PREceedings
# of the
Pre-Workshop on ILs and IL Approaches to MT

[1 October 1996, at Second Conference of the
Association for Machine Translation in the Americas, Montreal, Quebec, Canada]

****

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

WHAT IS AN INTERLINGUA AND WHAT INFORMATION SHOULD IT CONTAIN?

I. INTRODUCTION

The traditional defining characteristic of an Interlingua, according to Hutchins & Somers (p. 73), is that it "is neutral between two or more languages" so that, in principle, given a representation of an utterance in the Interlingua, the source language of the utterance cannot be determined from that representation. From this follow the other general features of an Interlingual MT system: the independence of analysis and generation, the use of language-independent knowledge sources, the attempt to represent the "meaning" of the text using the Interlingua, the claim to "universality," and the abstract nature of Interlingual representations.

In this position paper, I argue against this characterization and suggest that a better characterization is that an Interlingual representation is one geared to explicitly represent the intent of the text author(s). Consequences of this position for the information contained in an interlingua are also somewhat fleshed out.

II. IT IS IMPOSSIBLE TO ACHIEVE HIGH-QUALITY, AUTOMATED MACHINE TRANSLATION WITH A "LANGUAGE NEUTRAL" REPRESENTATION AS THE SOLE INPUT TO THE TARGET LANGUAGE GENERATOR.

The conclusion stated immediately above follows directly from Bar-Hillel's argument (1960) that, in principle, any bit of knowledge might be required in order to disambiguate some text. While this argument is usually made in support of the need for knowledge-based interlingual MT, it can also be used _against_ an Interlingual approach as characterized above. For surely, one of the possibly significant bits of information that might be necessary to adequately disambiguate a text is precisely the original language of the source text and the actual source language text itself. Thus, this information must be part of any Interlingual representation of the text, if an adequate target language text is to be generated from it. But then, of course, it is not interlingual in the sense described above, that is, containing no hint about the language of the original text.

A simple example might help. A cartoon sequence in the local paper was, for a while, printed with only Spanish text. (After protests from English-speaking readers it was then printed with English sub-titles as well, and then later dropped.) One sequence included the following dialogue. In this case the visual element is non-essential, but the background was a park filled with playground equipment.

?Por qui el pollo atraverss el parque?
!Para ir al otro tobogan!

This apparently inane dialogue makes sense if one hypothesizes an original English dialogue as follows:

Why did the chicken cross the park?
To get to the other slide!

Leaving aside for the moment the cultural specificity of the joke-type itself, given this reconstruction it is clear that even to begin to translate this dialogue appropriately, it is necessary to note the phonological closeness (rhyme) between the two English words "side" and "slide." Thus, this information must be available in the interlingual representation. But then it contains a reference to the source language itself.

It cannot even be plausibly argued that this information could be extracted and represented in the Interlingua in some non-language specific format. The aspects of the source language text that might be relevant are, in principle, unlimited, and could include information about the phonetics, phonology, morphology, syntax, or orthography of the source text.

In conclusion, to ensure high-quality translation at all times it is necessary to include language-specific information about the source language text in any representation that will serve as the generation input.

III. INTERLINGUA AS A MEANS OF REPRESENTING THE INTENTION OF THE AUTHOR(S).

Let me make it clear the claim in II. above does not imply support for a transfer-based approach to MT. Far from it. Indeed, in the example above, no transfer-based approach could begin to succeed. Critical to finding an appropriate translation for the dialogue is the recognition of the author's intent to tell a joke, the recognition of the genre of joke-telling (i.e., having a punchline) and then recognizing the subverting of that genre by having a non-punchline punchline as the key to the joke. In addition, it involves recognition of a particular joke as the underlying form on which this is a witty(?) takeoff.

It seems to me that the difference between an Interlingual approach and others is not on what the IL representation cannot contain (i.e., any reference to the source language), but rather on the depth of analysis contained in that representation.

And, as in the above example, it is not sufficient even to represent the "meaning" of the language involved. In order to provide an adequate translation, the intent of the author(s) must be ascertained. Therefore, I suggest, a better characterization of an Interlingua is as a representation of the source text which allows for an explicit representation the intent of the author(s). This requires analysis to a depth greater than that which even current IL/KB MT systems, such as Mikrokosmos, perform. It requires reasoning from the semantic propositions expressed plus relevant information in the Knowledge Bases in the system (plus, on occasion, information about the actual sounds, words, and structure used to express those propositions) to the intent of the author(s) of the text.

IV. CONSEQUENCES OF THIS POSITION.

Changing the characterization of an Interlingua from a formal categorization to a categorization based on content affects none of the the following characteristics of an Interlingua: knowledge-based processing, representation that includes the "meaning" of the text, universality or abstractness of representation. It should not even affect the independence of analysis and generation in that there do not need to be special rules in the generation procedure that specifically are cued to the source language (though they could not be ruled out by this definition).

In addition, the following features would seem to be, if not required, at least consonant with this definition.

A. Explicit representation of the beliefs of the author(s) and of other participants in the communicative event.

B. Explicit representation of various levels of author intent, e.g., locutionary, illocutionary, and perlocutionary intent.

C. Inferencing/control structure that is non-monotonic. Since the intent can only be deduced, not observed, often on the basis of default or other types of non-deductive reasoning, the intent can be determined only with a certain level of confidence.

D. Emphasis on the establishment of coherence of interpretation.

E. Explicit representation of the chains of inference used to determine the author's intent.

V. CONCLUSION.

It has been argued that an interlingua which cannot represent aspects of the actual source language text is insufficient as a basis for high-quality machine translation. Indeed, as the intent of the author(s) become more complex and their attention to the details of the text more conscious (as in poetry or other emotive uses of language), more and more of the physical aspects of the text will be vital clues to achieving even an adequate translation. It is suggested that an Interlingua be defined as one geared toward to the representation of all aspects of the author's intent.

* * ***

Interlinguas don't need language-specific information

Lori Levin
Carnegie Mellon University
Lori_Levin@prague.mt.cs.cmu.edu

The C-STAR consortium is currently in the process of designing an Interchange Format (IF). The partners of C-STAR (ATR Japan, ETRI Korea, CMU USA, U. of Karlsruhe Germany, Siemens Germany, IRST Italy, and ?? France) are collaborating on a multi-lingual speech translation demonstration scheduled for 1999. The current semantic domain for C-STAR is meeting scheduling with two dialogue participants. We are now moving on to a more general travel planning domain with multi-party dialogues.

A current point of negotiation among the C-STAR participants is whether the IF should reflect source-language syntax and phrasing. Arguments in favor of retaining source language structure include (1) source language noun phrases must be retained in the IF as antecedents of source language anaphors (I took a bath. It (the bath) was hot. vs. ?I bathed. It was hot.) and (2) it's hard enough to write an analyzer/generator for a language without having to worry about compatibility with other grammars.

My position is: Languages are different and you have to put in some effort to relate them to each other. You can put the effort into transferring between language specific representations or you can put the effort into identifying universal features of an interlingua. Either way you do the same work. In a multi-lingual system, it makes more sense to work with one language independent interlingua. That way each grammar developer has to learn only one system of meaning representation, instead of learning how to relate his/her language to several language-specific pseudo-interlinguas. (In a longer version of this position paper I can address the opposing position point by point, showing that having language-specific features in the IF will not save any time or effort.)

In moving to a larger domain, another issue we will have to deal with is whether to have one uniform interlingua or separate domain-specific interlinguas for the components of travel planning (scheduling, reserving, shopping, etc.) It looks like we are headed for several sub-domain interlinguas. This implies that part of the translation process is classifying utterances into sub-domains.

****

Interlinguas: natural languages, logics or arbitrary notations?

Yorick Wilks
University of Sheffield
yorick@dcs.shef.ac.uk

What are interlinguas, and does the answer have any practical effect on the usefulness, success, or otherwise of interlingual MT, a paradigm that still has life left in it, and some practical successes, certainly in Japan. In deference to the gathering, I will focus what I have to stay on the interlingual representation language, as you might expect to find it in an MT system—normally some pidginised version of English with a counterintuitive syntax—but what I have to say applies more generally to symbolic knowledge representations, including those applied to MT (e.g. KBMT) and those in the mainstream AI tradition.

If we take the view that they can be arbitrary notations then many more systems come within the class of interlingual machine translation systems than would normally be thought the case: certainly SYSTRAN in its earlier periods (before the point at which it was declared a transfer system) when the results of a source language analysis were stored in a complex system of register codes and, most importantly, this was done for more than one source language—thus giving the storage codings, which were largely arbitrary in conventional linguistic terms and certainly without comprehensible/readable predicates, a degree of linguistic "neutrality" that is thought to be part of what is meant by an interlingua.

Taken more strictly, SYSTRAN was never an interlingual system because its power came largely from its bilingual dictionaries, and, as a matter of definition, a bilingual dictionary is a language-pair dependent transfer device. At that level of strictness, there have been very few true interlingual MT systems, (i.e. without a bilingual dictionary). Probably the only candidate is Schank's MARGIE system of about 1972, which did some English-German MT via a conceptual dependency (CD) representation. Schank's CD representation is also much closer to the stereotype of an interlingual representation, having a range of English-like predicates (TRANS being the best remembered) within a specified syntax and diagrammatic notation.

My own small English-French MT system, contemporary with Schank's and in the same CS department was therefore not interlingual, even though our representations had much in common, because I believed a bilingual dictionary essential, and that the interlingual representation and its associated algorithms selected the correct equivalent from among a set of candidates the lexicon provided. Schank believed then, and I denied, that any such crypto-transfer mapping was needed, but only he Source-to-interlingua and interlingua-to-target translators. It was Charniak who later supplied argument's arguments whose force I already felt, that no level of coding at such a grain could be expected to distinguish in output: sweat, sneeze, dribble, spit, perspire, as well as a range of less attractive possibilities all associated with the Schankian primitive EXPEL taken along with a coding for LIQUID.

The issue of grain here was obviously a function of the richness of the interlingual vocabulary (Schank then had about 14 primitive actions and I about 100 primitives of different syntactic types). IF the interlingua had the resources of any natural language, then those distinctions could have been made and that of course focuses exactly the question of what it would mean for an interlingua to have the resources of a natural language as opposed to being a formal language with primitives that may appear to be language-like, as Schank's certainly did, but which their author's deny are in fact language items, let alone English words.

That position of denial is not to be found only in the 1970's: Schank's position is essentially that of Nirenburg and Raskin (1996), when supporting Mikrokosmos as a "language-neutral body of knowledge about the world" in contrast to the "recurring trend in the writings of scholars in the AI tradition..toward erasing the boundaries between ontologies and taxonomies of natural language concepts".

This dispute can take unhelpful forms such as "Your codings look like natural language to me"; "No they dont". Moreover, it is not clear that any settlement of this issue, for either side, would have effect whatever on the performance or plausibility of interlingual MT systems. One can can also filter out more full blooded versions of the NL-IL identity, such as the Dutch MT system based on Esperanto (an NL for the purposes of being dismissed from this argument) and the reported systems based on the S. American language Aymara, said to be without lexical ambiguity (see below) and therefore ideal for this role. These cases, whether or not they work in practice, fall under the criticism of Bar Hillel in the earliest discussions of interlingual MT that having an NL in the IL role would simply double the work to no benefit by substituting two MT tasks for one.

