La plate-forme Glozz : environnement d'annotation et d'exploration de corpus

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

Résumé. La nécessité d'une interaction systématique entre modèles, traitements et corpus impose la disponibilité d'annotations de référence auxquelles modèles et traitements pourront être confrontés. Or l'établissement de telles annotations requiert un cadre formel permettant la représentation d'objets linguistiques variés, et des applications permettant à l'annotateur de localiser sur corpus et de caractériser les occurrences des phénomènes observés. Si différents outils d'annotation ont vu le jour, ils demeurent souvent fortement liés à un modèle théorique et à des objets linguistiques particuliers, et 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 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.

Abstract. The need for a systematic confrontation between models and corpora make it necessary to have - and consequently, to produce - reference annotations to which linguistic models could be compared. Creating such annotations requires both a formal framework which copes with various linguistic objects, and specific manual annotation tools, in order to make it possible to locate, identify and feature linguistic phenomena in texts. Though several annotation tools do already exist, they are mostly dedicated to a given theory and to a given set of structures. The Glozz platform, described in this paper, tries to address all of these needs, and provides a highly versatile corpus exploration and annotation framework.

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

Keywords: Corpus Linguistics, Annotation, Software Framework.

1 Introduction

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 diffé-

rents 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 à 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 répond à ces différentes contraintes et propose un environnement d'exploration et d'annotation de corpus fortement configurable et non limité *a priori* au contexte discursif dans lequel elle a initialement vu le jour. Dans la présente introduction, nous préciserons les éléments essentiels devant être pris en compte pour l'établissement d'un environnement adapté à l'exploration et à l'annotation de corpus, dans des contextes théoriques variés, et utilisable notamment en analyse du discours. Nous montrerons alors que les outils disponibles dans la communauté ne satisfont pas pleinement les exigences ainsi mises au jour. Dans une seconde partie, nous mettrons en lumière certains des principes fondamentaux auxquels la plate-forme Glozz est adossée avant d'exposer, dans un troisième temps, les modalités de sa mise en œuvre. Dans une dernière partie, nous présenterons les campagnes d'annotation initiées à l'aide de cet environnement, et indiquerons certaines des orientations de nos développements actuels.

1.1 Problématique

La mise en place de la plate-forme d'annotation que nous présentons ici trouve son origine dans le projet ANR Annodis (Péry-Woodley *et al.*, 2009), qui vise notamment la mise en place d'un corpus de référence en matière d'analyse du discours. Cependant, si la plate-forme répond conséquemment à certaines contraintes propres à ce niveau d'analyse, elle a d'emblée été conçue pour ne pas s'y limiter. De cette double exigence - d'utilisabilité au niveau discours et de généricité - il ressort en particulier la nécessité de prendre en compte les éléments suivants.

Il importe tout d'abord de pouvoir annoter des objets linguistiques appartenant à des catégories structurelles variées, dont l'organisation discursive fait distinctement usage. Ainsi, il sera certes nécessaire de pouvoir délimiter des segments textuels (ensembles de sous-unités consécutives) tels que mots et syntagmes, mais également indispensable de rendre compte de l'organisation relationnelle opérant sur différents plans (relations syntaxiques, rhétoriques, de coréférence...).

D'autre part, nous ne souhaitons pas limiter l'annotation à une échelle particulière, mais permettre au contraire la prise en compte de phénomènes linguistiques opérant à des niveaux variés pouvant aller du local (caractères, mots, syntagmes...) au macroscopique (paragraphes, segments, sections...). Notons à cet égard que l'annotation discursive exige doublement une telle souplesse, d'une part parce que les phénomènes de cet ordre sont observables à différentes échelles, et d'autre part parce qu'ils reposent sur des mécanismes indiciaires impliquant des objets opérant à différents niveaux (connecteurs de discours, expressions indicatives, unités argumentatives...), que l'on souhaitera également annoter.

