# Natural Language Processing and Machine Translation
# Encyclopedia of Language and Linguistics, 2nd ed. (ELL2).
# Machine Translation: Interlingual Methods

**Bonnie J. Dorr**
Department of Computer Science and UMIACS
A.V. Williams Building
University of Maryland
College Park, MD 20742

**Eduard H. Hovy**
Information Sciences Institute
of the University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292

**Lori S. Levin**
Language Technologies Institute
Newell-Simon Hall
Carnegie Mellon University
Pittsburgh, PA 15213

**Keywords:** interlingua, machine translation, language independent representation, cross-language divergences, ontology, conceptual knowledge, thematic roles, lexical-conceptual structure, text-meaning representations, predicate-argument structure, semantic frames, semantic zones, compositionality, interlingual speech translation, approximate interlingua, semantic annotation

## Abstract

An interlingua is a notation for representing the content of a text that abstracts away from the characteristics of the language itself and focuses on the meaning (semantics) alone. Interlinguas are typically used as pivot representations in machine translation, allowing the contents of a source text to be generated in many different target languages. Due to the complexities involved, few interlinguas are more than demonstration prototypes, and only one has been used in a commercial MT system. In this article we define the components of an interlingua and the principal issues faced by designers and builders of interlinguas and interlingua MT systems, illustrating with examples from operational systems and research prototypes. We discuss current efforts to annotate texts with interlingua-based information.

## 1 Introduction

As described in the section on *Machine Translation Overview*, machine translation methodologies are commonly categorized as direct, transfer, and interlingual. The methodologies differ in the depth of analysis of the source language and the extent to which they attempt to reach a language-independent representation of meaning or intent between the source and target languages. Interlingual MT typically involves the deepest analysis of the source language.

Figure 1—the Vauquois triangle (Vauquois, 1968)—illustrates these levels of analysis. Starting with the shallowest level at the bottom, direct transfer is made at the word level. Moving upward through syntactic and semantic transfer approaches, the translation occurs on representations of the source sentence structure and meaning respectively. Finally, at the interlingual level, the notion of transfer is replaced with a single underlying representation—the Interlingua—that represents both the source and target texts simultaneously. Moving up the triangle reduces the amount of work required to traverse the gap between languages, at the

cost of increasing the required amount of analysis (to convert the source input into a suitable pre-transfer representation) and synthesis (to convert the post-transfer representation into the final target surface form). For example, at the base of the triangle, languages can differ significantly in word order, requiring many permutations to achieve a good translation. However, a syntactic dependency structure expressing the source text may be converted more easily into a dependency structure for the target equivalent because the grammatical relations (subject, object, modifier) may be shared despite word order differences. Going further, a semantic representation (interlingua) for the source language may totally abstract away from the syntax of the language, so that it can be used as the basis for the target language sentence without change.

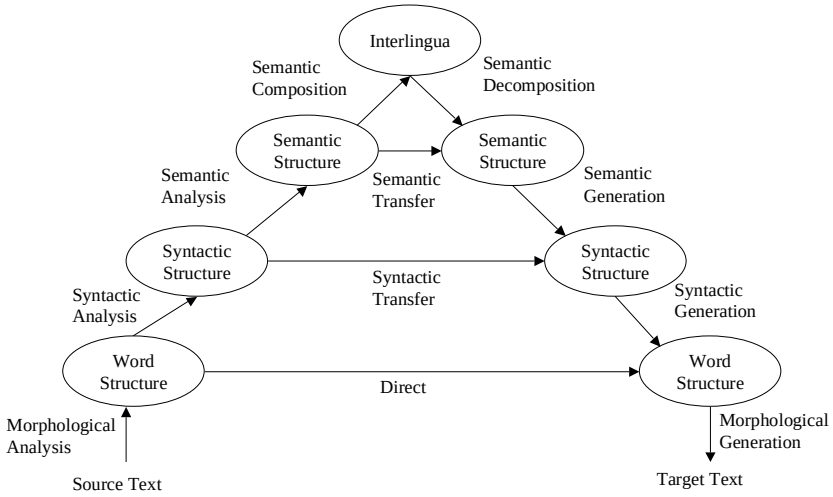


Figure 1: The Vauquois Triangle for MT

Comparing the effort required to move up and down the sides of the triangle to the effort to perform transfer, interlingual MT may be more desirable in some situations than in others. Because in principle an interlingual representation of a sentence contains sufficient information to allow generation in *any* language, the more (and the more different) target languages there are, the more valuable an interlingua becomes. To translate from one source into *N* target languages, one needs *(1+N)* steps using an interlingua compared to *N* steps of transfer (one to each target). But to translate pairwise among all the languages, one needs only *2N* steps using an interlingua compared to about $N^2$ with transfer—a significant reduction for the former. In addition, since in theory it is not necessary to consider the properties of any other language during the analysis of the source language or generation of the target language, each analyzer and generator can be built independently by a monolingual development team. Each system developer only needs to be familiar with his/her language and the interlingua.

Another advantage of the interlingua approach is that interlingual representations can be used by NLP systems for other multilingual applications, such as cross-lingual information retrieval, summarization, and question answering (see Figure 2). For example, it is a basic assumption of the semantic web that webpages will contain not only source text but also some interlingual representations thereof, against which queries issued in other languages and translated into the interlingua can be matched, and from which various target-language versions of the webpages can be generated. In all of these applications, there is a reduction in computation over approaches that tailor the underlying representation to the idiosyncrasies of each of the input/output languages. Absent an interlingual representation, all these multilingual applications require the insertion of a translation step at least once and often in two different places.

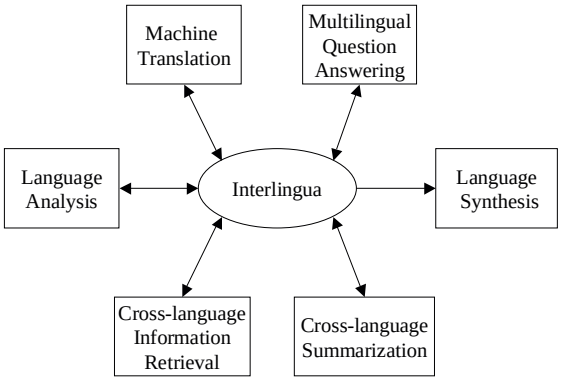


**Figure 2: Use of Interlingua in Multiple Applications**

Although interlinguas are a topic of recurring interest, only one interlingual MT system has ever been made operational in a commercial setting—KANT (Nyberg and Mitamura, 1992, 2000; Lonsdale et al., 1995)—and only a handful have actually been taken beyond the stage of research prototype. Interesting research prototypes are Pangloss (Frederking et al., 1994); CICC (CICC, 2003); NESPOLE! (http://nespole.itc.it, Lavie et al., 2001; Lazzari, 2000), and ChinMT (Habash et al., 2003a).

