# The European Union and the Semantic Web

**Neville Holmes,** University of Tasmania

**The European Union and the Semantic Web have a problem in common.**

A few months ago I got an e-mail from Alan Libert, a researcher engaged in writing a book about Esperanto and languages derived from it. He had come across the webpages I started putting up years ago in support of my essay "Languages and the Computing Profession" (*Computer*, Mar. 2004, pp. 102, 100-101). In the essay, I deplored the failure of the European Union to streamline the translation of their documents into the Union's various official languages and proposed E-speranto, a simplified dialect of Esperanto, as an intermediate language to form the basis for the streamlining.

To go by news reports from Brussels, the Union's language problems have been getting worse, which must threaten the long-term survival of minor languages and their rich cultures. Having an Estonian wife and Cornish maternal grandparents, I've fretted about this from time to time. Even major languages miss translation; the Italian Prime Minister has told his ministers to boycott EU meetings if documentation was not available in Italian (www.guardian.com.uk/world/2008/jul/14/italy).

The e-mail and, a little later, a closing remark—"Esperanto, anyone?"—by Robert Glass in his March/April 2008 *IEEE Software* column on the unsuitability of English as a lingua franca (pp. 96, 95) further stirred the coals with wild thoughts about the Semantic Web.

Just as the European Union has many languages, the Semantic Web has even more vocabularies (ontologies, in the jargon) as can be seen in, for example, its Wikipedia slot (en.wikipedia.org/wiki/Ontology_%28computer_science%29). Maybe they could both benefit from an intermediate language or vocabulary, and maybe both could exploit Esperanto's strict regularity.

## PHONEMIC SYNTHESIS

Esperanto is basically an Indo-European language. It aims to be uniform and consistent, and completely obedient to some 16 rules. However, I've found a simplification in Esperanto within its rules to be one of its most appealing features, a feature I proposed extending in my E-speranto. Esperanto uses synthesis at the level of individual sounds, its phonemes, to build or modify meanings systematically.

### Word endings

As an Indo-European language, E-speranto uses endings to express grammatical qualification of word stems. Thus, adverbs end in *–e*, infinitives in *–i*, and imperatives in *–u*. Synthesis starts showing in noun and adjective endings. Using a BNF-like notation, these are –[a|o](y)(n) where the required *a* or *o* signal an adjective or noun respectively, the optional *y* signals plurality, and the optional *n* signals accusativity.

Verb endings use the five vowels for a kind of temporal placement. Thus *a* signals here and now, *o* signals ahead or the future, *i* behind or the past, *u* conditional or propositional, and *e* perpetual or definitional. The full verb endings are *–as* for present tense, *–os* for future tense, *–is* for past tense, *–us* for conditional, and *–es* for perpetual. These verb endings can also be used as morphemes, so that *osa* is the adjectival *future* and *la aso* means *the present*.

The vowels are used as placers in other synthetic syllables that can be used as suffixes or ordinary morphemes. The formula [vowel](*n*)[*t*] yields 10 morphemes that qualify actions. With the *n* the action is active, without it passive, so *amanta* and *amata* are adjectives meaning *loving* and *loved*, and *ante* and *ate* are adverbs meaning *actively* and *passively*, all these set in the present.

This construction can be simply extended to specify horizontal spatial placement by [vowel](*m*)[*p*] and vertical by [vowel](*n*)[*k*], where the *m* and *n* specify placement relative to the sentence's subject. Also, *u* means *left* and *e* means *right*. Thus *la domompo* means *the house ahead* while *la domopo* means *the front of the house*. Such compounds can take some getting used to, but they are regular and powerful.

*Continued on page 106*

*Continued from page 108*

## Structural words

Synthesis at the phonemic level is even more expressive in its use when building the pronouns and correlatives essential to expressiveness. In Esperanto these center on the vowel *i*, with affixed phonemes to give the particular class of meaning.

The singular pronouns are *mi*, *ci*, *li*, *xi*, and *ji* for *I*, *thou*, *he*, *she*, and *it*, and *si* for reflexion. Here I use E-speranto spellings where *c* is pronounced like the *ts* in *tsunami*, and *x* as *sh*. There are also two plural pronouns: *ni* and *vi* for *we* and *you*. There are also two prefixes that can be used: *i-* to widen the scope and *o-* to abstract it. Thus *ili* means *he and his associates* or *they*, and *omi* means *someone like me* or *one*. The pronouns can also take grammatical suffixes such as *–n* for the accusative (*min* means *me*) and *–a* for the adjectival (*mia* means *my*).

The correlatives are two-dimensional, with one range of meanings given by a suffix, and an independent range by a prefix. The generic correlatives have no prefix. Having a simple vowel suffix, *ia*, *ie*, *io*, and *iu* mean respectively *some kind of*, *somewhere*, *something*, and *somebody*. Having a vowel+consonant suffix, *ial*, *iam*, *iel*, *ies*, *iol*, and *iom*, mean respectively *for some reason*, *at some time*, *in some way*, *someone's*, *in some number*, and *in some amount*.

The specific correlatives apply a variety of prefixes to the generic correlatives. The prefixes are *k–*, *t–*, *nen–*, and *q–* (pronounced *ch–*), and they give selective, indicative, negative, and inclusive meanings. For example, *kiam*, *tiam*, *neniam*, and *qiam* mean *when*, *then*, *never*, and *always*, respectively.

