has a very complicated internally defined structure and that is very difficult to handle for system development.

Also, the Japanese-funded parallel MMT (Multilingual MT) of CCIC (ODA funding, Japan, China, Thailand, Malaysia, Indonesia) produced IL-annotated multilingual corpora (Funaki 1993). This project started in 1987 and ended in 1994. The project produced a 50,000 basic-word dictionary and a 25,000 technical-word dictionary. It also produced 3,000 annotated sentences for testing multilingual translation systems. However, the results have not been made freely available for research.

More recently, the UNDL Foundation has produced UNL annotations of twenty-five documents in Arabic, French, Russian, Spanish, and English (Martins *et al.* 2000). However, these efforts are difficult to compare to our own study in that there has been no scientific study (with interannotator agreement across comparable documents) and it is difficult to obtain the results without paying a very high price. The current study is distinctive in its scale and novelty of the dimensions assessed and it serves as an example for future efforts, especially with respect to achieving and measuring interannotator agreement.

The closest comparisons to the present analysis are the OntoNotes project (Pradhan *et al.* 2007; Hovy *et al.* 2006) and a study using WordNet (Fellbaum

et al. 1998). In the former, annotators assigned the senses of nouns and verbs in several hundred thousand words of newspaper and transcribed broadcast news text in English, Chinese, and Arabic. By early 2009, all instances of the most frequent 900 English nouns and 800 English verbs had been sense-tagged, with annotators completing all instances of one word before proceedings to the next.

In contrast, in the latter study, annotators assigned a single WordNet sense for each closed class lexeme in approximately 660 words of running text. Approximately half the words were open-class lexemes, and the average number of senses was 6.6, with verbs being the most polysemous items. Instead of interannotator agreement coefficients, both studies used percent agreement. OntoNotes senses were carefully constructed to ensure that interannotator agreement would not drop below 90% on average (on the general assumption that if humans can agree at N%, then systems can be trained to agree at $N - 10$%). When the requisite agreement could not be reached, the word's senses were redefined, and in some cases made less fine-grained, until the results were satisfactory. In addition to agreement, the WordNet study also reported confidence ratings to assess the seventeen annotators. Overall agreement was 79% among novice taggers and 74% if experts were included. In both studies, agreement was lowest for verbs, which corresponds to our findings.

For our IAMTC data and in the WordNet task, novice annotators were instructed to examine specific lexical items for which they were to select concepts. In a similar annotation task with FrameNet (Fillmore, Johnson and Petruck 2003), but with expert annotators, annotators had to select which items to annotate from running text (Collin Baker, personal communication). Separate calculations were made regarding whether a word constituted a target, meaning whether FrameNet contained an appropriate concept (or Frame); this amounts to the question of whether annotators agreed on which words constitute Lexical Units (LUs; lexeme-sense pairs). For this stage of annotation, nine annotators had 86% agreement. Given cases of agreement on LUs, 94% agreed on the choice of frame. The higher figures for FrameNet compared with WordNet may be due to the smaller size of FrameNet in comparison with WordNet. The corresponding kappa value ranged from 0.65 for all annotators, to 0.74 for all but the worst annotator, to 0.87 for the two best annotators. These are quite comparable to our results.

In the role labeling task for PropBank (Palmer, Gildea and Kingsbury 2005b), individual annotators were presented with a specific class of items selected from corpora, rather than all words of running text. When annotators could not reach the desired level of agreement, the word senses and if necessary verb frames were refined and reformulated. For verbs from two corpora, percent agreement measures and kappa values were reported for verb role identification, verb role classification and the combination of the two, for pairs of annotations prior to an adjudication step. The verbs being annotated were less polysemous than the WordNet lexemes described above: 77% of the verbs had a single PropBank frame, with a maximum of twenty frames. In comparison, only 20% of the words in the WordNet text were monosemous, and an individual word could have as many as forty concepts or senses. For cases where annotators agreed on the verb roles, kappa values of 0.91 and 0.93 were reported on a per role basis (Palmer *et al.* 2005b), depending

on whether verb phrase adjuncts are included (e.g., temporal adverbs). However, neither the method of expected agreement nor the number of degrees of freedom is specified. In this work, annotators chose from a longer list of role names and somewhat lower kappa values of 0.84 were given for lexical items that at least one annotator classified as an argument.

The OntoNotes, PropBank, and FrameNet results indicate the benefit of using trained annotators, and in partitioning the task so that any annotation resembling semantic role labeling is done separately. The higher values found for PropBank show the benefit of assigning only selected lexical items to annotators, rather than having the same annotator do every word of running text.

Summarizing, our work takes additional steps beyond these earlier effors, annotating multiple translations of foreign-language texts with interlingual content, thus yielding multilingually induced parallel resources and also providing new agreement evaluation metrics.

3 Corpus

The target data set is modeled on, and is an extension of, the DARPA MT Evaluation data set (White and O'Connell 1994). It consists of six 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 six 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 150,000 and 200,000 words. Consequently, the size of the entire data set is around 1,000,000 words, including all translations.

As for corpus construction, the Spanish, French, and Japanese corpora are those compiled by White and O'Connell for the DARPA MT evaluations. A third translation, and in a few cases, a second and first translation, were provided by independent professional translators contracted by NMSU, MITRE, and CMU respectively. The Hindi corpus was compiled by Columbia, and consists of news reports in Hindi (Unicode) along with two human translations of each into English by different professional translation agencies. The initial Korean corpus of ten newspaper articles along with two careful manual translations for each was constructed at ISI. Finally, the Arabic corpus is a subset of the LDC's "Multiple-Translation Arabic (MTA) Part 1" (Walker *et al.* 2003) that has been selected to reflect the domain of focus (economics) and appropriate article length.

It is important to point out that there is a difference between monolingual annotation and the approach we are taking here, where multiple translations of the same source text are annotated. The translations ideally communicate the same information conveyed by the original source language text and, therefore, when they differ one of three states of affairs must hold:

(1) One translator simply made a target language error (a typographical mistake, a careless rendering of content, e.g., rendering *so high* as *so low*).

(2) The translators have used paraphrases – communicating essentially the same thing in different words – but the underlying (interlingual) representation would be the same.
(3) The translators are expressing differing but valid interpretations of the source language text.

It is an inevitable result of translation that target language texts may be at times more explicit and at times less explicit with respect to information content than the original source language text and, when both translators are more explicit, they may differ with respect to that information content because they might bring differing knowledge to bear and choose different terms and therefore express different information. In such cases, the translations should *not* have the same interlingual representation nor should either be the same as the source language interlingual representation. In any case, the process begins by comparing translations (which *a priori* should be the same) and identifying and categorizing each parallel constituent that differs as an error, a paraphrase or a case of differing interpretations. That process necessarily requires taking into consideration the original source language text. The annotation proceeds once the state of affairs is clear.

Although related to the problem of looking for monolingual paraphrases, this annotation process is different in that the translations are supposed to communicate the same information in the same way and, when they differ, the reason must be determined, with one possibility being that they are indeed paraphrases. For those who have attempted to establish paraphrastic relations on the basis of multiple (English) versions of classic literature, e.g., Don Quixote or the Bible, the activity is the same except that, to our knowledge, those studies have discounted translator error and more importantly differences in translator interpretation.

Consider 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 process is to identify the variations between the translations and then assess whether these differences are significant. In this case, the translations are the same for the most part 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's choice potentially represents an elaboration of the semantic content of the source expression and the question arises as to whether the annotation of these variants 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 concrete 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 a bias of a certain degree of strength and the second to a bias in a certain direction. Seemingly, then, the annotation of each of these expressions should differ.

As noted in Helmreich and Farwell (1998) and Farwell and Helmreich (1999), roughly 40% of the translation units for two conservative translations of the same text into the same language differ.[5] This increases when considering three translations. Of these differences only 15% are due to translator error, while 41% are paraphrases having a common interpretation and 44% reflect differing interpretations of the original text. It is specifically the need to classify and explain such variations across translations in order to annotate them that makes this multilingually induced parallel data set so unique and enriching.

We will see in Section 6 that the annotation effort for any given corpus involves assigning interlingual content to the source text and at least two parallel English versions of each source-language text.

4 Interlingual representation

Due to the complexity of an interlingual annotation as indicated by the differences described in the sections above, the IAMTC project focused on annotation of interlingual content at the first two levels of our the-tiered interlingual representation. Each level of representation incorporates meaning from a concept ontology (including both abstract concepts and specific word senses) and thematic information while removing existing syntactic information. We describe the three levels of representation and then present our approach to annotation of the first two levels.

4.1 Three levels of representation

In the IAMTC project, three levels of representation were developed, referred to as IL0, IL1, and IL2. These levels lie on a spectrum from the annotation of syntactic dependency structure to the representation of the meanings of interpretations; each higher-order level builds on the previous level.

IL0 is a deep syntactic dependency representation, constructed by hand-correcting the output of a dependency parser based on Connexor (`www.connexor.com`). Although this representation appears to be purely syntactic – similar in flavor to the Deep Syntactic Structures (DSyntS) of Meaning-Text Theory (Mel'čuk 1988) – it abstracts as much as possible away from surface-syntactic phenomena. For example, auxiliary verbs and other function words are removed from IL0. In addition, corresponding active and passive voice sentences receive the same representation in

[5] A *translation unit* may be as small as morphological unit (e.g., *pre-*) or as large as a multiword name or phrase (e.g., *United Arab Emirates*).

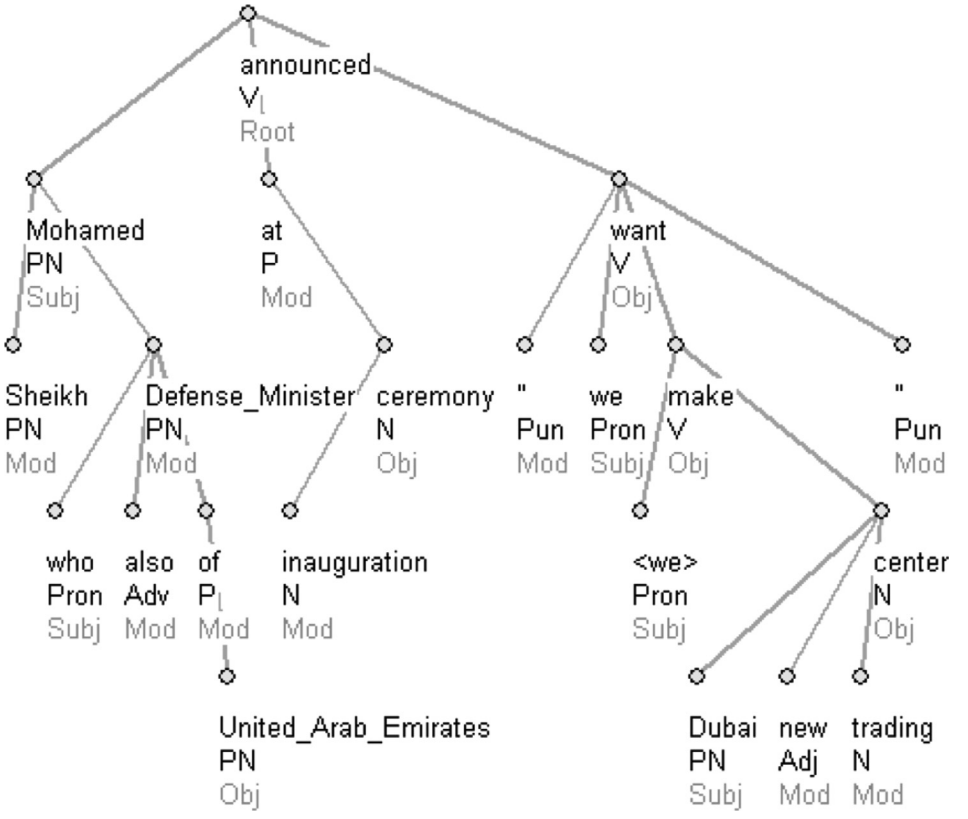

Fig. 1. An IL0 dependency tree.

IL0. For a discussion of the development of the annotation manuals for IL0 for the different languages, see Rambow *et al.* (2006).

In comparison to the annotations used in the Prague Dependency Treebank (Böhmová *et al.* 2003), IL0 is more abstract than the Analytic layer (which consists of superficial syntactic annotation) but more syntactic than the Tectogrammatical level (which aims to describe the linguistic meaning of a sentence). IL0 is a useful starting point for IL1 in that syntactic dependencies are often indicative of semantic dependencies. Figure 1 shows the IL0 representation for the sentence *Sheikh Mohamed, who is also the Defense Minister of the United Arab Emirates, announced at the inauguration ceremony that "we want to make Dubai a new trading center."*

IL1 is an intermediate semantic representation. Open class lexical items (nouns, verbs, adjectives, and adverbs) are associated with concepts drawn from the Omega ontology (Hovy *et al.* 2003b). Also at this stage, syntactic relations are replaced by semantic roles such as AGENT, THEME, and GOAL. However, IL1 is not a complete interlingua; it does not normalize over all linguistic realizations of the same semantics. Figure 2 shows the IL1 corresponding to the IL0 in Figure 1. Concept names (e.g., DECLARE) and thematic role names (e.g., AGENT) added by the annotators are in upper case. WordNet senses – encoded as concepts (rather

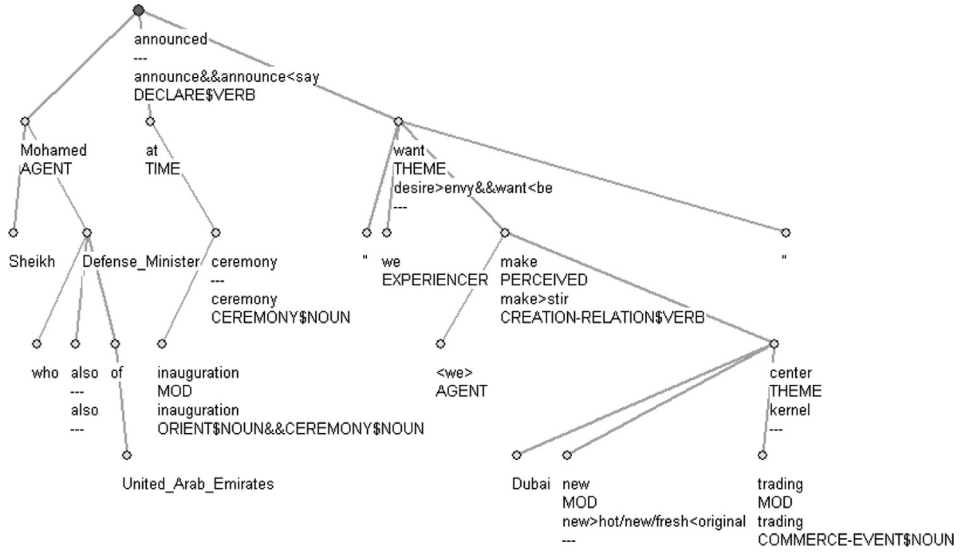

Fig. 2. An IL1 representation.

than synonym sets) – are indicated in lower case (e.g., announce<say). Some lexical items are associated with more than one concept, separated by two ampersands (&&). IL0 and IL1 have been documented with coding manuals and have been used by annotators to tag several texts (see Section 5.2).

