Interlingua Development and Testing through Semantic Annotation of Multilingual Text Corpora

Bonnie Dorr[1], David Farwell[2], Rebecca Green[1], Nizar Habash[1], Stephen Helmreich[2], Eduard Hovy[3], Lori Levin[4], Keith Miller[5], Teruko Mitamura[4], Owen Rambow[6], Florence Reeder[5], Advaith Siddharthan[6]

1: Institute for Advanced Computer Studies
University of Maryland
{bonnie,nizar,rgreen}@umiacs.umd.edu

2: Computing Research Laboratory
New Mexico State University
{david,shelmrei}@crl.nmsu.edu

3: Information Sciences Institute
University of Southern California
hovy@isi.edu

4: Language Technologies Institute
Carnegie Mellon University
{lsl,teruko}@cs.cmu.edu

5: Mitre Corporation
freeder@mitre.org
keith@mitre.org

6: Department of Computer Science
Columbia University
{rambow,as372}@cs.columbia.edu

Abstract

This paper describes a multi-site project to annotate the interlingual content of six sizable bilingual parallel corpora. The project addresses several principal problems in parallel: specification of interlingua content and notation, development of reliable annotation methods, and evaluation of annotated corpora. As a by-product, a growing corpus of annotated texts is being produced, which may eventually be useful for machine learning of semantics-based processing.

1 Introduction

Experience with the Penn Treebank has shown the immense value for NLP of a large corpus of annotated material. The Treebank sentences are annotated with syntactic trees. In this paper, we describe work on the next step: the creation of a *semantic* representation system and the development of a corpus of semantically annotated text, validated in six languages, and evaluated in several ways.

This one-year NSF-funded project is joint between CMU (who focus on Japanese), NMSU (Spanish), University of Maryland (Arabic), Columbia University (Hindi), MITRE (French), and USC/ISI (Korean). [1] Since October 2003 we have established a distributed and well-functioning research methodology, designed an interlingua notation, created annotation manuals and tools, developed a text collection in six languages with associated translations into English, annotated some 150 translations, and designed and applied various annotation evaluation metrics.

In this paper we present the background and objectives of the project. We then describe the data set that is being annotated, the interlingual (IL) representation language being used, auxiliary IL representation resources such as the symbol ontology and the theta roles, the annotation tools we have built, and the process of annotation itself. We then outline a preliminary version of our evaluation methodology and conclude with a summary of some issues that have arisen.

2 Project Goals and Issues

The objective of this work is to improve NLP by providing a practical, tested notation that embodies linguistically motivated levels of semantic representation, together with a corpus of annotated texts annotated. If we can demonstrate the feasibility of this framework and annotation methodology, the stage will be set for creating very large semantically annotated corpora suitable for most modern machine learning techniques.

The central goals of the project are:

- to produce a practical, commonly-shared system for representing the information conveyed by a text, or *interlingua*;

[1] See http://aitc.aitcnet.org/nsf/iamtc/.

- to develop a methodology and tools for accurately and consistently assigning such representations to texts across languages and across annotators, as well as evaluating the consistency of the annotations;
- to annotate a sizable multilingual parallel corpus of foreign texts and translations for IL content.

The tools and annotation standards are designed to facilitate further annotation of texts in the future.

A somewhat unique aspect of this project is its focus on multiple translations of the same text. By comparing the annotations of the source text as well as its translations, any differences indicate one of three general problems: potential inadequacies in the interlingua, misunderstandings by the annotators, or mistranslations. By analyzing such differences, we can sharpen the interlingua definition, and/or improve the instructions to annotators, and/or identify the kinds of translational differences that occur (and decide what to do about them).

For example, even within the first two paragraphs of K1E1 and K1E2 (two English translations of Korean text K1), several such issues become apparent:

> **K1E1**: Starting on January 1 of next year, SK Telecom subscribers can switch to less expensive LG Telecom or KTF. … The Subscribers cannot switch again to another provider for the first 3 months, but they can cancel the switch in 14 days if they are not satisfied with services like voice quality.
>
> **K1E2**: Starting January 1st of next year, customers of SK Telecom can change their service company to LG Telecom or KTF … Once a service company swap has been made, customers are not allowed to change companies again within the first three months, although they can cancel the change anytime within 14 days if problems such as poor call quality are experienced.