La contrainte de généricité impose par ailleurs que l'environnement d'annotation utilisé ne repose pas sur un modèle d'annotation limité à certains types d'objets linguistiques ou à un modèle théorique particulier. Il est toutefois nécessaire d'encadrer une campagne d'annotation en définissant les objets identifiables et en contraignant les caractérisations qui devront en être données. Aussi, l'environnement d'annotation devra pouvoir être paramétré par un modèle d'annotation, par rapport auquel il demeurera fondamentalement abstrait, grâce à l'utilisation d'un méta-modèle articulant des catégories structurelles suffisamment génériques.

De plus, il conviendra de maximiser l'expressivité dudit méta-modèle, en intégrant notamment la possibilité de faire cohabiter des « couches » d'annotation mutliples pouvant provenir d'origines diverses et porter sur des zones identiques, en autorisant chevauchement et imbrication entre les annotations, et en prévoyant des configurations « topologiques » assez peu classiques, mais fréquentes au niveau discours, qui permettront par exemple de rendre compte de l'incertitude sur la position des bornes d'un segment.

Aux objets identifiés, on souhaitera associer des caractérisations permettant de représenter formellement leur statut linguistique ou l'information pertinente (catégorie morphologique, valeur sémantique...) dont ils sont porteurs. On distinguera ainsi le *marquage* proprement dit (localisation d'un objet dans le texte), et cette *caractérisation*, l'ensemble constituant l'*annotation*.

Annoter un corpus, c'est aussi assurément prendre en considération l'information dont il est déjà porteur. Il est certes indispensable d'accéder à son contenu textuel brut, mais aussi à des indications additionnelles pouvant résulter de sa structuration typo-dispositionnelle ou d'un traitement préalable, manuel ou automatique, ayant permis d'identifier des éléments qui pourront jouer le rôle d'« indices » dans la phase d'annotation.

Enfin, le processus d'annotation requiert évidemment la possibilité d'explorer le corpus, en quête des indications pouvant révéler l'occurrence d'un phénomène étudié. Cela suppose en particulier la possibilité de mettre en évidence l'information indiciaire pertinente et de masquer celle qui demeure insignifiante, pour une tâche donnée. Cela rend également nécessaire la disponibilité de vues variées sur le texte annoté, qui révéleront les configurations indiciaires (visualisation *in situ*, concordanciers, représentation arborescente, indications statistiques...). D'autre part, la navigation dans le corpus devra s'appuyer sur des outils de recherche permettant d'accéder rapidement à des zones textuelles porteuses des indications attendues.

1.2 État de l'art

Différentes études et manifestations scientifiques révèlent l'importance de la question de l'annotation pour les communautés scientifiques auxquelles nous appartenons. Évoquons par exemple la tenue de l'atelier XBRAC (Witt *et al.*, 2004), et différentes tentatives de synthèse telles que (Dipper *et al.*, 2004) auxquelles il a donné lieu. Plus directement liés aux problématiques propres à l'analyse du discours, les travaux présentés à l'occasion de (Webber & Bryon, 2004) mettent également la question de l'annotation au premier plan.

Faisant face à la nécessité d'établir des corpus de référence, divers travaux ont d'ores et déjà permis la mise en place de méthodes et d'environnements logiciels permettant l'exploration et l'annotation manuelle des corpus. La plate-forme GATE (Cunningham *et al.*, 2002), bien connue de la communauté en tant qu'environnement de traitement automatique, intègre ainsi un module d'annotation manuelle. Au rang de ses avantages, on évoquera sa forte intégration dans l'environnement d'expérimentation, ainsi que la possibilité de guider l'annotation en définissant des schémas d'annotation. Au rang des limites, on notera que l'outil est prioritairement destiné à la délimitation d'unités textuelles, l'annotation de relations ne pouvant faire l'objet d'une saisie « graphique ». De plus, l'ergonomie de l'outil et les modalités de visualisation manifestent un clair privilège à l'annotation d'objets relativement locaux, à l'échelle du mot ou du syntagme.

