

Future Tense for Euro Systran?

The European Commission is officially phasing out support for its Systran machine-translation facility this year. Language International looks at what's ahead for the venerable EC MT workhorse.

REPORT BY COLIN BRACE

In the mid-1970s, at the instigation of a far-sighted EU official, Loll Rolling (now retired), the Commission made its first incursion into the world of machine translation (MT). It acquired a license to exploit Systran, and began beefing up the system's dictionaries with EU terminology and developing additional language pairs that the Commission required. While sharing its origins, the Commission's system, for practical purposes, has scarcely more than name in common with the commercial PC-based software developed and marketed by California software company Systran.

Since the beginning of the 1990s, use of MT at the Commission has soared, primarily due to the adoption within the organization of email, which simplifies submitting and retrieving texts, but partly thanks also to some judicious internal promotion. In 1996, some 220,000 pages were run through the system, making the Commission, volume-wise, faraway the most prolific user of MT in the world. The Commission's Translation Service (SdT) accounted for slightly less than a third of this volume.

with the remainder being nonlinguists in the many administrative departments.

Until now, funding for development and maintenance of Systran has been provided by DG XIII under the Multilingual Action Plans (MLAPs), but the success of the system within the organization placed the Commission in a quandary. DG XIII's *raison d'être* is funding research in telecommunications and language engineering, and it could no longer continue to justify subsidizing development of the system if it had truly passed from being an advanced research topic to a fully operational concern. So in 1995, DG XIII announced its intention to phase out its support by the end of 1997, and over the past two years the SdT has been contemplating the way forward.

In-House MT User Survey

One option is for the SdT to allocate funds from its own operational budget to support the use of MT within the Commission. To determine whether this would be appropriate, an extensive feasibility study was undertaken last year, encompassing a user survey, practical experiments with in-house translators, an

examination of legal issues, a market study, and a cost-benefit analysis.

More than 1,500 users—both translators and nonlinguists alike—responded to the survey, providing a very detailed snapshot of how machine translation is used within the Commission. With an eye on objective information, the SdT also surveyed a number of nonusers, people who for one reason or another do not avail themselves of the system. The results of the survey provide a unique picture of the use of MT within this vast organization.

Among its users in the administrative departments, the vast majority turn to MT for urgent translations that they might otherwise have sent to the SdT, for browsing, and for preparing draft versions of documents. Within the SdT, some translators consider that MT does not help them in their work, or remain opposed to the use of MT on principle. But many value the system's fast turnaround times, and find its vast terminological resources and its preservation of formatting to be important benefits. While post-editing machine output can



Dorothy Senez

be tedious, some translators, as Dorothy Senez of the MT Help Desk wryly notes, find consolation in the system's unintentional sense of humor.

Of course, the system is not cut from the same technical cloth in every detail, and the quality of its translations varies greatly among the language pairs. At the moment, the French-English, French-Spanish, French-Italian, and English-French language pairs are considered by users to be the best.

Feedback from the practical experiments carried out by the SdT shows that on average a translation time saving of 35 percent can be achieved, provided a number of conditions are met. Documents have to be of the appropriate type, post-editors should be experienced, and MT dictionaries prepared in advance. And the actual results will still depend on the quality of the language pair in question.

The study acknowledges that SdT users and administrative users have different requirements, but the general consensus is that the primary value of MT lies in its immediacy—MT is fast. They also perceive the need for improved linguistic

coverage as well as better promotion within the Commission.

So where does this leave the SdT with regard to the future of MT within the institution? While the exact details have yet to be hammered out, it appears that the SdT and DG XIII have reached a happy compromise. Now that the SdT's cost-benefit analysis has demonstrated to its satisfaction that the MT system both directly and indirectly benefits not just the Service itself but also the Commission as a whole, the SdT will now support the mature, operational language pairs.

DG XIII meanwhile has agreed to continue funding the development of other language pairs, under the famous subsidiarity principle—development will depend on co-financing by the relevant Member States. In other words, if a Finnish-English language pair is deemed a priority, the government of Finland will have to be prepared to partly underwrite the effort.