First, if the interlingua term repository contains different terms for *subscriber* and *customer*, then the single Korean source term will have given rise to different interpretations. Here, we face a choice: we can ask annotators to explicitly search for near-synonyms and include them all, or we can compress the term repository to remove these near-synonyms. Second, K1E1 contains *less expensive*, which K1E2 omits altogether. This is probably a translator's oversight. Third, is *voice quality* the same as *call quality*? Certainly *voice* is not the same as *call*. What should the interlingua representation be here—should it focus merely on the poor *quality*, skirting the modifiers altogether?

The researchers on the project participate in weekly phone meetings to discuss such issues as they arise, and to continue to develop the annotation procedures, manuals, and tools.

3 Corpus and Annotation Schedule

The data set consists of corpora in the 6 languages: listed above. Each corpus contains some 125 source language news articles together with at least two independently produced high-quality translations into English. (There is no overlap among the articles across languages.) Typically, each article has between 300 and 400 words (or the equivalent); each corpus has around 120,000 words, and the size of the entire data set is around 750,000 words.

The Spanish, French, and Japanese corpora are based on the DARPA MT evaluation data (White and O'Connell 1994), and the Arabic data is corpus is based on LDC's Multiple Translation Arabic, Part 1 (Walker et al., 2003). The Hindi and Korean texts were obtained from online news corpora and translated carefully into high-quality English by people independent of the project.

For any given subcorpus, the annotation effort involves assignment of IL content to sets of at least 3 parallel texts, 2 of which are in English, and all of which theoretically communicate the same information. Such a multilingual parallel data set of source-language texts and English translations offers a unique perspective and unique problem for annotating texts for meaning.

We have designed a round-robin annotation schedule in which each site's two annotators annotate both translated texts from their own site, one annotator annotates the source language text from his or her own site, and the other annotates a translated text from some other site (see Figure 1). In this way, we can compare across text source and its translations, across the two translations alone, across a site's annotators, across different sites' annotators, and (when everyone annotates the same text) across all the annotators.

In our first production run, lasting 3 weeks, annotators annotated 144 texts (6 source texts x 2 translations x 2 annotators x 6 sites). It takes approximately 5 hours to annotate a text of 350

words. Annotators are not allowed to communicate with one another until after the annotation is complete at all sites, following which they discuss problems in a countrywide telephone meeting.

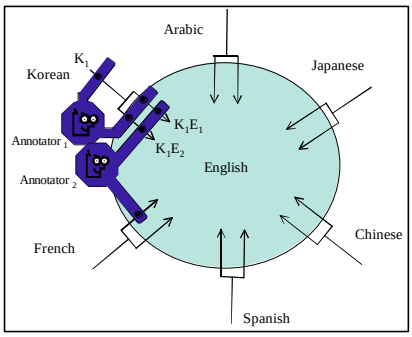

Figure 1. Annotator Rotation.

4 Interlingua

In order to ensure practicality and reliable annotation of the Interlingua, we are developing it in stages. For now, we omit many of the more complex phenomena, such as temporal and spatial notations, aspect, scoping of quantification, etc. Initially, we focus simply on annotating each major content-bearing word with its most appropriate semantic terms, and with linking together these terms using a small set of theta (case) roles.

We have developed a three-stage procedure for arriving at this admittedly still simple level of representation. Each stage of the representation moves progressively away from the surface and toward interlingual semantics. Should problems arise at any level, we can fall back to the previous level and start over. In this section we describe the current first three levels, *IL0*, *IL1*, and *IL2*, and then outline the IL symbol repository (the Omega ontology) and the theta roles used in the representations.

4.1 IL0

IL0 is a deep syntactic dependency representation. It includes part-of-speech tags for words and a parse tree that makes explicit the syntactic predicate-argument structure of verbs. The parse tree contains labels referring to deep-syntactic grammatical function (normalized for voice alternations). *IL0* does not contain function words (their contribution is represented as features) or semantically void punctuation. While this representation is purely syntactic, many disambiguation decisions, relative clause and PP attachment for example, have been made, and the presentation abstracts as much as possible from surface-syntactic phenomena. (Thus, our *IL0* is intermediate between the analytical and tectogrammatical levels of the Prague School (Hajicova et al., 2001).) *IL0* is constructed by hand-correcting the output of a dependency parser (see section 6), and allows annotators to see how textual units relate syntactically when making semantic judgments. Thus, it is a useful starting point for semantic annotation at *IL1*.