## 2 Interlingua Definition and Components

An *Interlingua* is a system for representing the meanings and communicative intentions of language. It can be defined as a triple (S,N,L) where

- S is a collection of representation symbols, often defined in an ontology, where each symbol denotes a particular aspect of meaning or intention (sometimes individually, and sometimes in concert with others according to specific rules of combination).
- N is a notation, within which symbols can be composed into meanings. The rules governing notational well-formedness reflect the compositional derivation of complex meaning out of 'atomic' symbols, an operation that is basic to the theory of the Interlingua.
- L is a lexicon, namely a collection of words of a human language such as English, in which each lexical element is associated directly or indirectly with one or more symbols from S. Interlingual MT systems typically include one lexicon for each language.

An *interlingua instance* is the representation of the meaning of a given fragment of text, such as a clause, sentence, or document. Such an instance is often written as a list of interconnected nested frame structures, where each proposition in the frame represents some 'atomic' component of the total meaning.

Details and examples of each of these components follow.

### 2.1 Representation Symbols

Typically, an interlingua comprises several kinds of symbols to represent meaning. The largest set can be thought of as the conceptual primitives; rather like the open-class words in a human language, these symbols stand for specific types of objects, events, relations, qualities, etc. Other, smaller, sets of symbols are defined to represent specific fields of meaning, and usually derive from a logical theory about the nature of some phenomenon. For example, the linguistic system of tense can be related to a theory of time, and time can be represented in an Interlingua according to a highly formalized subsystem; see (Reichenbach, 1947; Allen, 1984). Other typical subfields of meaning represent space (Hayes, 1985), causality (Hobbs, 2001), the epistemic status of events (actual, hypothetical, desired, etc.), etc.

These symbols are often arranged as taxonomies in which each node stands for a symbol, and information stored at higher-level symbols is inherited downwards and shared by lower ones. The contents and structure of the taxonomy thereby embody, to some degree, the Interlingua designer's conceptualization of

the world, making the taxonomy an ontology in the classical sense. Although ontologies are as old as Aristotle and are most commonly used in Artificial Intelligence systems to support complex reasoning, interlingua ontologies form a distinct type: they are generally large (comprising several thousands of symbols), contain relatively little information per symbol, and what information is contained is primarily devoted to interlingua instance composition or linguistic behavior instead of to inference.

It is not uncommon for an interlingual MT system to contain both an upper-level, very general, ontology and then one or more specific domain-oriented ones. The upper ontology contains notions that are shared over all domains in common language; the lower ones encode distinct subworlds, such as finances, sports, chemistry, etc. Usually, the higher-level symbols represent conceptual and linguistic abstractions for which there are no words, and the lower-level ones more concrete meanings for which words exist in the various languages' lexicons. (For example, the Penman Upper Model contains the nodes *NonDecomposableObject* and *DecomposableObject* to separate mass and count nouns.) One advantage of domain partitioning is ambiguity avoidance: the term "bond" in the financial domain has only one meaning, and in the chemistry domain another, enabling the MT system to proceed more expeditiously in each domain.

Ontologies developed for MT include ONTOS (ONTOS, 1989), SENSUS (Knight and Luk, 1994), and Mikrokosmos/OntoSem (Mahesh and Nirenburg, 1995; McShane et al., 2004; Nirenburg and Raskin, 2004). Ontologies developed and used for language technology applications in general include WordNet (Fellbaum, 1998), the Penman Upper Model (Bateman et al., 1989), and Omega (Philpot et al., 2003). Omega can be browsed using the DINO browser at http://blombos.isi.edu:8000/dino.

### 2.2 Notation

The notation is the vehicle by which the symbols' individual shades of meaning are assembled into a complex meaning. The notation is usually instantiated as a network of propositions represented as a set of nested frames, where each proposition employs the symbols of the interlingua, composed according to the specifications of the interlingua in general and of the symbols in particular.

Typically, a frame has a frame header, which may include a frame identifier, and one or more propositions, each being a relation-value pair that links the frame header to the value via the relation. Figure 3 provides an example from the KANT system, representing the meaning of *If the error persists, service is required*. The frame headers—each marked with an asterisk (*)—of the two clauses are BE-PREDICATE and QUALIFYING-EVENT. BE-PREDICATE has two arguments, an attribute and a theme. Each of these is headed by another frame, REQUIRED and SERVICE, respectively. The QUALIFYING-EVENT has a PERSIST event whose theme is ERROR.

```
(*BE-PREDICATE
  (attribute
       (*REQUIRED
          (degree positive)))
  (mood declarative)
  (predicate-role attribute)
  (punctuation period)
  (qualification
        (*QUALIFYING-EVENT
            (event
               (*PERSIST
                    (argument-class theme)
                    (mood declarative)
                    (tense present)
                    (theme
                        (*ERROR
                            (number  (:OR mass singular))
                            (reference definite)))))
             (extent   (*CONJ-if))
             (topic +)))
    (tense present)
    (theme
         (*SERVICE
             (number  (:OR mass singular))
             (reference no-reference))))
```

**Figure 3: KANT Representation of *If the error persists, service is required.***

In some sophisticated interlinguas, the notation contains separate zones for different kinds of meaning (Nirenburg et al., 1995); typically a zone for world semantics (the conceptual content of the text), a zone for interpersonal semantics (information in the text reflecting the writer, reader, their relationship, etc., which often affects the style of the text rather than the content), and a zone for meta-textual information (medium, such as spoken or written; genre, such as telegram, letter, report, article; situation, such as anonymous posting, personal delivery, etc.).

### 2.3 Lexicon

An interlingua lexicon includes information about the nature and behavior of each word in the language. For example, events and actions (usually expressed as verbs) include information about their preferred arguments (agents, patients, instruments, etc.). In some interlinguas, this information may reflect the verbal predilections of one language more than another; for example, "I swim across the river" is expressed in Spanish as "I cross the river swimmingly". Should the interlingual representation be anchored on "swim" or "cross"? The choice rests with the interlingua symbol set designer. To the degree such asymmetries in the interlingua prefer one language over another, it is said to deviate from true language-neutrality. A representation system reflecting one language closely is often called 'shallow semantics'.

Within a chosen representation system, the concepts on which events are anchored are called *Predicates* and the participants in the event are called *Arguments* following the formalism used in logical representations used in Artificial Intelligence systems. Predicate-argument structure (Grimshaw, 1992; Hale and Keyser, 2002) refers to the combination of an event concept and its participants—and a given predicate is said to have a certain number of potential participants—or *valency*. For example, the verb *load* has a valency of 3: the person doing the loading, the item that is loaded, and the place that the item is loaded.