Ouvertement dédié à l'annotation manuelle, Wordfreak (Morton & LaCivita, 2003) accorde également un net privilège à l'annotation d'objets locaux et révèle également une faible intégration de la notion de relation, dont la représentation se limite à des jeux d'identifiants. On appréciera

toutefois la variabilité des vues possibles sur le texte (représentation au fil du texte, concordancier, représentation arborescente...). L'absence d'évolution récente de cet outil prometteur pose cependant avec force la question de sa pérennité.

Notons également l'existence d'un *plugin* pour l'environnement Protégé[1] dédié à l'annotation manuelle (Ogren, 2006). Bénéficiant de la puissance de Protégé pour l'édition et la manipulation d'ontologies, Knowtator permet l'élaboration de schémas d'annotation complexes adossés à de telles ontologies. Explicitement dédié au traitement de l'information biomédicale, cet outil se limite, lui aussi, à l'annotation d'unités locales, et s'avère donc peu utile pour l'annotation d'objets macroscopiques et/ou de relations.

Intégrant clairement la notion de relation, MMAX (Müller & Strube, 2001) propose un environnement graphique et un modèle d'annotation génériques et configurables permettant, en théorie, une annotation multi-échelle. Il permet la délimitation de segments, la spécification de relations, et l'association d'attributs aux objets marqués, le tout pouvant être réalisé graphiquement, sur la base d'un schéma d'annotation défini par l'utilisateur. On notera toutefois que les modalités d'annotation et de représentation rendent difficile l'annotation à des échelles macroscopiques. Certes, des relations peuvent être tissées entre des objets assez fortement distants - c'est notamment le cas des chaînes coréférentielles, qui ont été particulièrement explorées à l'aide de MMAX -, mais ces objets demeurent en eux-mêmes préférentiellement locaux et peu emboîtés.

Fortement lié, lui aussi, à la problématique spécifique de l'annotation des chaînes de coréférence, PALinkA (Orăsan, 2003) reprend et généralise certains principes déjà utilisés dans CLinkA (Orăsan, 2000). Alors que ce dernier était irréductiblement lié à la question de la coréférence, PALinkA autorise la définition, par l'utilisateur, d'un schéma d'annotation. Il permet de plus la prise en compte d'annotations préexistantes, de différentes natures. S'il a effectivement fait l'objet d'utilisations sur des phénomènes linguistiques variés, reste que les relations ne donnent pas lieu à une visualisation graphique adéquate, et que les segments reliés demeurent inévitablement locaux, faute, là encore, de représentations idoines.

UAM Corpus Tool[2], héritier de Systemic Coder et initialement dédié à la linguistique systémique, offre un environnement d'annotation aisément configurable. Il permet à l'utilisateur de définir des schémas d'annotation hiérarchiques complexes qui orienteront les phases ultérieures d'annotation. On appreciera la disponibilité d'outils statistiques fort utiles pour l'exploration de corpus. Sa plus grande faiblesse, du point de vue qui nous occupe ici, est l'impossibilité d'annoter des relations. D'autre part, les modalités d'annotation et de visualisation des segments rendent certes possible la délimitation d'unités macroscopiques emboîtées ou se chevauchant, mais en exposant l'utilisateur à d'importantes difficultés de lecture et d'interprétation.

Intégralement dédié au modèle théorique de la *Rhetorical Structure Theory*, RSTTool (O'Donnell, 2000) permet bien entendu l'annotation des relations, qui sont au cœur de cette théorie. Il permet en outre la segmentation du texte en unités qui consitueront les termes de ces relations. Il s'appuie toutefois exclusivement sur un modèle d'annotation et une représentation graphique arborescente, conformes à la théorie et au formalisme RST, qui le rendent pour ainsi dire inutilisable dans une autre perspective théorique.

De la confrontation entre ces différents travaux et outils et les exigences présentées auparavant, il ressort distinctement que ces dernières ne sont pas pleinement satisfaites, et que la mise en place d'un nouvel environnement est justifiée et souhaitable.

[1] `http://protege.stanford.edu`.
[2] `http://www.wagsoft.com/CorpusTool/`.

2 Principes fondamentaux

La plate-forme Glozz repose essentiellement sur les principes suivants, qui visent à la fois à garantir sa généricité et à permettre l'annotation d'objets linguistiques, en particulier discursifs, difficiles à manipuler avec les outils existants.