The methodology for designing IL2 involves comparison of IL1s in the multi-parallel corpus in order to see how meaning equivalent IL1s can be reconciled or merged. IL2 is expected to normalize over:

- Conversives (e.g., *X bought a book from Y* versus *Y sold a book to X*), as does FrameNet (Baker *et al.* 1998) at the more general level of Commercial_transaction.
- Nonliteral language usage (e.g., *X started its business* versus *X opened its doors to customers*).
- Extended paraphrases involving syntax, lexicon, and grammatical features (see example in Section 3).

Figure 3 illustrates the relationship between IL1 and IL2 for the two sentences *Mary bought a book from John* and *John sold a book to Mary*. The IL1s for the two sentences are different because the verbs *buy* and *sell* use different participants as agents. However, the IL2 representation captures the common meaning of the buying and selling events, as has been suggested by many theories of meaning representation. This is not to say that *buy* and *sell* are identical (equivalent) lexical items, but that each provides a different perspective on the same event. The IL2 representation is intended to capture an underlying generalization, independent of the perspective that is chosen by a given language speaker/writer at the time that the source-language utterance/text is produced.

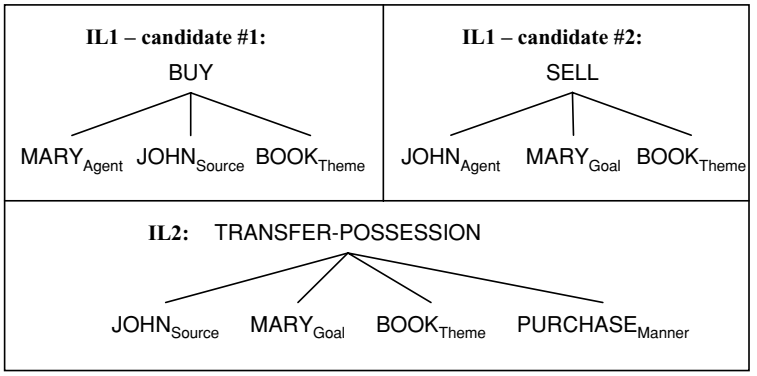

Fig. 3. IL1 and IL2 for conversives.

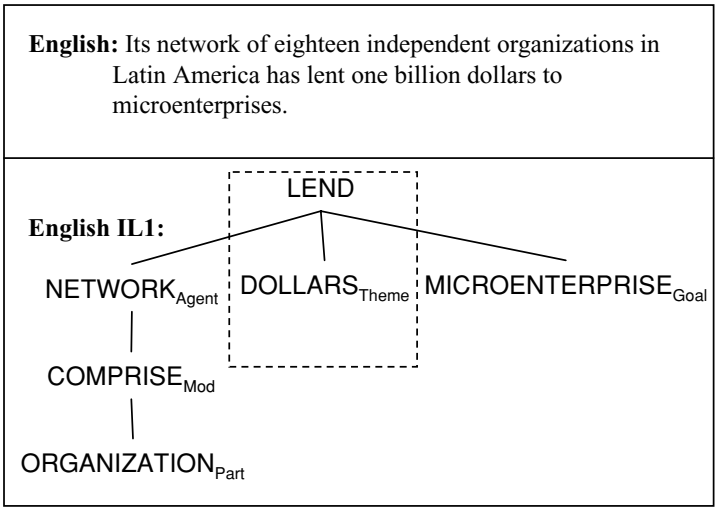

Fig. 4. English sentence and IL1.

Consider the following English/French extended paraphrases:

E: Its network of eighteen independent organizations in Latin America has lent one billion dollars to microenterprises.

F: Le réseau regroupe dix-huit organisations indépendantes qui ont déboursé un milliard de dollars.
'The network comprises eighteen independent organizations which have disbursed a billion dollars'

These sentences are taken from the January 1997 edition of the UNESCO Courier, which is available in twenty-nine languages. Figures 4 and 5 show the corresponding IL1 representations for these sentences. Note that the head of the English IL1 is the concept LEND, whereas the head of the French IL1 is the concept COMPRISE.

Figure 6 sketches the mapping rules that are needed to reconcile the IL1's from Figures 4 and 5 in order to produce the IL2 representation. The words *of* and *regrouper* are found to express the concept COMPRISE. The argument ORGANIZATION

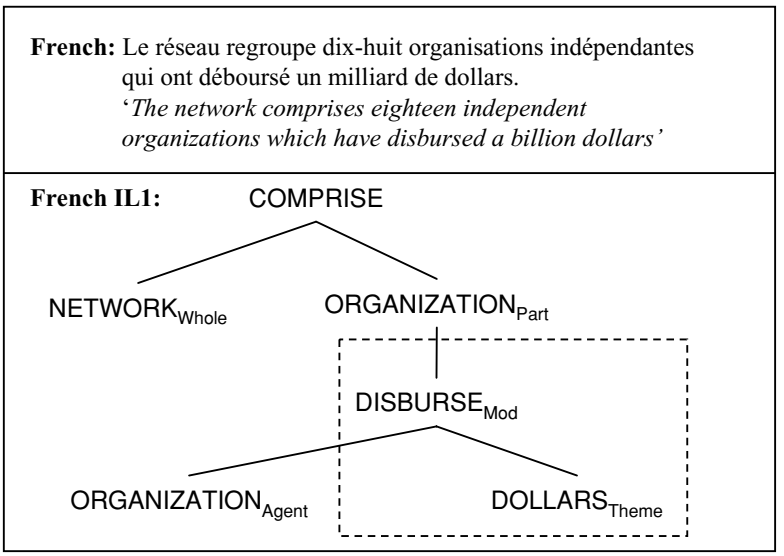

Fig. 5. French sentence and IL1.

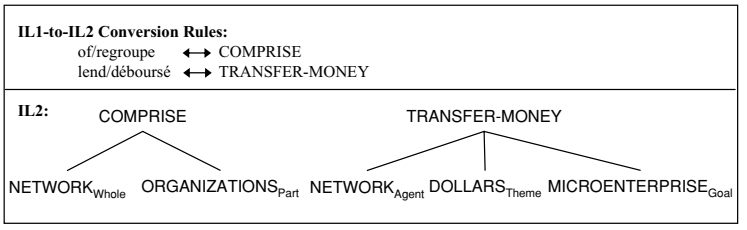

Fig. 6. IL1-to-IL2 conversion rules and resulting IL2 representation.

associated with both words confirms that *of* and *regrouper* describe the same relation. Similarly, the concept TRANSFER-MONEY is identified as a common concept for the words *lend* and *debourser*, which share two arguments, NETWORK and DOLLARS.

Note that the two components of IL2 are neither linearly nor hierarchically ordered, as the notion of ordering is language-dependent and idiosyncratic. Specifically, the "relative clause" rendering of the French *déboursé un milliard de dollars* and the "main clause" rendering of the English *lent one billion dollars* – corresponding to the IL1 components that are enclosed in dotted lines in Figures 4 and 5 – are language-dependent decisions that need not be captured in the IL2. In effect, the syntactic realizations associated with components of meaning are factored out in IL2, leaving behind the semantic notion of TRANSFER-MONEY. Similarly, the "main clause" rendering of the French *Le réseau regroupe dix-huit organisations indépendantes* and the "subject nominal phrase" rendering of the English *Its network of eighteen independent organizations* are language-dependent decisions that are ignored in the IL2 notion of COMPRISE.

We clarify further that, although there are conversion rules for mapping from the lexical units associated with IL1 to the semantic units in IL2, there is no direct mapping between the English IL1 and French IL1 (or vice versa). This means that there is no notion of restructuring a subordinate clause as a matrix clause, as one might find in a transfer system. Rather, the IL2 is intended to be a language-independent representation over which language-dependent mappings may be applied, but this mapping is outside of the scope of our work, which is focused primarily on the IL1 and its relation to IL2.

The range of paraphrase phenomena addressed by the different representation levels is summarized in Table 1, which is based on examples from (Hirst 2003), (Kozlowski, McCoy and Vijay-Shanker 2003) and (Rinaldi *et al.* 2003). The table indicates for which types our normalized representations would reflect the similarity in meaning between paraphrases of that type and at which level the normalization would take place.

As mentioned above, our evaluation studies focus on the first two levels (IL0 and IL1) of our three-tiered interlingual representation. It is not our intention to argue for the superiority of the meaning elements used in these representations (or, for that matter, in IL2) over those used in other interlingual representations. Our goal is to build a solid foundation for future investigation of the types of IL1–IL2 conversion rules necessary for generating deeper interlingua on a large scale. We expect that such rules would eventually be induced automatically using machine learning techniques that are similar to those that have been applied to problems involving other types of natural language equivalences, e.g., paraphrase extraction (Barzilay and Lee 2003; Pang *et al.* 2003; Dolan *et al.* 2004; Bannard and Callison-Burch 2005; Callison-Burch *et al.* 2006; Madnani *et al.* 2007).

Specifically, the focus of this study is on demonstrating the feasibility of the annotation process and, upon analysis of our results, arriving at a better understanding of the types of representational mappings required for the development of deeper semantic representations. Our view is that human annotation of IL0 and IL1 – coupled with the use of multiple translations of textual documents from different languages – is an important step toward discovering what would be needed for the IL2 representation. Additionally, we expect that our study will shed light on the types of algorithms that might be required to derive such representations automatically.

4.2 The Omega ontology

As mentioned in Section 4.1, the IL0 representation is semi-automatically produced through the process of applying Connexor, followed by manual corrections. To progress from IL0 to IL1, annotators selected semantic terms (concepts that represent particular types of objects, events, and states) for the nouns, verbs, adjectives, and adverbs in each sentence. These terms were selected from a 110,000-node ontology called Omega (Philpot, Fleischman and Hovy 2003) that includes a normalized form of Mikrokosmos concepts (Mahesh and Nirenburg 1995) integrated with most WordNet senses (Fellbaum 1998). (We henceforth refer to the WordNet component

Table 1. *Relationship types underlying paraphrase*

Relationship type	Example	Where normalized
Syntactic variation	The gangster killed at least three innocent bystanders. *versus* At least three innocent bystanders were killed by the gangster.	IL0
Lexical synonymy	The toddler sobbed, and he attempted to console her. *versus* The baby wailed, and he tried to comfort her.	IL1
Morphological derivation	I was surprised that he destroyed the old house. *versus* I was surprised by his destruction of the old house.	IL2
Clause subordination versus anaphorically linked sentences	This is Joe's new car, which he bought in New York. *versus* This is Joe's new car. He bought it in New York.	IL2
Different argument realizations	Bob enjoys playing with his kids. *versus* Playing with his kids pleases Bob.	IL2
Noun–noun phrases	She loves velvet dresses. *versus* She loves dresses made of velvet.	IL2
Head switching	Mike Mussina excels at pitching. *versus* Mike Mussina pitches well. *versus* Mike Mussina is a good pitcher.	IL2
Overlapping meanings	Lindbergh flew across the Atlantic Ocean. *versus* Lindbergh crossed the Atlantic Ocean by plane.	IL2
Comparatives versus superlatives	He's smarter than everybody else. *versus* He's the smartest one.	Not normalized
Different sentence types	Who composed the Brandenburg Concertos? *versus* Tell me who composed the Brandenburg Concertos.	Not normalized
Inverse relationship	Only 20% of the participants arrived on time. *versus* Most of the participants arrived late.	Not normalized
Inference	The tight end caught the ball in the end zone. *versus* The tight end scored a touchdown.	Not normalized
Viewpoint variation	The U.S.-led invasion/liberation/occupation of Iraq . . . You're getting in the way. *versus* I'm only trying to help.	Not normalized

of Omega as "WordNet concepts" and the Mikrokosmos component of Omega as "Omega concepts.") As an example of the information contained in Omega, one of the entries for the event-type LOAD is shown in Table 2.