Semantic roles—often called thematic roles—are by far the most common approach to represent the arguments of a predicate semantically. However, the numerous variant theories display little agreement

even on terminology (Fillmore, 1968; Foley and Van Valin, 1984; Jackendoff, 1972; Levin and Rappaport-Hovav, 1998; Stowell, 1981). A small set of examples is shown in Table 1. The reader is referred to the sections on *Logical and Lexical Semantics* for a more comprehensive set of examples.

| Role | Definition | Example |
|---|---|---|
| AGENT | An Agent should have the features of volition (able to make a conscious choice), sentience (having perception), causation (able to bring about an effect) and independent existence (existence not resulting from the action). | **John** *broke* the vase. |
| THEME | The Theme is causally affected, or is in a state or changes state, or is in a location or changes location, or comes into or out of existence. | John *broke* **the vase**. |
| INSTR | The Instrument has causation but no volition. Typically, an instrument appears with an agent and can be paraphrased with "using." | John *broke* the vase **with a hammer** |

**Table 1: Examples of Semantic Roles**

A number of Interlingua researchers have used semantic roles for interlingual MT (Dorr, 2001; Habash and Dorr, 2002; Nyberg and Mitamura 1992, 2000). More details are given in Section 4.

## 3 Issues in Interlingua

The notion of Interlingua appeals to many, but is a complex undertaking. In this section we examine the issues faced by designers of interlinguas and interlingual MT systems.

### 3.1 Problems with Representing Meaning

Probably the central problem of interlingua design is the complexity of "meaning". A great deal has been written about interlinguas, but no clear methodology exists for determining exactly how one should build a true language-neutral meaning representation, if such a thing is possible at all (Whorf, 1959; Nirenburg and Raskin, 2004; Hovy and Nirenburg, 1992; Dorr, 1994). It is always possible to add more detail to a meaning representation, but in order to implement an MT system, the details must end at some point. To date no adequate criteria have been found for deciding when to stop refining the meaning representation, although some preliminary attempts have been made in the NESPOLE! project (Levin et al., 2002, 2003) and in the IAMTC project (Section 5.2 below).

A basic design choice is granularity: the number of interlingual representation primitives. The parsimonious approach, exemplified by Conceptual Dependency (Schank and Abelson, 1977), declares that a small number of primitives are enough to compositionally represent all actions. This poses a daunting problem of meaning assembly that has never been seriously attempted. In contrast, the profligate approach, called 'Ontological promiscuity' (Hobbs, 1985), essentially allows a representation symbol for every shade of meaning (and certainly one for each lexical item). This poses a problem of representing the essential relatedness of notions such as *buy* and *sell*, *come* and *go*, etc. The ideal seems to have been to aim somewhere in between, seeking conceptual depth and coverage simultaneously. Many researchers (Nirenburg and Raskin, 2004) develop a deep semantic analysis that requires extensive world knowledge; the performance of deep semantic analysis (if required) depends on the (so far unproven) feasibility of representing, collecting, and efficiently storing large amounts of world and domain knowledge. This problem consumes extensive efforts in the broader field of Artificial Intelligence (Lenat, 1995).

We present an example. What, principally, are the primitive concepts of the meaning representation for *eat*? Do we also need more specific primitives like *eat-politely* and *eat-like-a-pig*? This distinction is required to distinguish between the verbs *essen* and *fressen* in German. In general, two strategies are possible (Hovy and Nirenburg, 1992). One is to adopt arbitrarily the conceptualizations of one language, and specify the variations of all others in terms thereof; the other is to multiply out all the distinctions found in any language. In the latter case one will obtain two interlingual items representing *eat* (because of German) and two for the object *fish* (because of the distinction between *pez* and *pescado* in Spanish). The situation worsens; in Japanese translation of the verb *wear* depends on where the object is worn, e.g., *head* or *hands*.

Ontologies greatly support the profligate approach, because they allow one to concisely represent systematic relationships between groups of concepts. However, building an ontology remains a problem. For example, the WordNet-based component of the Omega ontology (Philpot et al., 2003) mentioned above contains 110,000 nodes and often provides too many indistinguishable alternatives, whereas the Mikrokosmos-based component of Omega contains only 6,000 concepts and does not offer all the concepts needed to represent the full meaning of a word. Thus the word *extremely* contains four concepts in WordNet-based Omega, and sense is hard to distinguish from the others: (1) to a high degree or extent, favorably or with much respect; (2) to an extreme degree; (3) to an extreme degree, super; (4) to an extreme degree or extent, exceedingly. On the other hand, the Mikrokosmos-based part of Omega does not contain even one concept for the word *extremely*.

Another issue raised with respect to Interlinguas is that, because this representation is purportedly independent of the syntax of the source text, the target text generated reads more like a paraphrase than a strict translation (Arnold and des Tombe, 1987; Hutchins, 1987; Johnson et al., 1985). That is, the style and emphasis of the original text are lost. However, this is not so much a failure of the Interlingua as its incompleteness, caused by a lack of understanding of the discourse and pragmatics required to recognize and appropriately reproduce style and emphasis. In fact, in some cases it may be an advantage to ignore the author's style. Moreover, many have argued that, outside the field of artistic texts (poetry and fiction), preservation of the syntactic form of the source text in translation is completely superfluous (Goodman and Nirenburg, 1991; Whitelock, 1989). For example, the passive voice constructions in the two languages may not convey identical meanings. Taken overall, the current state of the art seems to confirm that it is possible to produce interlinguas that are reliably adequate between language groups (e.g., Japanese and Western European) for specialized domains only.

### 3.2 Divergences

An important problem addressed by interlingua approaches is that of structural differences between languages—*language divergences*—e.g., English *fear* vs. Spanish *tener miedo de*. Some examples from (Dorr et al., 2002) are:

- Categorial Divergence: The translation of words in one language into words that have different parts of speech in another language. For example, *to be jealous — tener celos (to have jealousy).*
- Conflational Divergence: The translation of two or more words in one language into one word in another language. Examples include *to kick — dar una patada (give a kick).*
- Structural Divergence: The realization of verb arguments in different syntactic configurations in different languages. For example, *to enter the house — entrar en la casa (enter in the house).*
- Head Swapping Divergence: The inversion of a structural dominance relation between two semantically equivalent words when translating from one language to another. For example, *to run in — entrar corriendo (enter running).*
- Thematic Divergence: The realization of verb arguments in syntactic configurations that reflect different thematic to syntactic mapping orders. For example, *I like grapes — me gustan uvas (to-me please grapes)*

Many divergences are caused by differences in language typology. For example, many verb serializing languages express the benefactive (e.g., write a letter *for me*) in a serial verb constructions (e.g., write letter *give me*). Some types of meaning are particularly susceptible to divergences. In English, sentences expressing the speech act of suggesting (*How about going* to the conference?, *Why not go* to the conference?) cannot be translated literally into most other languages. Divergences are also common in expressions of modality. For example, the expression of deontic modality in *You had better go* in English can be expressed in Japanese roughly as *Itta hoo ga ii*, literally *go(past form) way/option/alternative subj-marker good* or *(the) option (of) going (is) good*. Some authors have argued that divergences may be the norm rather than the exception (Levin and Nirenburg, 1994).