À la nécessité de prendre en charge des configurations structurellement variées, la plate-forme répond en s'appuyant sur un méta-modèle générique de l'organisation du discours, issu de (Widlöcher, 2008). Ce méta-modèle repose sur l'articulation entre les notions d'*unité*, de *relation* et de *schéma*, dont l'hyperonyme commun est le concept d'*élément*. Par *unité*, nous désignons, par simplification, une séquence d'éléments textuels adjacents, sans présupposé d'échelle. Mots, syntagmes, propositions, paragraphes, unités thématiques ou document répondent à cette définition. Par *relation*, nous désignons, par simplification, un rapport binaire entre deux unités, sans présupposé d'ordre ou de distance. Relations syntaxiques, rhétoriques ou de coréférence relèvent de cette notion. À ces catégories classiques, nous ajoutons la notion de *schéma*, qui désigne une configuration textuelle complexe récurrente impliquant unités et relations, dont la structure énumérative, composée d'une amorce, d'items et de relations hyperonymiques entre amorce et items est un exemple. Ce méta-modèle autorise également l'association, à chaque instance de l'une ou l'autre de ces catégories, d'une caractérisation exprimée par un type et une représentation symbolique, sous forme d'une structure de traits.

Glozz repose sur ce méta-modèle et ne fait aucune hypothèse *a priori* sur la nature linguistique des objets qui seront effectivement annotés, qui devront simplement pouvoir être rapportés à l'un ou l'autre de ces types d'éléments. La spécification du modèle linguistique particulier pour lequel on souhaite produire une annotation repose sur la définition d'un *modèle d'annotation*, pouvant paramétrer l'application et guider l'annotateur, mais par rapport auquel elle demeure fondamentalement abstraite. Pouvant être décrit de manière déclarative (*cf. infra*), un tel modèle d'annotation définit notamment, pour chaque catégorie structurelle (unité, relation et schéma), les types qui pourront faire l'objet d'une annotation, en précisant, pour chacun, le format des représentations symboliques qui en seront données.

Il est important de préciser que, du point de vue de l'environnement d'annotation, rien n'existe hors de ce méta-modèle abstrait et de ses réalisations particulières. Du point de vue de Glozz, « tout est annotation » ; toute information utilisable doit pouvoir être exprimée dans les termes du méta-modèle. Cela vaut, évidemment, pour tous les objets qui seront annotés par son truchement. Mais c'est également vrai de toutes les informations pouvant résulter d'un traitement préalable, manuel ou automatique, ainsi que des indications typo-dispositionnelles (structuration en sections, paragraphes...) qui ne font pas davantage l'objet d'un traitement *ad hoc*. Il en résulte en particulier que les différents objets pourront être manipulés et observés de manière uniforme, quelle qu'en soit la nature.

Nous insistons d'autre part sur le fait qu'une information adjointe au corpus n'y est pas rendue, *ipso facto*, fatalement observable. Elle doit être considérée comme principalement « disponible », sa visibilité n'étant rendue effective qu'à la faveur de la spécification d'un *point de vue* sur le texte, spécification indissolublement liée à une tâche d'annotation particulière, qui, seule, rend les informations disponibles signifiantes et leur observabilité conséquemment opportune.

Enfin, Glozz repose sur une annotation déportée (*stand-off*). La localisation des unités renvoie à la position de leurs bornes dans le texte et celle des relations et des schémas repose sur la référence aux identifiants des constituants impliqués.