4.2 IL1

IL1 is an intermediate semantic representation. It associates semantic concepts with lexical units like nouns, adjectives, adverbs and verbs. It also replaces the syntactic relations in *IL0*, like *subject* and *object*, with thematic roles, like *agent, theme* and *goal*. Thus, like PropBank (Kingsbury et al., 2002), *IL1* neutralizes different alternations for argument realization. However, *IL1* is not an interlingua; it does not normalize over all linguistic realizations of the same semantics. In particular, it does not address how the meanings of individual lexical units combine to form the meaning of a phrase or clause. It also does not address idioms, metaphors and other non-literal uses of language. Further, *IL1* does not assign semantic features to prepositions; these continue to be encoded as syntactic features of their objects, which may be annotated with thematic roles such as *location* or *time.*

4.3 IL2

IL2 is intended to be a simple interlingua, a representation of meaning that is (reasonably) independent of language. *IL2* is intended to capture similarities in meaning across languages and across different lexical/syntactic realizations within a language. For example, like FrameNet (Baker et al., 1998), *IL2* is expected to normalize over conversives (e.g., X *bought a book from Y* vs. Y *sold a book to X)* and also over non-literal language usage (e.g., *X started its business* vs. *X opened its doors to customers*). The exact definition of *IL2* is a major research contribution of this project. However, as mentioned above, *IL2* does not include more complex phenomena such

as discourse structure, pragmatic readings (of words such as *unfortunately* and *hello*), speech acts, cross-event semantic relationships such as causality, etc. These remain for *IL3* and beyond, to be developed in subsequent projects.

4.4 The Omega Ontology

In progressing from *IL0* to *IL1*, annotators select semantic terms (concepts) to represent the nouns, verbs, adjectives, and adverbs present in each sentence. These terms are selected from the 110,000-node Omega ontology (Philpot et al., 2003), under construction at ISI. Omega has been built semi-automatically from a variety of sources, including Princeton's WordNet (Fellbaum, 1998), New Mexico State University's Mikrokosmos (Mahesh and Nirenburg, 1995), ISI's Penman Upper Model (Bateman et al., 1989) and ISI's SENSUS (Knight and Luk, 1994). The ontology, which has been used in several projects in recent years (Hovy et al., 2001), can be browsed using the DINO browser at htp://omega.isi.edu; this browser forms a part of the annotation environment. Omega continues to be developed.

4.5 The Theta Grids

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

The theta grids used in our project were extracted from the Lexical Conceptual Structure Verb Database (LVD) (Dorr, 2001). The WordNet senses assigned to each entry in the LVD link the theta grids to the verbs in the Omega ontology. In addition to the theta roles, the theta grids specify syntactic realization information, such as Subject, Object or Prepositional Phrase, and the Obligatory/Optional nature of the argument. The set of theta roles used, although based on research in LCS-based MT (Dorr, 1993; Habash et al., 2002), has been simplified for this project.

5 Annotation Tool and Manuals

We have assembled and/or developed several tools to be used in the annotation process. Since we are gathering our corpora from disparate sources, we need to standardize the text before presenting it to automated procedures. For English, this involves sentence boundary detection, but for other languages, it may involve word segmentation and other operations. The text is then processed with a dependency parser. For English texts, dependency trees are produced by Connexor (Tapanainen and Jarvinen, 1997) and manually corrected by one of the team researchers. Dependency trees are viewed and corrected in TrEd (Hajicova et al., 2001), a graphical tree editing program, written in Perl/Tk[2]. The revised deep dependency structure produced by this process is *IL0*. We next describe the annotation tool and manuals used for creating the three levels of representation.

5.1 Tiamat

To create *IL1*, annotators use Tiamat, a tool developed specifically for this project. This tool facilitates viewing of the *IL0* tree and browsing of Omega (which includes the theta grids). Its principal window displays the current sentence to be annotated (in context), the current word being worked on, all directly accessible Omega terms that might represent the word, and other information. Tiamat provides the ability to annotate text via simple point-and-click selections of words, concepts, and theta-roles. When the annotator selects a lexical item to be annotated, the relevant options within Omega are displayed.