The interesting issue, given that we can all concede that ILs do not normally like quite like NLs, and do certainly have some superficial features of formal languages (brackets, capitalization, some non-linguistic connectives etc.) is whether they have any significant features of NLs, which I would take to mean above and beyond the simple appearance of a number of NL words in their formulas, normally drawn from English. English words appear in profusion in many programs particularly those that encourage arbitrary predicate naming, such as LISP and Prolog, from which one could not of course conclude that programs in those languages were IN ENGLISH. Appearance of words is not enough, or French could be declared English, or vice versa, or Roumanian Turkish.

The feature I would seize on in discussion is whether the primitives of interlingual formalisms suffer ambiguity and progressive extension of sense as all known NLs do (except perhaps Aymara, but that may reflect lack of study). Some formal languages can tolerate substantial ambiguity of symbols—early LISP, which functioned perfectly well, is now conventionally said to have had a key symbol (NIL) three-ways ambiguous. LISP programs presumably implicitly resolved this ambiguity in context, or did they?

It is widely believed that NLs have their ambiguities resolved in use, up to some acceptable level, and that extensions of sense take place all the time, whether rule governed (e.g. as in Pustejovsky's generative lexicon, deriving from Givon's early work in the 1960s) or, as in the old AI/NLP tradition by means of manipulations on lexicons and knowledge structures that were general procedures but not necessarily equivalent to lexical rules. It would it be like, and I have no clear answer, to determine that the primitives of an IL representation were in this position, too. Schank did, after all split the early TRANS into MTRANS, ATRANS and then others, so the suggestion has precedent.

An answer might require a trip through the more empirical aspects of recent NLP/CL and ask what evidence we have that any given symbol in its occurrences in corpora has more than one sense? Has this any empirical, non-circular answer, that does not require appeal to existing polysemous dictionaries or contrastive translations? I believe the answer is yes, and that an IL does have this key feature of an NL and that no disastrous consequence follows from thus viewing ILs as reduced NLs, rather than full ones. This has for me more intuitive plausibility than continuing to maintain that what seem to be NL features of ILs are not in fact but are language independent. That claim always seems to me one that it is impossible to defend in detail but is a clear residue of the hypnotic power of intuition in an age of empiricism and calculation.

****

The 'Lingua' in Interlingua
AMTA SIG-IL Workshop, 1996

Robert Belvin, Bonnie Glover Stalls, Christine Montgomery, Alfredo Arnaiz
Language Systems, Inc.
robin@lsi.com

An interlingua can be defined as a metalinguistic representation of the function of a linguistic object which is not dependent on the language-specific form of that object. In Language Systems, Inc. (LSI)'s multilingual Machine-Aided Voice Translation (MAVT) system, the interlingual representation consists of a set of event and object frames with slots that are filled with information derived from or associated with the text that is being processed. These slot fillers include information bearing on propositional content as well as communicative intent, pragmatics, and other kinds of information that are present to varying degrees of explicitness in the text. They also provide a means by which contextual and domain knowledge that has no realization or is ambiguous in the text can be used during the translation process. To a great extent the information filling these slots is not language-specific; however, there are some interesting ways in which some language-specificity is preserved and not only does not interfere in the translation process but actually facilitates it.

As David Farwell (ACL 1994) has pointed out, the goal of an interlingual representation is not "language-independence" but rather "language-neutrality". We essentially agree with this position, but would further suggest that it is not necessary to strip away from the textual representation all vestige of the source language, but rather to render it in a neutralized form that is easily mappable into any potential target language. An interlingua for an MT system, which must be capable of transmitting information from one language to another, is a kind of language, or representation of language, for which the requirement for neutrality need not limit its expressive power. In fact, as discussed at some length in Dorr (1993), preserving certain aspects of the linguistic structure of the text can help to minimize the need for a deeper level of conceptual representation and to construct the target language text. A case in point is the incorporation of lexical conceptual structure (LCS) into the interlingua in LSI's MAVT system as a means of ensuring that a verb or other predicate with an appropriate predicate-argument structure is selected and appropriately saturated in the target language.

Our position has been that predicate-argument relations are more efficiently represented AS predicate-argument structures than other kinds of representations. That is, since the information encoded in a P-A structure readily expresses the relations of interest, it seems unwise to transform that representation into some other type of structure. This seems especially true in employing an interlingua in an MT system. Our experience has led us to the belief that the use of a quasi-syntactic representation of eventuality (i.e. event or state) concepts has facilitated the translation process, especially w.r.t. lexical selection of target-language predicates, and generation of target-language syntactic structures. If one regards the creation of an interlingual representation as a series of steps which undo the language-specific packaging of the relevant information, the complete undoing of the syntactic organization of eventuality representations appears to be a step which may not only be unnecessary, but undesirable.

For example, when matching a concept in an interlingua to a lexical item in the generation stage of translation, the use of quasi-syntactic lexical-conceptual structures has allowed us to collapse a process which would otherwise require two steps into one. The characteristics of our LCSs usually allow us not only to find a lexical item (or items) in the target lexicon which matches the relevant concept (both in its core meaning and selectional restrictions), but we can simultaneously check if the candidate lexical item has an appropriate subcategorization frame.

It should be mentioned that we arrived at the decision to employ semantic structures modeled on Jackendoff's Lexical-Conceptual Structures in our MT system after considering several alternatives. Jackendoff type LCSs appeared more desirable than the alternatives because (i) they provided a way of representing concepts which partially solved word-sense disambiguation problems without relying on language-specific predicates and (ii) they facilitated mapping between interlingual concept representations and language-specific syntactic representations. This latter advantage seems to be due to the fact that they are structured in much the same way as a syntactic predicate. In addition, since our predicate concepts are quasi-syntactic, the possibility arises that the same kind of constraints can be applied to them as are known to apply to syntactic representations. This kind of parallelism between lexical-conceptual structure and syntactic structure has in fact been argued to reflect a genuine psychological reality based on various types of linguistic phenomena in research by Hale and Keyser (1993), Bouchard (1995), Jackendoff (1993), and others. While we are not necessarily committed to this premise, the possibility nonetheless provides additional support for the validity of the approach.

Our work in developing an interlingual MT system grew out of our Data Base Generation (DBG) system, which was developed over a number of years and which analyzed text and produced output for a variety of downstream applications, including information extraction and retrieval, and message fusion. The goal of the DBG system was to instantiate a set of event and object frames, called "templates", which represented the content of the text being processed. At the topmost level were meta-templates, which represented the meta-event of writing and sending the text being processed, and so could incorporate higher-level discourse features of the text (e.g., source of the text, time it was written, recipient, and so on).

One thing that we discovered in working with a variety of applications is that the content of an adequate representation "depends on the application." What is adequate for one type of application may be completely inadequate for another. For example, in highly regimented contexts such as written communications reporting flight activity of aircraft by military surveillance teams, there is virtually no need for anything but the scantiest information on communicative intent, since the communicative intent remains constant throughout the reports. Other factors, such as the degree of belief of the writer in the facts being reported, however, are highly significant and must be analyzed and represented. In a voice translation system designed for interrogation purposes, identifying communicative intent in the source speech and providing a reasonable approximation in the target speech is very important because the intent is highly variable, and the response of the hearer may be very sensitive to it; it must therefore be given some representation in an interlingual representation.

An interlingua is a kind of knowledge representation (KR), very similar in many ways to the KRs that we have worked with previously. One characteristic of an interlingual MT system such as the one we have developed which distinguishes it from many other interlingual MT systems is that it organizes concepts according to two different taxonomic schemes. One is a lexically oriented conceptual scheme (the LCS taxonomy), the other a typical (non-lexically oriented) KR scheme. In MT, using the interlingua as a means of preserving the structure of the source language sentence for use as a kind of filter or guide in selecting target language lexical items and syntactic structures, makes a good deal of sense. Conceptual information is certainly an inherent part of the process, and in the MAVT system is available when needed. Nouns in the lexicon are indexed to nodes in the conceptual hierarchy, and selection of the target language nouns is done by selecting the nouns associated with the same or related nodes (ontological entries) as those in the source language. However, for verbs the relations among verb concepts in the concept hierarchy are used primarily in cases where there is no exact LCS match. In that case, adjacent nodes are checked for possible near-matches that can incorporate the information in the interlingual representation.

This dual taxonomy strategy allows to take advantage of the virtues of lexically structured concepts where possible, but allows us to exploit non-lexically structured concepts when necessary. This organization, as we have suggested, is desirable in an interlingual MT application, but may be unnecessary in other types of applications. This is because the most obvious virtue of lexically structured concepts is that they facilitate target language generation. In a text understanding application, it may be preferable to bypass lexically structured concepts and map directly to a non-lexically structured knowledge base.

REFERENCES

Bouchard, D. (1995) The Semantics of Syntax, U. of Chicago Press, Chicago.

Dorr, B. J. (1993) Machine Translation: A view from the lexicon, MIT Press, Cambridge.

Hale, K. and S. J. Keyser (1993) "On Argument Structure and the Lexical Expression of Syntactic Relations," in Hale and Keyser, eds. The View from Building 20, MIT Press, Cambridge, pp. 53-109.

Jackendoff, R. (1993) "X-bar Semantics," in Pustejovsky, ed. Semantics and the Lexicon, Kluwer, Dordrecht, pp. 15-26.

****

Position on Interlingua:

John White
PRC
white_john@po.gis.prc.com

What really ought to represented in the intermediate representation?

I suspect that few people dispute the potential benefit of a language-independent representation in an MT model, and the interlingua that expresses the representation. Imagining what kinds of problems would have to be addressed there, it is easy to find tempting candidates for interlingual representation that create real problems for accepting the principle in the first place. Many of these problems live in the metaphysical, the knowledge-based, pragmatic, sociolinguistic, etc. trappings of language in use, and the possible different expectations of a native speaker of source and target. But there are some problems that have implications for plain old morphological or lexical forms, which give one pause. These problems have to do with how much contextual semantics work you must do in analysis (and represent in the IR) to cover really language-specific demands.

Here are two such problems, one with an apparently easy fix, which may suggest a way to fix the harder one.

The easy one:

Chinese dialects, Mayan languages, and, apparently Navajo, share a means of representing the lexical class of nouns by expressing a numeral classifier. These are like the English "round" (of ammunition) and "head" (of cattle), except that they are more fully specified over most or all of the count nouns in the language. Semantics are generally attributed to the selection of numeral classifiers; thus Tojolabal-Maya "wan" seems to express humanness, and may appear in expressions like "chab' wane? winik (two person-units man)", "osh wane? ishuk (three person-units woman)", and so on.

Given the apparent the observation that the classifier can be predicted by lexical semantics, and the fact that the phenomenon occurs in many languages, it is tempting to try to "handle" this somehow in the intermediate representation, by expressing the semantics for it in the interlingua. If the target language uses numeral classifiers, generation would make classifier selection based on that semantic representation.

What inserts some reality into the idealistic desire to represent the phenomenon in the interlingua is the fact that the numeral classifier languages do not organize the lexical universe in the same way. That is, the fact that Tojolabal-Mayan classifies "sky" as something wide and thin ("hun lame? sat k'inal" — one slice sky) does not in anyway predict that the other numeral classifier languages do. This means that in order to handle the phenomenon in the intermediate representation, the IR would have to express all of the semantic properties known to be sensitive in any classifier language, and hope that there are no contradictions (which there will of course be).