Table 2. *Omega information event-type LOAD*

Definition	WordNet concept	Mikrokosmos concept
Fill or place a load in	load>pack	DISPOSITIVE-MATERIAL-ACTION

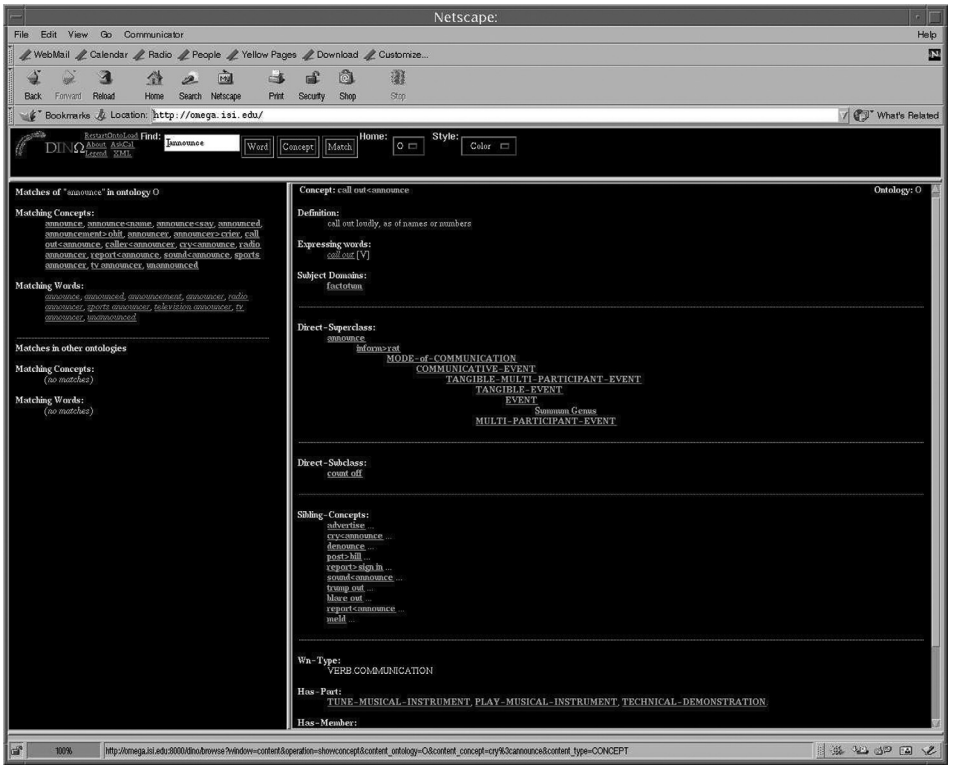

Fig. 7. Omega ontology browser.

Omega was assembled semi-automatically by researchers at ISI using a combination of Princeton's WordNet, New Mexico State University's Mikrokosmos, 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, Echihabi and Hovy 2003). The ontology, which has been used in several projects in recent years (Hovy, Marcus and Weischedel 2003a; Hovy *et al.* 2003c; Philpot, Hovy and Pantel 2005), can be browsed using the Omega ontology browsers at http://omega.isi.edu/doc/browsers.html. The earliest version of these browsers forms a part of the annotation environment for the IAMTC project. Figure 7 illustrates the Omega browser.

Table 3. *Theta roles in grid for event-type LOAD*

Role	Description	Type
Agent	The entity that does the action	OBLIGATORY
Theme	The entity that is worked on	OBLIGATORY
Possessed	The entity controlled or owned	OPTIONAL

In theory, Omega, WordNet (all WordNets), and Mikrokosmos are language-independent systems for representing word meaning (i.e., words in any language). For example, Mikrokosmos was developed from the outset for machine translation and used to define Spanish and French word meanings as well as English word meanings. WordNet has come a long way with respect to "neutralizing" its English bias given the some thirty different languages for which WordNets have been, or are being, developed. It is useful to remember that we are interested in WordNet synsets as semantic units, as defined by the synset gloss, without regard to which words from which languages might be found in the synset. Since these ontologies contribute symbols for constructing representations of *meaning*, not *words*, the source may be any given language as long as the symbols can be used for constructing coherent and consistent reasoning systems.

In later work, a new version of Omega was created as part of the OntoNotes project (Hovy *et al.* 2006; Pradhan *et al.* 2007). In this version, the new word senses, created for the OntoNotes corpus annotation and validated by achievement of 90% interannotator agreement, were clustered into semantically coherent groups and subordinated to a new version of the Omega Upper Model. These clusters of word senses (called 'sense pools' in Omega, and corresponding to WordNet's synsets) include senses drawn from English, Chinese, and Arabic nouns and verbs.

4.3 Theta grids and theta domains

In addition to Mikrokosmos concepts and WordNet concepts, annotators assigned one or more *theta grids* – groups of thematic roles – specifying the argument structure associated with event-types. The thematic roles are abstractions of deep semantic relations that generalize over event-types. While they are by far the most common approach in the field to representing predicate-argument structure, there are numerous variations with little agreement even on terminology (Fillmore 1968; Jackendoff 1972; Stowell 1981; Levin and Rappaport-Hovav 1998).

The theta grids used in the IAMTC project were extracted from the Lexical Conceptual Structure Verb Database (LVD) (Dorr *et al.* 2001). WordNet senses associated with each LVD entry were used to link the theta grids to the WordNet concepts in the Omega ontology. Theta grids included syntactic realization information for English, such as Subject, Object or Prepositional Phrase, and the Obligatory/Optional nature of these complements. For example, one of the theta grids for the event-type LOAD is listed in Table 3.

Although based on research in LCS-based MT (Dorr 1993; Habash, Dorr and Traum 2003), the set of theta roles used for this project was simplified. Table 4 shows the list of roles used in the Interlingua Annotation Experiment (Habash and Dorr 2003).[6]

For verbs, annotators were required to choose a sense and a theta grid, which constitutes a *theta-domain*. Linguists use the term theta-domain to refer to the verb-phrase complex consisting of a verb and its core grammatical arguments, realized semantically as thematic roles. We use the same term to refer to a specific sense of a verb and its thematic roles, or theta grid. Note that the same theta grid can apply to concepts associated with different verbs, thus the grid *Agent,Theme,Goal* applies to concepts associated with *donate* and *send*: "John donated (sent) a thousand dollars to his favorite museum." The same verb can have distinct theta grids associated with different senses, thus *Source,Theme,Proposition* applies to a different sense of *send*: "Back to the Future is a motion-simulator ride that sends you rocketing through avalanches, molten volcanoes and the jaws of a dinosaur."

5 Annotation processes

Recognizing the complexity of interlinguas, we adopted an incrementally deepening approach, which allowed us to produce some quantity of relatively stable annotations while exploring alternatives at the next level down. Throughout, we made as much use of automated procedures as possible. This section presents the tools and resources for the interlingual-annotation process and describes our annotation methodology.

5.1 Supporting tools and resources

We assembled and/or built a suite of tools to be used in the annotation process. Since our corpora were extracted from disparate sources, we had to standardize the text before presenting it to automated procedures. For English, this involved sentence boundary detection, but for other languages, it involved word segmentation, chunking of text, demorphing, or similar language-specific operations.

The text was processed using a dependency parser. For English, we use Connexor (Tapanainen and Jarvinen 1997). The Connexor output was converted into a form that was compatible with a visualization tool called TrEd (Hajič, Vidová-Hladká and Pajas 2001). This graphically based tree editing program, written in Perl/Tk, was used for viewing and hand-correcting the output of the dependency parser.[7] Figure 8 illustrates the TrEd Tree Editor.

The hand-corrected deep dependency structure produced by this process became the IL0 representation for that sentence. Already at this stage, some of the lexical

[6] Other contributors to this list are Dan Gildea and Karin Kipper Schuler.
[7] http://quest.ms.cuni.mff.cz/pdt/Tools/Tree_Editors/Tred/

Table 4. *List of Theta Roles with Examples*

Agent: Agents have the features of volition, sentience, causation and independent existence.
Examples: *Henry* pushed/broke the vase.

Instrument: Instruments have causation but no volition. Their sentience and existence are not relevant.
Examples: *The hammer* broke the vase; She hit him *with a baseball bat*.

Experiencer: Experiencers have no causation but are sentient and exist independently. Typically experiencers are realized as the subjects of verbs like *feel*, *hear*, *see*, *sense*, *smell*, *notice*, *detect*, etc. in English.
Examples: - **John** heard the vase shatter; **John** shivered.

Theme: Themes are typically causally affected or experience a movement and/or change in state. Themes appear as the complement of verbs like *acquire*, *learn*, *memorize*, *read*, *study*, etc. in English.
Examples: *John* went to school; John broke *the vase*; John memorized *his lines*; She buttered the *bread* with margarine.

Perceived: Perceived entities are the cause of what an experiencer experiences. They do not experience a movement or change in state. Their volition and sentience are irrelevant. Their existence is independent of an experiencer.
Examples: He saw *the play*; He looked *into the room*; *The cat's fur* feels good to John; She imagined *the movie* to be loud.

Predicate: Predicates are new modifying information about other thematic roles.
Examples: We considered him *a fool*; She acted *happy*

Source: Sources are where/when a theme starts its motion, or what its original state is, or where its original (possibly abstract) location/time is.
Examples: John left *the house*

Goal: Goals are where a theme ends 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 which results from an activity.
Examples: John ran *home*; John ran *to the store*; John gave a book *to Mary*; John gave *Mary* a book

Location: Locations are places where theme are or events take place- –as opposed to a source or goal.
Examples: He lived *in France*; The water fills *the box*; *This cabin* sleeps five people

Time: Times are periods of time or moments when events take place.
Examples: John sleeps *for five hours*; Mary ate *during the meeting*

Beneficiary: Beneficiaries are those that receive the benefit/result of the event/state.
Examples: John baked the cake *for Mary*; John baked *Mary* a cake; An accident happened to *him*

Purpose: Purposes are the intended outcome of an event/state.
Examples: He studied *for the exam*; He searched *for rabbits.*

Possessed: Possessed entities are those objects in someone's possession. They are the complements of verbs such as *own*, *have*, *possess*, *fit*, *buy*, and *carry* in English.
Examples: John has *five bucks*; He loaded the cart *with hay*; He bought it *for five dollars*

Proposition: Propositions are secondary events/states.
Examples: He wanted *to study for the exam*

Modifier: Modifiers are properties of things related to color, taste, size, etc.
Examples: The red book *sitting on the table* is old

Null: Null elements have no thematic contribution. Typical examples are impersonal *it* and *there* in English.
Examples: *It* was raining all morning in Miami

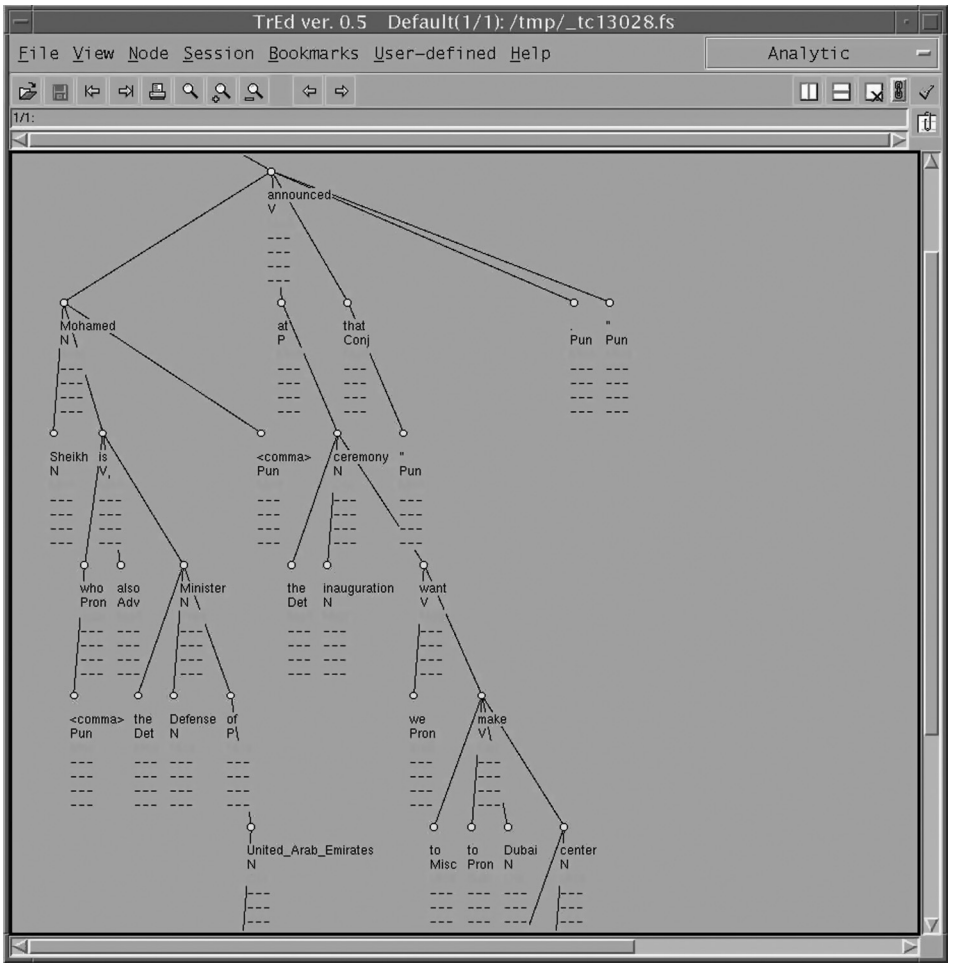

Fig. 8. TrEd tree editor.

items were replaced by features (e.g., tense), morphological forms were replaced by features on the citation form, and certain constructions were regularized (e.g., passive) with empty arguments inserted.

In order to derive IL1 from the IL0 representation, annotators used Tiamat, a tool developed specifically for this project. This tool displays background resources and information, including the IL0 tree and the Omega ontology. Drill-down into the ontology could be achieved, as needed, through the Omega browser shown earlier in Figure 7.

Principally, Tiamat was the annotator's workspace, showing the current sentence, the current word(s) to be annotated, the ontology's options for annotation, including theta roles (already connected to other parts of the sentence, as far as possible), etc. It provided the ability to annotate text via simple point-and-click selections of words, concepts, and theta roles. Figure 9 illustrates the Tiamat Annotation Interface.