Resolution of cross-language divergences is an area where the differences in MT architecture are most crucial. Many MT approaches resolve such divergences by means of construction-specific rules that map from the predicate-argument structure of one language into that of another. The interlingua approach to MT takes advantage of the compositionality of basic units of meaning to resolve divergences. For example, the conflational divergence above is resolved by mapping English "kick" into two components, the motional component (movement of the leg) and the manner (a kicking motion) before translating into a language like Spanish.

## 4 Interlinguas in Machine Translation

A typical interlingual system is illustrated schematically in Figure 4. Each language requires an analyzer and a synthesizer. The analyzer takes as input a source language sentence and produces as output an interlingual representation of the meaning. The synthesizer takes an interlingual representation of meaning as input and produces one or more sentences with that meaning. In theory, it is not necessary to consider the properties of another language during the analysis of the source language or generation of the target language. To translate from language $L_1$ to $L_2$, $L_1$'s analyzer produces an interlingual representation and $L_2$'s synthesizer generates an $L_2$ sentence with the same meaning.


**Figure 4: Interlingual MT System Architecture**

Below we illustrate several representative examples of interlingual representations used by developers of interlingual MT systems.

### 4.1 Pangloss

The Pangloss project (Frederking et al., 1994) started as an ambitious attempt to build rich interlingual expressions using humans to augment system analysis. As shown in Figure 5, the representation includes a set of frames representing semantic components (each headed by a unique identifier such as %*proposition_5*) and a separate frame with aspectual information (see %*aspect_5* at bottom) representing duration, telicity, etc. Some modifiers are treated as scalars and represented by numerical values; the phrase "active expansion" is represented in %*expand*_1 with an *intensity* of 0.75 (out of 1.0). Note also that all implicit arguments (for instance, the *agent* of %*expand*_1) are explicitly included.

```
%proposition_5
    head            %pursue_1
    time            %time_5
    aspect          %aspect_5
    polarity        positive

%pursue_1
    agent           %company_4
    theme           %policy_1
    purpose         %set-up_1
    means           %tie-up_2          ;coreference to %tie-up_1

%company_4
    name            $"Sezon Group"     ;coreference to %company_3

%policy_1
    policy-type     %expand_1

%expand_1
    agent           %pursue_1.agent
    destination     %overseas
    intensity       0.75

%tie-up_2
    tie-up-partner  %company_5         ;coreference to %company_2

%aspect_5
    phase           continue
    iteration       once
    duration        prolonged
    telicity        false
```

**Figure 5: Pangloss Interlingual Representation of *The Sezon Group will pursue an active overseas expansion policy by means of the tie-up with SAS***

### 4.2 Mikrokosmos/OntoSem

The focus of the Mikrokosmos project (Mahesh and Nirenburg, 1995)—more recently dubbed OntoSem (Nirenburg and Raskin, 2004) —is to produce semantically rich Text-Meaning Representations (TMRs) of unrestricted text that can be used in a wide variety of applications, including as an interlingua for MT. These representations provide the basis for addressing some of the most difficult problems of NLP, such as disambiguation and all aspects of reference resolution, from reconstructing elliptical utterances to linking textual referents to their real-world "anchors" in a fact repository.

TMRs (Ontosem's interlingua expressions) use a language-independent metalanguage compatible with that used to represent the underlying static knowledge resources—the ontology and ontologically-linked lexicons. A sample TMR for the input *He asked the UN to authorize the war*, is as shown in Figure 6. (Small caps indicate ontological concepts; the indices represent numbered instances of ontological concepts in the world model built up during this run of the system.)

| | |
|---|---|
| **REQUEST-ACTION-69** | |
| AGENT | HUMAN-72 |
| THEME | ACCEPT-70 |
| BENEFICIARY | ORGANIZATION-71 |
| SOURCE-ROOT-WORD | ask |
| TIME | (< (FIND-ANCHOR-TIME)) |
| **ACCEPT-70** | |
| THEME | WAR-73 |
| THEME-OF | REQUEST-ACTION-69 |
| SOURCE-ROOT-WORD | authorize |
| **ORGANIZATION-71** | |
| HAS-NAME | UNITED-NATIONS |
| BENEFICIARY-OF | REQUEST-ACTION-69 |
| SOURCE-ROOT-WORD | UN |
| **HUMAN-72** | |
| HAS-NAME | COLIN POWELL |
| AGENT-OF | REQUEST-ACTION-69 |
| SOURCE-ROOT-WORD | he ; *ref. resolution has been carried out* |
| **WAR-73** | |
| THEME-OF | ACCEPT-70 |
| SOURCE-ROOT-WORD | war |

**Figure 6: OntoSem Interlingual Representation of *He asked the UN to authorize the war***

This says that the word *ask* instantiates the 69th instance of the concept REQUEST-ACTION, whose agent is HUMAN-72 (the instantiation of *he*, which was resolved as *Colin Powell* using reference resolution procedures), whose beneficiary is ORGANIZATION-71 (the instantiation of UN, which was resolved to United-Nations using reference resolution procedures), and whose theme is ACCEPT-70 (the instantiation of 'authorize', whose theme is WAR-73 – the semantic representation of the meaning of the word *war*). One goal of recent work in the OntoSem environment has been to create TMRs for large amounts of text, populate a fact repository using a subset of information from the TMRs, and then use the fact repository as a language-independent search space for applications like question-answering and knowledge extraction.

### 4.3 JapanGloss

The Interlingua notation developed for the Japangloss MT system (Knight et al., 1995) and the Nitrogen sentence generator (Knight and Langkilde, 2000) used symbols from the SENSUS ontology (Knight and Luk, 1994), one of the precursors of Omega. In this notation, frame identifiers are symbols like *h1* and SENSUS symbols are delimited by bars; and in contrast to many other Interlinguas, modality predicates (e.g., likelihood and necessity) are represented as frame predicates, the same way other, normal, actions and events are. Thus in the example given in Figure 7, which represents "It is possible that you must eat chicken" (equivalently, "You might have to eat chicken"), *e4* is the eating by you of the chicken, which by *h2* is obligatory, which in turn by *h1* is possible.

```
(h1 / |possible<latent|
    :domain (h2 / |obligatory<necessary|
                :domain (e4 / |eat,take in|
                            :agent you
                            :patient (c1 / |poulet|))))
```

**Figure 7: Japangloss Interlingual Representation of *It is possible that you must eat chicken* or *You might have to eat chicken***

### 4.4 **KANT**

KANT is the only interlingual MT system that has ever been made operational in a commercial setting. The KANT system (Nyberg and Mitamura, 1992) is a knowledge-based, interlingual machine translation system. KANT is designed for translation of technical documents written in Controlled English to multiple target languages. The KANT Analyzer produces an interlingua expression for each sentence in the input document; an example appeared earlier in Figure 3. This interlingua is mapped into an appropriate target sentence by the KANT Generator. For each target language there is a separate lexicon and grammar.

