ACOLAD
un environnement pour l'édition de corpus de dépendances

Francis Brunet-Manquat, Jérôme Goulian

LIG-GETALP
BP 53
38041 Grenoble cedex 9, FRANCE
{Francis.Brunet-Manquat, Jerome.Goulian}@imag.fr

Résumé

Dans cette démonstration, nous présentons le prototype d'un environnement *open-source* pour l'édition de corpus de dépendances. Cet environnement, nommé ACOLAD (Annotation de COrpus Linguistique pour l'Analyse de dépendances), propose des services manuels de segmentation et d'annotation multi-niveaux (segmentation en mots et en syntagmes minimaux (chunks), annotation morphosyntaxique des mots, annotation syntaxique des chunks et annotation syntaxique des dépendances entre mots ou entre chunks).

Mots-clés : dépendances, chunk, édition

Intérêts

L'idée est de proposer dans un même environnement une palette de services permettant la création de corpus annotés pour les besoins classiques de développement d'outils linguistiques. Les intérêts de notre environnement sont les suivants :

- visualisation et édition graphique simple des annotations ;
- édition manuelle configurable :
  - choix entre différentes granularités d'unités d'annotations (dépendances entre mots et/ou entre chunks) ;
  - possibilité d'utiliser différents jeux d'étiquettes (grammaticales, syntaxiques au niveau des chunks et des dépendances) pour tenir compte des spécificités de chaque corpus (écrit, oral transcrit ou oral issu de la reconnaissance de parole par exemple) et des besoins en terme d'exploitation future du corpus [Valli et Véronis, 1999] ;
  - choix des contraintes structurelles de dépendances à associer à l'édition en cours (projectivité ou non projectivité, analyse totale ou partielle) ;
- possibilité d'éditer simultanément les ambiguïtés d'analyse tant au niveau mots, chunks, qu'au niveau des dépendances syntaxiques.

Aperçu de la démonstration

La démonstration vise à présenter les différentes étapes permettant la création d'analyses de dépendances : de la segmentation d'une phrase en mots (voir figure 1) à la création d'une ou plusieurs analyses de dépendances pour cette phrase (voir figure 2).