This problem is easy because the practical solution is also the right one, namely to generate the numerical classifier solely from the language specific requirements on the target side. Each lexical item will have some sort of specification about the classifier it expects. In the case where a noun may take different classifiers in different instances, the lexical description on the target side will contain variables instantiated from the normal compositional semantics of the intermediate representation.

Here is the hard problem:

English generally distinguishes between flesh on the hoof and meat on the plate (pig-pork, sheep-mutton, cow-beef, etc.). Spanish and German generally do not. We know from this that the intermediate representation must have done some work, possibly difficult work, in inferring from context whether a reference to animals or parts thereof is a reference to flesh or meat, so that the interlingua can express sufficient semantics for English generation.

If we do this, then we can rest assured that regardless of the source language, we will be able to generate the correct distinction in English. But what if I am translating from Spanish to German? My intermediate representation has done all of the inference work to generate the flesh-meat distinction, but in this pair I don't need it. Why should an MT system do substantial work that the pair doesn't need?

What makes this question hard is the fact that the English-required distinction can't be done on the target lexicon side, as it can for numeral classifiers. A great deal must be known about whether something is intended to be eaten, and what is doing the eating, etc., information that has to come from all over the place in the source language expression. So it appears that this distinction must indeed be handled in the IR.

But doesn't this mean that the IR has to express the semantics of all the lexical idiosyncrasies of all languages? Surely every language has some lexical phenomenon similar to English flesh-meat. Doesn't this leave us in the same trap from which we escaped in the numeral classifier case?

I don't know the answer to this particular problem. But I think the solution has to do with the level of generality at which we do our interlingual representation. It may well be that a great deal of the semantic work specific to a language must be done at generation time, possibly even after a round of lexical selection. In this model, a reasoning tool examines a partially or fully lexicalized target representation, and makes a judgment about its felicity (semantic, pragmatic, discourse-wise), choosing alternates in some cases and lexicalization of variables in others. This delegation of powerful reasoning to the generation component seems to violate our current sensibilities about the role of the interlingua, but the interlingua model of MT remains language independent, and in fact becomes more so by expressing only what is truly universal and not by trying to be all things to all languages.

****

Boyan Onyshkevych
Dept of Defense
baonysh@afterlife.ncsc.mil

Topic addressed: What information is captured by an adequate interlingual representation system?

Any examination of real corpora, especially in non-scientific domains, will reveal that metonymic expressions are pervasive in real language use. Although various definitions of metonymy may draw the distinction between metonymic and non-metonymic expressions differently, typically certain core metonymic expressions (such as "Moscow announced ...") which are pervasive in text will always be defined as metonymy.

Regardless of the strategy adopted for=12handling metonymy in the analysis phase of processing, the representation of metonymic expressions in the interlingua will be faced with one central decision: how to represent the metonymy, literally or as rendered in the source text? The position that I will argue for is that despite the processing overhead, it is beneficial to resolve the metonym in analysis and to represent the replaced entity in the interlingua explicitly.

Reasons for resolving metonymy and explicitly representing the replaced entity include:

1. In some cases it is necessary to resolve the metonymy before generation in MT because some different languages have different

inventories of metonymy, and word-for-word translation of some metonymies will result in anomalous translations. By representing the resolved metonymy, the generator can choose to either render the content literally or to produce an appropriate metonymy in the target language (assuming an appropriately generator is available...)

2. The replaced referent can provide context for use during word-sense disambiguation, whether by domain inference techniques, selectional restrictions (which would have been violated by the metonym), or other techniques.

3. One may need to make the replaced entity available as a referent. Metonymies such as the infamous "ham sandwich" example allow anaphora to the replaced referent: "The ham sandwich wants a cup of coffee. He also needs a new fork". The metonymy may in fact result in a full-fledged use of the referent, entering the entity into the "given" register; examples such as "I drive a Volvo, but the engine is shot" illustrate that the replaced entity (the car or truck) is available as if it had been used explicitly.

****

Interlingual representations, the MT triangle and good food

David Farwell
david@crl.nmsu.edu

The fact that two different translators can appropriately translate the Spanish expressions "[d]el tercer piso" and "el segundo piso" in:

... los 300 metros cuadrados del tercer piso estaban disponibles pero fueron aquilados ..., sslo queda el segundo piso ....

as, on the one hand, "the third floor" and "the second floor" respectively and, on the other, "the fourth floor" and "the third floor" respectively, demonstrates (1) that the representation of the semantics of the expressions uttered is insufficient for providing an appropriate translation and (2) that the representation of the translator's beliefs about the beliefs of the participants in the translation process (the SL speaker/author, the SL addressee and the TL addressee) which are needed for assigning an interpretation to the utterance are, in fact, necessary. That the seemingly contradictory translations provided above are both potentially appropriate is due to the fact that there are at least two floor naming conventions that are used around the world. Under the first, the ground level of a multi-story building is referred to as the ground or bottom floor while, under the second, it is referred to as the first floor. One of the translators above assumes that the source language author and addressee and target language addressee all follow the same convention. The second translator assumes that, while the source language author and addressee are following the first convention, the target language addressee is following the second convention.

The process of translation, then, consists of interpreting a speech act - someone's intentionally using some expression with a given semantics to communicate some message to someone else for some purpose - and then recreating that act to the degree possible using a different language and addressing a different audience - that is, the translator, adopting the relevant beliefs of author of the original act, intentionally uses a different expression with possibly a different semantics to communicate the same message to a different addressee for presumably the same purpose.

In the case above, if the semantics of the original expression used to convey the message is taken to be the (compositional) semantic representation of that expression, then the interpretation of the message is arrived at by inferencing from that representation in order to provide coherence within the context: the beliefs of the SL speaker and addressees, the speaker goals conveying that message at that point in the discourse. That is, while the semantics of say "el tercer piso" is some logical statement about a particular third story [of some building] as might be derived in combining the semantics of "el", "tercer" and "piso" appropriately, it remains to provide some convention for the ordering the storys of the particular building in order to provide the expression with a coherent interpretation, that is, to identify which floor of the building is being referred to. That convention is provided by the discourse context of the SL utterance for interpretation and for the discourse context of the TL utterance for producing a target language expression.

This model implies that expressions in different languages having the same (or to the degree possible similar) semantic representations may not be translations of each other if the contexts, the beliefs of the participants in the two interactions, that determine the interpretation are different. This in turn implies that (1) ILs must represent something more than the semantics of the SL expression (or TL) expression, that (2) it should include a representation of the relevant beliefs of the various participants needed for producing an appropriate translation and that (3) beliefs-based inferencing on the basis of knowledge of the world may operate on ILs after semantic analysis and before generation in order to maintain the coherence of the event reported with respect to the events that have preceded it within a shifting beliefs context from SL to TL utterances.

Finally, the traditional MT triangle, long known to be flawed in its representation of direct MT approaches, also appears to be flawed in terms of its representation of IL approaches. There need not be a single IL representation which is the result of SL analysis which serves as the input to TL generation. Rather, when SL context is swapped out for TL context, beliefs relevant to the interpretation process may be replaced by differing beliefs in the TL context triggering a revision of the IL. At best, the triangle now looks like the bottom of a cup.

In fact, this situation is more like following a cake recipe calling for flour, eggs, butter, water, sugar, etc. all baked at a certain temperature for a given amount of time. It produces excellent results in Moscow, Idaho and perhaps in Moscow, Russia but, because the wheat, the chickens, the cows, the water, the sugar, the altitude and the oven are all different, the cakes are not the same. And, in fact, to achieve as similar a cake as possible, proportions may have to be changed and baking methods varied.

**************************** *******************

Evelyne Viegas
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003

E-mail: viegas@crl.nmsu.edu
Fax: (505) 646 6218
Tel: (505) 646 5757

SIG-IL Pre-workshop Draft - Aug 28,1996
Do not quote - Very Rough Draft!

In this draft I will mainly address point 2), advocating that it takes an IL Text Meaning Representation (informed with planning techniques) to solve mismatches and divergences among various natural languages; and parts of point 3), in particular the different ways we experimented in MikroKosmos to scale up "static" knowledge sources, to provide coverage of Spanish and English.

I - Point 2) Solving Mismatches and Divergences through an IL: a Case Study

Statementl: It takes more than mere word sense disambiguation in
--------- analysis and lexical selection in generation to solve
mismatches and divergences: it takes an IL Text Meaning
Representation (TMR) informed by planning techniques.

In the following we will provide some empirical evidence from cross-linguistic data. We will first look at "simpler" cases of mismatches (such as "put" versus "polovit'" and "postavit'" in Russian) and then we will concentrate on the "continuum" that seems to exist between some mismatches and divergences as in "bake" and "cook" versus "cuire [+/- au four]" where only planning techniques seem to be able to generate the right lexeme or expression.

The following is brief, sketchy, and still needs argumentation...

Our interest for solving mismatches and divergences using an IL TMR along with planning, comes from noticing that all former enterprises (as described in Lindop et Tsujii, 1991; Door, 1990; Heid, 1993; Kameyama, 1991, Nirenburg and Levin, 1993, etc.) whatever the approach (or MT paradigm) seem to fail in solving (i.e., recognise and generate) divergences and mismatches. In terms of divergences (roughly speaking: same meaning but different syntactic structure) the problem seems to be linked to the impossibility to get an exhaustive typology of all the different types of divergences (cf Vandooren, 1993); moreover some cases seem difficult to classify, such as "wooden floor" -> "plancher" in French, similar to the conflation cases of Talmy (Talmy, 1985); or "bake" -> "cuire [+/- au four]" (where "au four" cannot be considered as a syntactic ellipsis). The case of mismatches (roughly speaking: the grammar and the lexicon of the SL do not make some distinctions which are required by the grammar and the lexicon of the TL) is even more problematic, as there is not only need for contextual knowledge but also for extra-linguistic knowledge, as discussed in (Kameyama, 1991).

Looking at real data from corpora, it seems that there are more examples which lay (still unexplained) in the continuum between divergences and mismatches than examples which can be classified as belonging to one case (clear example of predicative divergence: "he limped up the stairs" -> "il monta les marches en boitant") or the other (clear example of semantic underspecification "pez, pescado" ->

"poisson").

A big confusion wrt mismatches seems to arise from a largely shared belief that a language SL which has less lexical units to which correspond more lexical units in the TL (such as for "fish" in English -> "pez" and "pescado" in Spanish; or for "put" -> "polovit'" and "postavit'" in Russian; or for "cuire" in French -> "bake" and "cook"; ...) is ambiguous from a monolingual perspective.

To correct this supposed ambiguity one can decide there are two entries in the English dictionary for "fish" fish-N1 and fish-N2 corresponding to pez-N1 and pescado-N1 respectively. I believe a native English or American speaker to be very surprised to learn that where he had conceptualised one natural kind FISH he should now conceptualise two: FISH-living creature and FISH-food, without being able to make the link between the two, that is recognising the fact that what makes a fish a potential food, is the possibility of applying some cooking event to it in order to eat it (cf. Briscoe and Copestake, and their "grinding rule").

It rather seems to me that the word "fish" becomes ambiguous in Spanish while remaining unambiguous in English; same thing with polovit'/postavit' and put; or "bake/cook" and "cuire".

Isn't it rather the result of deliberate underspecification (elsewhere called vagueness) in some languages where inferences are sometimes preferred over short-cuts or fully specified meaning. Let me exemplified this with the Russian examples. I will consider the lexeme "put" as unambiguous in English but will have to consider it as underspecified wrt Russian.