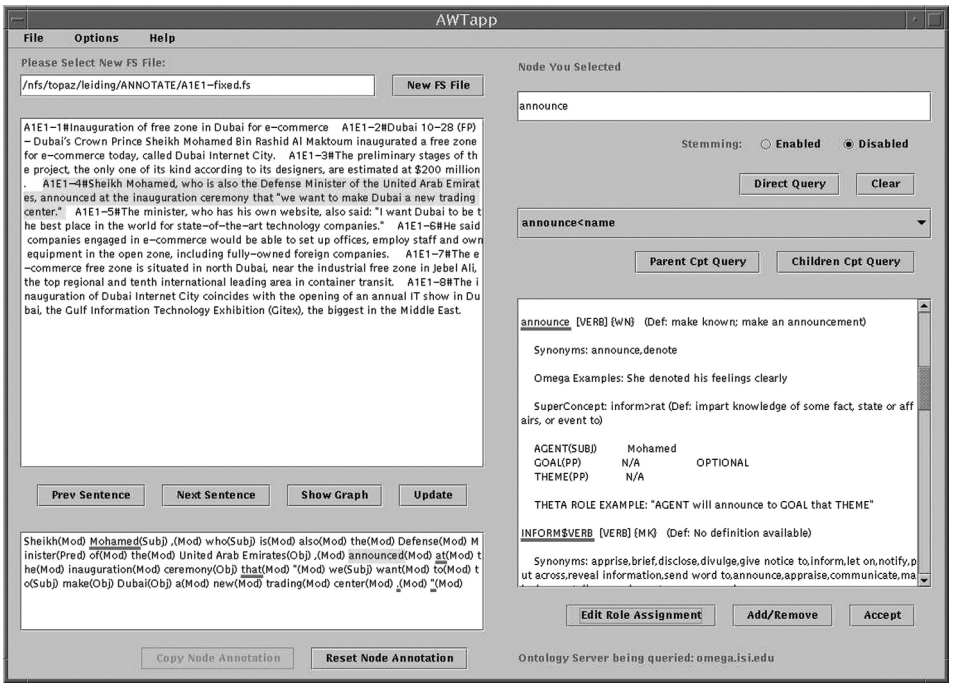

Fig. 9. Tiamat annotation interface.

Evaluation of the annotators' output would be daunting based solely on a visual inspection of the annotated IL1 files. Thus, an annotation agreement evaluation tool was also developed to compare the output and to generate various evaluation measures (see Section 5.2 below).

Appendix 7 provides an example of the internal representation resulting from the annotation of the sentence presented earlier in Figure 2 using the Tiamat Annotation Interface.

5.2 Annotation methodology

For the purpose of this discussion, we will restrict our attention to the annotations provided for six pairs of English translations, one pair per source language.[8] The English translations are referred to by their source- and target-language version numbers, e.g., A1E2 refers to the second English translation of the first Arabic text. The 12 texts considered in this study are: A1E1, A1E2, F2E1, F2E2, H2E1, H2E2, J2E1, J2E2, K1E1, K1E2, S1E1, and S1E2.

To derive IL1 from IL0, the annotators were instructed to annotate all nouns, verbs, adjectives, and adverbs with Mikrokosmos concepts, WordNet concepts, and theta grids. In order to determine an appropriate level of representational specificity

[8] We also annotated the source-language versions of some of the texts, but our inter-annotator reliability study focused on the annotation of English document pairs with the highest number of human annotations.

in the ontology, annotators were instructed to annotate each word twice – once with one or more WordNet concepts and once with one or more Mikrokosmos concepts; these two units of information are merged, or at least intertwined, in Omega.

Problem cases were flagged and brought to the attention of a guru (one of the PIs), who worked with the annotators to ensure that the guidelines adequately conveyed the approach to be taken, e.g., in ambiguous cases. Where necessary, annotators determined the set of roles or altered them to suit the text. Revised or new theta roles were also automatically routed to another guru for inspection. Annotators were also instructed to provide a thematic 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.

A novelty of this annotation procedure is that it allows multiple senses to be assigned to a word. Consider the word *proposed* in:

E: The Czech Minister of Transportation, Jan Strasky, proposed on Saturday that the State buy back the shares held by Air France.

Omega offers concepts for *propose* that are defined as *declare a plan*, *present for consideration* and *intend*. It can be argued that forcing the annotators to select only one of these senses will result in the loss of some aspect of the meaning of *propose* in this context. Indeed, one important question addressed in this project is that of whether multiple annotators can agree on finer grained sense distinctions like these (as opposed to coarse distinctions, e.g., the sense *ask to marry*) if they are allowed to select multiple concepts. (This point will be discussed further in Section 6.)

Markup instructions are contained in three manuals: a manual for creating IL0, a users' guide for Tiamat (including procedural instructions), and a definitional guide to semantic roles. Together these manuals allowed the annotator to understand (1) 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. All these manuals are available at http://aitc.aitcnet.org/nsf/iamtc/tools.html.

All annotators at all sites worked on the same texts. Every week, IL0 representations for two translations of each of two texts were created, with each site being responsible for IL0 creation for one of the six languages (including translations). These then served as the basis for IL1 creation over a three-week period, with 2 annotators at each site. Our goal was for annotators to produce 144 IL1-annotated texts (6 texts $\times$ 2 translations $\times$ 2 annotators $\times$ 6 sites). However, practical issues (e.g., availability of annotators, coordination of annotations across sites, etc.) constrained our study to a smaller number of annotators (11 instead of 12) and fewer annotators per document: 7 for one document (A1E1), ten for seven documents (A1E2, K1E1, K1E2, J2E1, J2E2, F2E1, and F2E2), and nine for four documents (S1E1, S1E2, H2E1, and H2E2), for a total of 113 annotated texts.

Each text annotation took about three hours. 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 other first. Half the sites annotated a full text at a time, and the other half annotated them sentence-interleaved (similar sentences ordered consecutively).

In a practice period leading up to the final annotation phase, annotators at each site were encouraged to discuss their annotations with each other and with researchers at their site. Intermittent conference calls allowed annotators to bring up various types of annotation problems and discuss alternative solution strategies. During the final annotation phase, such calls were restricted to consideration of completed texts, and adjustments to completed annotations were not permitted.

6 Evaluation of IL1 annotations

The annotation we have described here forges new ground; thus, we have taken an exploratory approach to its evaluation by looking at multiple cross sections of the data, and using multiple analytic tools. With the exception of early work with WordNet (Fellbaum *et al.* 1998), there has been little previous work on corpus annotation where all lexical items were annotated in context from an ontology, or of theta-role annotation in corpora; thus, we had little prior knowledge to shape our expectations about what annotators would do. In addition, the choices available for theta-role annotation were dependent on the annotators' choices for concept annotation, but we did not anticipate the degree to which this would reduce the number of cases all annotators considered in common. We report separate results for WordNet concepts, Mikrokosmos concepts, and theta roles.

The questions we address are how to quantify agreement given the prevalence of set-valued annotation categories; how to rate individual annotators in the absence of a ground truth or other standard to measure them against; and how to partition the IL1 data so as to discover which aspects of the annotation were more difficult for annotators, and conversely, which subsets of the data are the most reliable.

The next section describes our approach to partitioning data and our choice of agreement coefficients. Following this, we describe two agreement coefficients and the differences between them and introduce a new distance metric called MASI (for Measuring Agreement for Set-valued Items). We then overview the scope of the annotation task. Finally, we present overall interannotator agreement results, reliability scores for different parts of speech, and interannotator reliability for various combinations of annotators.

6.1 Data partitioning and choice of agreement coefficients

In addition to reporting separate results for WordNet concepts, Mikrokosmos concepts, and theta roles, we partition the annotation data in a variety of ways due to the highly distributed nature of the data collection, and the large number of annotators. To avoid obscuring site-dependent differences, we performed separate reliability assessments on the twelve datasets corresponding to the two translations from each of the six source languages. The differences across sites became particularly evident for theta-role annotation.

Note that each dataset has a different number of degrees of freedom, dependent in part on the number of lexical items to annotate and the number of annotators. The number of lexical items per dataset ranged from 76 (Hindi translation H2E2) to 136 (French translation F2E2). By and large, a pair of translations of the same source had roughly the same number of lexical items to annotate. However, this was more true for the two translations of Hindi (H2E1 with 77 and H2E2 with 76) and French (F2E1 with 130 and F2E2 with 136) than for the two translations of Arabic (A1E1 with 80 and A1E2 with 97), Korean (K1E1 with 92 and K1E2 with 112), and Spanish (S1E1 with 116 and S1E2 with 124).

The different sites were responsible for different subsets of the data (different source languages), and for training and supervising different annotators: CMU was responsible for Japanese, Columbia for Hindi, ISI for Korean, MITRE for French, NMSU for Spanish, and UMD for Arabic. There were eleven annotators who participated, with no more than ten and no fewer than seven who annotated the same dataset, meaning the same set of dependency parses from the same translation.

Agreement coefficients used in computational linguistics, such as Cohen's κ (Cohen 1960), Siegel and Castellan's K (Siegel and Castellan 1988), Scott's π (Scott 1955), Krippendorff's α (Krippendorff 1980), and so on, all compare the observed agreements among annotators to the amount of agreement that would be expected if annotators made random selections from the annotation categories. However, the coefficients differ regarding the computation of chance agreement, due to different assumptions about how to estimate the likelihood that an annotator would select a given annotation value, given i instances, j annotators and k annotation values. One common assumption, represented by Krippendorff's α and Scott's π, for example, is that there is a single probability distribution for the annotation values that can be estimated from the proportion of each $k \in \mathcal{K}$ in the entire dataset of $i \times j$ assignments. An alternative assumption, as represented by Cohen's κ (Cohen 1960), is that for each annotator there is a distinct probability distribution for the annotation values, estimated from the proportion of each value k in the i assignments made by annotator j. The latter type can detect differences in bias across annotators, so for ease of reference, we refer to the two types as bias-free and bias-sensitive. Each type has strengths and weaknesses, as we note below in a separate section.

We require an agreement coefficient that handles multiple annotators, and as discussed below, we require a weighted coefficient to handle the set-valued annotations. Both Krippendorff's α and κ^3 (Artstein and Poesio 2005b) fit the bill. The latter is an extension of Cohen's weighted κ (Cohen 1968) to accommodate numbers of annotators greater than two. Like κ, it is bias-sensitive. We began by using both coefficients, primarily to illustrate that the quantitative differences are slight. We report both metrics in our initial results on the twelve datasets for the concept annotation tasks. After showing that the two metrics yield virtually the same results, we report the more fine-grained results using α alone. In part, this is because the treatment of missing values, which turns out to be important in this dataset, is a much simpler computation for α than for κ^3.

One of the more specific questions we address is how to compare annotations when annotators are free to choose multiple categories if they cannot decide between

them, or if they feel a single choice is insufficient. For a typical annotation task in computational linguistics, annotators choose a single label from a predetermined set of labels. The comparison of the choice made by one annotator for a given item with the choice made by another annotator is made on a categorical basis, meaning the annotators agree if they have chosen the same label and disagree otherwise. Such data is referred to as nominal data. However, some tasks ask annotators to choose labels from an interval scale (e.g., a Likert scale), or other type of data scale. Both κ and α can be weighted in order to assign partial agreemment or disagreement among data values using the appropriate scale, but the scales discussed in previous literature do not include set-valued data.

Elsewhere, we have proposed to scale the comparison of set-based values (Passonneau 2004, 2006, 2010; Passonneau *et al.* 2005; Passonneau, Habash and Rambow 2006). A weighting, or distance metric was designed to handle coreference annotation and it was argued that this same weighting, referred to as MASI (Measuring Agreement for Set-valued Items) is appropriate for a wide range of semantic and pragmatic annotations, including the IAMTC data (cf. discussion of MASI in Artstein and Poesio 2008). Here we have expanded the analysis of the IL1 data presented in (Passonneau *et al.* 2006) in two ways. We have added the computation of agreement using κ^3: when compared with α values this allows us to determine whether annotators have different biases. Another change is that here we factor out missing values.

6.2 Methods

This section gives a general explanation of agreement coefficients, describes the interpretation of agreement results, and defines the MASI distance metric.

6.2.1 Background on bias-free and bias-sensitive agreement coefficients

Annotation data can be tabulated in an i by j matrix that represents all decisions made by all annotators. If the i rows represent the i annotators, and the j columns represent the j items being annotated, then each cell (i, j) contains the kth value (or category) that the ith annotator chose for the jth unit. Table 5 gives a hypothetical example of concepts selected for four lexical items by five annotators. There are twenty cells in the table, and ten unique values. The observed agreement is calculated by counting all observed pairs of values within each column that agree, and dividing by the total number of such pairs. In column one, for example, Ann1 agrees with Ann2 and Ann5, Ann2 agrees with Ann1 and Ann5, and so on. In any one column, each annotator can be compared with four other annotators, yielding twenty pairs per column, or eighty for the table. Of these eighty pairs, thirty agree, so the observed agreement is $\frac{30}{80}$, or .375, and the observed disagreement is therefore .625. An agreement coefficient factors out how much of the observed agreement (or disagreement) would be given by a chance distribution, given some assumption about the probability of each annotation category.

Table 5. *A sample reliability matrix with five annotators, four items, ten values*

	1. Cost	2. Zone	3. Launch	4. Make
Ann1	COST	ZONE<PART	FOUND<OPEN	MAKE<BECOME
Ann2	COST	ZONE>ISLAND	FOUND<OPEN	MAKE<BECOME
Ann3	MONETARY_VALUE	ZONE>ISLAND	LAUNCH<MOVE	MAKE<BECOME
Ann4	TOLL<VALUE	ZONE	LAUNCH<MOVE	MAKE<BECOME
Ann5	COST	ZONE	FOUND<OPEN	MAKE<HAVE

Agreement coefficients all use the same general formula to factor out chance agreement, whether they compare expected agreement to observed agreement, or expected disagreement to observed disagreement. Where A_O and A_E represent observed and expected agreement, the general formula for measuring agreement is:

$$\frac{A_O - A_E}{1 - A_E} \tag{1}$$

and where D_O and D_E represent expected and observed disagreement, the equivalent formula that uses disagreement is:

$$1 - \frac{D_O}{D_E} \tag{2}$$

In principle, values range from one to minus one, although data with more than two annotators and categories raises the lower bound. A value of one represents perfect agreement, zero represents no deviation from chance distribution, and minus one represents the systematic disagreement that would occur if two annotators always used the opposite of two annotation values.