The KANT system was integrated with the ClearCheck document checking interface (built by Carnegie Group) and deployed in the Caterpillar document workflow during the middle 90's. The work for Caterpillar involved development of a Caterpillar Technical English (CTE), a corresponding KANT Analyzer, and KANT Generators for French, Spanish and German. The system delivered to Caterpillar represents the first large-scale deployment of controlled language checking integrated with machine translation. The interlingua used in the KANT system is based on research on the generation of additional target languages, such as Portuguese, Italian, Russian, Chinese and Turkish.

### 4.5 **Interlingual Systems for MT of Spoken Language**

The interlingua approach to machine translation has been implemented in several demos and prototypes for translation of spoken language. MT for spoken language begins with speech recognition. The output of the speech recognizer is then passed to the source language analysis module of the MT system. In addition to the problems faced by MT for text, MT for spoken language must deal with disfluencies in speech and imperfect output from a speech recognizer. For this reason, most spoken language MT systems are restricted to task-oriented domains such as travel planning or doctor-patient interviews.

Interlinguas for spoken, task-oriented dialogue typically focus on the dialogue act that the speaker intends to accomplish with his/her utterance. Examples of dialogue acts are suggesting, accepting, rejecting a time or price. In interlinguas for spoken language, less emphasis is placed on predicate argument structure, with the exception of (Lee et al., 2002). The emphasis on speaker intent means that the same interlingual representation will be used for sentences that have very different syntactic structures. For example, the following sentences all carry out the dialogue act of giving information about the price of a room. The concept of costing is expressed by the verb (*cost*) in the first sentence, and the subject (*price*) in the second sentence. In the third sentence, the concept of costing is implicit in the predicate nominal (*one hundred dollars*).

The room costs one hundred dollars per night.
The price of the room is one hundred dollars per night.
The room is one hundred dollars per night.

The JANUS system was the earliest spoken language MT system using an interlingua in the early 1990's (Waibel, 1996). JANUS was part of the C-STAR consortium (Consortium for Speech Translation Advanced Research), many of whose members adopted the interlingua approach for an international demo in 1999 (Woszczyna, 1998; Levin et al., 2000). Other interlingual speech translation systems include Enthusiast (Qu et al., 1997), CCLINC (Lee et al., 2002), MARS (Gao et al., 2002), NESPOLE! http://nespole.itc.it, Speechalator (Waibel et al., 2003), Carnegie Mellon University's Thai speech translation system (Schultz et al., 2004), and FAME http://isl.ira.uka.de/fame/index.html.

Figure 8 provides an example from the NESPOLE! project, in which both sentences are represented by the given interlingua instance. The NESPOLE! interlingua is based on an annotated corpus of transcribed dialogues in English, German, Italian, and Japanese. It has also been applied to Chinese, Spanish, and French. Its precursor, the C-STAR interlingua, has also been applied to Korean.

*"I want to know what time the flight leaves Pittsburgh."*
*"What time does the flight leave Pittsburgh?"*

request-information+departure
(time = (clock = question),
transportation-spec = (flight, id = yes),
origin = name=Pittsburgh)

**Figure 8: Two Sentences and Corresponding Interlingua Instance from the NESPOLE! Project**

### 4.6 Universal Networking Language

The Universal Networking Language (UNL) is a formal language designed for rendering automatic multilingual information exchange (Martins et al., 2000). It is intended to be a cross-linguistic semantic representation of sentence meaning consisting of concepts (e.g., "cat", "sit", "on", or "mat"), concept relations (e.g., "agent", "place", or "object"), and concept predicates (e.g., "past" or "definite"). The UNL syntax supports the representation of a hypergraph whose nodes represent "Universal Words" and whose arcs represent "Relation Labels". An example is shown in Figure 9 for the sentence *The cat sat on the mat.*

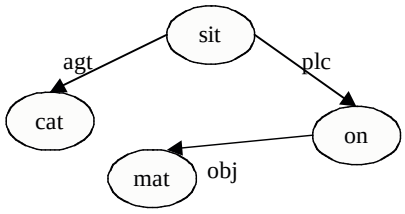


**Figure 9: UNL Representation of *The cat sat on the mat***

Several semantic relationships hold between universal words (synonymy, antonymy, hyponymy, hypernymy, meronymy, etc.) which compose the UNL Ontology.

### 4.7 Lexical Conceptual Structure

Lexical Conceptual Structure (LCS) is an interlingual representation used as part of a Chinese-English Machine Translation (MT) system, called ChinMT (Habash et al., 2003a) that has also been used for many other MT language pairs (e.g., Spanish and Arabic) and other natural language applications (e.g., cross-language information retrieval). The LCS-based approach focuses on the types of divergences described in Section 3.2. Consider, for example, the case of a conflational divergence between Arabic and English:

Arabic: ارسل الصحفي ايميلا الى محطة الجزيرة
Gloss: 'The-reporter sent email to Al-Jazeera.'
English: The reporter emailed Al-Jazeera.

The LCS representation for this example is shown in Figure 10, glossed as "The reporter caused the email to go to Al-Jazeera in a sending manner". Here, the primary components of meaning are the top-level conceptual nodes *cause* and *go*. These are taken together with their arguments, each identified by a semantic role (agent, theme, and goal), and a modifier (manner) *send+ingly*.

```
(event cause
    (thing[agent] reporter+)
    (go loc
        (thing[theme] email+)
        (path to loc
            (thing email+)
            (position at loc (thing email+) (thing[goal] aljazeera+)))
        (manner send+ingly)))
```

**Figure 10: LCS Representation of *The reporter emailed Al-Jazeera***

### 4.8 Approximate Interlingua

One response to the MT divergence problem (discussed in Section 3.2) is the use of an *approximate interlingua* (Dorr and Habash, 2002). In this approach, the depth of knowledge-based systems is approximated by tapping into the richness of resources in one language (often English) and this information is used to map the source-language (SL) input to the target-language (TL) output.

The focus of the approximate-interlingua approach is to address the types of divergences covered by the LCS-based approach, but with fewer knowledge-intensive components. Thus, a key feature of an approximate interlingua is the coupling of basic argument-structure information with some, but not all, components the LCS representation. Only the top-level primitives and semantic roles are retained. This new representation provides the basis for generation of multiple sentences that are statistically pared down so that the most likely sentence is generated according to the constraints of the TL.