I will assume a knowledge-based approach semantics based, and a conceptual world or ontology where i have a concept labeled PUT, which contains the following relevant information:

PUT
AGENT: HUMAN
THEME: PHYSICAL-OBJECT
SOURCE: PLACE
DESTINATION: PLACE

The semantics for "put", "polovit'" and "postavit'" should minimally have the following information:

```
put(X,Y,Z)
    sem: PUT(X,Y,Z), AGENT(X), THEME(Y), DESTINATION(Z)
```

polovit'(X,Y,Z)
sem: PUT(X,Y,Z), AGENT(X), THEME(Y), DESTINATION(Z),
DIRECTIONALITY(Y,FLAT)

postavit'(X,Y,Z)
sem: PUT(X,Y/Z)/ AGENT(X), THEME(Y), DESTINATION(Z),
DIRECTIONALITY(Y,UPRIGHT)

Now let us assume the following concepts GLASS and PLATE in the ontology with their associated conceptual relevant information:

GLASS
ISA: ARTIFACT
DIRECTIONALITY: UPRIGHT
CONTAINS: LIQUID

PLATE
ISA: ARTIFACT
DIRECTIONALITY: FLAT
CONTAINS: FOOD

Relevant extracts of an IL TMR for the simplified English sentence (a) John put the glass on the table, should look like:

PUT
AGENT: John
THEME: GLASS
DESTINATION: TABLE

Translating the above sentence into Russian does require some processing as there are two entries ("polovit'" and "postavit'") which can lexicalise the concept PUT. However, "polovit'" requires its theme to have a DIRECTIONALITY FLAT, which is the case of the word glass mapped to GLASS. Therefore mismatch viewed as specialisation (cf Kameyama, 1991) of lexical units is clearly a generation problem, not an analysis one.

Now if we look at the examples for "cook" and "bake" which translate into "cuire [+/- au four]", then here we seem to be confronted to a "generalisation" problem (cf. Kameyama, 1991). Here too we claim that we are confronted with a generation problem and not an analysis one as there is no reason to consider "cuire" as ambiguous in French. Now, let us consider the data below to illustrate the point that it takes an IL TMR to solve mismatches and divergences.

Let us look in this draft at some isolated sentences, for the sake of simplicity:

b) Cuis le pain -> Bake the bread

c) Cuis les pa^tes al'dente -> Cook the pasta (al'dente)

d) Cuis les pa^tes au four -> d1) Bake the pasta
-> d2) Cook the pasta in the oven

e) Cuire les pa^tes au gratin
pas plus de 20mns -> e1) Bake the pasta au
gratin no longer than 30mns
-> e2) Cook the pasta au
gratin no longer than 30mns

f) I prefer baked meals to meals
cooked on the stove top -> Je preferre les plats au four aux plats
(cuisines) sur le feu

g) Cuire le pain et les pa^tes -> bake the bread, then cook the pasta

I said that "cuire" was not ambiguous in French. What remains to be seen is whether or not we get two concepts BAKE and COOK to which maps "bake" and "cook" respectively, with "cuire" mapping to COOK; therefore, going from English to French would be a question of generalisation whereas going from French to English would be a question of specialisation, as mentioned by (Kameyama, 1991). The problem with this approach is that it seems difficult in example f), which is a case of generalisation, to avoid to generate "je preferre des plats cuis a' des plats cuis sur le feu" (i preferred cooked meals to cooked meals on the stove)! Moreover, if we now want to specify, we have to rely on the semantics of the noun which sometimes is ambiguous, such as in example e) where although there is a preference for generating el) rather than e2), it is still acceptable to have e2). Finally, example g) shows that generating a mismatch requires more than lexical selection, it does require a planning of the sentence, as the conjunction "et" in French might be interpreted as a temporal-succession in which case it is necessary to develop the ellipsis. Moreover, contextual constraints present in the TMR will help to eventually generate bake the pasta if in the linguistic context we are told that "pasta" is a reference for "lasagna". The point remains that it is impossible to "freeze" the meanings of "bake" and "cook" as equivalent to "cuire au four" and "cuire" respectively, this is why I advocate an IL TMR along with planning to solve cross-linguistic problems of this kind.

Information to be included in the knowledge sources:

```
COOK
  AGENT: HUMAN
  THEME: PHYSICAL-OBJECT
  INSTRUMENT: COOKING-EQUIPMENT
  LOCATION: PLACE
```

```
cook(X,Y)
          sem: COOK(X,Y), AGENT(X), THEME(Y)

bake(X,Y)
          sem: COOK(X,Y), AGENT(X), THEME(Y), INSTRUMENT(OVEN)

cuire(X,Y)
          sem: COOK(X,Y), AGENT(X), THEME(Y),
INSTRUMENT(COOKTNG-
EQUIPMENT)
```

(to be developed; compare with "i started cooking at 18" -> cuisiner)

II - 3) Scaling up the "static" knowledge sources

Statement 2: Scaling up static knowledge sources to perform coverage ----------is doable within a contemplative view of the lexicon: we did it!

The most difficult task seems to get started, namely get the core lexicon. In Mikrokosmos we developed a computational semantic lexicon for Spanish; each entry containing in the semantic zone an "unsaturated piece of IL-TMR". A core lexicon of about 7000 entries (lexemes) have been acquired by hand, with the use of computational tools to accelerate acquisition (lextool interface for acquisition; corpora search; on-line dictionary search; ontology browser; ontology request...). Then, we extended the core lexicon using derivational morphology applied to verbs, reaching around 35,000 entries-lexemes (for which we can produce the POS, the syntax and the semantics).

The big advantage of using an IL representation to encode the meanings of words is that the analysis lexicon can be reversed or indexed on concepts; this allowed us to perform many "exercises" as varied as:

- use the "reversed" lexicon as a pivot lexicon in a multilingual generation environment, by lexicalising in different languages the semantic zone. For instance, the conceptual frame:

```
INGEST
        AGENT(X)
        THEME(Y)
        EDIBLE(Z)
```

coming from the Spanish verb "comer-V1", can be lexicalised as "manger" in French, "eat" in English, etc... it can also serve as the basis for the lexicalisations of "close synonyms" "avaler, ingurgiter", ..., in French. Note that parallel corpora could also be used to see how "comer" translates into other languages; however, there still will be a need for human checking, but this should be faster than developing another lexicon from scratch, as our experience showed.

- we can generate from the TMR the text in Spanish and then analyse the gaps between the original source text and the text generated, this could enhance a lot the issue of what to put and what to omit in the IL and also how good our lexicons are.

Statement 3: Before scaling up for coverage there is still many work
---------- to be investigated if we adopt an inquisitive view of the
lexicon (how useful it is wrt a particular task).

In the previous statement, I claimed it is doable to get coverage in a fairly small amount of time (it took us about a year with 4 person/year to develop a Spanish lexicon of about 35,000 roots, from scratch).

Here I would like to defend the position that the advantage of using an IL TMR lays in the power it gives us to capture meanings across languages. From the point of view of the "static" knowledge sources, the trade-offs between the lexicon and the ontology, calls sometimes for procedures (such as specialisation or specification or planning) not yet fully understood; however, I do believe that IL has more than any other approach to give us to capture the meaning(s) of words. The question of scaling up for coverage is not particular to IL approaches it is a common problem faced by any symbolic approach, and as such I do not think we should spend too much time on it. I guess we should rather try to explain unsolved phenomena, recognising which procedures to use to solve them. From the lexicon point of view, work on "underspecification" might well be the way to reconcile the contemplative view with the inquisitive view of the lexicon.

****

Semantic Frame:
A Flexible Interlingua for Machine Translation and Human/Machine Interaction

Young-Suk Lee and Clifford Weinstein
ysl@sst.ll.mit.edu
MIT Lincoln Laboratory

To serve as a practical tool in a machine translation system, an interlingua must be straightforwardly derivable from the analysis module, and represented in a form from which a well formed sentence of any language can be easily generated. In addition, categories/features in an interlingua has to be easily manipulable so that the system developer can add/delete some of the categories until it takes the form of a truly language-independent meaning representation expressed by language universal categories/features.

The English/Korean machine translation system developed by MIT Lincoln Laboratory under DARPA sponsorship, which is based on the understanding/generation (TINA/GENESIS) system developed by MIT Laboratory for Computer Science, produces an interlingua called SEMANTIC FRAME. SEMANTIC FRAME satisfies the above mentioned conditions for an ideal interlingua in a practical machine translation system, and has proven to be effective in multilingual Human/Machine Interaction systems.

A semantic frame is directly derived from the parse tree. All major parse tree constituents (regardless of whether they are semantic or syntactic) are reduced into one of the three language neutral categories in a semantic frame, namely, clause-type, topic and predicate. All clause-level categories such as statement, infinitives are mapped onto "clause." All noun phrase expressions are mapped onto "topic." All modifiers as well as verb phrases are mapped onto "predicate." Reduction of all major parse tree categories into one of the three semantic frame categories enables the generation system to easily produce a syntactically well-formed sentence of any language, especially the well-formed word order of the target language. In principle, however, there is no limit to the kind/number of categories/features which can be expressed in a semantic frame. Linguistic features like 'tense,' 'number,' etc. are easily added or deleted depending on the need.

With SEMANTIC FRAME as an interlingua, the system produces high quality translation output of naval operational messages of a highly telegraphic nature (see Weinstein et al. 1996 for details) as well as other more natural texts of English. In the presentation, we will discuss capabilities and limitations of the system in detail, and some informal ideas about how to overcome the system limitations.

References

James Glass, Joseph Polifroni and Stephanie Seneff. 1994. Multilingual Language Generation Across Multiple Domains. In "Proceedings of the 1994 International Conference on Spoken Language Processing," Yokohama, Japan.

Stephanie Seneff. 1992. TINA: A Natural Language System for Spoken Language Applications. Computational Linguistics, 18-1, pages 61-88.

Clifford Weinstein, Dinesh Tummala, Young-Suk Lee and Stephanie Seneff. 1996. Automatic English-to-Korean Text Translation of Telegraphic Messages in a Limited Domain. In "Proceedings of COLING '96," Copenhagen, Denmark.

Victor Zue. 1996. Research and Development of Multilingual GALAXY: A Status Report. In "Proceedings of C-STAR II 96 ATR International Workshop on Speech Translation," Kyoto, Japan.

****

Nili Mandelblit
University of California, San Diego
Dep. of Cognitive Science,
La Jolla, CA 92093-0515
e-mail: mandelbl@cogsci.ucsd.edu

GENERAL TOPIC: What sort of information is to be captured by an adequate interlingua?

In my work, I investigate instances of linguistic expressions which provide only PARTIAL information about a conceived event (i.e., only some aspects of a conceived event are expressed explicitly, while others are understood implicitly). I argue that languages differ in the type of information they most often express explicitly (i.e., each language explicitly communicates DIFFERENT aspects of the same generic event).

In such cases, a direct transfer of the source text into the target language cannot provide a correct translation, and it is the role of the interlingua to COMPLETE the missing (implicit) information from the source text, before the target text can be generated.

A crucial aspect of my research is in identifying how languages actually differ in the information they explicitly communicate. I propose that a primary factor in defining which aspects of generic event-types are commonly expressed in a particular language is the inventory of CONSTRUCTIONS available in the language.