3 Réalisation

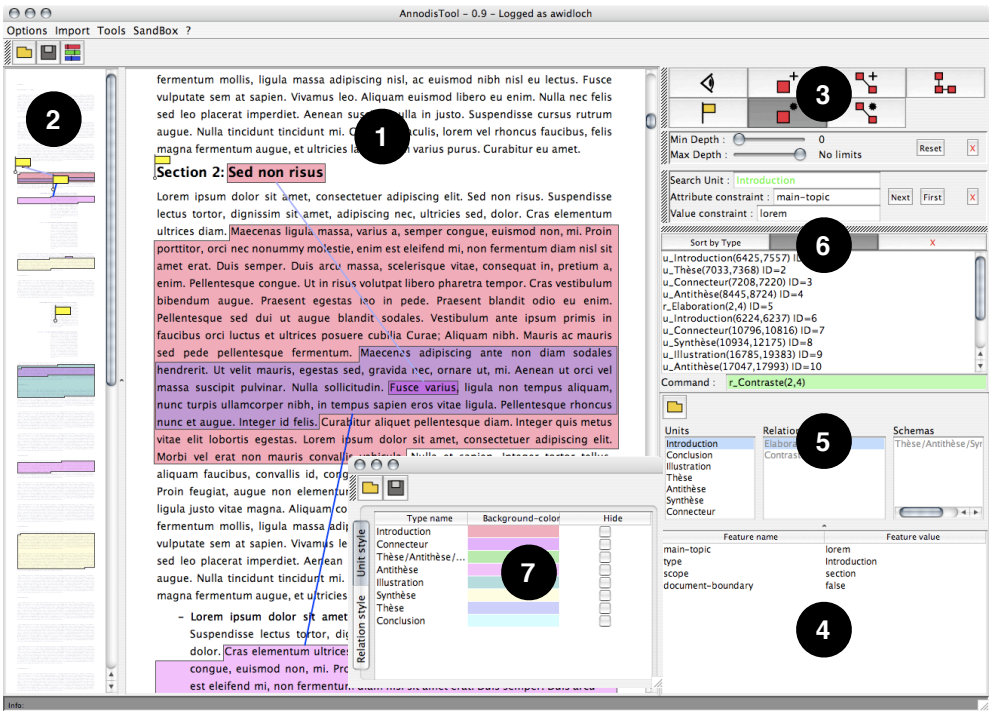

FIG. 1 – Interface principale de Glozz

3.1 Généricité, méta-modèle et modèle d'annotation

La plate-forme Glozz peut être adaptée à un modèle d'annotation quelconque, pour peu que ce dernier puisse se conformer au méta-modèle dont nous avons fait état dans la partie précédente : pour une campagne d'annotation donnée, la première tâche consiste à définir les unités, les relations et les éventuels schémas que l'on souhaite manipuler. La définition de ce modèle se fait de manière déclarative, conformément à un schéma XML relativement simple, comprenant notamment les éléments `units`, `relations` et `schemas`, qui permettent de définir les différentes catégories d'objets disponibles, ainsi que le format de la structure de traits attendue pour chacune d'entre elles. À titre d'exemple, la définition d'un modèle d'annotation utilisable en matière argumentative pourrait ressembler à ceci :

```
<annotationModel>
        <units>
                <type name="Introduction">
                        <featureSet>
                                <feature name="scope"><value type="free" default="section" /></feature>
                                <feature name="document-boundary"><value type="boolean" default="false" /></feature>
                                <feature name="main-topic"><value type="free" default="" /></feature>
                        </featureSet>
                </type>
                <type name="Conclusion">[...]</type>
                 [...]
        </units>
         [...]
<annotationModel>
```

Deux catégories d'unités, *Introduction* et *Conclusion*, sont définies. Les instances de la première catégorie possèderont les traits *scope*, *document-boundary* et *main-topic*. Certains de ces traits se voient attribuer une valeur par défaut telle que *section* pour le trait *scope* ou *false* pour le trait *document-boundary*. Parmi les différents types de valeurs possibles, qui permettront d'adapter les masques de saisie offerts à l'annotateur, citons `free` pour le texte libre, `boolean` pour les valeurs booléennes, `enum` pour les ensembles finis de valeurs possibles, ou encore `int` pour les valeurs entières. La suite du document décrit les types de catégories et les structures de traits attendues pour les relations et les schémas. L'écriture de ce descripteur XML se fait actuellement de façon manuelle, mais une interface graphique d'édition est en cours de réalisation.

