# An Application of the TEI Guidelines: Encoding Linguistic Corpora

Nancy Ide
Jean Veronis

Department of Computer Science
Vassar College
Poughkeepsie, New York USA

Laboratoire Parole et Langage
CNRS/Universite de Provence
Aix-en-Provence, France

The TEI Guidelines establish an intellectual foundation for text encoding and provide a set of encoding conventions comprehensive enough to serve the fundamental needs of many large-scale projects. Because the TEI scheme is intended to be maximally applicable across a wide range of disciplines and applications, it is necessary to build upon the TEI scheme to develop several sets of inter-dependent, customized text encoding standards, each suited to a specific domain and built upon the foundation of mechanisms and principles provided by the TEI Guidelines.

This poster describes a project to develop a TEI-based and TEI conformant markup standard optimally suited for use in language engineering, which can serve as a widely accepted set of encoding standards for corpus work in the field of computational linguistics. The overall goal is the identification of a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited for use in a text database). It is also intended to provide other levels of more extensive encoding specifications depending on different types of use and applications.