When computing interannotator reliability for multiple annotators, it is necessary to calculate observed and expected agreement relative to all annotators at once in order to get correct estimates for expected agreement, rather than averaging the results for all pairwise combinations. For example, if annotators i_1, i_2, and i_3 each assign k_1 at a rate of $\frac{1}{2}$, then the likelihood that all three assign k_1 to the same instance is $\frac{1}{8}$, whereas the likelihood that any pair will do so is $\frac{1}{4}$.

Consider how one would compute the probability that an annotator will assign COST as a value, using the data in Table 5. If it is assumed that all annotators should more or less make the same choices, and any one annotator will deviate more or less from a hypothetical norm, then the proportion of COST in the entire dataset is taken as an estimate of the probability that any annotator would use COST; here COST occurs in three out of twenty cells, thus p(COST) = 0.15

Krippendorff's α is an example of what we refer to as a bias-free coefficient because it is based on the assumption that all annotators are considered to have an equal likelihood of using a particular annotation category. The first agreement coefficient to make this assumption was probably Scott's π (Scott 1955); α is equivalent to π for large samples in the case of two annotators and nominal categories. Siegel and Castellan's K makes the same assumption.

As noted above, the formula for Krippendorff's α explicitly incorporates a distance metric for comparing annotation values, which is useful for noncategorical data scales. With categorical data, values assigned by two annotators are either the same or different. If annotators i_1 and i_2 both assign k_1 for a given instance, nothing is added to the overall measure of disagreement. This is equivalent to assigning a zero value to the distance between k_1 and k_1: $d(x, y) = 0$ if $x = y$. If two annotators assign different values to the same instance, then the overall measure of disagreement is incremented, which is equivalent to assigning a distance value of 1 $d(x, y) = 1$ if $x \neq y$. Here we motivate use of a set-based distance metric designed for semantic and pragmatic annotation tasks (Passonneau 2004, 2006, 2010).

Let us take a to be the number of annotators, i the number of items, k the number of annotation categories (or values), n_{ik} the number of coders who assign item i to category k, n_{ak} the number of items assigned by annotator a to category k, and n_k the number of items assigned by all coders to category k. Let d be the distance between two assignments to item i, k_n and k_m. For categorical scales, d is 0 when $k_n = k_m$, and 1 when $k_n \neq k_m$; nothing is added to the summation of disagreements when two annotators' choices on an item i is the same. This is the unweighted distance metric which says that any disagreement is full disagreement, and any agreement is full agreement. Then formulas for computing observed disagreement (D_O) and expected disagreement (D_E) for α are:

$$D_O = \frac{1}{ia(a-1)} \sum_{i \in I} \sum_{j=1}^{k} \sum_{l=1}^{k} n_{ik_j} n_{ik_l} d_{k_j k_l} \tag{3}$$

$$D_E^{\alpha} = \frac{1}{ia(ia-1)} \sum_{j=1}^{k} \sum_{l=1}^{k} n_{k_j} n_{k_l} d_{k_j k_l} \tag{4}$$

D_O (the numerator of the term in 2) is a summation over the product of counts of all pairs of values k_n and k_m, times the distance metric d, across columns. D_E^{α} (the denominator of the term in 2) is a summation of agreements and disagreements in rows.

The formula for computing D_O for κ^3 is the same as for α (3); the formula for D_E for κ^3 is:

$$D_E^{\kappa^3} = \frac{1}{i^2 \binom{a}{2}} \sum_{j=1}^{k} \sum_{l=1}^{k} \sum_{m=1}^{a-1} \sum_{n=m+1}^{a} n_{a_m k_j} n_{a_n k_l} d_{k_j k_l} \tag{5}$$

Di Eugenio and Glass argue that a more complete picture of inter-annotator reliability is given by presenting the results for both bias-free and bias-sensitive metrics. The more alike the values are, the lower the degree of bias in the data. The more disparate the values, the greater the bias. However, as noted in Artstein and

Poesio (2005a), bias decreases as the number of annotators increases. For the data given in Table 5, $\alpha = 0.32$ and $\kappa^3 = 0.31$.[9]

For a more detailed discussion of these agreement coefficients, the underlying assumptions that affect qualitative interpretation, and examples showing a range of quantitative differences they yield, the reader is referred to di Eugenio and Glass (2004) and Artstein and Poesio (2005b, 2008).

6.2.2 Interpreting agreement results

Krippendorff (Krippendorff 1980) proposed that interannotator agreement of 0.67 supports "highly tentative and cautious conclusions," and that 0.80 represents "solid results." At the same time, he argued against "ad hoc" standards that would apply across the board and that in regard to the question of how reliable is reliable enough, "there is no set answer" (p. 146). It should be noted that Krippendorff's discussion pertains to the requirements of Content Analysis. As discussed in Artstein and Poesio (2008), varying agreement thresholds have been proposed in different disciplines; for computational linguistics, they "doubt that a single cutoff point is appropriate for all purposes." Passonneau (Passonneau 2010) takes this point a step further to argue that more research is needed regarding the impact of different agreement threshholds for different situations. She recommends that interannotator agreement results be reported along with the results of any independent uses of the same data. For example, in an assessment of the Pyramid method, a content analysis method used to evaluate summarization systems, Passonneau reports interannotator agreement results on pyramid annotations from pairs of annotators, combined with analysis of variance (ANOVA) of the performance of sixteen summarization systems derived from the two sets of annotations. Interannotator agreement on five pyramids ranges from 0.68 to 0.80. Using Tukey's Honest Significance Difference method for comparing pairs of systems, only two comparisons out of 120 (16 choose 2) differ, supporting the conclusion that the interannotator agreement values are sufficiently high for system evaluation. Reidsma and Carletta (Reidsma and Carletta 2008) argue in a similar spirit that interannotator agreement values are not strong predictors of how useful a dataset is for machine learning. Through a simulation study, they show that machine learning performance does not degrade if the disagreement among human annotators looks like noise.

As with statistical inference in general, interpretation of results depends on how the data will be used. If there is no single use, which is the intended situation for the IAMTC data, there is no single answer. In a study of data from the 2005 Document Understanding Conference (Passonneau *et al.* 2005), it is pointed out that datasets for computational linguistic applications are often assembled independent of a specific application, or are intended for multiple applications. As a consequence, it is necessary to resort to general criteria, such as those proposed by Krippendorff, to begin addressing the question of whether annotations are reliable. But the reliability

[9] For a detailed illustration of the step-by-step computation of α (see Krippendroff 1980), and for computation of κ^3 (see Artstein and Poesio 2005b).

Table 6. *Three monotonically increasing sets*

Annotators	Set label	WordNet concepts
1–5	A	{COST}
6	B	{COST, MONETARY_VALUE}
7–9	C	{COST, MONETARY_VALUE, TOLL<VALUE}

analysis should not stop there. A reliability value that is sufficient for one task might not be for another, so it is always useful to compare one annotation against another within the context of a specific task. We hope that users of the IAMTC data will do so.

Interannotator reliability metrics support other types of inference besides evidence for the reliability of annotators. In the data presented here, we use reliability metrics in a contrastive manner, to identify subsets of the data that are more or less reliable. We find, unsurprisingly, that some annotators are more consistent, that some datasets were coded more reliably, and that different subtasks were more reliably executed than others.

6.2.3 MASI

MASI is a distance metric for comparing two sets, much like an association measure such as Jaccard (Jaccard 1908) or Dice (Dice 1945). In fact, it incorporates Jaccard, as explained below. When used to weight the computation of interannotator agreement, it is independent of the method in which probability is computed, thus of the expected agreement. It can be used in any weighted agreement metric.

As explained above, annotators were allowed to select multiple concepts or theta roles if a single selection seemed insufficient. Table 6 shows an example of an assignment of WordNet concepts by nine annotators to the lexical item *cost*. Five annotators selected a singleton set, one selected a superset with two members, and three selected a larger superset with three members. If no weighting is applied to the comparison of values for these annotations, then annotators one through five agree with each other, but disagree with annotators six through nine. Further, all disagreements are treated equal, meaning that {COST} is as dissimilar from {COST, MONETARY_VALUE} as it is from {COST, MONETARY_VALUE, TOLL<VALUE}. In our view, this fails to capture important relations among the annotation values. We offer an alternative method using MASI that treats the dissimilarities differentially, depending on two factors: whether the sets are in a monotonic relation, and the difference in size of the sets.

Given two sets, A and B, the formula for MASI is:

$$1 - J_{A,B} \times M_{A,B}. \tag{6}$$

where J is the Jaccard (Jaccard 1908) metric for comparing two sets: a ratio of the cardinality of the intersection of two sets to their union. M (for monotonicity) is a four-point scale that takes on the value 1 when two sets are identical, 2/3 when

one is a subset of the other, 1/3 when the intersection and both set differences are non-null, and 0 when the sets are disjoint. MASI ranges from zero to one. It approaches 0 as two sets have more members in common and are more nearly equal in size. Referring to Table 6, $\text{MASI}_{A,B} = 1 - \frac{1}{2} \times \frac{2}{3} = .67$; $\text{MASI}_{A,C} = 1 - \frac{1}{3} \times \frac{2}{3} = .78$; $\text{MASI}_{B,C} = 1 - \frac{2}{3} \times \frac{2}{3} = .56$. By this measure, sets B and C are most alike, and sets A and C are most dissimilar.

6.3 Scope of the annotation task

Here we provide a concrete sense of the number of decisions annotators were faced with, and the degrees of freedom for each decision, across all three annotation tasks. As we noted in the preceding section, the concept annotation and theta-role annotation tasks were quite distinct. All open-class lexical stems were assigned concepts, whereas theta-role annotation was performed only for verb arguments (since these were based on the verb-centric LCS framework described in Section 4.3), including explicit pronouns and elided pronouns. Also, the choices made during concept annotation determined whether there were theta-role annotation decisions that could be compared. As with most lexical semantic representations (cf. FrameNet, PropBank), we assume theta-role selection is governed by the verb sense. As noted in section 4.3, we refer to the combination of a specific verb sense and the associated theta grid as a theta-domain. Annotators were required to select a theta role for each verb argument; thus, we evaluate agreement by comparing all roles within a theta domain across annotators.[10]

Overall, there were 1,268 lexical items to which the annotators were asked to assign concepts, in contrast to 628 theta domains. For concept annotation, all annotators were faced with the same decision, thus we can compare the 1,268 lexical items across all annotators, including those who made no selection. In contrast, for annotation of theta domains, we first have to identify sets of annotators who made theta-role decisions for the same theta domains.

Table 7 illustrates how the 628 theta domains become further partitioned. It lists how many verb concepts were agreed on by N annotators for all values of N up to nine; there were no cases where ten annotators agreed on the concept for a governing verb. By subtracting the case of 1 annotator, we get only 378 cases where at least two annotators agreed on the concept, hence on the theta domain. It is for these 378 cases that we can measure interannotator agreement. Columns three and four show a further distinction relevant to assessing agreement that we will discuss further below. Many of the verb concepts had at least one theta grid (289), but more than half had no associated grid (339). We assess agreement on theta roles separately for verb concepts with and without grids.

The concept annotation and theta-role annotation tasks also differed regarding the number of values annotators could select. In principle, the full set of concept selections made by annotators was unbounded given that annotators could make

[10] This is in contrast to the evaluation of agreement on theta roles given in Passonneau *et al.* (2006), where agreement was assessed per theta role, rather than per theta domain.

Table 7. *Number of theta domains agreed upon for each cardinality of annotators*

#Annotators	Total concepts	Concepts Plus grid	Concepts minus grid
9	14	3	11
8	29	5	24
7	41	16	25
6	31	13	18
5	40	20	20
4	40	18	22
3	61	24	37
2	122	58	64
1	250	182	68
TOTALS	628	289	339

multiple selections of concepts. Note that the totals for all types of annotations include *None* and no selection (a missing value) as possible annotation labels. The 1,268 lexical items were labeled using 1,864 distinct WordNet concept sets composed from 1,509 distinct WordNet senses; annotators used 569 distinct Mikrokosmos concept sets composed from 438 distinct concepts. In contrast, for labeling theta roles, annotators still had a large number of choices, but far fewer than for concepts. There were fifteen theta-role labels, plus the two additional responses of *None* and no selection, yielding seventeen items that could be combined in various ways for an entire theta grid. The number of combinations is quite large, but many would never occur. The number that actually occur is 108; eliminating *None* and no selection gives 58 observed theta-grid selections. The breakdown for the 108 grids by number of roles is 45 with 3 roles, 47 with 2 roles and 16 with 1.

6.4 Results

Below we present overall interannotator results, results for four parts of speech, and (for subsets of annotators) the maximal subsets of annotators that achieve reliability greater than 0.70. Results from the two metrics used in our evaluations differ very little. Overall interannotator agreement rates for three types of annotation (on the 12 datasets) indicate that reliability is quite good for concept annotation, and even better when missing values are handled appropriately. Reliability for theta-role annotation is better than chance only when the annotator selects from an existing theta grid; theta-role annotation is clearly more difficult for annotators than concept annotation.