Une fois chargé dans la plate-forme, le modèle d'annotation apparaît en zone (5) (*cf.* figure 1), sous forme de listes de catégories. L'annotateur dispose alors d'une *palette* de types, et peut annoter le texte selon le paradigme imposé par ce modèle. La zone (4) fait apparaître la caractérisation de l'objet annoté sélectionné, et en permet l'édition.

3.2 Accès multi-échelle au contenu textuel

Le processus d'annotation nous conduit à considérer le texte à différentes échelles, et ce pour au moins deux raisons. La première concerne la *navigation* : il est essentiel de disposer d'une vision synthétique (à grande échelle) du texte, permettant un accès rapide aux zones porteuses d'indices, dont le détail doit pouvoir être exploré à une échelle plus restreinte. La seconde raison, plus fondamentale, renvoie à la fois à la granularité des phénomènes observés et au processus même d'annotation, qui requiert souvent un point de vue et un grain de visualisation particuliers, pour qu'émergent les éléments signifiants. Par exemple, si une relation de contraste peut apparaître à un grain assez fin, entre deux propositions, elle peut aussi se manifester à un niveau macroscopique, entre le premier et le dernier paragraphe du texte, qui devront pouvoir être appréhendés d'un même regard. Glozz propose donc une vue double et simultanée sur le texte. La vue principale, en zone (1), permet la lecture du texte et la saisie d'annotations. Une seconde vue, nommée « ruban » et située en zone (2), propose une vue globale du corpus, simplifiant la navigation entre ses différentes zones, et au sein de laquelle les indices annotés disponibles pourront être rapidement identifiés, grâce aux paramètres de représentation visuelle fixés par la feuille de style (*cf. infra*). Les vues locale et macroscopique peuvent être utilisées conjointement. Si l'on souhaite par exemple établir une relation entre deux unités distantes, on pourra repérer les deux unités en zone (2), et les sélectionner précisément en zone (1).

3.3 Annotation visuelle

Glozz propose différents outils permettant de voir, mais aussi de créer directement sur le texte, à la souris, les différents types d'annotations. Les unités apparaissent sous forme de blocs colorés, conformément à la feuille de style choisie (*cf. infra*), encadrant la zone textuelle délimitée. L'une des forces de la plate-forme est d'autoriser chevauchement et imbrication entre unités et d'en proposer un rendu visuel : lorsqu'une unité est comprise dans une autre, le cadre qui la représente se trouve graphiquement emboîté dans le cadre matérialisant l'unité englobante. De ce fait, même en présence d'un fort taux d'imbrication, les unités restent toujours visibles et accessibles. Les relations apparaissent pour leur part sous forme de lignes tracées entre les unités reliées, dont la couleur est donnée, ici encore, par la feuille de style utilisée. Enfin, les schémas sont représentés par un sur-encadrement des unités et relations qui les composent.

Une palette présente en zone (3) permet de choisir un mode d'action parmi création ou édition d'unités, de relations, ou de schémas. Une fois le mode sélectionné, le travail s'effectue principalement en zone (1), les zones (5) et (4) offrant pour leur part un aperçu actualisé et éditable des types de catégories et des structures de traits. Un mode « *glue note* » permet l'adjonction de commentaires en texte libre, positionnés dans le texte et ancrés visuellement à l'aide d'un drapeau.

3.4 Représentation formelle des annotations

Parallèlement à cette édition graphique, une représentation de tous les objets annotés par des expressions formelles d'inspiration logique, proche des notations utilisées par exemple en RST ou en SDRT, est donnée en zone (6). Dans cette notation, un élément est ancré dans le texte, soit par la position de ses bornes (dans le cas des unités), soit par les identifiants de ses termes ou constituants (dans le cas des relations et des schémas). Ainsi, une relation d'*élaboration* entre les unités d'identifiants respectifs 2 et 4 sera notée par exemple `r_Elaboration(2,4)`. Ce second paradigme s'avère plus adapté et plus rapide pour certaines tâches d'annotation (création de schémas, annotation à un niveau très local...). Par ailleurs, même quand l'édition graphique est préférée, il est souvent utile de disposer de la liste des objets créés, triés par type ou par date de création, liste permettant du reste de naviguer, dans le texte, entre ces objets. Une ligne de commande permet la saisie directe de ces expressions, qui se traduit par la création de l'objet textuel correspondant. Cette saisie est contrainte par une complétion automatique supervisée, qui interdit l'expression d'une formule inconsistante. La plate-forme permet l'utilisation simultanée des deux paradigmes de création et de visualisation, graphique et formel.