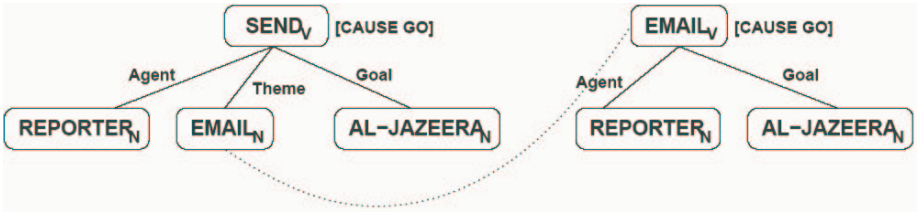


**Figure 11: Approximate Interlingua for English-Arabic Example**

Consider, for example, the conflational divergence example given above (Section 4.7) between Arabic to English. Figure 11 illustrates the approximate-interlingua approach to translation for this example. The top-level conceptual nodes are first checked for possible matches. Following this, unmatched thematic roles are checked for 'conflatability', i.e., cases where semantic roles are absorbed into other predicate positions. As long as there is a relation between the conflated argument ($EMAIL_N$) and the new predicate node ($EMAIL_V$), disregarding part-of-speech. (This relation is derived from a database of categorial variations (Habash and Dorr, 2003b).)

## 5 Annotating Text with Interlingual Information

The success of corpus-based language technology over the past decade has shown the value of systems that automatically learn their processing from large collections of annotated examples. Although no one has yet created an Interlingua-annotated corpus to parallel the 1 million sentences plus syntax trees of the Penn Treebank (Kingsbury et al., 2002), several efforts to annotate important parts of an Interlingua are underway. Principally, these efforts focus on verbs and their arguments. We list these and then describe one initiative—IAMTC—in more detail to illustrate the issues involved in annotation.

### 5.1 Semantic Annotation Initiatives

WordNet (Fellbaum, 1998), see http://www.cogsci.princeton.edu/~wn/, provides a terminology taxonomy for English containing over 100,000 terms. Several ontology-building efforts have used this resource as a starting point. Focusing on the creation of wordnets for other languages, the Global WordNet Association http://www.globalwordnet.org/ lists EuroWordNet http://www.illc.uva.nl/EuroWordNet/, GermaNet http://www.sfs.nphil.uni-tuebingen.de/lsd/Intro.html, BalkaNet http://www.ceid.upatras.gr/Balkanet/, and many others.

Term taxonomizing and ontologizing efforts include the Chinese HowNet (Dong, 2000) and the Mimida multilingual semantic network http://www.gittens.nl/SemanticNetworks.html.

Focusing on verbs alone, the FrameNet project (Baker et al., 1998), see http://www.icsi.berkeley.edu/~framenet/, is classifying all verbs into groups according to the case roles (thematic roles) they support. The SALSA project (Erk et al., 2003), see http://www.coli.uni-sb.de/lexicon/index.phtml, parallels FrameNet, working on German verbs. Other FrameNet-related projects http://www.nak.ics.keio.ac.jp/jfn/ for Japanese and http://gemini.uab.es/SFN/ for Spanish.

The PropBank project (Kingsbury et al., 2002), http://www.cis.upenn.edu/~ace/ or http://www.cis.upenn.edu/%7Empalmer/project_pages/ACE.htm, resembles FrameNet in that it focuses on verbs, but it does not employ a fixed set of case roles, preferring instead a more neutral set of labels with no overall semantics. VerbNet http://www.cis.upenn.edu/group/verbnet/ is an associated effort to assign FrameNet-like case roles to verbs. The list http://www.cis.upenn.edu/%7Edgildea/Verbs/ combines VerbNet and FrameNet. The NomBank Project (Meyers et al., 2004), see http://nlp.cs.nyu.edu/meyers/NomBank.html, closely parallels PropBank, but focuses on nouns (such as nominalized verbs and relational nouns) with argument structure.

The Interlingual Annotation of Multilingual Text Corpora (IAMTC) project (Farwell et al., 2004), http://aitc.aitcnet.org/nsf/iamtc/, is an ambitious attempt to investigate interlingual semantics by annotating and comparing semantic phenomena across six languages. Having prepared bilingual corpora pairing English texts with corresponding text in Japanese, Spanish, Arabic, Hindi, French, and Korean, annotators are assigned to each text impairs to select semantic representation symbols from the Omega ontology (Philpot et al., 2003) for all nouns, verbs, adjectives, and adverbs. We describe this project in more detail below.

### 5.2 Interlingual Annotation of Multilingual Text Corpora (IAMTC)

The IAMTC project has the following goals:

- Development of an interlingual representation framework based on a careful study of text corpora in six languages and their translations into English.
- Development of a methodology for accurately and consistently assigning such representations to texts across languages and across annotators.
- Annotation of a corpus of six multilingual parallel subcorpora, using the agreed-upon interlingual representation.
- Development of semantic annotation tools which serve to facilitate more rapid annotation of texts.
- Design of new metrics and evaluations for the interlingual representations, in order to evaluate the degree of annotator agreement and the granularity of meaning representation.

The IAMTC project is radically different from those annotation projects that have focused on morphology, syntax, or even certain types of semantic content (e.g., for word sense disambiguation). It is most similar to PropBank (Kingsbury et al., 2002) and FrameNet (Baker et al., 1998). However, IAMTC places an emphasis on: (1) a more abstract level of mark-up (interpretation); (2) the assignment of a well-defined meaning representation to concrete texts; and (3) issues of a community-wide consistent and accurate annotation of meaning.

The data set consists of 6 bilingual parallel corpora. Each corpus is made up of 125 source language news articles along with three independently produced translations into English. (The source news articles for

each individual language corpus are different from the source articles in the other language corpora.) The source languages are Japanese, Korean, Hindi, Arabic, French and Spanish. Typically, each article contains between 300 and 400 words (or the equivalent) and thus each corpus has between 150,000 and 200,000 words. The Spanish, French, and Japanese corpora are based on the DARPA MT evaluation data (White and O'Connell, 1994). The Arabic corpus is based on LDC's Multiple Translation Arabic, Part 1 (Walker et al., 2003).

The interlingual representation comprises three levels and incorporates knowledge sources such as the Omega ontology (Philpot et al., 2003) and thematic roles (Dorr, 2001). The three levels of representation are referred to as *IL0, IL1* and *IL2*. The aim is to perform the annotation process incrementally, with each level of representation incorporating additional semantic features and removing existing syntactic ones. *IL2* is intended as the interlingual level that abstracts away from (most) syntactic idiosyncrasies of the source language. *IL0* and *IL1* are intermediate representations that are useful stepping stones for annotating at the next level.

### 5.3 Issues in Interlingual Annotation

A preliminary investigation of intercoder agreement on multiple annotations shows that the more annotators learn the process, the better they become, resulting in an improvement of intercoder agreement (Mitamura et al., 2004). Two assumptions may be made regarding the training of novice annotators in order to improve intercoder agreement. One is that novice annotators may make inconsistent annotations within the same text, but these may be reconciled through a process of *intra-annotator consistency checking*, in which annotators go over their results to find any inconsistencies within the text. Another assumption is that, if two annotators at the same site discuss their annotation results after their annotation tasks are completed, their judgments may be reconciled through a process of *inter-annotator checking*, in which each annotator votes, they discuss the differences, and then vote again.

