
Une application de la TEI aux industries de la langue : *le Corpus Encoding Standard*

Nancy IDE^{a,b} et Jean VÉRONIS^b

^a*Department of Computer Science
Vassar College
Poughkeepsie
New York, NY 12601, USA
ide@cs.vassar.edu*

^b*Laboratoire Parole et Langage
Université de Provence et CNRS
29, avenue Robert Schuman
13621 Aix en Provence Cedex 1, France
veronis@univ-aix.fr*

1. Motivation

La *Text Encoding Initiative (TEI)* couvre un ensemble de types de textes extrêmement large (prose, poésie, théâtre, dictionnaires, bases terminologiques, etc.) et vise les domaines les plus variés (publication électronique, analyse littéraire et historique, lexicographie, traitement automatique des langues, recherche documentaire, hypertexte, etc.). Le jeu de balises qu'elle offre, ainsi que la Définition de Type de Document (DTD) est donc nécessairement à la fois trop riche et trop pauvre pour une application donnée. Trop riche, car la plupart des balises ne sont pas utiles dans l'application en question et risquent d'être source d'erreurs ; trop pauvre parce que, aussi minutieux que soit l'effort des différents comités de travail, il restera toujours des besoins ponctuels de codage de phénomènes particuliers propres à une application et pour lesquels aucun mécanisme général de codage n'aura été fourni.

Les concepteurs de la TEI, conscients de ces difficultés irréductibles, ont bâti les jeux de balises et la DTD autour de principes modulaires qui d'une part permettent de sélectionner de grands sous-ensembles du jeu de balises par catégorie de texte (les jeux de base : prose, poésie, etc.) et par type de codage désiré (les jeux additionnels : liens hypertextuels, etc.) et, d'autre part, permettent d'étendre à volonté la DTD par de nouvelles balises, voire de redéfinir les balises existantes. La TEI offre ainsi les moyens de personnaliser la DTD tout en conservant un certain nombre de traits communs, tels que les balises du « noyau », dont l'indispensable « en-tête » décrivant le document (*TEI header*). Chacune des personnalisations de la DTD, accompagnée le cas échéant

de restrictions et de recommandations précises d'utilisation des balises, peut être vue comme une «application» de la TEI, comme la TEI est elle-même une «application» de SGML.

Nous avons développé une telle application, le *Corpus Encoding Standard* (CES) ou «Standard de Codage des Corpus», dans le cadre du projet MULTTEXT, en collaboration avec le sous-groupe *Text Representation* du projet EAGLES. En effet, l'ingénierie linguistique utilise de plus en plus de grands corpus de textes pour la mise au point de modèles de langage (par exemple probabilistes) servant à la construction ou à l'amélioration d'outils logiciels pour l'étiquetage morpho-syntaxique de textes, la construction semi-automatique de terminologie, l'assistance à la traduction, etc. Le CES fournit un ensemble de balises et des DTD qui sont spécifiques au codage des corpus de textes pour les besoins de l'ingénierie linguistique, ainsi qu'un ensemble détaillé de recommandations pour l'usage des balises, et leur sémantique précise dans le contexte des corpus.

2. Principes

Le CES introduit une distinction entre données primaires (c'est-à-dire les textes originaux) et les annotations linguistiques (étiquetage morpho-syntaxique, alignement de texte multilingues, etc.) rajoutées aux données primaires.

Un des grands principes préconisés par le CES est la séparation des données primaires et des annotations : les annotations sont enregistrées dans des documents séparés qui pointent vers les données primaires par des liens de type hypertexte. La raison sous-jacente en est que les annotations linguistiques sont potentiellement infinies pour un même texte primaire : on peut rajouter divers étiquetages morpho-syntaxiques, une lemmatisation, une désambiguïsation du sens des mots, divers alignements avec des traductions du texte, etc., et il devient rapidement difficile (et probablement inintéressant) de conserver la totalité de l'information dans un document unique.

Le CES fournit ainsi trois DTD qui constituent des personnalisations de la DTD de la TEI :

CesDoc pour le codage des données primaires,

CesAna pour les annotations linguistiques (étiquetage morpho-syntaxique, découpage en phrases, etc.),

CesAlign pour l'alignement de textes parallèles multilingues.

De plus, les efforts actuels de collecte de données textuelles pour l'ingénierie linguistique aboutissent souvent à la création de corpus de très grande taille et contenant des textes de types très divers (juridiques, administratifs, littéraires, journalistiques, etc.), obtenus sous des formes électroniques préexistantes et variées (bandes de photocomposition, etc.). La transformation de ces données en un format de type TEI de type fin peut-être très coûteuse. Le CES définit donc trois niveaux de codage des documents :

Niveau 1 Codage de la structure grossière (divisions) jusqu'au niveau du paragraphe.

Niveau 2 Codage grossier des éléments internes au paragraphe.

Niveau 3 Codage fin des éléments internes au paragraphe.

Ces trois niveaux de codage ont un coût croissant, puisque seul le premier peut-être (dans la plupart des cas) complètement automatisé. Le second peut généralement être partiellement automatisé, mais le troisième requiert une grande quantité de codage manuel.

Une première version du CES a été rendue publique (voir information ci-dessous), et est en cours de test dans plusieurs projets internationaux (MULTEXT, MULTEXT-EAST, PAROLE, les Actions de Recherche Partagée de l'AUPELF-UREF, le serveur SILFIDE du CNRS et de l'AUPELF-UREF, etc.).

3. Information

Le *Corpus Encoding Standard* est disponible en version électronique navigable sur le *World Wide Web* aux adresses :

<http://www.lpl.univ-aix.fr/projects/multext/CES/CES1.html>

<http://www.cs.vassar.edu/CES/>

L'information sur le projet MULTEXT se trouve à l'adresse :

<http://www.lpl.univ-aix.fr/projects/multext/>

et celle sur le projet EAGLES à l'adresse :

<http://www.ilc.pi.cnr.it/EAGLES/home.html>

Référence bibliographique

IDE, N., VÉRONIS, J., « MULTEXT (Multilingual Tools and Corpora) », *Proceedings of the 14th International Conference on Computational Linguistics, COLING'94*, Kyoto, Japan, 90–96, 1994.