This description shows how phonemic synthesis can yield dramatic richness simply. Further, the correlatives can also take the *i–* and *o–* scoping prefixes, and both the pronouns and correlatives can be grammatically suffixed.

## THE SEMANTIC WEB

How is E-speranto relevant to the Semantic Web? E-speranto would be a simple and systematic language suited to being an intermediary between the various languages of the European Union to facilitate translation of both speech and text by professional translators and maybe by computers. An intermediary language following similar principles could facilitate translation of vocabularies or semantic ontologies between specialties and between natural languages for Web 3.0, the Semantic Web.

**Achieving a unified Semantic Web requires a unified standard vocabulary.**

### The need

The proliferation of formal vocabularies for the Semantic Web seems to result from various interest groups each focused on supporting their own needs. This is rather like a library divided into interest sections, each with its own independent topic classification. This makes cross-disciplinary research difficult and adventitious discovery through browsing less likely.

Achieving a unified Semantic Web requires a unified standard vocabulary. This aim faces many obstacles from political or marketing issues. The natural languages used in text being semantically indexed also pose problems.

First, any natural language, because it is a social artifact, would include homonyms and homophones, metaphors and jests, allusions and circumlocutions. Word meanings would overlap and vary in different ways for different users and would vary as time passes. These factors all make it hard to establish and enforce a standard vocabulary for broad use in any such language.

Second, developers must cater to more than one natural language, and mismatches between languages are much more severe and frequent than mismatches between dialects. It is unrealistic to expect English to persist as a lingua franca, even though machine translation will become more practical.

To make as much Web content accessible as possible to everyone, machine translation would indeed be needed, but here the use of a full intermediary language like E-speranto, with a grammar as well as a vocabulary, would be essential. However, only the vocabulary is needed for semantic indexing, and E-speranto's vocabulary is unsuitable for this. This requires a formally structured vocabulary, somewhat along the lines of the Dewey or Library of Congress classification schemes, but much more detailed and extensible—one that could have a grammar like Esperanto's built around it to make machine translation of text practical.

### The vocabulary

Because this vocabulary is intended primarily for use as an intermediary tool, researchers can build it without any need for borrowing from natural languages. This will free up many possible word forms. However, because at least some people will need to work with and discuss the vocabulary, the words should be briefly pronounceable. This rules out using numbers, as in the Dewey classification, or a mixture of letters and digits, as in the Library of Congress scheme.

These considerations point to using a synthetic scheme of the kind used for pronouns and correlatives in Esperanto, but in such a way as to specify a tree structure or hierarchy of meaning for the universal vocabulary. As a practical measure, the synthetic scheme suggested here will have fixed-length components.

At the topmost level, items of the vocabulary will have five phonemes: initial, prefix, vowel, suffix,

and final. Assuming the plain Latin alphabet as a basis, with each letter having a single distinct pronunciation, there are five vowels—*a*, *e*, *i*, *o*, and *u*; the prefix and suffix are chosen from eight semivowels or fluent consonants—*l*, *m*, *n*, *r*, *s*, *w*, *y*, and *z*; and the initial and final letters derive from the other 13 consonants. This arrangement will make all items pronounceable—easily by digital speech technology, if somewhat awkwardly at first for people.

At this level, the potential is for around 50,000 items in the five-level structure. Closing codes would be needed at each level to provide the tree structure, say *–lalb*, *–alb*, *–lb*, and *–b* to mark the level of the item in the tree. Thus, the 13 most general items would all end in *–lalb* and, for example, *plalb*, *pralb*, *prulb*, *prumb*, and *prump* would be a path through the vocabulary's tree. A weak point is that the second level down can only branch in five ways, but this could be remedied by adding long vowels, ideally marked by the traditional macron, a provision that would also double the possible items.

A tree of five levels would provide enough vocabulary items for general purposes, though it would be impractical to insist on even half the possibilities being put to use. Developers could then construct specialist vocabularies by adding an extra syllable.

The idea of using a single vocabulary to encode all knowledge must be at the heart of any conversion of the present World Wide Web to a single Semantic Web. To suggest that this vocabulary should be entirely artificial might seem wildly impractical, yet it is perhaps more practical than embracing all the world's more popular writing systems within one coding scheme that seems to need frequent subsetting and modification.

The difference is that meaning must be expressed at the level of words, not graphemes, and the words must be arranged in a hierarchy of meanings. Further, the vocabulary is not intended for general use, but only as an intermediary between texts in different languages and fields. Translation of terms into the intermediary vocabulary would remove ambiguities and idiosyncrasies, so that translation into another language would be better and easier. Also, for ordinary use, the intermediary vocabulary itself could be translated into other natural languages.

Ultimately, establishing an intermediary universal vocabulary could make automatic indexing and searching of text on the Web more effective and independent of the source language. Indeed, it could eventually make general text and speech translation by machine much more effective than at present, even unto Babel fish. Maybe the European Union would then pick it up. Ah, this is the stuff that dreams are made of.

***Neville Holmes*** *is an honorary research associate at the University of Tasmania's School of Computing and Information Systems. Contact him at neville.holmes@utas.edu.au.*