Commercial Arena

The Commission will be issuing calls for tenders for the maintenance of the most promising language pairs as well as for systems or services for languages not covered by the Commission's system or for language pairs which are of lesser quality. Keeping an eye on developments in the commercial arena, the Commission could conceivably license a language pair

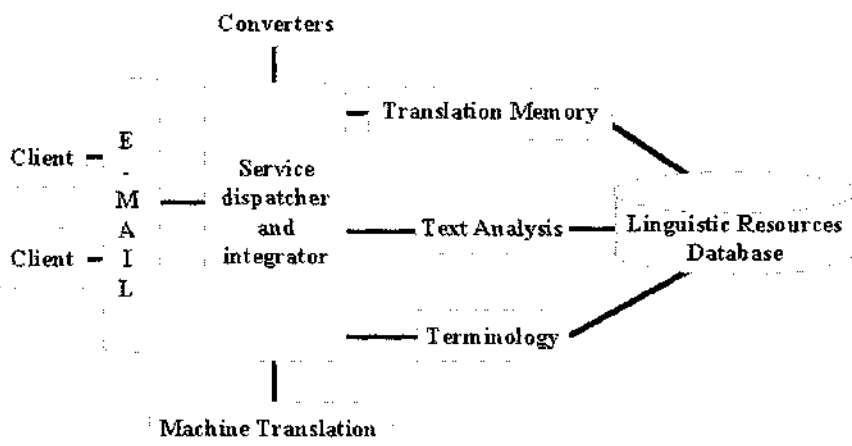
(continued on next page) ❖

Fact file: Reverse-Engineering Babel at the EC

With 11 official languages to contend with, the collective institutions at the helm of the European Union are probably the largest producer and consumer of translation in the world today. The most familiar of these is the European Commission, the enormous administrative wing of the EU divided between Luxembourg and Brussels. Others include the European Parliament, the Council, the Court of Justice, the Court of Auditors, the Economic and Social Committee, and the European Investment Bank.

Because the EU is committed to "linguistic equality" among its member states, all of the legislation of the Union needs to be translated into the 11 official languages. But it doesn't stop there. Countless calls for proposals, internal reports, and intermediary drafts need to be translated on an ad-hoc basis, not to mention the need for live interpreting of meetings. Moreover, as new countries join the Union and additional languages need to be reckoned with, the translation burden expands exponentially.

The major EU institutions all have their own internal translation departments, the largest of which is the Commission's Translation Service (SdT). With an army of some 1,500 full-time professional translators divided between Luxembourg (one-third) and Brussels (two-thirds), the SdT is the largest single translation organization in the world, producing over one million pages a year.



EURAMIS General Architecture

not covered by Systran from a third-party developer, should such a product become available.

By virtue of both its substantial internal translation requirements and its commitment to linguistic diversity, the European Commission is in many ways an exemplary test bed for language technology such as MT. As such, it is in the unique position of playing the roles of both user *and* mover. What lessons can be drawn from the Commission's experience by other, albeit smaller organizations?

Integrating MT into Workflow

For one, the MT development team and the MT user base (the SdT in particular) have enjoyed close proximity; feedback from the latter to the former has ensured practical results. Since few organizations can justify development of their own MT system (the Pan American Health Organization being an exception that comes to mind), this is admittedly an exceptional albeit pertinent factor.

In addition, the Commission has striven to integrate MT within the document flow of the organization. That means a substantial investment in the software-engineering side of things—such as document format filters and integration with email—admittedly prosaic matters which have all too often been given short

shrift by language technologists in the past. This task has been neither easy nor trivial (see "EURAMIS Tools Up" box).

The EC's implementation of machine-translation system Systran—the most sophisticated in the world—will likely remain locked within the EC for the foreseeable future.

Not least of all, the Commission has also expended tremendous effort building up the Systran dictionaries. The four top-rated language pairs boasted nearly 700,000 entries—and that was before the Eurodicautom data were imported. Currently, Systran has more than four million entries distributed across the 16 extant language pairs. This is a lesson that applies to both small and large MT users alike.

