

# Interlingua in Machine Translation

Author: Andrew Lampert (Andrew.Lampert@csiro.au)

Date: December 2004

Over many years of research, a number of different approaches have been made to tackle the problem of machine translation (MT). Generally, these approaches fall into one of the following three categories:

- Direct machine translation, which performs a word-for-word based substitution (with some local adjustment) between language pairs;
- Transfer based machine translation, which parses source text into a syntactic structure representation, then maps that using transfer rules to a syntactic structure representation for the target language; and
- Interlingua based machine translation, which converts source text into a language neutral, abstract meaning representation, then uses that representation to generate the target text.

Relative to the other main approaches, two principal advantages have always been claimed for the interlingua based approach. Firstly, because each language has an independent mapping to and from the interlingua, any number of source and target languages can be connected, without the need to define explicit rules for each language pair in each translation direction. A system to translate between each of  $N$  languages would require  $N$  components to translate each language into the interlingua, and  $N$  components to translate each language from the interlingua, for a total of  $2N$  components. Using transfer based machine translation, a separate component is required to translate in each direction for every pair of languages for a total of  $N(N-1)$  components. As a concrete example, extending a transfer based MT system for 3 languages into one for 5 languages would require 14 new components to be developed. Going from a 5 language transfer based system to a 7 language system would require the number of components to more than double - from 20 to 42 components. Thus, interlingua based systems have the potential to both save development time and reduce system size, especially for multilingual systems involving more than two languages. Examples of such highly multilingual systems are especially common in arenas such as the European Union.

The second claimed advantage of interlingua based systems, is that an intermediate language-neutral representation of meaning should be able to provide a neutral basis of comparison for equivalent texts that differ syntactically, but share the same meaning. This would be of great use in related fields such as information retrieval, where the current state of the art relies largely on syntactic matching for the gathering of relevant information. If natural language could be easily transformed into a semantic based interlingua, our ability to search for and find information could be dramatically improved.

Despite these apparent advantages, interlingua based machine translation has been used far less widely than other approaches. Perhaps the reason for this lies in the heinous difficulty of defining a universal, or even widely applicable, interlingua. Indeed, as noted

in the topic outline, whether or not building such a universal interlingua is possible, is an open philosophical question.

In general terms, an interlingua language must be designed to satisfy two primary constraints:

- It must be easier to accurately translate from the source natural language into the interlingua than directly into another natural language; and
- It must be easy to accurately translate from the interlingua into the target language.

Mapping between natural languages and the interlingua must be both accurate and computationally tractable, in order for the interlingua to be useful for machine translation.

Many properties of language act to impede such accurate and simple transformation. Some of these properties include:

- The form of text often under-determines the content, which causes problems when attempting to analyze the source text (e.g., in order to convert it to some interlingua); and
- The content often under-determines the form, which causes problems when synthesizing the target text.

Examples of where content under-determines form include sentences in English such as “*You’ll never recognize our David – he’s grown another foot!*”. Clearly, anything close to a literal interpretation of this sentence does not preserve the intended meaning. Similarly, sentences with structural, lexical or other forms of ambiguity cause great problems for all forms of machine translation. One of the problems behind this is that any fact at all can be crucial to correct interpretation and translation. Such facts may be derived from anything in the wide sphere of context of the utterance.

On the other side of the translation, there are problems of form under-determining content. An example proposition could be: “*David has a black dog*”. Depending on the context, there are many ways in which this meaning could be realized in text. Examples include: “*David has a dog. It is black.*”, “*There is a black dog. It belongs to David.*”, “*The dog which David has is black*”, or “*David has a dog which is black*”. A correct realization of the “meaning” can only be chosen if the relevant context of the original realization has been understood and modeled correctly.

These and other issues in performing machine translation mean that defining an interlingua that can unambiguously capture the appropriate meaning from any language, and can explicitly preserve the appropriate semantic, pragmatic and other contextual information is a hugely formidable, if not impossible task. Browsing through the efforts of Alexander Gode and colleagues at the International Auxiliary Language Association [2] serves only to reinforce the difficulty of this task.

When examining the work of Gode et al., we see that they have aimed at providing a standard international vocabulary. They clearly state that they believe that there are

enough “common elements” in the “speech forms of huge segments of civilized mankind” to create an international language. While this is a bold claim, it does not seem to be validated by the scope limitations they have needed to impose to complete their work.

