[*Language Industry Monitor* 18, November-December 1993]

Multext: Multilingual Corpora and More

With a budget of ECU3,210.000 for 238.5 person-months, MULTEXT is by far the largest of the LRE shared-cost projects (the industrials pay half of their own way). In brief, a two-tier group of academic and industrial partners from six countries will be testing and extending the TEI specifications to encompass multilingual corpora. They will also design and develop modular, largely language-independent software tools for corpora creation and analysis. In addition, the consortium will assemble a multilingual corpus of English. French. Spanish, German, Italian and Dutch for test purposes. Like several other LRE II projects, MULTEXT boasts a built-in evaluation mechanism as well; at a later stage in the project, its industrial participants will demonstrate the usefulness of these corpus-based technologies in appropriate applications. While an increasing number of corpus tools and English-language corpora are becoming available, there is a dearth of such resources for other languages, and this is one of the problems MULTEXT will be addressing.

If everything goes according to plan, MULTEXT's academic partners will obtain funding for support and subsequent dissemination of their research, its industrial partners (DEC, SNI, Cap debis, SITE, and Rank Xerox) will gain tools, materials, expertise which they could not otherwise develop or acquire, and other interested parties will have access to the results, which will be placed in the public domain. In the view of ISSCO's Susan Armstrong, Community funding in the form of MULTEXT is essentially the only remaining support in Europe for such vital R&D activities in the field of language processing. She suggests that Europe is at a comparative disadvantage to the US because it lacks strong R&D centers supported by the IT industry.