

GTS: NEW AND AFFORDABLE MACHINE TRANSLATION

Remember a machine translation system called TWP, marketed by TII Translation Services? Well, here's the news: TII has been renamed Globalink Language Services after being acquired by Hadron Inc., and TWP has re-emerged under the name Globalink Translation System (GTS). Since TWP had only been sold to a handful of clients before 1989, GTS is now making a fresh start with a version labelled 1.0.

Chief designer of the system is Czech-born Bedrich Chaloupa, a former member of the old Georgetown MT project which succumbed to the ALPAC Kiss of Death in 1968. Chaloupa wouldn't let go, however, and pursued his design efforts through the seventies, first on mainframes and then on microcomputers. The result is a reasonably sophisticated system which runs on PCs and sells for the surprisingly low price of US\$2000 per language pair.

Five language pairs are now available: French-English and reverse, Spanish-English and reverse, and German-English. Other pairs are under development: English-German, Italian-English, Russian-English, and Chinese-English. This review is based on the French-English and English-French pairs.

All packages require a PC-XT (minimum) with 15 Mb of free disk space. This means that for most users a 20 Mb harddisk will not be big enough. Adding the opposite direction for the same two languages, however, may require only an additional 1 Mb on top.

LINGUISTIC PROCESSING

What do you get for \$2000? Certainly more than GTS's former name (the modest "Translating Word Processor!") would suggest. The full range of basic linguistic processing is supported. GTS detects all inflected forms, so that you only need to enter canonical forms in the machine dictionary. Homograph choices are made. Most agreements, word shifts, word insertions, and word deletions are handled.

GTS fits into the tradition of the so-called "direct" MT systems, the doyen of which is Systran. These systems feature rather "flat," unscholarly linguistic processing, operating on wordstrings rather than on syntactic trees, with the emphasis on dictionary coding and ad hoc rules. The rationale here is that working with sophisticated syntactic trees can easily become unwieldy both computationally and linguistically.

The price to pay for this is local analysis. That is: parsing (analysis of the source language) is based on applying local rules, looking right and left for context through a "window" which is often rather narrow. A complete parsing of the sentence as a whole is not attempted, which certainly puts a ceiling on output quality.

GTS uses no semantic codes (like ANIMATE, CONCRETE, LIQUID) on which

to base decisions. Parsing is purely syntactic. This can lead to unresolved ambiguities. In the phrase "metallic safety bulkhead," for example, the program has no way of determining which word "metallic" refers to. If "safety" were coded ABSTRACT and "bulkhead" CONCRETE, the system could give preference to the concrete noun for this adjective – provided it had been coded accordingly.

LET'S TALK DICTIONARY

The GTS dictionary, as in most commercial MT systems, is divided into two subdictionaries: one for single-word entries (including grammatically critical words, which are not user-modifiable), and one for multi-word entries (called the Semantic Unit Dictionary, although there is nothing semantic about its coding – the term simply refers to the common notion that the global meaning or "semantics" of a phrase is often different from that of its individual words). Multi-word entries must still be entered in all inflected forms, although this will be changed in the program's next upgrade.

The GTS user does not have access to fancier dictionary coding, such as conditional entries or phrases with gaps for variables.

A rather unusual feature is that GTS bilingual dictionaries are in fact made up of pairs of monolingual dictionaries, related by pointers. On the one hand, this design provides for full reversibility – and thus economy in storage space and coding effort if you work in both directions. On the other hand, it makes things confusing when you first update a dictionary.

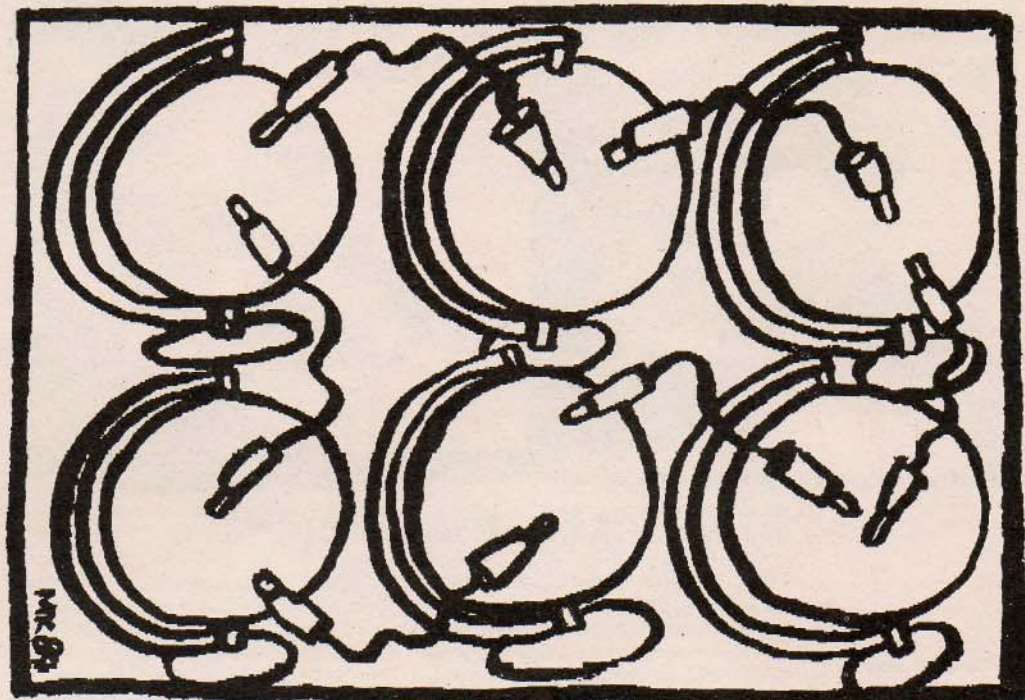
Another aspect of this peculiar structure is that the multi-word subdictionary has provision for stacking synonyms. All synonyms are used as source language entries, but only the top one in the stack can be active as a target-language entry. The order in the stack can be switched manually.