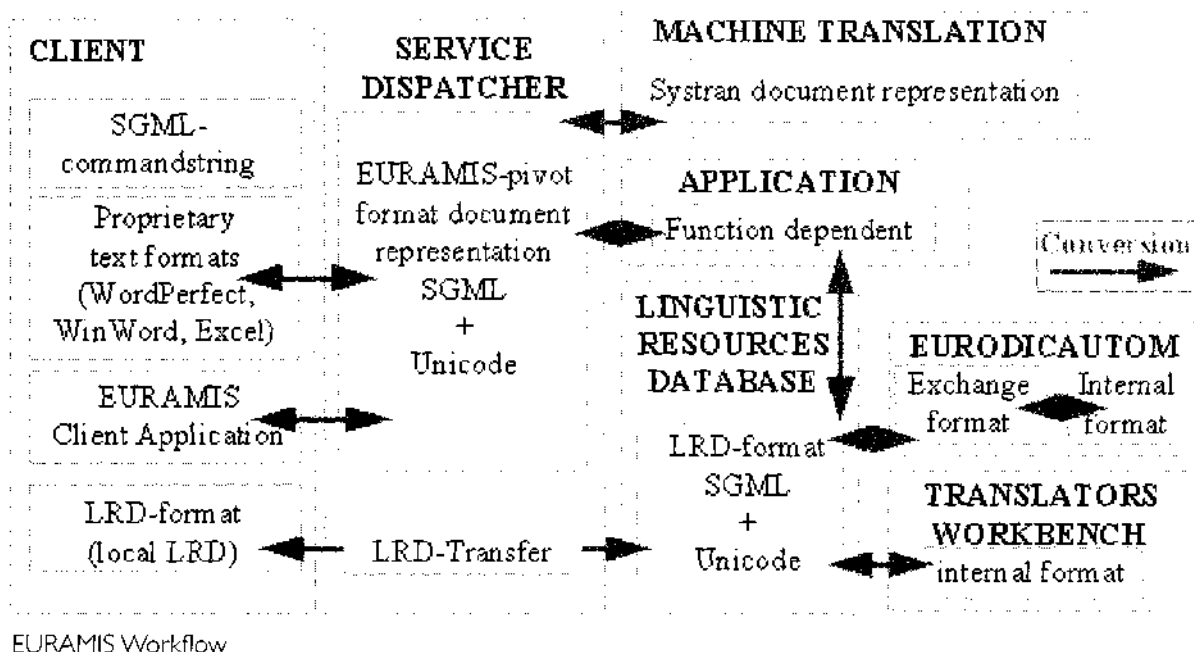
As such, given the Commission's need and resources, any inability to leverage this technology would have sent out a warning message to the MT community in general. As Dorothy Senez puts it, "If we can't make it work, who can?"

Insiders Only

Unfortunately, for interested parties outside the Union's institutions, the Commission's Systran is likely to remain an attractive suitor but out of reach for the foreseeable future. In accordance with deep-seated neo-liberal economic philosophy, access to the Commission's MT system is restricted to users within the EU institutions. The rationale for this is that the Commission should at all cost avoid distorting the competitiveness of the free market.

In any event, Systran, which originated from research at Georgetown University (Washington, DC) in the 1950s, looks poised to enjoy a rosy future well into the 21st century. If language is the soul of a culture, then the soul of a unified but multicultural Europe lies in its multilinguality, symbolized imperfectly yet impressively by Systran.

Colin Brace is a language-technology writer and consultant based in Amsterdam.



EURAMIS Tools Up

One way in which Systran is expected to integrate into the EC's SdT is via the EURAMIS project (short for European Advanced Multilingual Information System). Launched in 1994, and due to become semioperational for a pilot user group by the end of this year, EURAMIS aims to provide a single client interface to a panoply of server-based translation tools and resources for the Commission's translators, but also other end-users of translated documents. If it reaches its long-term goals, EURAMIS will very likely be the most comprehensive attempt ever to integrate 25 years of natural-language processing and tool development into a heavy-duty workflow architecture for translators and their "customers."

The key components of EURAMIS include a translation memory (TM) facility (presumably based on the Trados Workbench—Trados recently confirmed that it is providing their product to the European Union institutions for use by some 2,000 in-house translators)

(see News, page 4), a terminology extraction tool, the Systran MT facility, and a Linguistic Resources Database containing the system's complete linguistic entries, with links to Eurodicautom, among other term bases. Users with a document to translate will send an email request to EURAMIS which will automatically convert the document into SGML format with a Unicode character set, process it using selected TM and MT resources and email back a text with automatic candidate translations in the target language. Translators will then post-edit the result to deliver a final version.