A Construction is a syntactic (or morphological) pattern which is independently associated with a semantic structure (cf., Fillmore & Kay, ms.; Goldberg, 1995). A specific construction is used by language speakers to express a novel conceived event only if correlation is found between the semantic structure of the conceived event and the semantics associated with the grammatical pattern (the construction). Since constructions vary not only in the semantic structures associated with them but also in the partial information they highlight, variations in the inventory of constructions across languages also imply variations in the type of information explicitly communicated in each language (Mandelblit 1995a).

Example:

Goldberg (1995) analyzes the Caused-Motion construction in English. Its syntactic form is [NP V NP PP], and the semantic structure associated with it (according to Goldberg) is a generic Caused-Motion event (i.e., X causes Y to move in direction Z). Sentence (1-3) below are instances of the Caused-Motion construction. Note that the caused-motion semantics does not exist in any of the lexical items independently, and is hence assumed to exist in the syntactic structure itself.

(1) Martha trotted the horse into the stable.
(2) The wind blew the ship off course.
(3) The audience laughed the poor actor off the stage.

An important point to note is that only PARTIAL information about the conceived caused-motion event is actually expressed in examples (1-3).

In (1), the event being communicated is one in which Martha is causing the horse to trot (and move) into the stable. However, nothing is said explicitly about HOW Martha made the horse trotting (what was the

CAUSING event). Did Martha lead the horse into the stable, or did she hit the horse, thereby causing the horse to trot in the direction of the stable?

In (2), the event being communicated is one in which the wind blowing causes the ship to move away from its original course. In this example, the sentence provides explicit information about the causing force that made the ship change its location (the wind blowing). However, no explicit information is given about the resulting motion event (i.e., in what manner and where did the ship move: was the ship being SHIFTED into another course? or was it drowning down into the sea?). In both examples (1-2), a default scenario is commonly imposed by the listener to complete the missing information.

Example (3) provides again explicit information about the causing event (the audience laughing), but the resulting motion event is left implicit. In what manner did the actor move off the stage? Was the actor passively SHIFTED off the stage (like the ship in example 2), or was the actor voluntarily RUNNING AWAY from the stage? Again, background knowledge of default scenarios imposes a specific interpretation.

What happens when we try to translate English Caused-Motion sentences into other languages (Hebrew or French, for example)? Hebrew and French do not have an independent Caused-Motion construction. Rather to express a caused-motion event as in sentences (l)-(3), Hebrew and French speakers make use of a GENERIC CAUSATIVE construction that exists in the language (i.e., the *faire* construction in French, or the morphological *hifUil* construction in Hebrew). However, while the main verb in the Caused-Motion construction in English may express either the resulting motion event (as in example 1), or the causing event (as in example 2-3), the main verb in the causative *faire* construction in French and the *hifUil* construction in Hebrew ALWAYS denotes the RESULTING event of a causal sequence of events (and the CAUSING event is left implicit). Hence, clearly a translation of sentences (2)-(3) into Hebrew and French cannot be a direct function of the main verb in the source text.

Below are the Hebrew and French translations for sentences (1-3). The English examples (i) are followed by an Hebrew translation (ii), a word-to-word transfer of the Hebrew version into English (iii), and a French translation (iv).

(1) (i) She *trotted* the horse into the stable.
(ii) Hi hidhira(d.h.r-hifUil) et hasus letoch haurva.

(iii) She TROT-CAUSE(past) the-horse into the-stable.
(iv) Elle a fait trotter le cheval dans 1 ecurie.

(2) (i) The wind *blew* the ship off course.
(ii) Haruax hesita(n.s.t-hifUil) et hasfina mimaslula.
(iii) The wind SHIFT-CAUSE(past) the-ship off-its-course,
(iv) Le vent a ecarte le navire de sa trajectoire.

(3) (i) The audience *laughed* the actor off the stage.
(ii) Hakahal hivrix(b.r.x-hifUil) et hasaxan min habama (besxoko).
(iii) The audience RUN-AWAY-CAUSE(past) the-actor off the-stage.

The main verb in the Hebrew (and French) translations in example (2)-(3) is not a function of the main verb (or any other lexical item) in the source sentence. To perform the translation of sentences (2)-(3), a translator (human or machine) must first reconstruct the original causal sequence of events communicated in the source language. The representation of the whole causal sequence of events (the causing event and the implicit effected motion event) forms the INTERLINGUAL representation. From the interlingual representation, a target text can be generated (the translation) using available grammatical constructions in the target language (i.e., the *faire* or *hifUil* constructions in French and Hebrew respectively, both explicitly communicating only the effected motion event).

What kind of information is needed to construct an interlingual representation from the source text, and to generate a target sentence from the interlingual representation?

In addition to language-specific and interlingual lexicons, a translation system must have information about:

(1) The inventory of constructions available in the source and target languages. For each construction we need to specify the generic event type associated with the construction, and the partial information most commonly highlighted (or explicitly expressed) by the construction.

(2) Background knowledge - common scenarios of causally related events.

REFERENCES:

For further information about this work, please refer to:

(1) Mandelblit, N. (1995a). RCognition, Translation, and NLPS.

Technical Report, Cognitive Science, University of California, San Diego.

(2) Mandelblit, N (1995b). "Beyond Lexical Semantics: Mapping and Blending of Conceptual and Linguistic Structures in Machine Translation", In Proceedings of the 4th Int. Conf. on the Cognitive Science of Natural Language Processing, Dublin, Ireland, July 1995.

(3) Blending: Creative Aspects in Grammar and Translation. Presentation at the Computing Research Laboratory, New Mexico State University, April 1996.

****

First Steps toward Building Interlinguas of Scale

Eduard Hovy
USC Information Sciences Institute
hovy@isi.edu
August 1996

In the construction of an Interlingua for Machine Translation a system, two principal challenges stand out:

1. the design of a representation approach simple enough to be manageable by human representers, yet sophisticated enough to be able to capture meaning in a truly interlingual manner;

2. the construction of an Interlingua 'lexicon', a set of representation terms enough to provide adequate coverage over a real-world domain yet consistent enough internally to be manipulated by automated processes.

The former challenge is the domain of ongoing efforts in lexicography and knowledge representation [Nirenburg et al., Copestake, Dorr], among others. It is a complex endeavor with little assistance from automated procedures — the crucial design work must be done by humans, and the development of methodologies, rigorous performance procedures, and criteria of evaluation is ongoing.

The latter challenge is the focus of the proposed presentation. I will (if accepted) discuss the creation and testing of an Interlingua 'lexicon' of large enough scale to support open-domain translation (as well as of various other NLP tasks). It may seems strange—some would say impossible —to separate the creation of an IL 'lexicon' from the actual representation. However, by IL 'lexicon' I mean here a termbank, a set of symbols containing only a very sparse semantics, that serves as a pivot structure between the lexica of the various languages. I am not talking about an Interlingua Lexicon in the full, rich, sense.

Again, some may question the purpose of constructing such a set of 'empty' symbols. The answer is simple: practical experience has shown that it is possible to achieve wide-domain MT (and a variety of other NLP tasks such as IR) with such symbols. Such a set provides a baseline for IL-based MT performance. Admittedly, the quality is not always very good; that is the reason for continuing work on enriching the symbols' contents. But proceeding "from the outside in" (large set of symbols, gradually enriched), as opposed to "from the inside out" (small set of richly annotated symbols, gradually grown in number), is an important approach, for it tells us several things:

1. roughly, the semantic 'regions' in which symbols are required (for example, the 'region' of speech act representation, or the region of attributes);
2. a sense of the levels of representational delicacy required in each such 'region' relative to other 'regions' (measured by, say, the approximate numbers of such symbols relative to the numbers in other 'regions');
3. an indication of the organization of such symbols and the underlying interrelationships that prove most useful (for example, ISA vs. PARTOF vs. SYNONYM);
4. an overall framework of anchor points into which we can embed more delicately articulated symbol sets, whenever they are forthcoming;
5. a way of helping to bring different, competing, enriched symbol sets (or domain theories) into correspondence and to compare them.

Some of these ideas may be controversial, and I will not belabor them. Instead I will describe some work I have recently performed in service of creating a new, large-scale, IL termbank, organized as a property inheritance taxonomy. The goal of this work is to establish a kind of 'standard' Reference Ontology (R.O.) which would be available to anyone via the Web, for the following uses (among others):

- to allow interprocess communication for systems developed at different sites (when both systems' terms are translated into the IL terms),
- to allow the comparison of different domain models for the same domain,
- to identify areas of shortcoming in the R.O. and invite suggestions on how to fill them.

This effort is a collaboration of the following researchers:
- IBM San Jose (Bob Spillers, Andras Kornai) (lead organization),
- USC/ISI (Eduard Hovy),
- CYCorp (Doug Lenat, Fritz Lehmann),
- Conceptual Graphs (John Sowa),
- Stanford University (Bob Engelmore, Adam Farquhar)

and more. It is part of an ANSI Standards group working on representations.

The work I will describe represents the first steps in creating an R.O. These steps bring together the uppermost regions of several large ontologies and relate their terms to one another, to the extent this can be done. The result is a taxonomy, viewable either as a single integration of the terms from each of the constituent ontologies, or as each constituent alone, with pointers to the others as appropriate.

The constituent ontologies (with, parenthesized, the number of concepts under current consideration) are:

1. The Pangloss Ontology Base from ISI (approx. 300 concepts)
2. The 'top' concepts from CYC (approx. 1500)
3. The top concepts from EDR (approx. 100)

The result is represented in SENSUS, a simple KR system analogous to ART, KEE, Loom, FrameKit, and the like.

How one goes about integrating (or at least, finding correspondences and linking together) such disparate symbol sets is not exactly clear. I will describe the six-step process I followed, providing in a handout the inputs and results of each stage, and run some of the transformation programs as part of the discussion. I will highlight the easy and the difficult aspects of the task, and point out some problems caused by the idiosyncrasies of the various symbol sets.

If there is a discussion period, I would be extremely interested in hearing opinions as to how this work should continue, what results (if any) would be useful to researchers, and (if useful) in what form.

****

Topic: HOW CAN INTERLINGUAL SYSTEMS BE SCALED UP?
Title: Use of Syntax-Semantics Relation for Automatic Construction of
Interlingual Lexicons

Bonnie J. Dorr
University of Maryland

bonnie@umiacs.umd.edu

Our research at the University of Maryland has focused on the construction of dictionaries for interlingual applications. One of the central questions we have addressed is that of how to build automatic procedures for scaling up these dictionaries. We believe that answering this question is the first step toward building serious, large-scale (and completed) systems for use in tasks such as machine translation, foreign language tutoring, and other multilingual information processing tasks. We take the components of meaning in our dictionary representations to be interlingual and have used these as the basis of dictionaries for languages such as Arabic, Spanish, French, and Korean.

While our emphasis appears to lie on the "supply" side of the equation (i.e., construction of large dictionaries), we are well aware that these representations must be applicable to the "demand" side of the equation (i.e., large, working systems). Nirenburg (1996) describes these two sides and categorizes the work of lexicon researchers accordingly. We view our position in this categorization to be much more fuzzy than described, falling across the supply-demand boundary.

