TALN 2009, Senlis, 24–26 juin 2009

La plate-forme d'annotation Glozz

Antoine Widlöcher Yann Mathet

Laboratoire GREYC, CNRS UMR 6072, Université de Caen
{antoine.widlocher,yann.mathet}@info.unicaen.fr

Mots-clés : Linguistique de corpus, Annotation, Plate-forme logicielle.

Keywords: Corpus Linguistics, Annotation, Software Framework.

1 Une plate-forme d'annotation générique

De plus en plus de travaux en linguistique, en linguistique computationnelle et en Traitement Automatique des Langues manifestent un intérêt croissant pour les études sur corpus. À travers une importante diversité d'approches, la nécessité d'une interaction systématique entre modèles, traitements et corpus rend nécessaire la disponibilité - et donc l'établissement - d'annotations de référence auxquelles les modèles et les traitements pourront être confrontés, pour leur élaboration ou pour leur évaluation. Or la mise en place de telles annotations est un processus complexe qui requiert à la fois un cadre formel permettant la représentation d'objets linguistiques variés, et des applications dédiées à l'annotation manuelle proprement dite, permettant à l'annotateur de localiser sur corpus et de caractériser les occurrences du phénomène observé. Or, si différents outils d'annotation ont conséquemment vu le jour, il convient de remarquer, d'une part qu'ils demeurent souvent fortement liés à un modèle théorique et des objets linguistiques particuliers, et d'autre part qu'ils ne permettent que marginalement d'explorer certaines structures plus récemment appréhendées expérimentalement, notamment à granularité élevée et en matière d'analyse du discours.

La plate-forme Glozz, qui est l'objet de notre démonstration, répond à ces différentes contraintes et propose un environnement d'exploration de corpus et d'annotation fortement configurable et non limité *a priori* au contexte discursif dans lequel elle a initialement vu le jour. Elle repose sur un méta-modèle générique qui permet la représentation et la caractérisation de structures linguistiques variées, observables à différents niveaux de granularité, qui pourront reposer sur des segments, des relations ou des dispositifs plus complexes, et auxquelles des représentations symboliques pourront être associées, sous forme de structures de traits. Pour une campagne d'annotation donnée, un modèle d'annotation pourra être défini, qui spécifiera les objets linguistiques disponibles, et le format des caractérisations dont ils devront faire l'objet. Contraint par ce modèle, l'annotateur pourra produire graphiquement, ou décrire formellement, de nouvelles annotations, en les localisant dans le texte et en y associant les indications attendues. Dans ce but, il pourra s'appuyer sur les différentes informations portées par le corpus, qu'il s'agisse du contenu textuel initial ou d'informations linguistiques (typographiques, morphologiques, syntaxiques, sémantiques...), qui pourront résulter d'un traitement préalable, manuel ou

automatique. Glozz simplifie l'accès à ces informations, par le biais de différents paramètres et outils de navigation. Il offre notamment une vue synthétique du corpus donnant un accès rapide aux zones porteuses d'indices, permet l'emphase graphique des objets linguistiques pertinents et le filtrage des objets insignifiants pour une tâche donnée. De plus, Glozz intègre des outils de recherche avancée pouvant guider l'exploration du corpus. L'utilisateur peut notamment restreindre une recherche plein-texte à des contextes spécifiques correspondant à des environnements linguistiques particuliers, ou naviguer rapidement entre les différentes instances d'un même type d'objet linguistique, en exprimant éventuellement des contraintes sur leurs propriétés antérieurement spécifiées. Toute annotation s'inscrit ainsi dans un processus d'enrichissement incrémental. Elle s'appuie sur des informations préexistantes et produit de nouveaux objets, qui pourront eux-mêmes, à leur tour, devenir « indices », lors de l'annotation d'éléments d'ordre supérieur, l'ensemble des objets identifiés (indices et induits) étant représenté de manière uniforme et exploitable par d'autres outils.

2 Déroulement de la démonstration

La démonstration débute par l'installation et l'ouverture du logiciel, puis l'ouverture d'un corpus annoté. Nous exposons ensuite les 7 points relatifs à la figure suivante, que nous explicitons ci-après.