From an MT perspective, issues include evaluating consistency in the use of the annotation language, given that any source text can result in multiple, different, legitimate translations (Farwell and Helmreich, 2003). Along these lines, there is the problem of annotating texts for translation without including in the annotations inferences resulting from the source text. The IAMTC effort described above is the only initiative, to date, that addresses issues of this type in large-scale annotation of data for use in interlingual MT.

## References


Allen, J.F. 1984. A General Model of Action and Time. *Artificial Intelligence* 23(2).

Arnold, D. and L. des Tombe. 1987. Basic Theory and Methodology in Eurotra. In S. Nirenburg (ed.) *Machine Translation: Theoretical and Methodological Issues*. Cambridge University Press, Cambridge, England, 114-135.

Baker, C., C.J. 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, CA. A version of this paper appears in 1990 as: Upper Modeling: A Level of Semantics for Natural Language Processing. *Proceedings of the 5th International Workshop on Language Generation*. Pittsburgh, PA.

Carletta, J.C. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22(2), 249–254

CICC. 2003. http://www.cicc.or.jp/english/kyoudou/top.html.

Erk, K., A. Kowalski, S. Pado, and M. Pinkal. 2003. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. *Proceedings of the ACL 2003 conference*, Sapporo, Japan.

Dong, Z. 2000. HowNet Chinese-English Conceptual Database. Online Software Database released at ACL, http://www.keenage.com.

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

Dorr, B.J. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics* 20(4), 597–633.

Dorr, B.J. 2001. LCS Verb Database, Online Software Database of Lexical Conceptual Structures and Documentation, University of Maryland. http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

Dorr, B.J., L. Pearl, R. Hwa, and N. Habash. 2002. DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment. *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas (AMTA-2002),*Tiburon, CA, 31–43.

Dorr, B.J. and N. Habash. 2002. Interlingua Approximation: A Generation-Heavy Approach. *Proceedings of Workshop on Interlingua Reliability*. Workshop at the Fifth Conference of the Association for Machine Translation in the Americas (AMTA-2002), Tiburon, CA, 1–6.

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

Farwell, D., S.n Helmreich, B.J. Dorr, N. Habash, F. Reeder, K. Miller, L. Levin, T. Mitamura, E.H. Hovy, O. Rambow, and A. Siddharthan. 2004. Interlingual Annotation of Multilingual Text Corpora. *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, 55–62.

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

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 (eds), *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, 1–88.

Foley, W. and R. D. Van Valin Jr. 1984. Functional Syntax and Universal Grammar. Cambridge: Cambridge University Press.

Frederking, R., S. Nirenburg, D. Farwell, S. Helmreich, E.H. Hovy, K. Knight, S. Beale, C. Domashnev, D. Attardo, D. Grannes, and R. Brown. 1994. The Pangloss Mark III Machine Translation System. *Proceedings of the 1st AMTA Conference*. Columbia, MD.

Gao, Y., B. Zhou, Z. Diao, J. Sorensen, and M. Picheny. MARS: A Statistical Semantic Parsing and Generation-Based Multilingual Automatic tRanslation System. *Machine Translation* 17(3), 185–212.

Grimshaw, J. 1992. Argument Structure. Cambridge: MIT Press.

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

Habash, N., B.J. Dorr, and D. Traum. 2003. Hybrid Natural Language Generation from Lexical Conceptual Structures. *Machine Translation*, 18(2), 81–128.

Habash, N. and B.J. Dorr. 2003. A Categorial Variation Database for English. *Proceedings of North American Association for Computational Linguistics Conference (HLT-NAACL)*, Edmonton, Canada, 96–102.

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

Hale, K. and S. J. Keyser. 2002. *Prolegomenon to a Theory of Argument Structure*. Linguistic Inquiry Monographs. Cambridge: MIT Press.

Hayes, P.J. 1985. The Second Naive Physics Manifesto. In J.R. Hobbs and R.C. Moore (eds.), *Formal Theories of the Common-sense World*. Norwood: Ablex, 1–36.

Hobbs, J.R. 1985. Ontological Promiscuity. *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 61–69.

Hobbs, J.R. 2001. Causality. *Proceedings of the Fifth Symposium on Logical Formalizations of Commonsense Reasoning.* NYU. 145–155.

Hovy, E.H. and S. Nirenburg. 1992. Approximating an Interlingua in a Principled Way. *Proceedings of the DARPA Speech and Natural Language Workshop.* Arden House, NY.

Hutchins, W.J. 1987. Prospects in Machine Translation: Proceedings of MT Summit I. *Proceedings of Machine Translation Summit*, Japan.

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

Johnson, R., M. King, and L. des Tombe. 1985. EUROTRA: A Multilingual System Under Development. *Computational Linguistics*, 11:2-3, 155-169.

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

Kipper, K. and M. Palmer. 2000. Representation of Actions as an Interlingua. *Proceedings of the Third AMTA SIG-IL Workshop on Interlinguas and Interlingual Approaches*, Seattle, WA.

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

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

Lavie, A., L. Levin, T. Schultz, C. Langley, B Han, A. Tribble, D. Gates, D. Wallace, K. Peterson. 2001. Domain Portability in Speech-to-Speech Translation. *Proceedings of HLT 2001*. Human Language Technology Conference. San Diego, California.

Lazzari G. 2000. Spoken Translation: Challenges and Opportunities. In B. Yuan, T. Huang, and X. Tang (eds.) *Proceedings of 6th International Conference on Spoken Language Processing [ICSLP 00]*, vol IV, pp. 430-435. Beijing, China.

Lenat, D.B. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38(11), 32–38.

Lee, Y.-S., D.J. Sinder, and C. Weinstein. Interlingua-based English-Korean Two-way Speech Translation of Doctor-Patient Dialogues with CCLINC. *Machine Translation* 17(3), 2002. 213–243.

Levin, B. 1993. English Verb Classes and Alternations: A Preliminary Investigation. Chicago: University of Chicago Press.

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

Levin, L., D.Gates, D. Wallace, K. Peterson, and A. Lavie. 2002. Balancing Expressiveness and Simplicity in an Interlingua for Task Based Dialogue. *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, Association for Computational Linguistics. Philadelphia, USA.

Levin, L., C. Langley, A. Lavie, D. Gates, D. Wallace, and K. Peterson. 2003. Domain Specific Speech Acts for Spoken Language Translation. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan.

Levin, L., A. Lavie, M. Woszczyna, D. Gates, M. Gavaldà, D. Koll, and A. Waibel, 2000. The Janus III Translation System. *Machine Translation*.