In comparing the several translation subsets, the data show a trend of highest reliability in the middle of the project. We hypothesize that the increase relative to the beginning of the project is due to a training effect, and the slight degradation at the end of the project to time pressure. We also report results for each dataset partitioned by part of speech, which shows that assigning concepts to verbs is more difficult for annotators than for other parts of speech. Finally, we summarize the results of

Table 8. *α and* κ^3 *Values using the MASI distance metric for WordNet concept annotations*

Docset	# Annotators	# Lexical Items	# Values	α	κ^3	α_{Miss}
A1E1	7	80	151	0.54	0.55	0.69
A1E2	10	97	249	0.45	0.45	0.66
K1E1	10	112	211	0.44	0.45	0.68
K1E2	10	92	163	0.40	0.41	0.68
J2E1	10	111	286	0.60	0.60	0.72
J2E2	10	117	294	0.55	0.56	0.74
S1E1	9	116	287	0.60	0.60	0.69
S1E2	9	124	278	0.60	0.60	0.71
F2E1	10	130	237	0.49	0.49	0.68
F2E2	10	136	267	0.42	0.42	0.64
H2E1	9	77	129	0.56	0.56	0.68
H2E2	9	76	148	0.58	0.58	0.68
Mean	9.4	105.7	225	0.52	0.52	0.69

Table 9. *α and* κ^3 *Values using the MASI distance metric for mikrokosmos concept annotations*

Docset	# Annotators	# Lexical items	# Values	α	κ^3	α_{Miss}
A1E1	7	80	63	0.26	0.29	0.60
A1E2	10	97	95	0.29	0.30	0.67
K1E1	10	112	78	0.40	0.41	0.75
K1E2	10	92	75	0.34	0.35	0.72
J2E1	10	111	115	0.40	0.41	0.73
J2E2	10	117	118	0.38	0.39	0.74
S1E1	9	116	106	0.35	0.37	0.72
S1E2	9	124	108	0.37	0.38	0.74
F2E1	10	130	71	0.26	0.28	0.68
F2E2	10	136	83	0.18	0.21	0.61
H2E1	9	77	63	0.40	0.41	0.75
H2E2	9	76	74	0.41	0.42	0.75
Mean	9.4	105.7	87.4	0.34	0.35	0.71

comparing all combinations of annotators, and show how this comparison yields information about individual annotators, as well as which subsets of annotators produce the best annotations.

6.4.1 General results

Overall interannotator results are shown in Table 8 for WordNet concepts, Table 9 for Mikrokosmos concepts, and Tables 11–12 for theta roles. Because different sets of annotators were involved in each dataset, the tables present results separately for

each of the English translations from Arabic, French, Hindi, Japanese, Korean, and Spanish. The double lines separate document sets that were annotated earlier versus later.

Tables 8 and 9 have the same format. The number of annotators is shown in column two, but not the identity. The ten annotators who coded French2 (F2E2), for example, were not the same ten annotators who coded Arabic1 (A1E2). Column 3 gives the number of lexical items, which is a rough measure of the relative scope of the task across document sets. Column 4 gives the number of distinct annotation values, or concepts, that annotators selected. Columns 5 and 6 give the α and κ^3 values using the MASI distance metric; for these computations, missing values were treated as another annotation choice. Column 7 gives the α values using MASI, but where we use a computation presented in (Krippendorff 2007) for handling missing values. The principle behind this computation is that instead of considering all comparisons within the i by j table of k values, only those cells are considered that have nonmissing values. Observed disagreement, for example, is the ratio of disagreements to the total choices made by annotators, not to the total number of possible choices. (For the precise method, see Krippendroff 2007.)

Agreement values in Tables 8–9 differ very little between the two metrics for computing interannotator reliability, α and κ^3. This is due in part to the very large number of distinct annotation values. This leads to relatively low probabilities for any value, thus smaller differences in probability for different methods of computing the probability that a given value will occur. In subsequent tables, we will no longer report κ^3. When missing values are factored out (column 7 of Tables 8 and 9), interannotator agreement increases to a quite respectable level for the concept annotations. Due to the clear difference when missing values are factored out, we report all subsequent results using this metric (α_{Miss}).

Both types of annotations yielded roughly the same reliability. The primary difference between the two concept annotations is that, since Mikrokosmos is much smaller, it is much less complete than WordNet and misses many of the concepts required for annotation. Annotation of words without appropriate concepts are shown as missing values in the tables. The larger number of missing values for Mikrokosmos annotation can be seen indirectly in the large difference in the mean of the number of values (sets of concepts) assigned per translation set: 225 for WordNet concepts versus 87.4 for Mikrokosmos concepts. This result is not unexpected, given that WordNet has much broader coverage.

Whatever the reason for the missing values for the concept annotations, it apparently does not reflect annotator confusion about the task, given that the consistency among annotators is more apparent after factoring out the missing values. The means for α_{Miss} are 0.69 and 0.71 for WordNet concepts and Mikrokosmos concepts, respectively. Note that means are reported only for convenience and are not meant to represent a global reliability measure. In fact, we performed separate analyses for the different translation sets because the topics discussed are somewhat different, which could potentially lead to more or less difficulty in the semantic task of selecting concepts. We do see clear differences in reliability across translation sets for all three tables, a point we return to below.

Table 10. *Illustration of annotator assignment of concepts to one verb*

Sentence-verb-id-concept	# Grids	# Annotators	Which annotator
A1E2-8.consider.200.see<think	3	8	1,2,5,6,8,9,10,11
A1E2-8.consider.200.consider	0	2	2,7
A1E2-8.consider.200.moot	0	2	2,7
A1E2-8.consider.200.study>liken	2	2	2,7
A1E2-8.consider.200.think>see	1	2	5,6
A1E2-8.consider.200.take>play	0	1	7
A1E2-8.consider.200.regard<look	0	1	7

In contrast to the concept annotation task, the theta-role annotation task produced far fewer data points and exhibited less agreement among annotators. It is important to keep in mind that for each theta domain, there were several dimensions of freedom. As noted above (Section 6.3), each annotator was free to select any concept for the governing verb. Second, for any concept selection, the ontology might associate one or more theta grids with the concept, or in the case of incomplete information in the ontology, no theta grids. Third, if there were any theta grids associated with the concept, the annotator was free to select one or not. However, in contrast to the concept annotation, annotators could not make multiple selections. Thus we report interannotator agreement without using a distance metric. In the course of presenting our results for theta-role assignment, we discuss how much freedom annotators actually took with respect the above dimensions.

To assess agreement on theta domains, we must identify those cases where the same annotators were presented with the same theta domains for which to choose theta roles. Given the large number of annotators (a maximum of ten per dataset), for any verb-concept pair that any one annotator identified, there are a large number of combinations of other annotators that might identify the same verb-concept pair. Table 10 shows the number of assignments of seven different concepts (and associated theta grids) to a single verb token. The first three fields of the entries in the first column give the sentence id, the verb root, and a numeric identifier for the lexical item; the last field gives the concept that was assigned. One concept (see<think) was selected by eight annotators, four by two, and two by one. One annotator made five assignments (7), one made four (2), and so on.

One of the key points to note from Table 10 is that while there were four concepts selected by exactly two annotators, in three cases, it was the same pair of annotators (2,7), and in the fourth case it was a different pair (5,6). We create one agreement matrix for all the theta domains examined by the same set of annotators, thus one for 2 and 7, and a separate, analogous matrix for annotators 5 and 6.

Table 10 also illustrates how the number of theta grids seen by annotators varies (column 2), depending on the concept. There were 339 theta domains that had no grid associated with the concept selection, and 289 that had grids, with an average of 2.2 grids per concept (the maximum was 7). We computed interannotator agreement separately for the two cases of zero grids versus one or more, on the view that if annotators were presented with a theta grid, they were more likely to agree on

Table 11. *Interannotator agreement for theta domains with theta drids*

Docset	#Theta domains	#Annotators	Which annotator	α_{Miss}
A1E1	4	5	5, 7, 9, 10, 11	0.24
	2	6	5, 7, 8, 9, 10, 11	0
	2	5	5, 7, 8, 9, 10	1.0
	2	5	5, 8, 9, 10, 11	0.22
A1E2	5	9	1, 2, 5, 6, 7, 8, 9, 10, 11	0.26
	3	8	1, 2, 5, 6, 7, 9, 10, 11	0.06
	3	8	1, 2, 5, 6, 8, 9, 10, 11	0.37
	2	4	1, 2, 6, 9	−0.25
	2	3	1, 5, 6	1.0
F2E1	3	3	1, 2, 6	0.33
	2	5	1, 2, 6, 9, 11	0.23
F2E2	2	2	2, 6	0.40
H2E1	3	7	1, 2, 5, 6, 9, 10, 11	0.34
	2	7	1, 2, 5, 6, 7, 9, 10	−0.03
H2E2	3	8	1, 2, 5, 6, 7, 9, 10, 11	0.16
	3	7	1, 2, 5, 6, 9, 10, 11	0.36
J2E1	2	2	2, 6	1.0
	2	2	6, 7	−0.20
J2E2	5	2	2, 6	0.69
	2	7	2, 5, 6, 7, 9, 10, 11	−0.04
	2	3	2, 6, 7	0.55
K1E1	18	2	2, 6	0.74
	6	3	5, 6, 8	0.58
	4	3	2, 6, 9	0.71
	3	2	6, 8	0.29
	2	7	1, 2, 3, 6, 8, 10, 11	0.23
	2	6	1, 2, 5, 6, 8, 10	0.21
	2	6	1, 2, 6, 8, 9, 10	0.52
	2	5	1, 6, 8, 10, 11	0.56
	2	2	5, 6	0.40
K1E2	7	2	2, 6	1.0
	2	8	1, 2, 5, 6, 8, 9, 10, 11	0.42
	2	3	5, 6, 8	1.0

theta-role choices for a given theta domain. This turns out to be the case. However, agreement varied widely, and was often no better than chance.

Table 11 gives the interannotator agreement figures for theta domains where a theta grid was associated with the concept selected from the ontology, and where the same annotators agreed on at least two theta domains. The annotator data for theta roles is quite sparse in comparison to the number of theta domains; note the absence of agreement matrices here for S1E1 or S1E2. This is due in part to the experimental design, where annotators selected concepts and theta roles in one step. To compute agreement among a given set of annotators on theta roles, there must first be a fixed set of annotators greater than one, all of whom chose the same verb concept for the same verb in the case of at least two verbs. It sometimes happens, for example, that while there may be three annotators who select the same concept

Table 12. *Interannotator agreement for theta domains without theta drids*

Docset	#Theta domains	#Annotators	Which annotators	α_{Miss}
A1E2	2	3	2, 6, 8	−0.20
	6	2	7, 10	1.00
F2E1	2	7	1, 2, 5, 6, 9, 10, 11	0.14
H2E2	2	2	2, 7	−0.20
J2E1	2	2	2, 7	1.00
J2E2	2	9	1, 2, 5, 6, 7, 8, 9, 10, 11	−0.01
	2	5	1, 7, 9, 10, 11	−0.08
	2	3	5, 9, 10	0.38
K1E1	3	7	1, 2, 5, 6, 8, 9, 10	0.20
	2	2	2, 8	0.40
K1E2	2	7	1, 2, 5, 6, 8, 9, 10	0.04
	2	7	1, 2, 6, 8, 9, 10, 11	0.14
S1E1	2	2	2, 6	0

for verb V_i, and three annotators who select the same concept for verb V_j, it is not the same three annotators. This happened throughout the Spanish data.

The first two columns, the number of theta domains and the number of annotators, gives the size of the agreement matrix. As Table 11 shows, when a given translation set has only a few agreement matrices, the values often vary widely; A1E1, for example, has four matrices, three of which include only two theta domains, and agreement ranges from perfect to little or no different from chance.

Table 11 also illustrates that most translation sets have three or fewer agreement matrices, typically with only a few theta domains, or only a few annotators. K1E1 presents the most unusual case in that there are nine agreement matrices, and in about half these cases, the agreement is moderately good. In one case where agreement is above 0.70, two annotators agreed on theta-role selections for eighteen theta domains ($\alpha = 0.74$). In the other case, three annotators agreed on theta roles for four theta domains ($\alpha = 0.71$).

Table 12 gives the results for the set of theta domains where annotators saw no grid. Here the data is even more sparse than for Table 11. Despite the fact that there were more theta domains without grids overall (cf. Table 7), there were far fewer that multiple annotators had in common. Table 12 has thirteen rows in contrast to thirty-four in Table 11, and all but two of the thirteen rows have only two theta domains per agreement matrix. Agreement was perfect for two matrices, was about 0.40 for two, but otherwise was close to chance.

The differences between the three subgroups we identified on the basis of the time period at which the annotation was performed (Arabic/Korean, Japanese/Spanish, French/Hindi) reaffirm the motivation for reporting reliability separately for each data set. In Table 11, we see a striking difference for one data set: K1E1. However, we have too little information to explain the larger number of agreement matrices, and higher agreement values. In Tables 8 and 9, the three interannotator agreement columns all show the greatest interannotator reliability for the English translations of

H2E1-6 Last year, due to a famine, the growth rate had been 1 per cent lesser than the estimated growth rate.
H2E2-6 Last year due to drought conditions, India's economy grew at a rate 1% less then estimated.

Fig. 10. Two translations of the same source sentence from Hindi.

Japanese and Spanish. We have too little information to account for this difference, but we consider three possibilities. First, it is possible that the sites that performed the Japanese and Spanish annotations provided more support for the annotator task, either through better training, a more consistent schedule, or other types of human factors. In an earlier paper (Passonneau *et al.* 2006), where we reported only the column five metric (α using MASI, no special treatment of missing values), we suggested there was a possible training effect, with lower interannotator reliability while annotators were still learning the task, then lasting improvement once the project was truly underway. Here, where we can see a consistent pattern for α whether we address missing values or not, we can see that there seems to be some degradation towards the end of the project, which could be a result of greater time pressure.

