Interface Web pour l'annotation morpho-syntaxique de textes

Thierry Hamon[1,2]
(1) LIMSI, CNRS, Université Paris-Saclay, Bat 508, rue John von Neumann, Campus Universitaire, 91405 Orsay, France
(2) Université Paris 13 - Sorbonne Paris Cité, 99 avenue J.B. Clément, 93430 Villetaneuse, France
hamon@limsi.fr

RÉSUMÉ

Nous présentons une interface Web pour la visualisation et l'annotation de textes avec des étiquettes morphosyntaxiques et des lemmes. Celle-ci est actuellement utilisée pour annoter des textes ukrainiens avec le jeu d'étiquettes Multext-East. Les utilisateurs peuvent rapidement visualiser les annotations associées aux mots d'un texte, modifier les annotations existantes ou en ajouter de nouvelles. Les annotations peuvent chargées et exportées au format XML TEI, mais aussi sous forme tabulée. Des scripts de conversion de format et de chargement dans une base de données sont mis à disposition.

ABSTRACT

Web interface for the morpho-syntactic annotation of texts

We present a Web interface for visualizing and annotating texts with POS tags and lemma. This interface is currently used for the annotation of Ukrainian texts with the Multext-East POS tagset. The users have a fast access to the annotations associated with words from a text. They can also modify existing and add new annotations. The annotations can be loaded or exported in the TEI XML format and tabular separated format. Several scripts for loading and converting the data are also available.

MOTS-CLÉS : Annotation morphosyntaxique, Lemmatisation, Multext-East, Ukrainien.

KEYWORDS : Morpho-syntactic annotation, Lemmatization, Multext-East, Ukrainian.

1 Introduction

La mise au point de méthodes de TAL sur des langues peu outillées nécessite de constituer des corpus annotés manuellement. Ainsi, dans le cadre du développement d'outils pour l'ukrainien, et notamment un étiqueteur morphosyntaxique, nous souhaitons disposer de textes annotés morphosyntaxiquement avec le jeu d'étiquettes Multext-East (Erjavec, 2012). Ainsi, nous avons développé une interface Web pour l'annotation des mots d'un texte avec des étiquettes morphosyntaxiques et des lemmes, mais aussi la correction de ces informations lorsque le texte est pré-annoté.

Notre objectif est d'une part, de faciliter l'utilisation du jeu d'étiquettes Multext-East, et d'autre part, de limiter les actions de l'utilisateur afin de réduire le temps d'annotation. En effet, ce jeu d'étiquette morphosyntaxique est complexe puisqu'il propose 12 catégories grammaticales, mais aussi jusqu'à 10 traits morphologiques pour les adjectifs et 11 valeurs de traits possibles pour décrire les types de pronom. Il s'agit donc de ne présenter que les traits morphologiques et les valeurs pertinentes pour

une catégorie donnée. L'utilisateur doit donc également avoir une vue synthétique des annotations et la visualisation des annotations associées à un mot doit être rapide.

2 L'interface d'annotation

Afin de répondre aux objectifs présentés ci-dessus, nous avons développé une interface d'annotation s'appuyant sur des technologies Web (XHTML, PHP, AJAX). Une base de données est également utilisée pour le stockage des annotations et la description du jeu d'étiquettes. L'adaptation de l'interface à un autre jeu d'étiquettes ne nécessite donc que la modification de la description du jeu d'étiquettes.

Les textes au format tabulé ou XML conforme à la TEI[1] (Wittern *et al.*, 2009) pré-annotés ou non peuvent être chargés dans la base à l'aide de scripts Perl. Les textes annotés peuvent exportés dans les mêmes formats.

La figure 1 présente une capture d'écran de l'interface d'annotation. Le document à annoter est affiché une seule fois dans l'interface et les annotations déjà associées aux mots apparaissent dynamiquement. Une vue synthétique des annotations est proposée lorsque l'utilisateur passe avec la souris sur un mot (par exemple, **використані** à la figure 1). Les annotations peuvent être modifiées ou ajoutées en cliquant sur le mot concerné (par exemple **приймаєте** visible à droite sur la figure 1).

L'interface est téléchargeable à l'adresse suivante : `https://perso.limsi.fr/hamon/Rada/index.php`.

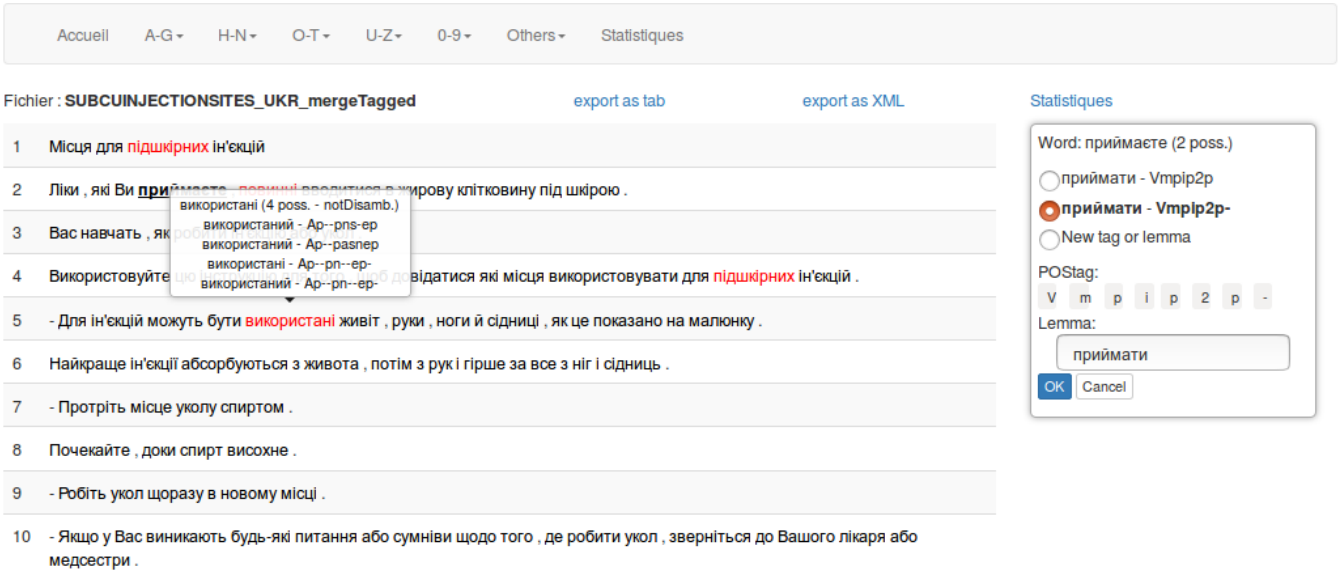

FIGURE 1 : Exemple de visualisation d'un document en cours d'annotation.

Remerciements

Ce travail a été financé par l'action incitative *Outiller l'Ukranien* du LIMSI-CNRS. Nous remercions également Natalia Grabar et Anastasiia Kuznietsova pour leur commentaires sur les premières versions de l'interface.

[1] `http://www.tei-c.org`

Références

ERJAVEC T. (2012). Multext-east : Morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, **46**(1), 131--142.

WITTERN C., CIULA A. & TUOHY C. (2009). The making of tei p5. *Literary and Linguistic Computing*, **24**(3), 281--296.