Our approach to building large dictionaries relies on a number of techniques based on the notion that there exists a basic relation between the semantics of a verb and its corresponding syntactic behavior. Of course, we need to provide convincing evidence for his underlying assumption—the central thesis of Levin (1993)—i.e., we need to show that the semantics of a verb and its syntactic behavior are predictably related. A large part of our work has focused on demonstrating the validity of this hypothesis (Dorr and Jones, 1996a). In our experiments, we provided theoretical justification for the bases upon which we proceeded for our lexical-acquisition work, i.e., we have demonstrated that 98% of Levin's semantic classes have uniquely identifying syntactic signatures (i.e., clusters of syntactic behaviors). We view these experiments as a necessary step for proceeding with further experimentation for construction of verb classes, i.e., we want to ensure that our starting point is solid before undertaking large-scale acquisition based on the syntax-semantics relation.

Upon completion of these experiments, we have begun a long process of verb categorization of "novel" (previously unseen) verbs in English. This work has resulted in a database of verbs, classified semantically based on a system similar to that of Levin (1993). We are currently developing syntax-semantics tests for other languages, e.g., Arabic, Spanish, French, and Korean, so that we can similarly classify verbs for those languages.

A common point of confusion (e.g., during the presentation of this work at ACL, COLING, and related workshops in summer of 1996) concerns the nature of the semantic classes upon which we have built our dictionaries. Several researchers (Saint-Dizier, 1996, among others) have pointed out that these classes are not universal, and thus cannot serve as the basis of an interlingua. What should be kept in mind is that it is not the CLASSES that are intended to be universal, but the COMPONENTS OF MEANING that underlie these classes. By their very nature (i.e., that they are based on English-specific syntactic "alternations") the English semantic classes do not hold cross-linguistically. However, it was not the intention to classify translation equivalents identically, but to isolate the meaning components associated with semantic classes, and to then find a relation between these meaning components. The meaning components, not the syntactic behaviors, are expected to be language-independent.

For example, the "Motion/Impact" verbs, but not the "Change-of-state" verbs participate in the conative:

Motion/Contact
  She tapped at the window
  She banged at the door

Change of State
  She broke at the window
  She smashed at the door

Although the conative does not exist in other languages (e.g., French), there is clearly some meaning component associated with "tap" and "bang" (contact, but no no change in structural integrity) that is not associated with break and smash (contact and change in structural integrity). While we wouldn't use the conative in French, clearly the notions of contact and structural integrity can be expressed in French, and so the isolation of these meaning components (through application of the conative test in English) is clearly of cross-linguistic value. These meaning components are what should then be included as part of the interlingual structure.

While we have justified the use of the syntax-semantics relation as the basis for building an interlingua, we are still faced with the problem of scaling up our database of lexical representations. We have addressed this problem by using automatic procedures based on syntactic tests (such as the ones above) for mapping verbs onto lexical-semantic representations. One of the major difficulties we were faced with in automatic classification of unknown verbs is that of "polysemy" (word sense ambiguity), which, in previous work (Dorr et al. 1995) resulted in very low "precision" (i.e., a high percentage of verbs assigned incorrectly to semantic classes)—13%. (We use precision as our primary metric for judging the effectiveness of our acquisition technique. Details are given in (Dorr and Jones, 1996b).) As an attempt to address the polysemy problem, we used a WordNet based filter for classification of unknown words. We tested the filter on three different proportions of the original 2813 Levin verbs: (a) 50%, (b) 70%, and (c) 90%, chosen randomly. We then checked whether the "unknown" verbs (those not used to construct the semantic filter) were assigned to their correct classes. The result was a drastic improvement in precision—64% (for the 90% case) in contrast to the 13% precision of Dorr et al. (1995).

Our experiments indicate that, not surprisingly, but not insignificantly, the syntax-semantics relationship is very clear, particularly in our later experiment where we accounted for word sense ambiguity. These experiments served to validate Levin's claim that verb semantics and syntactic behavior are predictably related and also demonstrated that a significant component of any lexical acquisition program is the ability to perform word-sense disambiguation. We have used the results of our experiments to aid in the construction and augmentation of online dictionaries for novel verb senses and we are currently porting these results to new languages using online bilingual lexicons.

REFERENCES

Dorr, Bonnie J. and Douglas Jones (1996a). "Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues," Proceedings of the International Conference on Computational Linguistics, Copenhagen, Denmark.

Dorr, Bonnie J. and Douglas Jones (1996b). "Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision," Proceedings of the Workshop on Breadth and Depth of Semantic Lexicons, 34th Annual Conference of the Association for Computational Linguistics, Santa Cruz, CA.

Dorr, Bonnie J., Joseph Garman, and Amy Weinberg (1995). "From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT," Machine Translation, 9:3-4, pp.~71-100.

Nirenburg, Sergei (1996). "All the Lexical-Semantic Flowers Bloom, Each by Itself", Proceedings of the Workshop on Breadth and Depth of Semantic Lexicons, 34th Annual Conference of the Association for Computational Linguistics, Santa Cruz, CA, 1996.

Levin, Beth (1993). "English Verb Classes and Alternations: A Preliminary Investigation", University of Chicago Press, Chicago, IL.

Saint-Dizier, Patrick (1996). "Semantic Verb Classes based on 'Alternations' and on WordNet-like Semantic Criteria: a Powerful Convergence", Proceedings of the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases, Toulouse, France, 1996.

****

Michelle Vanni
Dept of Defense

Pre-Workshop on Interlinguas
Position Paper

In this paper, I would like to address the second and third issues presented in the announcement, namely, "What information is captured by an adequate interlingual representation system?" and "How can interlingual representation systems be built or scaled up?" My basic position is that while interlinguas are comprehensive and well-designed, there is a heavy reliance on the lexicon to carry meaning feature information to the interlingua. Too little investment is made in processing the interaction either among overlapping lexical feature values or between such values and the output of analyzers at other levels. Investigation of new approaches to such processing holds great promise for improving the efficiency and effectiveness of MT systems using an interlingual design.

It seems that we have managed to consult our semantics and pragmatics books and design into our systems nearly all those aspects of meaning which may be relevant to text understanding. Interlinguas are quite well-developed. In ULTRA'S Intermediate Representation (IR), referential, rhetorical, and intentional aspects of communicative acts are fully described with up to 52 possible fillers for some slots! Robustness of design is not a problem. But, how does one determine where the values for the features are encoded in the source language? Where is the research being done to accomplish this?

One way in which the assignment of feature values can be optimized is through investigation of overlapping values and the provision of defaults in the case of overgeneration. During preliminary stages of the Pangloss project, analyses were performed on candidate IRs output by the ULTRA parser for a set of 25 sentences. In the course of this work, it was noticed that most of the variations between IRs for a particular sentence were binary alternations in the filler (value) of a particular feature slot.

For example, since frequently in Spanish there is no difference between the written representation of a declarative sentence and that of an interrogative one (except for punctuation), almost all sets of candidate IRs for a sentence would include an alternation based on the choice of declarative v. interrogative. Add to this the general v. specific and the existential v. unique alternations for articles in a given sentence and you already have at least eight possible combinations. Analysts invoked the work of Halliday and Hasan (1976) suggesting that principles of textual cohesion which predict the values of particular slots or a high probability for a value might be integrated into the internal processing of IRs as constraints to limit their overgeneration.

The interaction between analyzer output and the lexicon is another area which could benefit from further research. Much of the information represented in one major interlingual design, the Mikrokosmos Text Meaning Representation (TMR), is initially coded in the ten zones available for each lexical entry: grammatical category, user information, orthography, phonology, morphology, syntactic features, syntactic structure, semantics, lexical relations, and pragmatics. I would like to suggest that values for some of these zones are misplaced in the lexical structure and that they can be better obtained at other levels of processing.

For example, in the Mikro lexical structure, indications of irregular morphological formation are supposedly noted in Zone 5, morphology. But, it is unclear what purpose this zone actually serves. In order to perform lexical search in a reasonable way, the lemma to be searched needs to be determined, even if the form from which it is derived is irregular, prior to lexical processing. Irregular forms, usually finite in number for a given language, are easily identified when listed in a morphological analyzer which returns a lemma with morphological features attached. It is only then that the form is efficiently searched in the lexicon.

A more effective use of the proposed morphology zone might be the identification of a unique sense of a form, that is, an attachment to an ontological node of a lemma associated with a particular set of morphological features which cannot be assigned to another form of the same lemma. An example might be be the different tense and aspect feature values associated with forms of the verb, _conocer_, in Spanish which point to different places in the ontology, as indicated in Guillen-Castrillo (1996).

Now it may be that the designers of this particular lexical structure had a specific linguistic issue or phenomenon in mind in creating their zones and it is not my intention to take issue with the features of any single interlingual design. In fact, in this case, there is surely a "microtheory" planned to address the issue. I only want to point out that, if we want these designs to be viable, we need to take a closer look at how we plan to populate them and seek to align them with what occurs in actual textual data.

Accounting for linguistic phenomena occurring in actual text has long been a goal of the computational corpus linguistics (CCL) community. In fact, the value of rational models for parsers designed to access linguistic data in corpora, given that those models account for a small percentage of the phenomena which actually occur in text, is regularly assessed. Once the limitations of the existing models are understood, issues revolve around the extent to which the models are valuable, the quantity of additional phenomena to be represented, and the criteria for determining which phenomena are appropriate to represent.

The IL MT and the CCL communities share the imperative to provide a measure of linguistic coverage in their system designs. Once it is realized that current IL models beg the question of how to access the value information with which to fill the feature slots, or at least address it only minimally at present, the real work of prioritizing such "microtheory"-type investigations and performing linguistic corpus analysis to determine how the values can be derived, can begin. One promising approach is the development of "Construction Grammars" proposed by Levin and Nirenburg (1994) for augmenting current lexicons with information to be gleaned from identification of the construction in which the lexical item occurs. It is my belief that extensive corpus analysis will lead to the discovery of more than a few construction types which can then be inventoried in a system to serve as a feeder to the IL of semantic feature-value information.

In summary, my interest in the pre-workshop regards deriving from the input text the information provided for in proposed IL structures (ULTRA IRs, Pangloss TMRs, and UNITRAN LCSes, etc). Integration of morphological information, corpus-based linguistic analysis, and work on developing construction grammars can be effective. While it is true that such work is provided for within the "microtheory" concept for Mikro, few substantial results or attempts at implementation of these ideas have been forthcoming possibly because no standard vehicle or way of talking about such progress has been established.

I think we need to (1) prioritize "microtheory"-type work, (2) work with corpora to find sense differences triggered by different morphological forms and grammatical constructions, and (3) progress toward development of a common language for reporting on findings. Already-robust IL designs can be enhanced with results of investigations of semantic phenomena occurring in actual textual data.

****

Structuring a Multilingual Multipurpose Lexical Database Using a Simple Interlingual Approach

Rémi Zajac
rzajac@crl.nmsu.edu

For structuring a Multilingual Multipurpose Lexical Database, we advocate the use of a simple interlingua based on word senses where concepts have no internal structure. This type of interlingua can be used for developing NLP lexicons from Machine-Readable Dictionaries and can serve as the foundation of more elaborated interlingual lexicons.

Background

CRL had and has several multilingual projects concerning multilingual machine translation, multilingual tools for translators and multilingual information retrieval and extraction. The languages concerned include: Arabic, Chinese, English, German, Japanese, Spanish, Russian, and Serbo-Croat. From the breadth of lexical work being pursued at CRL, the need for a multipurpose multilingual database should be obvious. Let me explain more precisely what is meant by multipurpose in the context of the lexical work at CRL. The Mikrokosmos project is a multilingual machine translation project using an interlingua (the "Text-Meaning Representation") linked to an ontology; the Corelli project is a multilingual machine translation project using a glossary-based translation approach and lexical transfer; the Norm project has built a translator's tool-set including on-line electronic bilingual dictionaries; the information retrieval and extraction projects (part of the Tipster and TREC programs) use bilingual dictionaries and thesauri for generating multilingual queries.