Other possible sources of difference between the document set reliability scores could depend on differences in the semantic complexity of the concepts expressed, or to differences in the translation quality. Figure 10 shows two translations of the same sentence. The first translation is a less fluent sentence of English: the repetition of the NP *the growth rate* is somewhat awkward, and the word *less* would have been more correct instead of *lesser*. This conceivably could affect annotators' sense judgements. However, it is difficult to imagine how to control for these conditions, apart from conducting a very large scale study with randomized materials.

6.4.2 Comparing annotators

To compare annotators, we computed reliability for all combinations of annotators from 2 to N, where N is the total number of annotators. From this, we identified the minimum number of annotators that could be dropped in order to achieve an interannotator agreement of 0.70 or higher. The column labeled *N* in Table 13 indicates the number of annotators remaining after dropping this minimum number. Column WHO indicates the identity of the best subset of annotators that achieves this threshold, while AVG_N gives the average reliability over all combinations of annotators of the same cardinality N. Table 13 illustrates that very good reliability can be achieved by dropping relatively few annotators, and that individual annotators differ in reliability.[11]

[11] Variability in reliability across annotators is a common experience, although this is not so obvious. Krippendorf advocates a methodology where the written guidelines and annotator selection process are designed to make all annotators interchangeable. This clearly should be a major focus area for future annotation studies of this type.

Table 13. *Maximum number (N) of annotators to achieve agreement of .70 or above versus average across all combinations of N annotators (AVG_N)*

DocSet	Full reli	Subset reli	N	WHO	AVG_N
A1E1	0.69	0.71	6 of 7	4 5 7 9 10 11	0.52
A1E2	0.66	0.70	7 of 10	1 2 4 6 8 9 11	0.40
K1E1	0.68	0.71	9 of 10	2 3 4 5 6 8 9 10 11	0.42
K1E2	0.68	0.70	9 of 10	2 3 4 5 6 8 9 10 11	0.38
J2E1	0.72	0.74	9 of 10	2 4 5 6 7 8 9 10 11	0.56
J2E2	0.74	0.74	10 of 10		
S1E1	0.69	0.70	8 of 9	2 4 5 6 7 8 9 10 11	0.57
				1 2 4 5 6 7 8 9 10	0.57
S1E2	0.71	0.72	8 of 9	2 4 5 6 7 8 9 10 11	0.56
				1 2 4 5 6 8 9 10 11	0.56
F2E1	0.68	0.73	8 of 10	2 3 4 5 6 9 10 11	0.47
F2E2	0.64	0.70	7 of 10	2 3 4 5 6 10 11	0.40
H2E1	0.68	0.71	7 of 9	1 2 4 6 9 10 11	0.55
H2E2	0.69	0.71	8 of 9	1 2 4 5 6 9 10 11	0.56

Table 14. *Number of times each annotator occurs in the subset achieving tnterannotator reliability of 0.70 (in 11 of the 12 translation subsets)*

Coder	N of 11 times
2	11
4	11
6	11
11	11
9	10
10	10
5	9
1	6
8	5
3	4
7	4

We also were able to identify groups of individual annotators with more consistently high interannotator agreement, and to determine which selection of annotators would yield the most consistent annotations. If we restrict our attention to the most consistent pair of annotators per dataset, reliability ranges from 0.75 to 0.83. Table 14 indicates how often each annotator occurred in the subset that achieves a reliability of 0.70 or greater in the eleven translation subsets excluding A1E1, which had only seven annotators. As shown, annotators 2, 4, 6, and 11 were always in the subset of the most reliable annotators; annotators 9 and 10 were in the most reliable subset all but one time. Thus half the annotators were consistently very reliable. A similar analysis of the Mikrokosmos concept and theta-role reliability results indicates that relatively more annotators would need to be dropped to reach the same threshold of 0.70.

Table 15. α_{Miss} *Partitioned by part-of-speech*

	All	Noun	Adj	Verb	Adv
A1E1	N = 80	N = 40	N = 22	N = 16	N = 2
	0.69	0.67	0.69	0.67	1.00
A1E2	N = 97	N = 49	N = 17	N = 24	N = 7
	0.66	0.67	0.59	0.67	0.58
K1E1	N = 112	N = 65	N = 9	N = 31	N = 7
	0.68	0.71	0.77	0.52	0.81
K1E2	N = 92	N = 56	N = 9	N = 21	N = 6
	0.68	0.66	0.74	0.62	0.72
J2E1	N = 111	N = 57	N = 22	N = 22	N = 10
	0.72	0.78	0.66	0.54	0.85
J2E2	N = 117	N = 58	N = 23	N = 27	N = 9
	0.74	0.80	0.73	0.54	0.81
S1E1	N = 116	N = 66	N = 25	N = 14	N = 11
	0.69	0.72	0.64	0.62	0.63
S1E2	N = 124	N = 75	N = 30	N = 14	N = 5
	0.71	0.75	0.64	0.59	0.54
F2E1	N = 130	N = 71	N = 37	N = 15	N = 7
	0.68	0.66	0.72	0.56	0.69
F2E2	N = 136	N = 76	N = 41	N = 12	N = 7
	0.64	0.61	0.73	0.51	0.54
H2E1	N = 77	N = 42	N = 13	N = 16	N = 6
	0.68	0.70	0.58	0.65	0.60
H2E2	N = 76	N = 40	N = 16	N = 15	N = 5
	0.68	0.71	0.63	0.68	0.26
Mean	0.69	0.71	0.68	0.59	0.64

Consistency across annotators is desirable to the degree that it confirms that different annotators interpret the annotation task and guidelines in the same way. It should be noted, however, that interannotator agreement metrics cannot rate the quality of annotations unless the annotation is compared to a gold standard of known quality. An annotator who is an outlier can be either inferior or superior to other annotators.

6.4.3 Partitioning the data by part-of-speech

We computed separate reliability scores for the four parts of speech that were annotated: noun, verb, adjective and adverb. In general, the reliability scores by part of speech were distributed similarly to the full set, with nouns having somewhat higher reliability on average. Verbs, however, had significantly lower scores, a finding which has often been replicated (Fellbaum *et al.* 2001; Palmer, Dang and Fellbaum 2005a; Passonneau, Salleb-Aouissi and Ide 2009). Table 15 shows the $\alpha_{MASI/miss}$ values for the full dataset in column two, and for each part of speech in columns three through six. The last row of the table gives the mean of each column (again, for illustrative purposes only), showing that verbs had the lowest mean reliability.

6.5 Evaluation summary

Annotators made use of multiple WordNet concepts often, at an average rate of two per lexical item for all annotated words, thus it is important to have a method for measuring interannotator agreement that addresses set-valued data. We have presented MASI, an association measure for semantic and pragmatic annotation, and illustrated its use in handling set-valued data when incorporated as a distance metric in Krippendorff's α. This metric is bias-free and thus suited to a new task where there are no known factors that introduce an a priori bias, such as gender differences in sense interpretation. In any case, the large number of annotators minimizes the difference between bias-free α and bias-sensitive κ^3 (Artstein and Poesio 2005b).

The more significant benefit of Krippendorff's α for the IL1 data is the large number of missing values in the dataset. If missing values are excluded from the computation of expected and observed agreement instead of treated as a distinct value (null), agreement on WordNet and especially Mikrokosmos concepts goes up significantly. The resulting agreement values on concepts are in the range of 0.60–0.75, which is quite respectable, particularly given the large number of annotators and very large number of annotation values. In contrast, interannotator agreement on theta roles was low for both cases of verbs with and without theta grids. However, this was partly due to the experimental design. The fact that one site achieved relatively high reliability on theta-role annotation (for the Korean datasets) indicates that under some conditions, even novice annotators can perform theta-role annotation reliably. Future attempts to annotate theta roles should present this as an entirely separate task for verbs whose sense has already been selected.

There are differences in level of agreement across the twelve datasets that coincide with differences in time and site. While we cannot say definitively that there is a causal relation, such differences have been observed in other annotation studies (Cerrato 2004; Gut and Bayerl 2004). The lesson for future multisite annotation projects would be to impose similar training regimens and schedules, to recruit annotators from the same population, and if possible, to test for cross-site differences by documenting all of these factors. We also observe different levels of agreement for different parts of speech, with verbs having the lowest agreement. This finding is also consistent with prior studies (Fellbaum *et al.* 2001; Palmer *et al.* 2005a; Passonneau *et al.* 2009).

A large number of annotators participated in our project, with clear and consistent differences in their ability to annotate reliably. As none of the annotators were identifiable *experts* whose annotations could be considered more trustworthy on independent grounds (e.g., long experience and training), the most conservative interpretation is that annotators who are consistently more reliable are the best annotators. Another option for future work is to pre-test annotators and select the most reliable ones before engaging in large-scale annotation.

In the NLP literature, interannotator agreement measures are usually presented in order to make the claim that an annotation is reliable. We have demonstrated that the WordNet concept and Mikrokosmos concept annotations are relatively

reliable, given the application of a distance metric for measuring interannotator reliability on set-valued items (MASI). We have also illustrated another use of reliability measures, namely to examine variations in reliability along different dimensions.

7 Conclusions and future work

We have demonstrated the feasibility of applying a methodology for interlingual annotation of parallel corpora. The scientific interest of this research lies in the definition, and annotation feasibility of, a level of semantic representation for natural language text – the interlingua representation – that captures important aspects of the meaning of different natural languages.

To date, corpora have been annotated at a relatively shallow (semantics-free) level, forcing NLP researchers to choose between shallow approaches and hand-crafted approaches, each having its own set of problems. Although we have not constructed an annotated corpus large enough for heavy-duty machine learning algorithms, we see our work as paving the way for developing solutions to representational problems and thereby for enabling other, larger annotation efforts.

Indeed, showing the validity of our representational system has proven important for determining which representational systems are most reliable for building NLP systems that make use of this information. For example, (Habash, Dorr and Monz 2009) demonstrate that a generation-heavy hybrid MT system based on an "approximate interlingua" that is similar to the representational formalism studied in this annotation project produces output with improved grammatical processes such as verb–subject realization and long-distance dependency translation. The resulting system outperforms both the symbolic and the statistical approaches in an evaluation of these grammatical processes.

More recently, twenty members of an eight-week JHU Human Language Technology Center of Excellence *Summer Camp for Applied Language Exploration* (SCALE-2009) further demonstrated the use of deeper representations (e.g., modality) for improving translation quality in the face of sparse training data Baker *et al.* (2009, 2010). Our study was among the first to make it possible to formulate a framework within which such investigations have been conducted and will continue to be conducted in the future.

We have encountered a number of difficult issues for which we have only interim solutions. Principal among these is the granularity of the interlingual terms to be used. Omega's WordNet symbols, some 110,000, afford too many alternatives with too little clear semantic distinction, negatively impacting agreement rates. On the other hand, the number of Mikrokosmos concepts – 6,000 – is too small to capture many of the distinctions people deem relevant. An evaluation of interannotator agreement shows that a high degree of agreement is possible when an appropriate number of alternative entries is present in the ontology. It is on this basis that the OntoNotes annotation project created the sense pools that constitute the new version of Omega; senses that were simply impossible for the annotators to handle were grouped together or discarded.

Similarly, the theta roles in some cases appear hard for annotators to understand. While we have considered following the example of FrameNet and defining idiosyncratic roles for almost every process, the resulting proliferation does not bode well for later large-scale machine learning. Additional issues to be addressed include: (1) personal name, temporal and spatial annotation (Ferro *et al.* 2001); (2) causality, co-reference, aspectual content, modality, speech acts, etc; (3) reducing vagueness and redundancy in the annotation language; (4) interevent relations such as entity reference, time reference, place reference, causal relationships, associative relationships, etc; and (5) cross-sentence phenomena.

Acknowledgements

This work is supported, in part, by the National Science Foundation under Grants No. IIS-0326553, IIS-0705832, and IIS-0531176, and in part, by the Johns Hopkins Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

References

Artstein, R., and Poesio, M. 2005a. Bias decreases in proportion to the number of annotators. In *Proceedings of FG-MoL 2005*, Edinburgh, UK, pp. 141–150.

Artstein, R., and Poesio, M. 2005b. Kappa Cubed = Alpha (or Beta). Technical Report NLE Technote 2005-01, University of Essex.

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* **34**: 555–596.

Baker, C. F., Fillmore, C. J. and Lowe, J. B. 1998. The Berkeley FrameNet project. In C. Boitet, and P. Whitelock (eds.), *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pp. 86–90. San Francisco, CA: Morgan Kaufmann Publishers.

Baker, Kathryn, Bloodgood, Michael, Dorr, Bonnie J., Filardo, Nathaniel W., Levin, L., and Piatko, C. 2010. A modality lexicon and its use in automatic tagging. In *Seventh Language Resources and Evaluation Conference (LREC-2010)*. University of Malta, Malta.

Baker, K., Bethard, S., Bloodgood, M., Brown, R., Callison-Burch, C., Coppersmith, G., Dorr, B., Filardo, W., Giles, K., Irvine, Ann, K., Mike, L., Lori, M., Justin, M., Jim, M., Scott, P., Aaron, P., A., Piatko, C., Schwartz, L., and Zajic, D 2009. Semantically informed machine translation. Technical Report 002, Human Language Technology Center of Excellence, Summer Camp for Applied Language Exploration, Johns Hopkins University, Baltimore, MD.

Bannard, C., and Callison-Burch, C. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI, pp. 597–604.

Barzilay, R., and Lee, L. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*, Edmonton, Canada, pp. 16–23.

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. Technical Report Unpublished research report, USC/Information Sciences Institute, Marina del Rey. ISI-TR-85-029.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. 2003. The prague dependency treebank: three-level annotation scenario. In A. Abeillé (ed.), *Treebanks: Building and Using Syntactically Annotated Corpora*, pp. 103–128. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Callison-Burch, C., Koehn, P., and Osborne, M. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL*, New York, pp. 17–24.

Cerrato, L. 2004. A coding scheme for annotation of feedback phenomena in conversational speech. In *Proceedings of the LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisbon, Portugal, pp. 25–28.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1): 37–46.