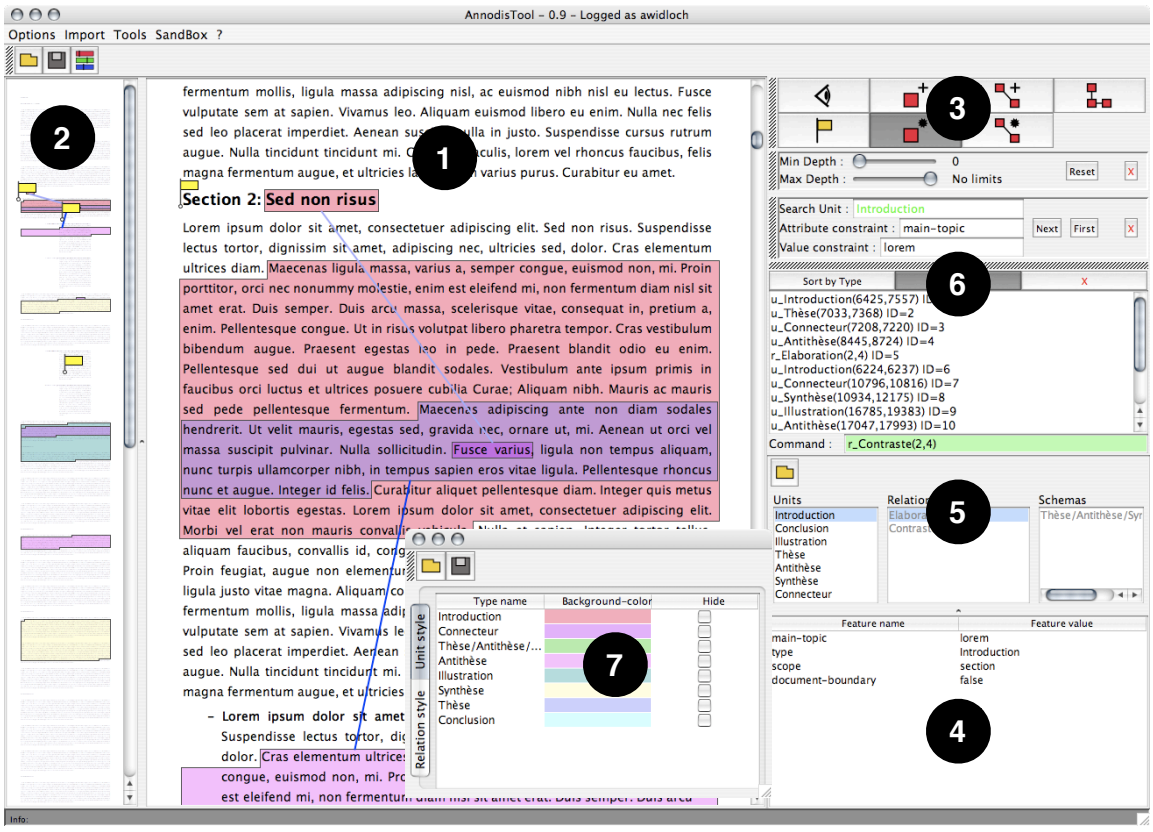

FIG. 1 – Interface principale de Glozz

L'utilisateur dispose de deux « vues » sur le corpus annoté. La vue principale, en (1), permet de lire et d'annoter le texte. Simultanément, la vue synoptique, en (2), propose un niveau de grain permettant d'embrasser l'essentiel du texte d'un coup d'œil. Nous montrons comment exploiter efficacement ces deux vues qui fonctionnent en synergie.

Nous indiquons ensuite la marche à suivre pour créer et modifier des éléments (unités, relations et schémas), via les différents modes dédiés, qui sont accessibles par les boutons présents en (3). Nous détaillons chacun de ces modes sur des exemples concrets, en créant et modifiant quelques unités, puis en en mettant certaines en relation (en évoquant la création de relations sur des unités distantes grâce à la vue synoptique), et enfin en montrant l'intégration de quelques unités et relations au sein d'un schéma. En complément, nous présentons les commentaires en texte libre qu'il est possible de disposer au fil du texte.

Nous montrons alors qu'il est possible d'associer une structure de traits à chaque entité créée. Tout d'abord, les structures par défaut attribuées automatiquement en début de démonstration sont modifiées sur quelques exemples, en (4). Puis nous montrons qu'il est possible de choisir différents modèles d'annotation, et en donnons deux exemples, en (5).

Une fois les principes de l'annotation et de la caractérisation (attribution de structures de traits) acquis, nous présentons les styles qu'il est possible d'associer aux unités et aux relations, via la fenêtre dédiée en (7).

Enfin, nous présentons les modules optionnels complémentaires, qui apparaissent à la demande dans la zone (6) et qui permettent :

– de faire des recherches « plein texte », ou des recherches par types d'unités, en imposant éventuellement des contraintes sur les attributs/valeurs,
– de paramétrer des seuils d'imbrication afin de rendre visibles ou non certaines unités,
– d'accéder à toutes les annotations via une représentation logicoïde permettant non seulement de parcourir rapidement l'ensemble des annotations, mais aussi d'en créer de nouvelles par une saisie textuelle se conformant à une syntaxe simple.

3 Informations complémentaires

Cette démonstration de Glozz, ainsi qu'un guide d'utilisation du logiciel plus général, sont accessibles sur le web au format vidéo, à l'adresse `http://www.glozz.org`.

Par ailleurs, l'article intitulé « La plateforme Glozz : environnement d'annotation et d'exploration de corpus », publié dans les actes de TALN 2009, présente le contexte scientifique dont Glozz est issu, propose un état de l'art, indique les principes fondamentaux auxquels il est adossé et précise les conditions de sa mise en œuvre.