Levin, L. and S. Nirenburg. 1994 Construction-Based MT Lexicons. A. Zampolli, N. Calzolari, and M. Palmer (eds.) *Current Issues in Computational Linguistics: In Honour of Don Walker*. Pisa/Dordrecht: Giardini editori e stambatori and Kluwer publishers. P ages 321-338.

Lonsdale, D., T. Mitamura. and E. Nyberg. 1995. Acquisition of Large Lexicons for Practical Knowledge-Based MT. *Machine Translation*, 9:3-4.

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

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

McShane, M., S. Beale, and S. Nirenburg. 2004. OntoSem Methods for Processing Semantic Ellipsis. *Proceedings of Computational Lexical Semantics Workshop at HLT-NAACL.*

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young and R. Grishman. 2004. The NomBank Project: An Interim Report. New York University Computer Science Department technical report.

Goodman, K. and S. Nirenburg (eds). 1991. The KBMT Project: A Case Study in Knowledge-Based Machine Translation, San Mateo: Morgan Kaufmann.

Nirenburg, S., V. Raskin and B. Onyshkevych. 1995. Apologiae Ontologia. *Proceedings of the International Conference on Theoretical and Methodological Issues* (TMI). Leuven, Belgium.

Nirenburg, S. and V. Raskin. 2004. *Ontological Semantics*. Cambridge: MIT Press.

Nyberg, E. and T. Mitamura. 1992. The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains. *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France.

Nyberg, E. and T. Mitamura. 2000. The KANTOO Machine Translation Environment. In J. S. White (ed.) Envisioning Machine Translation in the Information Future. 4th Conference of the Association for Machine Translation in the Americas (AMTA 2000). Lecture Notes in Artificial Intelligence, Vol. 1934. Berlin: Springer Verlag.

Ogden, B., J. Cowie, E. Ludovik, H. Molina-Salgado, S. Nirenburg, N. Sharples and S. Sheremtyeva. 1999. CRL's TREC-8 Systems: Cross-Lingual IR and Q&A, *Proceedings of the Eighth Text Retrieval Conference (TREC-8).*

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

Qu, Y., B. DiEugenio, A. Lavie, L. Levin, and C. P. Rosé. 1997. Minnimizing Cumulative Error in Discourse Context. In E. Maier, M. Mast and S. LuperFoy (eds.), *Dialogue Processing in Spoken Language Systems*, Lecture Notes in Artificial Intelligence 1236, Berlin: Springer Verlag.

Reichenbach, H. 1947. The Tenses of Verbs, chapter in *Elements of Symbolic Logic*. London: Collier Macmillan.

Schultz, T., D. Alexander, A.W. Black, K. Peterson, S. Suebvisai, and A Waibel. 2004. A Thai Speech Translation System For Medical Dialogs. *Proceedings of the conference on Human Language Technologies (HLT-NAACL*), Boston, MA.

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 Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, DC.

Vauquois, B. 1968. A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation. *Proceedings of the IFIP Congress-6*. 254–260.

Waibel, A. 1996. Interactive Translation of Conversational Speech. *Computer*, 19(7), pages 41-48.

Waibel, A., A. Badran, A.W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin , K. Lenzo, L. Mayfield Tomokiyo, J. Reichert, T. Schultz, D. Wallace, M. Woszczyna, and J. Zhang. 2003. Speechalator: two-way speech-to-speech translation on a consumer PDA. *Proceedings of the 8th European Conference on Speech Communication andTechnology (Eurospeech-2003)*, Geeva, Switzerland.

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

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

Whitelock, P. 1989. Why Transfer and Interlingua Approaches to MT are Both Wrong: A Position Paper. *Proceedings of the MT Workshop: Into The 90's*, Manchester, England.

Whorf, B.L. 1959. *Language, Thought, and Reality*. Cambridge: MIT Press.

Woszczyna, M., M. Broadhead, D. Gates, M. Gavaldà, A. Lavie, L. Levin, and A. Waibel. 1998. A Modular Approach to Spoken Language Translation for Large Domains. In *Machine Translation and the Information Soup: Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA '98)*, Lecture Notes in Artificial Intelligence 1529, Berlin: Springer-Verlag.

## Biographies

***Bonnie Dorr*** is an associate professor at the University of Maryland, with a joint appointment in Computer Science, UMIACS, and Linguistics, and is co-director of the Computational Linguistics and Information Processing laboratory. She graduated from the Massachusetts Institute of Technology in 1990 with a Ph.D. in computer science. Her research focuses on several areas of broad-scale multilingual processing, e.g., machine translation, summarization, and cross-language information retrieval. She has investigated the problem creating new statistical models that are linguistically informed, leading to higher quality output for a wide range of languages while still being practical to train and use. Dr. Dorr is the recipient of a NSF Presidential Faculty Fellowship Award, Maryland's Distinguished Young Scientist Award, the Alfred P. Sloan Research Award, and a NSF Young Investigator Award. She has served on numerous editorial boards and executive committees and is the author of *Machine Translation: A View from the Lexicon*.

URLs:
http://www.umiacs.umd.edu/~bonnie
http://www.umiacs.umd.edu/research/CLIP/

***Eduard Hovy*** leads the Natural Language Research Group at the Information Sciences Institute of the University of Southern California. He is also Deputy Director of the Intelligent Systems Division, as well as a research associate professor of the Computer Science Departments of USC and of the University of Waterloo in Canada. He completed a Ph.D. in Computer Science (Artificial Intelligence) at Yale University in 1987. His research focuses on automated text summarization, question answering, text planning and generation, the semi-automated construction of large lexicons and ontologies, and machine translation. Dr. Hovy is the author or co-editor of five books and over 140 technical articles. In 2001 Dr. Hovy served as President of the Association for Computational Linguistics (ACL) and in 2001-03 as President of the International Association of Machine Translation (IAMT). Dr. Hovy regularly co-teaches a course in the Master's Degree Program in Computational Linguistics at the University of Southern California, as well as occasional short courses on MT and other topics at universities and conferences. He

has served on the Ph.D. and M.S. committees for students from USC, Carnegie Mellon University, the Universities of Toronto, Karlsruhe, Pennsylvania, Stockholm, Waterloo, Nijmegen, Pretoria, and Ho Chi Minh City.

URLs:
http://www.isi.edu/natural-language/nlp-at-isi.html
http://www.isi.edu/~hovy

***Lori Levin*** is an Associate Research Professor at Carnegie Mellon University's Language Technologies Institute. She received a B.A. in Linguistics from the University of Pennsylvania in 1979 and a Ph.D. in Linguistics from M.I.T. in 1986. She has co-directed several machine translation projects covering both spoken and written language.

URLs:
http://www-2.cs.cmu.edu/afs/cs.cmu.edu/user/lsl/www/home.html
http://www.lti.cs.cmu.edu/index.html