For *IL2*, annotation involves selecting *all* relevant concepts from Omega—those originating from WordNet synsets and those from Mikrokosmos (these sources of information are intertwined in Omega). As described below, this duplication will help with a later decision about the optimal size of the ontology. In choosing a set of appropriate ontology concepts, annotators are encouraged to consider the name of the concept and its definition, the name and definition of the parent node, example sentences, lexical synonyms attached to the same node, and sub- and superclasses of the node.

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

All dependents of the selected word are automatically underlined in red in the sentence view. Annotators can view all information pertinent to the process of deciding on appropriate ontological concepts in this view. They can save decisions, undo them later, flag problematic cases for later inspection by one or more of the researchers assigned to the particular phenomenon (case roles, ontology problems, etc.).

Tiamat runs locally, on the annotator's own machine (on Unix and Windows platforms), and communicates with Omega, which runs centrally (on two servers at ISI) for global consistency.

5.2 Annotation Manuals

Markup instructions are contained in three manuals: a users' guide for Tiamat (including procedural instructions), a definitional guide to semantic roles, and an extensive manual for creating *IL0*. Together, these manuals allow the annotator to understand the intention behind aspects of the dependency structure; how to use Tiamat to mark up texts; and how to determine appropriate semantic roles and ontological concepts. The manuals are under constant revision, as new problem cases are encountered by annotators, discussed in the phone meetings, and brought to the attention of the relevant researcher.

6 Evaluation

Evaluation is of course a complex undertaking. We are still in the process of determining the most useful measures and procedures. A preliminary investigation of intercoder agreement on multiple texts indicates an improving trend over time as annotators learn more about and become more comfortable with the task. A complete study of intercoder agreement will be present in the final version of this paper.

We have developed these methods and tests:

Inter-translator consistency: We compare the two (or more) translations made of each text, listing the different choices for nouns, verbs, etc. We classify these for how they affected the semantic term choices of the annotators.

Inter-annotation reconciliation: After having made their own annotations, annotators see all the other selections made by annotators, and vote (privately) whether they find each of them acceptable or not. After a session of open annotator discussion they vote again (privately).

Inter-annotation agreement: The annotation decisions for each word and each theta role are recorded on a web page, tagged with the number of annotators that selected them. Figure 2 shows, that the sentence fragment "you can change your cell" elicited quite a lot of variation among the 10 annotators who addressed this sentence. In the column *Theta role*, for the first word "you", one annotator abstained ("1=---") and 9 annotators chose AGENT, and no-one chose either an Omega-WordNet or Omega-Mikrokosmos concept (annotators were instructed to skip pronouns).

The verb "change", however, elicited one abstention and 6 Omega-WordNet concepts as possible interpretations (including *alter*, *change<replace*, and so on, for a total of 17 votes

Word id	Word	Dep tree role	Theta role	Omega: WordNet concept	Omega: Mikro concept
50	you	Subj	1=---, 9=AGENT	10=---	10=---
60	can	Mod	1=MODIFIER, 1=NIL, 8=---	1=can==be_able_to<comma> _have_the_ability_to, 9=---	10=---
70	change	Root	10=---	1=---,1=alter, 1=change<replace, 2=change>gel, 3=change>add, 3=change>utilize, 7=switch>surf	1=---, 1=ADJUST$VERB, 1=DummyConcept, 8=CHANGE-EVENT$VERB
80	your	Mod	1=MODIFIER, 9=---	10=---	10=---
90	cell	Mod	1=MODIFIER, 9=---	1=DummyConcept, 1=DummyConcept:mobile_phone, 1=cell<compartment, 1=cell>fuel_cell, 2=---, 4=cell	5=---, 5=DummyConcept

Figure 2. Portion of integrated annotators' results, for evaluation

overall—annotators were allowed to choose more than one concept), as compared to 3 Omega-Mikrokosmos concepts.

We have developed several procedures and tools to compare annotations and to generate a series of evaluation measures. The reports generated by the evaluations allow the researchers to study both gross-level phenomena, such as inter-annotator agreement, and more detailed points of interest, such as lexical items on which agreement was particularly low, possibly indicating gaps or other inconsistencies in the ontology.