Cohen, J. 1968. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**: 213–220.

di Eugenio, B., and Glass, M. 2004. The Kappa statistic: A second look. *Computational Linguistics* **30**(1): 95–101.

Dice, J. L. R. 1945. Measures of the amount of ecologic association between species. *Ecology* **26**: 297–302.

Dolan, W., Quirk, C., and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of COLING 2004*. Geneva, Switzerland.

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

Dorr, B. J., Green, R., Levin, L., Rambow, O., Farwell, D., Habash, N., Helmreich, S., Hovy, E., Miller, K. J., Mitamura, T., Reeder, F., and Siddharthan, A. 2004. Semantic annotation and lexico-syntactic paraphrase. In *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora (LREC-2004)*. Portugal.

Dorr, B. J., Olsen, M., Habash, N., and Thomas, S. 2001. LCS verb database. Technical Report Online software database, University of Maryland, College Park, MD. http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html [2010, March 29].

Farwell, D. and Helmreich, S. 1999. Pragmatics and translation. *Procesamiento de Lenguaje Natural* **24**: 19–36.

Farwell, D., Helmreich, S., Reeder, F., Miller, K., Dorr, B., Habash, N., Hovy, E., Levin, L., Mitamura, T., Rambow, O., and Siddharthan, A. 2004. Interlingual annotation of multilingual text corpus. In *Proceedings of the Workshop on*

Frontiers in Corpus Annotation. Workshop at the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Boston, MA, pp. 55–62.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press. http://wordnet.princeton.edu/ [2010, March 29].

Fellbaum, C., Grabowski, J., and Landes, S. 1998. Performance and confidence in a semantic annotation task. In C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, pp. 217–239. Cambridge, MA: MIT Press. http://wordnet.princeton.edu/ [2010, March 29].

Fellbaum, C., Palmer, M., Dang, H. T., Delfs, L., and Wolf, S. 2001. Manual and automatic semantic annotation with wordnet. In *Proceedings of the Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA.

Ferro, L., Mani, I., Sundheim, B., and Wilson, G. 2001. TIDES temporal annotation guidelines, Version 1.0.2. Technical Report MTR 01W0000041, Mitre, McLean, VA.

Fillmore, C. 1968. The case for case. In E. Bach, and R. Harms (eds.), *Universals in Linguistic Theory*, pp. 1–88. New York: Holt, Rinehart and Winston.

Fillmore, C., Johnson, C., and Petruck, M. 2003. Background to FrameNet. *International Journal of Lexicography* **16**(3): 235–250.

Fleischman, M., Echihabi, A., and Hovy, E. H. 2003. Offline strategies for online question answering: answering questions before they are asked. In *Proceedings of the ACL Conference*. Sapporo, Japan.

Francis, W. N., and Kucera, H. 1982. *Frequency Analysis of English Usage*. Boston, MA: Houghton Mifflin.

Funaki, S. 1993. Multi-lingual machine translation (mmt) project. In *Proceedings of the MT Summit IV*. Washington, DC.

Garside, R., Leech, G., and McEnery, A. M. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Addison Wesley Longman.

Gut, U., and Bayerl, P. S. 2004. Measuring the reliability of manual annotations of speech corpora. In *Proceedings of Speech Prosody*, Nara, Japan, pp. 565–568.

Habash, N., and Dorr, B. J. 2003. Interlingua annotation experiment results. In *Proceedings of AMTA-2002 Interlingua Reliability Workshop*. Tiburon, CA.

Habash, N., Dorr, B., and Monz, C. 2009 Symbolic-to-statistical hybridization: extending generation-heavy machine yranslation. *Machine Translation* **23**(1): 23–63.

Habash, N., Dorr, B. J., and Traum, D. 2003. Hybrid natural language generation from lexical conceptual structures. *Machine Translation* **18**(2): 81–128.

Hajič, J., Vidová-Hladká, B., and Pajas, P. 2001. The prague dependency treebank: annotation structure and support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pp. 105–114. University of Pennsylvania, Philadelphia, PA.

Helmreich, S., and Farwell, D. 1998. Translation differences and pragmatics-based MT. *Machine Translation* **13**(1): 17–39.

Hirst, G. 2003. Paraphrasing paraphrased. In *Keynote address for The Second International Workshop on Paraphrasing: Paraphrase Acquisition and*

Applications. Association for Computational Linguistics ACL 2003, Sapporo, Japan. http://ftp.cs.toronto.edu/pub/gh/Hirst-IWP-talk.pdf

Hovy, E. H., Marcus, M., Palmer, M., Pradhan, S., Ramshaw, L., and Weischedel, R. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology/North American Association of Computational Linguistics conference (HLT-NAACL 2006),* New York.

Hovy, E., Marcus, M., and Weischedel, R. 2003a. OntoBank. In *Presentation at Darpa PI Meeting*. Arden House, Harriman, New York.

Hovy, E. H., Philpot, A., Ambite, J. L., Arens, Y., Klavans, J., Bourne, W., and Saroz, D. 2003c. Data acquisition and integration in the DGRC's energy data collection project. In *Proceedings of the NSF's dg.o 2001 Conference*. Los Angeles, CA.

Hovy, E., Philpot, A., Klavans, J. L., Germann, U., and Davis, P. T. 2003b. Extending metadata definitions by automatically extracting and organizing glossary definitions. In *Proceedings of the National Conference on Digital Government Research*. Boston, MA.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles* **44**: 223–70.

Jackendoff, R. 1972. Grammatical relations and functional structure. In *Semantic Interpretation in Generative Grammar*. Cambridge, MA: The MIT Press.

Kingsbury, P., and Palmer, M. 2002. From treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Spain.

Kingsbury, P., Snyder, B., Xue, N., and Palmer, M. 2003. PropBank as a bootstrap for Richer annotation schemes. In *Sixth Workshop on Interlinguas: Annotations and Translations, MT Summit IX*. New Orleans, LA.

Kipper, K., Palmer, M., and Rambow, O. 2002. Extending PropBank with VerbNet semantic predicates. In *Workshop on Applied Interlinguas (AMTA-2002)*. Tiburon, CA.

Knight, K., and Luk, S. K. 1994. Building a large-scale knowledge base for machine translation. In *Proceedings of AAAI*. Seattle, WA.

Kozlowski, R., McCoy, K. F., and Vijay-Shanker, K. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, Sapporo, Japan, pp. 1–8. ACL 2003.

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage Publications.

Krippendorff, K. 2007. Computing Krippendorff's alpha-reliability. http://www.asc.upenn.edu/usr/krippendorff/webreliability.doc [2010, March 29].

Levin, B., and Rappaport-Hovav, M. 1998. From lexical semantics to argument realization. In H. Borer (ed.), *Handbook of Morphosyntax and Argument Structure*. Dordrecht: Kluwer Academic Publishers.

Madnani, N., Ayan, N. F., Resnik, P., and Dorr, B. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*. Prague, Czech Republic.

Mahesh, K., and Nirenburg, S. 1995. A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*. Montreal, Canada.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. 1994. Building a large annotated corpus of english: the Penn treebank. *Computational Linguistics*, **19**(2): 313–330.

Martins, T., Rino, L. H. Machado, Nunes, M. G. Volpe, Montilha, G., and Novais, O. O. 2000. An interlingua aiming at communication on the web: how language-independent can it be? In *Proceedings of Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP, ANLP-NAACL*. Seattle, WA.

Mel'čuk, I. A. 1988. *Dependency Syntax: Theory and Practice*. New York: State University of New York Press.

Mitamura, T., Miller, K. J., Dorr, B. J., Farwell, D., Habash, N., Levin, L., Helmreich, S., Hovy, E., Levin, L., Rambow, O., Reeder, F., and Siddharthan, A. 2004. *Semantic Annotation of Multilingual Text Corpora*. Portugal.

Miyoshi, H., Sugiyama, K., Kobayashi, M., and Ogino, T. 1996. An overview of the edr electronic dictionary and the current status of its utilization. In *Proceedings of the 16th conference on Computational Linguistics*, Copenhagen, Denmark, pp. 1090–1093.

Moore, R. C. 1994. Semantic evaluation for spoken-language systems. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Princeton, NJ.

Palmer, M., Dang, H. T., and Fellbaum, C. 2005a. Making fine-grained and coarse-grained sense distinctions. *Journal of Natural Language Engineering* **13**: 137–163.

Palmer, M., Gildea, D., and Kingsbury, P. 2005b. The proposition bank: a corpus annotated with semantic roles. *Computational Linguistics* **31**(1): 71–106.

Pang, B., Knight, K., and Marcu, D. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *Proceedings of HLT-NAACL*. Edmonton, Canada.

Passonneau, R. 2004. Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal.

Passonneau, R. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy.

Passonneau, R. J. 2010. Formal and functional assessment of the pyramid method for summary content evaluation. *Natural Language Engineering* **16**: 107–131.

Passonneau, R., Habash, N., and Rambow, O. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy.

Passonneau, R., Nenkova, A., McKeown, K., and Sigelman, S. 2005. Applying the pyramid method in DUC 2005. In *Proceedings of the Document Understanding Conference (DUC) Workshop*. Vancouver, Canada.

Passonneau, R. J., Salleb-Aouissi, A., and Ide, N. 2009. Making sense of word sense variation. In *Proceedings of the NAACL-HLT 2009 Workshop on Semantic*

Evalutions: Recent Achievements and Future Directions (SEW-2009), Boulder, CO, pp. 2–9.

Philpot, A., Fleischman, M., and Hovy, E. H. 2003. Semi-automatic construction of a general purpose ontology. In *Proceedings of the International Lisp Conference*. New York.

Philpot, A., Hovy, E., and Pantel, P. 2005. The omega ontology. In *Proceedings of IJCAI*. Edinburgh, Scotland.

Pradhan, S., Hovy, E. H., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. 2007. OntoNotes: a unified relational semantic representation. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC-07)*, Irvine, CA, pp. 517–524.

Rambow, O., Dorr, B., Farwell, D., Green, R., Habash, N., Helmreich, S., Hovy, E., Levin, L., Miller, K. J., Mitamura, T., Reeder, F., and Advaith, S. 2006. Parallel syntactic annotation of multiple languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*. Genoa, Italy.

Reeder, F., Dorr, B., Farwell, D., Habash, N., Helmreich, S., Hovy, E., Levin, L., Mitamura, T., Miller, K., Rambow, O., and Siddharthan, A. 2004. *Interlingual Annotation for MT Development*. Georgetown University, Washington, DC.

Reidsma, D., and Carletta, J. 2008. Reliability measurement without limits. *Computational Linguistics* **34**: 319–326.

Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., and Moll, D. 2003. Exploiting paraphrases in a question-answering system. In *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, Edmonton, Canada, pp. 25–32. ACL 2003.

Scott, W. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* **17**: 321–325.

Siegel, S., and Castellan, N. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

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

Tapanainen, P., and Jarvinen, T. 1997. A non-projective dependency parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing and Association for Computational Linguistics*. Washington Marriott Hotel, Washington, DC.

Véronis, J. 2000. From the Rosetta stone to the information society: a survey of parallel text processing. In J. Véronis (ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora*, pp. 1–24. London: Kluwer Academic Publishers.

Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C., and Doddington, G. 2003. Multiple-translation arabic corpus, Part 1. Technical Report catalog number LDC2003T18 and ISBN 1-58563-276-7, Linguistic Data Consortium (LDC).

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

Appendix Example of IL1 internal representation

This appendix shows an example of the internal representation resulting from the annotation of the sentence presented earlier in Figure 2 using the Tiamat Annotation Interface. Information stored in each "node" of the tree structure includes a unique node identifier, the surface lexeme (e.g., *announced*), part of speech (e.g., V), underlying form (e.g., *announce), features (e.g., past), position of IL1 tree node (e.g., Root), Thematic Role (e.g., THEME), WordNet sense(s) (e.g., announce<say), and Mikrokosmos concept(s) (e.g., DECLARE, INFORM).*

```
([140,announced,V,announce,feat:past,Root,announce&&announce<say,
  DECLARE$VERB&&INFORM$VERB]
  ([20,Mohamed,PN,mohamed,feat:nom_num:sg,Subj,AGENT]
   ([10,Sheikh,PN,sheikh,feat:nom_num:sg,Mod],
    [90,Defense_Minister,PN,defense_minister,feat:nom_num:sg,Mod]
    ([40,who,Pron,who,rel:+_feat:wh_feat:nom,Subj],
     [60,also,Adv,also,Mod,also],
     [100,of,P,of,Mod]
     ([120,United_Arab_Emirates,PN,unitedarabemirates,feat:nom,Obj])),
      [150,at,P,at,Mod]
      ([180,ceremony,N,ceremony,feat:nom_num:sg,Obj,ceremony,CEREMONY$NOUN]
       ([170,inauguration,N,inauguration,feat:nom_num:sg,Mod,inauguration,
         ORIENT$NOUN&&CEREMONY$NOUN])),
        [220,want,V,want,tense:pres,Obj,THEME&&desire>envy&&want<be]
        ([200,",Pun,",Mod],
         [210,we,Pron,we,feat:pers_feat:nom_num:pl_per:1,Subj,
          EXPERIENCER&&EXPERIENCER],
         [250,make,V,make,sform:inf,Obj,PERCEIVED&&PERCEIVED,make>stir,
          BUILD$VERB&&CREATION-RELATION$VERB]
         ([240,<we>,Pron,<we>,Subj,AGENT],
          [300,center,N,center,feat:nom_num:sg,Obj,THEME,kernel]
          ([260,Dubai,PN,dubai,feat:nom_num:sg,Subj],
           [280,new,Adj,new,feat:abs,Mod,new>hot&&new&&fresh<original],
           [290,trading,N,trading,feat:nom_num:sg,Mod,trading,
            COMMERCE-EVENT$NOUN])),
         [320,",Pun,",Mod])))
```