A bare-bones interlingua

In order to build a multilingual multipurpose lexical database with limited resources, a rational choice is to use an interlingua structure which limits the number of mappings between the various languages described in the database. As I will argue thereafter, the use of a bare-bones interlingua, like the one advocated in Sérasset [94a, 94b] does not prevent the definition of lexical transfer relations (for transfer-based MT systems for example) and moreover, it is entirely compatible with more sophisticated versions of interlinguas, such as TMRs.

With drastic constraints on the resources available for building such a database, reuse of existing dictionaries developed at CRL, mostly from Machine-Readable versions of paper Dictionaries (MRDs), is the only approach we can use, and a sensible approach is to use a simplified version of the interlingua defined in the Ultra project. In this project, a concept of the interlingua has a one-to-one correspondence with a word sense of the Longman Dictionary of Contemporary English (LDOCE), and has a structure (which includes for example, the arguments for a predicative concept). In order to accommodate the constraints mentioned above, I advocate two changes to the definition of the interlingua.

It must accommodate various interlingual theories as well as transfer-based relation: the concept of the Corelli interlingua will not have any structure in itself but various theories can be defined and grafted on the interlingua, enhancing the database.

It must accommodate a wide variety of languages and be open to new languages as well as new lexical material: the concepts of the interlingua will not be restricted to the set of word senses from LDOCE, but will be the union of word senses found in all bilingual dictionaries used to build the database.

There are of course well-known problems associated with the proliferation and the management of concepts in this approach, problems that I will qualify, since they are certainly not of conceptual nature, as engineering problems. It must be noted that all interlingual approaches must solve this problem in some way, and they can, for example, choose to do so by limiting the number of concepts in the interlingua, with a trade-off: augment the complexity of the internal of a concept to be able to represent all sense distinctions in all languages.

```
                               /—> (TMR)
                              /
                             /
 <w_xl,w_x,#cx.l,#cx> <—\              / uK
                  \         /
                   \       /
                  transfer   #cx < ----------------w_x
                    \     !                  LANGUAGE B
      [InterToken] <—\    \      !
                   \   \     !
                   Ultra \     O____!  ___
                    \   !        !
                     \ !         !
                      \!         !
      w_xl ----------- > #cx.l        !
                              !
                              !
            w_x2 -------------> #cx.l
     LANGUAGE A
```

Figure 1 : Translation mismatch and multi-theoretical approach.

Figure 1 shows the relationships between different components of the lexical database for the case of a translation mismatch between a word w_x in language B that can be translated as w_xl or w_x2 in language A. Each word sense has a concept (represented by an arbitrary symbol) in the interlingua: #c_x and two 'sub- concepts' #c_x.l and #c_x.2 (I will come back to the notion of sub-concept or sense refinement in the next section). In the Mikrokosmos approach, each word sense is related to one TMR; in this case, according to the guidelines specified to mapping a sense to a TMR, the predicate of all three TMRs would probably be exactly the same concept of the Ontology (which has a rather different structure than the interlingua structure as presented here), the difference being expressed as a difference in some attribute [Meyer et al., 90, Nirenburg 94, Mahesh 96]. In Ultra, each concept (represented as a Prolog predicate) would correspond to an English word sense and all concepts *have* a translation in each language [Farwell et al., 93]. The 'super-concept', #cx, would simply not exist except in cases of true hyperonymy in English. Within the proposed interlingua structure, assuming we want to use an approach similar to Ultra's, it would be necessary to specify conditions and transformations on the mapping from one sense (concept) to another since not all concepts are linked to words in all languages. This structure would, however, support more directly a mixed interlingua and transfer-based approach such as the one adopted by EDR [90] which define contrastive relations between two lexical entries by referring to the associated concepts.

Formal properties

From a mathematical point of view, the interlingua has no existence of its own and is no more than a convenient trick to represent in a compact graphical notation, a relation between word senses. In our approach, a word sense has no structure, it is merely a symbol in some set which is defined as the set of word senses in a given language, a convenient way of referring to a lexical sub-entry describing this word sense. Similarly, a concept has no internal structure, it is only a way of relating synonymous word senses between various languages: it simply defines a tuple <t1, t2,-, tn> of word senses t with ti being a word sense in language i. Cases like the one shown in Figure 1 add some interest to this otherwise rather boring structure. To simplify the notations, suppose that we have only 2 languages A and B: Figure 1 pictures the relation defined by the couples <w_xl, w_x> and <w_x2, w_x>, it is a compact graphical representation of the translation relation between these three word senses, factorizing the tuple notation by representing each element of a relation only once and using a disjunctive notation to represent ‘sub-senses' of the interlingua. This view suggests that we can derive simple formal properties on the interlingua from the relational model, for example, that in a given interlingua sub-graph, there must be a link to a word sense in each of the language, otherwise the translation relation is not well-formed.

From a linguistic point of view, however, the interlingua has classically a lattice structure representing hyperonymy, hyponymy and (true) synonymy relationships. If the monolingual parts of the lexical database contain also these relations in the lexical entries, these relationships can be used either for deriving similar relationships in the interlingua or for checking the coherence between the interlingua and the various monolingual dictionaries (all relationships in a given language must also hold -modulo transitivity of relations- in the interlingua). The process of creating the interlingua is then essentially the merging of monolingual lattices of word sense relationships.

Road-map

The construction of a multilingual multipurpose lexical database is not unrelated to the approach of Knight and Luk [94]. The emphasis, however, is not on building an ontology but on defining translation relations between word senses in various languages by pairing these word sense through the mediation of a simple interlingua directly derived from these word senses. This interlingua can then be used for supporting mappings to some ontology.

Using MRDs to build NLP lexicons is now a well understood and well documented process, especially in the initial phases of parsing, restructuring and complementing the dictionaries to build electronic versions [Véronis and Ide 92, Farwell et al., 93, Bauer et al., 94]. This is, however, only a preliminary step even with a bilingual dictionary such as the Collins Spanish-English dictionary. Since we want our bilingual lexicon to be reversible, we need to complement the target side by a monolingual dictionary, e.g., the LDOCE [Sanfilippo et al., 92] or use the reverse version of the dictionary (English-Spanish). Adding new dictionaries should be done with bilingual (or monolingual) dictionaries where one of the languages is already present in the database [Chua et Amat 94, Tanaka and Umemura 94]. This processes clearly involves many steps and a lot of meticulous work that must be carefully planned and for which an appropriate toolkit must be available.

References

Daniel Bauer, Frédérique Segond and Annie Zaenen. 1994. "Enriching an SGML-tagged bilingual dictionary for machine-aided comprehension". Technical Report MLTT-011, Rank Xerox Research Centre, Grenoble, France, October 1994.

Choy-Kim Chua and Salina A. Amat. 1994. "From a bilingual non-electronic dictionary to a ready-to-print bilingual/trilingual electronic dictionary". Proc. of the International Conference on Linguistic Applications, 26-28 July 1994, UTMK-USM, Penang, Malaysia. pp178-191.

EDR. 1990. Proceedings of the International Workshop on Electronic Dictionaries, November 8-9 1990, Oiso, Japan. EDR Technical Report TR-031.

David Farwell, Louise Guthrie and Yorick Wilks. 1993. "Automatically creating lexical entries for ULTRA, a multilingual MT system". Machine Translation, 8(3), pp127-145.

Kevin Knight and Steve K. Luk. 1994. "Building a large-scale knowledge base for machine translation". Proc. of the 12th National Conference on Artificial Intelligence, AAAI'94.

Kavi Mahesh. 1996. "Ontology Development for Machine Translation: Ideology and Methodology". Memorandum in Computer and Cognitive Science, MCCS-96-292, Computing Research Laboratory, New- Mexico State University, Las Cruces, NM.

I. Meyer, B. Onyshkevych and L. Carlson. 1990. "Lexicographic principles and design for knowledge-based machine translation". Technical report CMT-CMU-90-118, Carnegie Mellon University, August 13, 1990.

Sergei Nirenburg. 1994. "Lexicon Acquisition for NLP: A Consumer Report". In B.T.S Atkins and A. Zampolli (eds.), Computational Approaches to the Lexicon. Clarendon Press, Oxford, UK. pp313-347.

Antonio Sanfilippo, and Victor Poznanski. 1992. "The acquisition of lexical knowledge form combined machine-readable dictionaries". Proc. of the 3rd Conference on Applied Natural Language Processing, 31 March - 3 April 1992, Trento, Italy. pp80-87.

Gilles Sérasset. 1994. "Interlingual lexical organization for multilingual lexical databases in Nadia". Proc. of the 15th International Conference on Computational Linguistics - COLING'94, August 5-9 1994, Kyoto, Japan. pp278-282.

Gilles Sérasset. 1994. "Sublim: un système universel de base lexicales multilingues et Nadia: sa specialisation aux bases lexicales interlingues par acceptions". Ph.D. Dissertation, December 1994, Universite Joseph Fourier, Grenoble, France.

Komati Tankage, and Kyoji Umemura. 1994. "Construction of a bilingual dictionary intermediated by a third language". Proc. of the 15th International Conference on Computational Linguistics - COLING'94, August 5-9 1994, Kyoto, Japan. pp297-303.

Jean Véronis and Nancy Ide. 1992. "A feature-based model for lexical databases". Proc. of the 14th International Conference on Computational Linguistics - COLING'92, August 23-28 1992, Nantes, France. pp588-594.

Rémi Zajac. 1990. "A Relational Approach to Translation". Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, June 1990, Austin, TX, USA.

****

Developing Ontological Foundations for Interlingua

Kavi Mahesh
CRL, NMSU
mahesh@crl.nmsu.edu

1. What is an interlingua?

I would like to reiterate the position that an interlingua must be grounded in an ontology, especially in a multilingual context (i.e., for MT of more than one source or target languages). It is not possible to determine what an interlingua is without considering the purpose(s) it is intended to serve. MT is too nebulous a task; in the following, let us assume that interlingual representations must in fact support both interpretation and transfer:

- extraction of source meanings including word sense disambiguation, disambiguating literal interpretations from metonymic, metaphoric, and other non-literal interpretations, resolving syntactic and semantic ellipsis, coreference, etc.

- transfer of syntax, stylistics, non-literal usage, ambiguity, etc. when possible: e.g., word sense ambiguities can be carried along when the target language provides an equivalent ambiguous word

2. What information is captured by an adequate interlingual representation system?

Consider a related question: "what information must be captured in the ontology to support interlingual representation?"

- a mere taxonomy of primitives (or concepts) is NOT sufficient

- a rich set of inter-concept relationships must be present, especially to support meaning extraction in the presence of ambiguity, non-literal usage, semantic ellipsis, etc.

- default knowledge must be present: selectional constraints on inter-concept relationships must be "tight" to be useful. It is not prohibitively expensive to acquire precise constraints; it is much more practical to acquire constraints at two different levels: a default constraint that is as tight as possible and an overall constraint that is largely inclusive but still useful.

- uniform coverage of any kind of knowledge is of utmost importance. For example, if default values of attributes are included for a concept, they must be available for all other concepts that could have the same attribute. Without such uniformity of coverage and uniform grain-size of representation, knowledge acquired at great cost turns out useless during processing.