The GTS dictionary is monolithic; only one global dictionary is used at runtime. The only way you can customize your system is by modifying existing entries and adding new ones in this global dictionary. You cannot create smaller additional dictionaries and use them "on top of" the general one, as is the case with some other MT systems.

Thus, a word of solid advice: do plan your dictionary customization ahead, or you're headed for trouble. If you don't see the point, suppose, for example, that you have two translation applications: general and pharmaceuticals. If you want to enhance your general dictionary at all – as indeed you should do – you have to do it first. Then you make a copy of it, which will become your pharmaceuticals dictionary, and add your pharmaceutical terms to this.

As soon as you start using separate copies, any improvement on vocabulary common to both dictionaries has of course to be entered separately into each one. And if, for instance, Globalink later brought out a pharmaceuticals "microdictionary" – which would in fact be the full

WHILE THE WORLD IS STILL PATIENTLY WAITING FOR AN MT SYSTEM TO YIELD IMPRESSIVE QUALITY, USERS ARE SLOWLY TURNING AWAY FROM PURE LINGUISTIC PERFORMANCE ISSUES AND ADOPTING A MORE PRACTICAL APPROACH.



**TRANSLATION
SPEED IS ABOUT
6,000 WORDS
PER HOUR ON A
PC-XT, AND
20,000 WORDS
PER HOUR ON A
386.**

current general dictionary with pharmaceutical terms added – you could only use it instead of your own dictionary. This would mean that you would lose all the customization you had put into it.

Globalink is currently preparing several specialized microdictionaries, which will sell for about \$500 apiece. Hopefully, the company will also come up with a dictionary merge facility. Otherwise, this monolithic design looks to me like a time bomb for multi-domain users.

Current packages are shipped with a single-word subdictionary of about 20,000 canonical forms, and a multi-word subdictionary with only 1,000 entries – which is definitely too thin. System upgrades will include dictionary enhancements (in this particular case, merging with user-customized dictionaries should be possible).

USABILITY FEATURES

Translation speed is about 6,000 words per hour on a PC-XT, and 20,000 words per hour on a 386. The system can translate sentence by sentence “interactively.” You can update the dictionary and post-edit between sentences, but this is awkward and not recommended. The alternative is the usual batch mode.

Coding the dictionary is not too difficult, but not altogether easy. It does require guidance to supplement user literature, which I feel is neither complete nor explicit enough. User literature is avail-

able in all languages covered: English, French, German, and Spanish – a rather unusual (and very welcome) effort in an Anglo-dominated MT industry.

As for peripheral features, GTS prepares a list of words not found while running a translation, but offers no other text pre-analysis tools, such as searching for potential noun phrases (possible technical terms). It cannot import vocabulary lists – which would be useful for massive dictionary-making. Finally, GTS does not handle formatting codes: texts submitted for translation must be pure ASCII files.

SHOULD I BUY IT?

All the GTS packages currently available were launched in 1989, and the current version 1.0 still shows signs of hasty coding – as is the case with any MT system when it hits the market. The raw output is of reasonable quality but reveals youthful errors both in the dictionary and the linguistic rules, which I feel are readily correctable. As I mentioned earlier, the effects of the multi-word subdictionary’s small size are quite noticeable.

Wisely, Globalink offers upgrades up to version 2.0 free of charge. These should iron out those qualitative bugs. Indeed, running a test on pre-release version 1.1, I could see noticeable improvements. So despite the abovementioned shortcomings of version 1.0, my impression is that

the developers know what they’re doing and may get to make the best of their “direct,” no-semantics design.

GTS does not change the world MT picture one bit as far as output quality is concerned. All you can expect from current commercial MT systems is limited quality output (depending on the difficulty of your texts and on dictionary customization) which demands postediting to achieve publication-level quality (which again may vary according to your standards), and GTS is no exception.

But from the dictionary coding features and the output quality, it is clear that GTS is in the same class as, say, the WCC MicroCAT, which sells at a much higher price. This is definitely good news.

While the world is still patiently waiting for an MT system to yield impressive quality, users are slowly turning away from pure linguistic performance issues and adopting a more pragmatic approach. If a system supports corporate terminology and phraseology, while not making a mess of the target language – and is easy enough to use and not too pricy – it might prove to be useful and cost-effective. A system like GTS might just do the trick.

– Claude Bédard.

*Globalink Language Services Inc., 9990 Lee Highway, Fairfax, VA 22030, USA.
Tel: +1 (703) 359 6270*