Firstly, it should be emphasized that the primary focus of developing such an international vocabulary is not so much on the names that are given to concepts, but rather how they should be represented. Of course, there is a need to ensure uniqueness of names, but these names could be entirely arbitrary - even using numbers to represent concepts could work, though it is doubtful that such a vocabulary would be intuitive to work with! More challenging are decisions of which concepts to capture, and at what level. As an example, one might choose to incorporate a concept ‘CORNER’, which we can take as an interlingual representation of the English noun ‘corner’. While this works just fine for English, from the point of view of other languages, this concept is insufficient. Spanish, for example, has different words for inside corners (*rincón*) and outside corners (*esquina*). As a result, it seems necessary to refine our representation of the CORNER concept to be more specific – we might, for example, choose to have both OUTSIDE-CORNER and INSIDE-CORNER representations. But where do we stop?

Because different languages and cultures conceptualize the world in completely different ways, settling on a single vocabulary for a truly universal Interlingua would involve either arbitrary decisions about which language to base the conceptualization on, or a requirement to multiply-out all the many varying distinctions found in every single known language. Obviously, in the second case, there is the potential for an explosion in the number of concepts required. This could be addressed by reducing the set of primitive concepts, and defining more complex concepts in terms of those primitives. For example, the representation of the English word *kill* might be along the following lines:

[CAUSE [ BECOME [ NOT [ ALIVE ] ] ] ]

There are significant doubts in general about how feasible this process of lexical decomposition into semantic primitives might be [1]. In the example shown, potential problems lies in the fact that there are small but significant difference in meaning between *kill* and *cause to become not alive*. For example the concept of ‘killing’ is a single event, whereas ‘causing to become not alive’ involves at least two events (‘causing’, and ‘dying’). If the causal chain that links a particular event to dying is long enough, it may be correct that the event caused the dying, but it may not be true that there has been a ‘killing’. Such problems as these plague the selection of a universal set of language-neutral semantic primitives. If a universal set of semantic primitives were possible, the creation of an interlingua would also be possible. It remains an open question as to what a set of primitives would include, and even whether a set which captures all the required concepts can indeed be defined in practice.

Returning to our examination of Gode et al’s Interlingua-English dictionary, it should be noted that the authors have defined an international vocabulary, but have specifically avoided proposing a grammatical system with which to operate this vocabulary. This in

itself is a telling omission, as without some grammar system defining how to combine these words to represent meaning, the vocabulary is far from an “international language”. Similarly, the authors are tackling only the mappings from English into the international vocabulary, and so ignore or avoid many of the issues involved in a real, language-neutral interlingua.

As already hinted at, there are very significant restrictions in the scope of inclusions for this international vocabulary. Any word that does not occur in more than one national language is automatically discarded. The requirements for inclusion in their vocabulary are further expanded by the decision that “only those international words need be considered which have a fairly wide range of occurrence throughout regions of the world inhabited by peoples who participate in international intercourse”. In other words, the authors are gleefully simplifying their task by ignoring many thousands of languages that may express concepts that do not fit neatly into their model of an international vocabulary. Gode and his colleagues are also very open about their focus on science and technology as the “most important group of international words”. As a result, much of their focus is on words from science and technology.

The authors go even further and constrain their “restricted sphere of languages” to include only English and the Romance languages. The authors do claim that their narrow focus of what constitutes an international vocabulary does not preclude the possibility of shifting the boundaries in the interest of capturing further words and concepts. However, they state this, subject to a very significant proviso, namely that any additional languages must “still be held together by a common basis, which means, that their center of gravity remains in the Anglo-Romance sphere”. This reinforces the exclusion of many of the world’s languages from the “international” vocabulary. One must then ask, how international or representative can the vocabulary be, if it excludes such wide tracts of language families, including languages like Chinese and Japanese.

It seems undeniable that the interlingua vocabulary defined and collated by the International Auxiliary Language Association falls short of being a useful and practical interlingua for universal machine translation between human languages. While choosing the vocabulary is one of the central problems in defining an interlingua, the scope of their work is so narrow that it seems reasonable to question whether they actually believe their own claim that an “international language exists potentially in the common elements of the speech forms of huge segments of civilized mankind”.

In summary, while the goal of defining a universal interlingua is both intellectually stimulating and has many potential advantages, it seems highly unlikely that we will ever be able to create a truly universal interlingua. Indeed, the quest for a universal interlingua seems closely entwined with the linguistic theory of the universal “Ur” language, from which all other languages derive. Just as no “Ur” language has yet been shown to have ever existed, we are yet to see evidence of the possibility of a universal interlingua. What we may see, however, are interlinguas between specific sets of languages, or within clearly defined domains, where the set of primitive concepts is much more easily defined.

## References

1. Arnold, D., Balkan, L., Meijer, S., Humphreys, R. and Sadler, L., (1994) *Machine Translation: an Introductory Guide*, Blackwells-NCC, London.
2. Gode, A et al. <http://www.interlingua.org/> , Website for the Interlingua-English Dictionary from the National Auxiliary Language Association. Accessed on 1/12/2004.