On the other hand, what information is not needed for MT?

- formal definitions are not needed; intuitive descriptions of concepts and their properties are sufficient. Formal definitions (in the form of necessary and sufficient conditions for each concept, for instance) are prohibitively expensive to acquire on a large scale.

- there need not be a well-defined distinction between every pair of siblings: e.g., the real difference between WALK and RUN is not useful for MT.

- moderate granularity and limited expressiveness of interlingual representations are indispensable virtues in practice. Almost any meaning can be decomposed into arbitrarily detailed and complex representations; we must limit this tendency and live with a coarse interlingual and ontological representation for practical reasons.

3. How can interlingual representation systems be built or scaled up?

Ontological (and lexical) knowledge is best acquired by following a situated development methodology: that is, every piece of knowledge acquired must be required for solving a real problem in a real MT situation and it must be put to use and tested immediately upon acquisition. Close cooperation among lexicographers, ontologists, domain experts, MT system developers and testing teams is inevitable for successful knowledge acquisition. An ideal situation is one where the ontology and lexicons (for both analysis and generation) for at least two different languages are being acquired simultaneously.

From our experience in building the Mikrokosmos ontology, we can claim that:

-10-20 person years of effort is sufficient to acquire a sufficiently broad ontology (about 50,000 concepts) with sufficiently rich inter-concept relationships and constraints

- the cost of acquiring such an ontology is NOT significantly greater than the cost of acquiring a lexicon with sufficiently rich semantic information, i.e., it does not introduce an unsurmountable bottleneck any more than what we already have in lexicon acquisition for interlingual MT

- ontologies are much more reproducible than many people think. There are striking similarities in concept organization and classification across all major ontologies (Cyc, Mikrokosmos, Wordnet, Sensus, etc.). It is not unthinkable to agree upon a common ontology for MT or merge previously acquired ontologies to build a broader foundation for interlingual MT.

****

Sergei Nirenburg
Computing Research Laboratory
New Mexico State University
sergei@crl.nmsu.edu

How can interlingual representation systems be evaluated?

In the final analysis, only through evaluating the success of applications based on it. Some partial evaluations can be attempted before, by estimating the combination of size, depth and breadth of coverage of the knowledge sources (see, e.g., Nirenburg, Beale and Mahesh, Measuring Semantic Coverage, Proceedings of COLING-96).

****

Topic #5: Apart from their role in support of MT,
what can IL representations be used for?

Using a Multi-Level Approach and Lexical Interlingual Forms in the NL Component of a Virtual Reality System

Clare R. Voss
Army Research Laboratory (ARL)
Adelphi, Maryland
voss@umiacs.umd.edu

The field of MT research lacks a consensus on what an interlingua (IL) is and how it is defined [Dorr and Voss (1993)]. MT system developers in building their individual interlinguas have drawn on a variety of semantic formalisms and have made quite distinct assumptions concerning the overall design of the MT systems in which their formalisms are embedded [Voss (1996)].

Vanderlinden and Scott (1995) point out that, even given the variation that currently exists among individual ILs, the variation has been bounded indirectly: the current IL-based MT paradigm assumes a content invariance in sentence-by-sentence translation, in effect creating an IL "ceiling" above which variation in content selection for the IL does not occur.

Thus, any argument for using IL representations beyond the MT application—in "non-MT" environments—must be made narrowly, in terms of the MT system design where the IL was defined and the ceiling or level of representation at which the IL's content was established.

In this brief paper, I take the general position that MT researchers need to make available their IL definition, development and evaluation for re-use outside of MT. Specifically below I take the narrow position that two aspects of a "working IL" in the MT research of Dorr and Voss (1993,1996) and Voss (1996),

(i) the multi-level system design of PRINCITRAN, in which distinct representational languages are used for different types of knowledge, and

(ii) lexical interlingual forms, in which the NL semantics of English lexical items is represented,are directly relevant to "non-MT" applications. The support for this argument comes from current research developing a natural language (NL) processing system for a virtual reality (VR) environment, a "non-MT" application under construction at the Army Research Lab (ARL) [Gurney, Klipple, and Voss (1996)].

Both PRINCITRAN and the NLVR system have been developed with special attention to the same semantic domain, namely representing spatial expressions, NL sentences that describe locational relations between physical objects in 3-dimensional space (e.g., a helicopter at the airport). The difficulties that arise in MT in identifying the range of interpretations for as simple a sentence as,

"the mouse ran between the chairs"

also arise in the NLVR system, albeit typically with different objects, as in,

"drive the tank between the buildings".

Consider, for a moment, several possible meanings for these sentences. Does the mouse/tank move TO a place between the chairs/building and stop? Or does the mouse/tank move PAST such a place on its way elsewhere? Or ABOUT in some path at such a location?

With respect to (i), the comparable multi-level design of the MT system and the NLVR system—i.e., where one can identify comparable levels of representation—makes it possible to designate where the ambiguity in the sentences above arise: namely the same places in the lexicon pre-runtime and in the parse trees at runtime.

Furthermore contributions from a discourse level of representation, not present in PRINCITRAN but in the NLVR system design (Gurney, Perlis, and Purang, 1995), are more readily assessed for integration back into MT, given the comparable system designs.

With respect to (ii), PRINCITRAN and, in due course, the NLVR system share a decompositional lexical semantics that distinguishes the semantic structure and the semantic content of its lexical entities. This also will make it possible to extend to both systems the cross-linguistic insights from research in spatial relations as well as measure phrases and aspect by Klipple (1991). It also leaves open the possibility that recent work of Asher and Sablayrolles (1995) can be tested within an NLVR system first and then, as relevant, brought to bear for translating spatial expressions in the MT system.

Ultimately the extension of IL research to "non-MT" applications ought to enable the MT community to both offer and take advantage of a wider range of software systems. Haller and Mark (1990), as just one example from the GIS (geographic information systems) community, report the significant need for an interlingua—to them, "a neutral yet expressive core of concepts"—that will support multiple representations of the same geographic object, arising both from multiple conceptualizations and lexicalizations of these objects cross-linguistically.

References

Asher, N. and P. Sablayrolles (1995) "A Typology and Discourse Semantics for Motion Verbs and Spatial PPs in French." Journal of Semantics, 12, pp. 163-209.

Dorr, B. and C. Voss (1993) "Machine Translation of Spatial Expressions: Defining the Relation between an Interlingua and a Knowledge Representation System." In Proceedings of the 12th Conference of the AAAI, pp. 374-379. Washington, D.C.

Dorr, B. and C. Voss (1996) "A Multi-Level Approach to Interlingual

MT: Defining the Interface Between Representational Languages." International Journal of Expert Systems, to appear.

Gurney, J., E. Klipple, and C. Voss (1996) "Talking about What We Think We See: Natural Language Processing for a Real-Time Virtual Environment." In Proceedings of the Second IEEE Symposium on Image, Speech, and Natural Language Systems [ISNLS], Washington, D.C.

Gurney, J., D. Perlis, and K. Purang (1995) "Active Logic and Heim's Rules for Updating Discourse Context." In Proceedings of the IJCAI Workshop on Context in Natural Language Processing. Montreal, Canada.

Haller, S. and D. Mark (1990) "Knowledge Representation for Understanding Geographical Locatives" In Proceedings of the 4th International Symposium on Spatial Data Handling, Zurich, Switzerland, pp. 465-477.

Klipple, E. (1991) "The Aspectual Nature of Thematic Relations: Locative and Temporal Phrases in English and Chinese." Ph.D. thesis, Department of Linguistics and Philosophy, M.I.T.

Vanderlinden, K. and D. Scott (1995) "Raising the Interlingual Ceiling with Multilingual Text Generation." In Proceedings of the IJCAI Workshop on Multilingual Text Generation. Montreal, Canada.

Voss, C. (1996) "Interlingua-based Machine Translation of Spatial Expressions." Ph.D. thesis, Department of Computer Science, University of Maryland, College Park, MD.

****

Topic Addressed: What Are Other Uses of Interlinguas?

Will Computers in the Future Speak English to Each Other?

Kevin Knight
knight@isi.edu

Computer programs enjoy artificial, unambiguous languages. That's how they talk to each other, and to us. Right now we have thousands of such languages and protocols. People can only master a few of these, and programs too. This heavily restricts who can talk to whom. If you call up an airline computer, you have to know exactly what to type.

Unfortunately, this restrictiveness lets a lot of air out of the promise of software agents. These agents are supposed to be autonomous and communicate freely with one another. When communication is restricted by language barriers, we wind up with impoverished hierarchical models (like the so-called "food chain"). I talk to the software travel agent, who talks to the airline computer. If an airline agent wants to coordinate with a hotel agent, they have to learn each other's languages.

People solve this problem with shared natural language. Interlinguas may solve this problem for software agent society. If the whole travel industry settled on a common set of terms, relations, and speech acts, then any agent could talk to any other one.

That would still leave people out in the cold, at least those who don't learn interlinguas. If our enterprise is modestly successful, there may be "interpreter" agents that translate between English and various interlinguas. Then again, it may be useful for each software agent to have its own "personal" interpreter (for example, one that knows that "plane" means "airplane" in its context). Then every program will have English capabilities. The question is: will they speak English to each other? To the extent that broad, shared interlinguas can be developed, they won't. But as long as language barriers exist, English may find a niche in software agent society. (Or, Chinese: which is where machine translation comes in.)

****

Martha Palmer
University of Pennsylvania
mpalmer@linc.cis.upenn.edu

Moving towards applications by augmenting verb classes with information structure

How, apart from their role in support of machine translation, might interlinguas be applied to various other information processing tasks (e.g., text summarization, information extraction, query systems, information retrieval, tutoring, multimodal communication, and the like)?

This depends very much on what the "interlingua" is. If it is a canonical "English" semantic network representation that includes coreference and links to the domain model(GRAIL <example>, Mikrokosmos <example>), then it functions very much like a standard deep semantic/pragmatic representation (a la Pundit <example>). As such it would provide an appropriate basis for performing the tasks that comprise MUCK evaluations: named entity, template elements, coreference, and finally scenario templates <example>. But that does not necessarily mean that it does a good job of capturing cross-linguistic generalizations.

Is it possible to include the hooks for cross-linguistic generalizations in the canonical "English" semantic network?   Yes, Mikrokosmos, also Pundit used LCSs, and Dorr bases an interlingua on LCSs, so it has to be possible - the primitive LCS predicates can correspond to supertypes of the verbs (and nouns). But that means that these supertypes must be carefully chosen to be universal or at least multilingual. Then they could be represented as either predicates or features that can map onto both source and target languages.

If the interlingua focusses more on cross-linguistic generalizations, (a la Dorr's LCSs <example>, a la verb classes in STAGs <example>) then it would need to be augmented with an information structure (a la Doran and Stone <example>), that would include the semantic and pragmatic information necessary for building an application. Does this then look any different from the canonical network mentioned above? Not necessarily, although it could allow two different languages to have different underlying predicate argument structures (a la STAG).

The two languages would need to share a discourse model and a domain model, and a set of common verb and noun supertypes that could be co-indexed in order to capture cross-linguistic generalizations. As long as the entities that are referred to by the arguments can be co-indexed by the source language rep and the target language rep, and the important cross-linguistic supertypes can be shared, then the representation can function as both an interlingua and a basis for applications requiring semantics/pragmatics.

****