3.5 Outils de recherche avancée

Deux moteurs de recherche sont intégrés à la plate-forme, en zone (6), l'un portant sur le contenu textuel, l'autre sur les annotations. Le premier fonctionne en mode *plein texte*, les différentes occurrences d'une suite de caractères étant recherchées. Il est de plus possible de restreindre la recherche à des contextes particuliers, en spécifiant le type des unités délimitant ces contextes, unités choisies parmi les annotations disponibles. Par exemple, on pourra rechercher la séquence « amertume » dans des passages annotés en tant qu'*Introductions*. L'outil de recherche d'unités s'appuie pour sa part exclusivement sur les annotations disponibles et permet de naviguer entre les objets d'un certain type. La portée de la recherche peut être restreinte en fixant des valeurs auxquelles les structures de traits des unités recherchées devront se conformer. On pourra ainsi, par exemple, naviguer entre les unités de type *Introduction* dont l'attribut *main-topic* possède la valeur *amertume*.

3.6 Définition d'un point de vue sur le texte

Le processus d'annotation étant souvent incrémental, au sens où une certaine étape d'annotation s'appuie non seulement sur le texte, mais aussi sur les annotations déjà produites, il est intéressant de pouvoir paramétrer la représentation de ces dernières, pour mettre en lumière les informations pertinentes pour une tâche d'annotation donnée et occulter les autres. Ce réglage de la vue doit du reste pouvoir être modifié à tout moment, les différentes phases d'annotation

pouvant nécessiter de faire varier le point de vue sur le matériau textuel. Glozz propose deux outils complémentaires de configuration du rendu visuel. L'éditeur de style, en zone (7), permet de choisir la couleur de chaque type d'unité et de relation, et permet par ailleurs de déterminer, pour chacun d'entre eux, si ses occurrences sont visibles ou non. D'autre part, le sélecteur de profondeur, disponible en zone (6), permet de régler les niveaux d'imbrication minimal et maximal des unités que l'on souhaite visualiser. Par exemple, en réglant ces seuils respectivement sur 2 et 3, on ne voit que les unités qui sont doublement ou triplement imbriquées. Ce réglage s'avère particulièrement utile dans les zones de texte fortement emboîtées, afin de ne plus révéler que certaines unités localement pertinentes, pour un grain d'imbrication donné.

4 Exploitation et perspectives

La conception de Glozz a débuté en janvier 2008, et son développement à l'été 2008. La plate-forme est arrivée à un niveau de maturité et de stabilité suffisants pour être d'ores et déjà exploitée. Comme il a été dit, son élaboration a été initiée dans le cadre du projet ANR Annodis, pour faire face à l'absence d'outils répondant à son cahier des charges[3]. Ce projet implique linguistes et informaticiens, et a pour particularité de procéder à une annotation discursive à des niveaux de granularité très variés. C'est donc une mise à l'épreuve tant théorique que pratique qui a commencé dès le mois de décembre 2008, et qui implique désormais une vingtaine de chercheurs dans une phase d'annotation manuelle débutée en mars 2009. Insistons sur le fait que l'outil est exploité non seulement par des utilisateurs aux profils différents (linguistes et informaticiens), mais aussi selon deux approches radicalement différentes, l'une, *ascendante*, partant d'unités de grain fin, et l'autre, *descendante*, partant de structures discursives de haut niveau. Glozz semble répondre aux attentes des deux approches, au sein d'une même plate-forme, et permettra donc à ces deux dernières de converger dans un corpus conjointement annoté.

