

## **XLT: Ready for Liftoff**

**One of the world's premier industrial NLP labs is launching its technology on the market.**

With the unveiling of Xerox Lexical Technology (XLT), Xerox is the most recent newcomer to the increasingly competitive OEM market for linguistic software. XLT is a toolbox which offers morphological reduction and generation, morphological derivation, part of speech disambiguation, and tokenization functions for a number of languages. XLT will be licensed in packs of three languages, with English-French-German available now, and other European languages, including Spanish, Italian, and Portuguese, available shortly. Eventually, Xerox plans to extend support to include Japanese, Korean, and Chinese as well.

XLT was developed at Xerox's renown PARC laboratories in Palo Alto, California, where researchers Martin Kay, Ron Kaplan, Lauri Karttunen, and others have virtually dictated the direction of computational linguistics over the past fifteen years. Based on the insight that many lexical processes, such as morphological analysis and generation, could be described in terms of finite-state mechanisms, Xerox PARC's researchers have spent the past decade building an array of so-called finite state transducers, bidirectional programs (they handle both analysis and generation) which represent all possible morphological mutations of a word in terms of mathematical relations. Xerox says these linguistic modules are clean, compact, efficient, and uniform, and the code which runs these is very small and language-independent.

"Implementing a two-step model is extremely tricky," points out Lee Humphreys, a computational linguist at Site Eurolang. "It's not for the faint of heart." The morphology modules in most of today's commercial software are based on simpler methods: large look-up and exception tables and algorithms which employ what Xerox calls "cut-and-paste" methods. Such techniques are workable, points out Xerox computational linguistic Ken Beesley, but they are not portable across many languages, highly inflectional languages in particular. "It is almost impossible to do Finnish morphology using conventional techniques," explains Beesley. "You really need a more sophisticated model like ours."

XLT is being marketed by the Desktop Document Services (DDS) division of Xerox, which Keith Loris, vice president of technology, says has the character of "a startup within Xerox." Loris claims that with XLT DDS is offering the linguistic engineering world software which has been "tested and hardened" by, among others, Xerox itself. Loris points out that Xerox linguistic technology is an integral part of TextBridge, Xerox Information System's OCR package. TextBridge has met with critical and commercial success, capturing twenty percent of the Windows OCR market in the past year. The underlying OCR engine has also been licensed to other software companies on an OEM basis for incorporation in, for example, fax software. TextBridge linguistic smarts comes in the form of a Lexifier, a function which discerns non-word strings embedded in text, such as social security numbers, telephone numbers, mathematical symbols, and multi-character Roman numerals, thereby improving the recognition rate of the program.

Loris also points out that XLT certainly isn't Xerox's first incursion into the commercial linguistic software market. Microlytics was a Xerox spin-off which enjoyed a measure of success in the 1980s by exploiting Xerox's advanced text compression technology in a variety of commercial applications and OEM products. However, it failed to achieve

lasting commercial success and recently licensed most if not all of its linguistic technology to InfoSoft.

**M**orphological processing is one of the corner-stones of language processing and it is likely to be found in virtually all mainstream software within a few years. Despite getting a late start, will Xerox nonetheless be able to move quickly and aggressively enough to capture a piece of this action? While few will quibble that Xerox's technology is fundamentally superior, Xerox must also offer commensurate breadth, both in terms of lexical coverage and the number of languages it supports, and this means lots more not so exciting lexicon coding. Xerox will also be faced with convincing potential takers that the Xerox technology has more than just first-class theoretical pedigree; the company is entering an environment where ad hoc solutions dominate. The prevailing attitude towards linguistics among software engineers who have had to deal with it can best be summed up as: "less is more."

Xerox Corporation, Advanced Office Document Services, 3400 Hillview Ave, Palo Alto, CA 94304, USA; Tel: + 1 415 813 6804, Fax: + 1 415 813 6792

COPYRIGHT © 1994 BY LANGUAGE INDUSTRY MONITOR