We have identified several metrics for evaluation of intercoder agreement on annotations. Two measures are Kappa (Carletta, 1993) and our own Wood Standard (Habash and Dorr, 2002). For expected agreement in the Kappa statistic, *P(E)* is defined as *1/(N+1)*, where *N* is the number of choices at a given data point. In the case of Omega nodes, this means the number of matched Omega nodes (by string match) plus one for the possibility of the annotator traversing up or down the hierarchy. The Wood Standard is the category chosen by the most annotators. In cases of no agreement, a random selection is picked from the annotators' selections. Multiple measures are used because it is important to have a mechanism for evaluating inter-coder consistency in the use of the IL representation language that does not depend on the assumption that there is a single correct annotation of a given text.

In addition to intercoder agreement, we are also designing an extrinsic measure of the quality of an IL expression derived by annotation. Given the project goal of generating an IL representation useful for MT (among other NLP tasks), we measure the ability to generate accurate surface texts from the IL representation. At this stage, we plan to use an available generator, Halogen (Knight and Langkilde, 2000). A tool to convert the representation to meet Halogen's requirements is being built. Following the conversion, surface forms will be generated and compared with the originals through a variety of standard MT metrics (ISLE, 2003). This will indicate whether the IL representation elements are sufficiently well-defined to serve as a basis for inferring interpretations from semantic representations.

7 Issues

We have encountered a number of difficult issues for which we have only interim solutions. Principal among these is the granularity of the IL terms to be used. Omega's WordNet symbols, some 100,000, afford too many alternatives with too little clear semantic distinction, resulting in large inter-annotator disagreement. On the other hand, Omega-Mikrokosmos, containing only 6,000 concepts, is too limited to capture many of the distinctions people deem relevant. We plan to manually prune out the extraneous terms from Omega. Similarly, the theta roles in some cases appear hard 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 (e.g., Ferro et al., 2001); (2) causality, co-reference, aspectual content, modality, speech acts, etc; (3) reducing vagueness and redundancy in the annotation language; (4) inter-event relations such as entity reference, time reference, place reference, causal relationships, associative relationships, etc; (5) and finally, cross-sentence phenomena remain a challenge.

8 Conclusions

The scientific interest of this research lies in the definition and annotation feasibility testing 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, no such level of representation has been defined complete with an associated annotated corpus of any size. As a result, 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 cannot construct an annotated corpus large enough for heavy-duty machine learning algorithms, we see our work as paving the way and developing solutions to the representational problems and thereby enable other, larger annotation efforts.

References

Baker, C., C. Fillmore, and J.B. Lowe, 1998. The Berkeley FrameNet Project. *Proceedings of the ACL Conference*.

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

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

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

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

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

Farwell, D. and S. Helmreich. 2003. Pragmatics-based Translation and MT Evaluation. *Proceedings of Towards Systematizing MT Evaluation. MT-Summit Workshop*. New Orleans, LA.

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

Ferro, L., I. Mani, B. Sundheim and G. Wilson. 2001. TIDES Temporal Annotation Guidelines. Version 1.0.2 MITRE Technical Report, MTR 01W0000041

Fillmore, C. 1968. The Case for Case. In E. Bach and R. Harms (editors), *Universals in Linguistic Theory*, 1–88. Holt, Rinehart, and Winston.

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

Habash, N. and B.J. Dorr. 2002. Interlingua Annotation Experiment Results. *Proceedings of the AMTA-2002 Interlingua Reliability Workshop*. Tiburon, CA.

Habash, N., B.J. Dorr, and D. Traum, 2002. Efficient Language Independent Generation from Lexical Conceptual Structures. *Machine Translation* 17:4.

Hajicova, J., B. Vidová-Hladká, and P. Pajas. 2001. The Prague Dependency Treebank: Annotation Structure and Support. *Proceedings of the IRCS Workshop on Linguistic Databases*. University of Pennsylvania, 105–114.

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

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

Kingsbury, P., M. Palmer, and M. Marcus. 2002. Adding Semantic Annotation to the Penn TreeBank. *Proceedings of the Human Language Technology Conference (HLT 2002).*

Knight, K., and I. Langkilde. 2000. Preserving Ambiguities in Generation via Automata Intersection. *Proceedings of AAAI.*

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

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

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

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

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

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

White, J. and T. O'Connell. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. *Proceedings of the AMTA Conference.*

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