Étant conçu autour d'un méta-modèle à vocation générique, Glozz devrait cependant trouver un usage dans des tâches d'annotations très diverses, à différents niveaux de granularité et au-delà du projet Annodis. Citons par exemple l'annotation de structures temporelles itératives, récemment initiée dans le cadre d'un projet caennais, ou encore l'annotation d'expressions évaluatives, envisagée dans le cadre d'un projet consacré à la veille d'opinon.

Cependant, nombre d'améliorations et d'évolutions sont prévues, qui devraient nécessiter encore six à douze mois de développement dans un premier temps. En premier lieu, le modèle de relation actuellement implémenté est une simplification de notre modèle théorique ne permettant de mettre en relation que deux unités. Les développements actuels visent à en faire une implémentation complète qui décuplera le pouvoir expressif de la plate-forme, en permettant de mettre récursivement en relation, non plus seulement les unités, mais tous les types d'éléments (unités, relations et schémas). Si cela répond à un besoin exprimé par les linguistes travaillant dans le cadre de la SDRT, ce gain en expressivité sera également important pour d'autres approches théoriques. En second lieu, les deux paradigmes d'appréhension des annotations actuellement présents dans Glozz sont le texte graphiquement annoté d'une part et la représentation formelle des annotations d'autre part. Un troisième paradigme, reposant sur une représentation sous forme de graphe, viendra bientôt enrichir la plate-forme, pour permettre une édition des annotations conforme aux notations habituellement utilisées dans certaines théories

[3]Pour une présentation précise des objectifs de ce projet et de l'utilisation de Glozz dans son contexte particulier, on pourra se reporter à (Péry-Woodley *et al.*, 2009).

(RST, SDRT...). Certes, d'autres outils proposent déjà une telle représentation, mais sa déclinaison dans Glozz bénéficiera d'un triple avantage. Tout d'abord il sera possible d'exploiter conjointement ces trois paradigmes complémentaires de représentation, en fonction des phénomènes observés. De plus, cette représentation sera toujours basée sur une sélection faite par l'annotateur et non sur la totalité du texte, ce qui permettra d'obtenir des graphes aisément interprétables. Enfin, le travail d'annotation effectué dans l'un des paradigmes sera toujours répercuté dans les autres.

Références

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. & TABLAN V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, p. 168–175, Philadelphia, USA.

DIPPER S., GÖTZE M. & SETDE M. (2004). Simple annotation tools for complex annotation tasks : an evaluation. In (Witt *et al.*, 2004).

MORTON T. & LACIVITA J. (2003). WordFreak : An Open Tool for Linguistic Annotation. In *Proceedings of Human Language Technology (HLT) and North American Chapter of the Association for Computational Linguistics (NAACL)*, p. 17–18, Edmonton, Canada.

MÜLLER C. & STRUBE M. (2001). MMAX : A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, WA, Etats-Unis.

O'DONNELL M. (2000). RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, p. 253 – 256, Mitzpe Ramon, Israel.

OGREN P. V. (2006). Knowtator : A Protégé plug-in for annotated corpus construction. In *Proceedings of Human Language Technology (HLT) and North American Chapter of the Association for Computational Linguistics (NAACL)*, New-York, États-Unis.

ORĂSAN C. (2000). CLinkA a coreferential links annotator. In *Proceedings of LREC'2000*, p. 491–496, Athens, Greece.

ORĂSAN C. (2003). PALinkA : a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, p. 39–43, Sapporo, Japan.

PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09), session poster*, Senlis, France.

B. WEBBER & D. BRYON, Eds. (2004). *Proc. of the ACL 2004 Workshop on Discourse Annotation.*, Barcelone, Espagne.

WIDLÖCHER A. (2008). *Analyse macro-sémantique des structures rhétoriques du discours - Cadre théorique et modèle opératoire*. PhD thesis, Université de Caen Basse-Normandie.

A. WITT, U. HEID, H. S. THOMPSON, J. CARLETTA & P. WITTENBURG, Eds. (2004). *Workshop on XML-based richly annotated corpora (XBRAC)*, Lisbonne, Portugal. Conférence LREC 2004.