THE RETURN OF LOW-LINGUISTICS MT

PC-TRANSLATOR: CAN A LOW COST MACHINE TRANSLATOR DO THE JOB?

Linguistic Products, a Texas-based company, has been developing and selling PC-based machine translation systems since 1984. The company's current software, PC-Translator, is offered in seven language pairs covering Spanish, French, English, Swedish, and Danish. Each pair has English as its source or target language. The following evaluation was conducted by our resident MT expert Claude Bédard on both the English-Spanish and French-English packages.

Throughout the fifties, the general approach in machine translation was high on engineering and low on linguistics. Then came Chomsky followed by ALPAC – and the "Engineer's Song" fell into disfavor. The *linguists* picked up the gauntlet, and for thirty years, they've been spending millions teaching natural language to computers – to this very day without coming up with really convincing solutions.

In recent years, though, the engineers have had their second chance. While computational linguistics has been making interesting though slow progress, computer technology has exploded. Within the past decade, inexpensive micros have landed on everyone's desk and been intensively applied in office automation.

This has made low-linguistics MT once more thinkable – this time based on microcomputers. Compared to high-linguistics mainframe systems, these systems are user-controllable, relatively simple to use, less costly to develop, run on inexpensive hardware, and have their import/export operations considerably facilitated by the new office automation technology, which also permits onscreen postediting.

These advantages can now make MT an attractive proposition for hard-pressed document-cranking operations.

SEARCH AND REPLACE

PC-Translator is an eloquent example of a low-linguistics MT package. In fact, its linguistic features can be summarized in a few words.

Some words are shifted during processing; essentially, this is limited to adjectives and nouns. Regular adverbs don't have to be in the dictionary; if a word has a regular adverb ending and if the corresponding adjective is in the dictionary, the system will supply a translation based on the adjective. Inflected forms of nouns and adjectives – if quite regular – don't have to be in the dictionary; the system will find the root form. Adjectives agree with directly adjacent nouns.

The rest is essentially search-and-replace.

Now what exactly are the implications of a search-and-replace MT system from the user's point of view? The answer depends very much on the particular language pair considered.

On the *input* (dictionary coding) side, it means that the system has limited morphological capabilities – morphology is not only cute, it also provides for dictionary *economy*. Main result: user fatigue if the source language is highly inflected.

For instance, you have to enter each inflected form of every verb. This is bearable when *English* is the source language ("play, plays, played, playing"). But it can get out of hand in Spanish or French, with dozens of different forms per verb.

On the other hand, search-and-replace has the advantage that coding each entry has never been easier. Syntactic information is reduced to the word class – like V for verb. Nouns require gender and/or number as applicable.

PC-Translator has three types of dictionary: a "core dictionary" and several modifiable "user dictionaries," both for single-word entries; and a modifiable "phrase dictionary" for two-or-more-word entries. There's no morphology for the latter, so noun compounds, for example, have to be entered in both their singular and plural forms.

The number of entries in the core dictionaries is 21,000 for English and 72,000 for French and Spanish as source languages. This clearly reflects how many more entries have to be made for the more inflected languages. (Note that these figures refer to *forms*, not root words, and can't therefore be compared directly with the number of entries quoted for systems with full morphology.)

RAW OUTPUT

On the *output* (raw translation) side, a search-and-replace MT package such as PC-Translator has no sense of context and therefore no basis for making decisions. The rule is simple: any given word can have only one translation. Main result, you guessed it: crude translation – especially if the target language is highly inflected and if word order is quite different. Here are five aspects of crudity:

The first is word agreement. Since verbs have no morphology, you can easily imagine what happens with conjugation: *I work = Yo trabajar, they work = ellos trabajar*, etc. You do get *trabajamos* for *we work*, provided you write an entry for it in the phrase dictionary.

Though adjectives and nouns are inflected, articles, surprisingly, don't agree with their accompanying nouns. The masculine article is used as the default and is applied in all cases. *El* mujer?!? Yes-sirree, bob . . . unless you code *the woman* as *la mujer* in the phrase dictionary.

Not surprisingly, PC-Translator's designers insist that their system works better *into English*, because of its relative lack of inflection.

Second, the system does not attempt to insert any new words (such as articles or basic prepositions).

Thirdly – and conversely – the system has no rules for merging words. So *will + verb* is translated in Spanish as *va a + verb*, which conveniently avoids the future tense. But then, *would + verb* translates less successfully into *quisiera + verb*. Again, you can get a true future or conditional — if you code the phrase as such.

Fourth, the system does not deal with homographs, or words that belong to more than one word class. For a word such as *light*, for instance, you have to choose the verb, the adjective, or the noun translation – and kiss the rest goodbye.

This can get disturbing for frequent function words, especially in French:

J'ai reçu des fleurs des champs. = I have received of the flowers of the fields.
Le marchand le lui a dit. = The merchant the him has said.

Of course, the same goes for multiple-meaning words. The cure, says PC-Translator's manual, is – as for homographs – to code your own choices in the user dictionary. These new entries then override corresponding ones in the core dictionary.

Fifthly and lastly, word shifts are strictly limited to nouns and adjectives – though this does not apply to noun strings, which are left untouched. For example:

emergency exit door =
urgence sortie porte
porte de sortie d'urgence = door of exit of emergency

If you want these two translated right, welcome back to the phrase dictionary. By this time it may be dawning on you that the main asset of a search-and-replace MT system is indeed its phrase dictionary. The designers make no secret of this, encouraging the user to code as many phrases as s/he need.

LET'S GET PRAGMATIC

What can you expect from a search-and-replace MT system? Repetitiveness is the key word here. If all those phrases you pour into your dictionary occur many times in your text – which must be severely restricted in style and vocabulary – then they will be a good investment.

And if your quality requirement is moderate to low, coding can be less exhaustive. Also, despite my comments above on verb forms, for some technical documentation you may only need verbs in the third persons (singular and plural) of the present tense or the infinitive. Even so, in all cases, brace yourself for some hefty postediting.

(continued on page 57)