Utiliser les outils CORLI de conversion TEI pour l'analyse de corpus de langage oral

Christophe Parisse, Loïc Liégeois

Utiliser les outils CORLI de conversion TEI pour l'analyse de corpus de langage oral

Parisse Christophe[1] Loïc Liégeois[2, 3]
(1) Modyco, Nanterre, France
(2) LLF, Paris, France
(3) CLILLAC-ARP, Paris, France
cparisse.parisnanterre.fr, loic.liegeois@univ-paris-diderot.fr

RÉSUMÉ

Le consortium CORLI développe des outils pour faciliter le dépôt, l'interrogation et la réutilisation des corpus oraux. Ces outils libres et open source sont basés sur la TEI comme format commun de partage. Nous présenterons deux outils différents : un outil pour la saisie et l'édition de fichiers de métadonnées et un outil permettant d'intégrer et d'utiliser des corpus de différentes sources de données transcrits dans différents logiciels.

ABSTRACT

The CORLI consortium develops tools to facilitate sharing, interrogation, and reusing of spoken language corpora. These free tools are based on TEI as a sharing format. We present two tools: a tool for metadata edition, and a tool to aggregate and use corpora transcribed in different formats and coming from different sources.

MOTS-CLÉS : TEI, PRAAT, ELAN, Transcriber, TXM, Metadonnées, Corpus.
KEYWORDS: TEI, PRAAT, ELAN, Transcriber, TXM, Metadata, Corpora.

1 Le partage de corpus de langage oral et la TEI

L'utilisation d'outils pour la recherche sur les corpus oraux est incontournable, que ce soit pour la transcription, l'édition, l'étude qualitative (PRAAT, ELAN, Transcriber, CLAN, etc.) ou pour l'analyse, soit grammaticale, soit exploratoire (outils de textométrie). Ces diverses applications à partir de formats et de conventions presque tous incompatibles requièrent des interventions manuelles lourdes et réduisent les possibilités de recherche scientifique. CORLI cherche à faciliter le travail des chercheurs en mettant à disposition des outils accessibles sans qualification informatique pour permettre de consacrer plus de temps aux analyses elles-mêmes. Les logiciels libres que nous avons développés sont tous basés sur un format commun en TEI, qui a l'avantage

d'être connu, libre, valide pour la pérennisation et assez ouvert pour s'adapter à la variété des corpus. Les outils disponibles à ce jour portent sur trois grandes fonctionnalités : l'édition des métadonnées des corpus et des enregistrements ; la conversion de formats pour l'édition des corpus ; la conversion de format pour l'exploration des corpus et l'utilisation d'outils syntaxiques sur les données convertis en TEI.

2 Trois exemples d'utilisation

L'espace de démonstration que nous proposons contiendra un poster qui présentera l'ensemble des activités de CORLI et en particulier celles qui sont liées à la problématique de l'usage de la TEI. Par ailleurs nous présenterons trois démonstrations basées sur des cas pratiques. La première concerne l'outil TEIMETA (http://ct3.ortolang.fr/teimeta/), les deux suivantes l'outil TEICORPO (http://ct3.ortolang.fr/teicorpo/). Chaque présentation permettra de préciser les compétences requises et les liens que l'on peut faire avec les sites de dépôt de corpus.

2.1 TEIMETA : création d'un fichier de métadonnées

Pour cette partie de la démonstration, nous nous mettrons dans la position d'un chercheur ayant à renseigner un ensemble de métadonnées liées à un fichier de transcription. Souvent long et fastidieux, nous verrons comment ce travail peut être facilité par l'outil au moyen par exemple de l'usage de vocabulaires contrôlés et comment le résultat s'intègre à tout fichier TEI.

2.2 TEICORPO : utilisation de PRAAT et la TEI

Il est possible de convertir des fichiers d'origine différente (PRAAT, ELAN, CLAN, Transcriber) vers la TEI et uniformisant autant que possible les transcriptions. Une analyse syntaxique basée sur TreeTagger ou sur la bibliothèque Stanford NLP (format CONLL) est également possible, avec un résultat peut alors être converti vers PRAAT pour permettre par exemple une utilisation dans PRAAT de données comprenant transcription et annotations grammaticales.

2.3 TEICORPO : utilisation de TXM et la TEI

Il est possible de rassembler par conversion des corpus issus d'origines variées pour construire une base au format TEI. Pour une utilisation scientifique pertinente, il est souvent nécessaire d'enrichir les métadonnées des corpus. En dehors de l'utilisation de TEIMETA, il est possible d'insérer des métadonnées éditées au format CSV, puis d'exporter les corpus pour leur utilisation dans TXM, avec des métadonnées requêtables et la possibilité d'écouter le son dans TXM. Un exemple basé sur le grand corpus ESLO sera présenté.