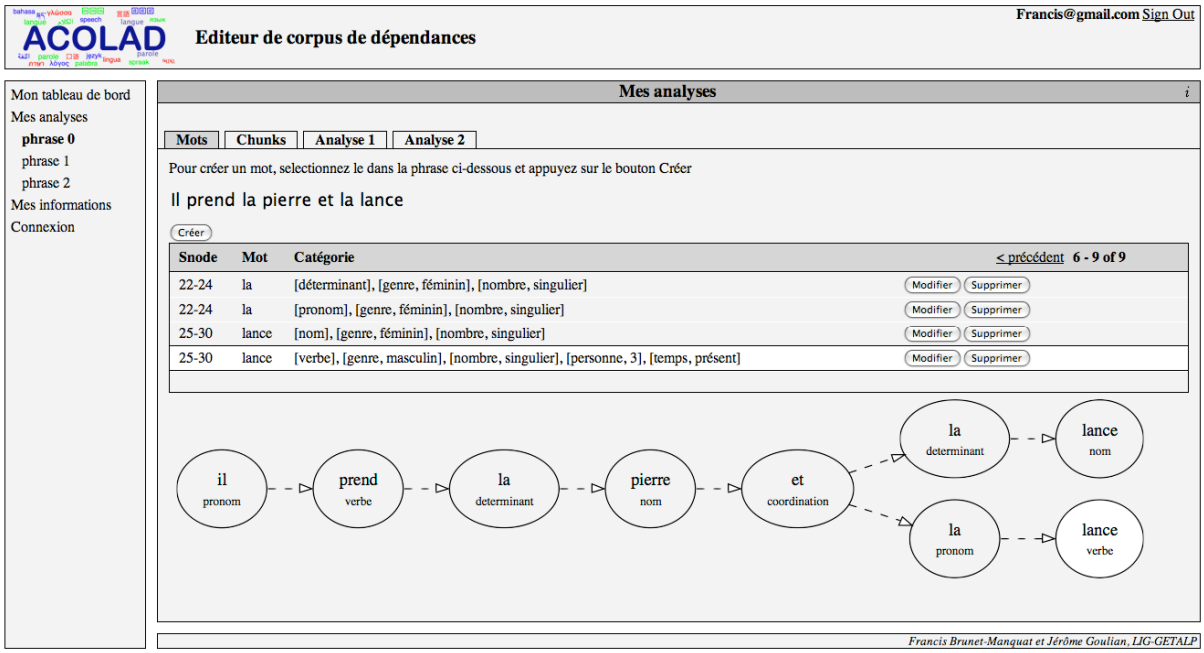

Figure 1 : segmentation d'une phrase en mots

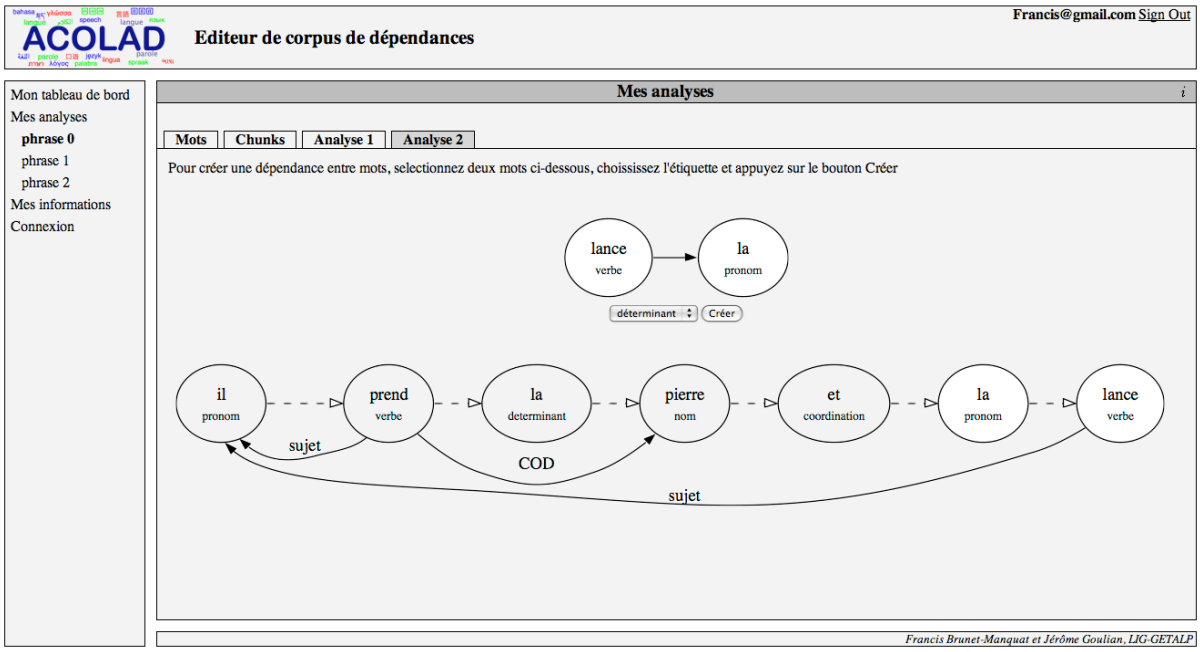

Figure 2 : création d'une analyse de dépendances

Nous présenterons également les services permettant d'importer des segmentations et/ou des annotations préexistantes et la possibilité d'exporter nos résultats sous forme XML.

Équipement

La démonstration ne demandera pas d'équipement particulier. Les participants auront le matériel nécessaire et les logiciels installés pour présenter l'environnement d'édition de corpus de dépendances ACOLAD.

Disponibilité

Notre environnement est proposé à la communauté sous licence publique générale limitée GNU (GNU Lesser General Public License).

Perspectives

ACOLAD propose de nombreux services manuels permettant de segmenter une phrase ou de produire des analyses de dépendances. Cet environnement est actuellement utilisé pour produire des analyses de dépendances pour le français. Mais cet outil, par son formalisme de dépendances et son aspect configurable (choix des jeux d'étiquettes, choix des contraintes structurelles de dépendances -- en particulier pour tenir compte de la variabilité faible ou forte de l'ordre des mots selon la langue [Holan, 2000]), a un potentiel multilingue.

Des expérimentations sont actuellement menées pour intégrer les premiers résultats d'ACOLAD dans nos travaux sur l'analyse syntaxique automatique et la production de documents auto-explicatifs [Blanchon et al., 2006].

Les services proposés par ACOLAD pourront également être intégrés à des environnements tel que Sectra_w, un système collaboratif permettant d'évaluer, de présenter, d'exploiter et de réviser des corpus de traduction automatique [Huynh et al., 2008], par exemple pour proposer l'ajout de dépendances syntaxiques.

Enfin, ACOLAD pourra être proposé dans le cadre de campagne d'évaluation d'analyseurs syntaxiques de dépendances pour aider à fabriquer les analyses références.

Références

BLANCHON, H., BOITET, C. AND CHOUMANE, A. (2006). Traduction automatisée fondée sur le dialogue et documents auto-explicatifs: bilan du projet LIDIA. *in TAL*. vol. 47(3): 30 p.

HOLAN T., KUBON, OLIVA K., PLATEK M. (2000). On complexity of word order, *in TAL*., 41(1), pp. 273-300, Hermès, Paris, France.

HUYNH C.-P., BOITET C. & BLANCHON H. (2008). SECTra_w : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. *Proc. LREC-08*, Marrakech, 27-31/5/08, ELRA/ELDA, ed., 8 p.

VALLI, A. ET VERONIS, J. (1999). Etiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée*, IV(2), 113-133. (dossier : l'oral spontané)