Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples

Nicolas Mazziotta, Ph.D.

Université de Stuttgart

Résumé

Nous présentons le logiciel libre NotaBene. Il s'agit d'un logiciel destiné principalement à l'annotation linguistique de ressources textuelles écrites en ancien français, mais dont les principes se veulent suffisamment génériques pour servir à l'exploitation d'autres types de ressources et éventuellement dans d'autres disciplines. NotaBene est basé sur des standards (XML, RDF et OWL) et permet l'affichage simultané d'analyses de nature différentes (p. ex. : lemmatisations vs arbres syntaxiques ou de progression thématique) ou simplement concurrentes (effectuées par deux chercheurs différents). Le logiciel permet également à chaque utilisateur de définir son propre modèle d'annotation et son jeu d'étiquettes personnalisé. Le modèle est enregistré de manière formalisée pour permettre les échanges et les traitements automatiques.

Mots clés

LINGUISTIQUE, VISUALISATION MULTIPLE, ONTOLOGIES, RDF, LOGICIEL LIBRE

Introduction

Nous voudrions présenter ici le logiciel libre NotaBene (disponible à l'adresse https://sourceforge.net/projects/notabene/), dont nous avons commencé la conception et le développement d'un prototype en Python entre janvier et juin 2008 dans le cadre d'un post-doctorat au laboratoire ICAR (Université Lyon 2, UMR 5191, http://icar.univ-lyon2.fr/). Il s'agit d'un logiciel destiné principalement à l'annotation linguistique de ressources textuelles écrites en ancien français, mais dont les principes se veulent suffisamment génériques pour servir à l'exploitation d'autres types de ressources et éventuellement dans d'autres disciplines. Le logiciel est actuellement développé dans le cadre du projet nommé *Syntactical Reference Corpus of Medieval French* – projet subventionné par l'Agence nationale de la recherche (ANR, France), sous la responsabilité de Sophie Prévost (LaTTiCe, UMR 8094, Paris), et la Deutsche

RECHERCHES QUALITATIVES – Hors Série – numéro 9 – pp. 83-94.
LOGICIELS POUR L'ANALYSE QUALITATIVE: INNOVATIONS TECHNIQUES ET SOCIALES
ISSN 1715-8702 - http://www.recherche-qualitative.qc.ca/Revue.html

Forschungsgemeinschaft (DFG, Allemagne), sous la responsabilité d'Achim Stein (Université de Stuttgart), de janvier 2009 à janvier 2012.

NotaBene est basé sur des standards (XML, RDF et OWL, voir la section 3 ci-dessous) et implémente un modèle de visualisation multiple des ressources.

Abordons l'interface du logiciel au travers d'un exemple très simple pour commencer (nous montrerons ci-dessous d'autres emplois correspondant à une utilisation réelle). Nous voyons, dans la Figure 1, l'analyse de la phrase forgée *Demain, Jean ira à Paris* (les principales zones de l'interface sont délimitées par des rectangles noirs).

Comme le montre la Figure 1, l'environnement de travail est subdivisé en deux parties. Le panneau de gauche contient de multiples annotations d'un même document, visualisées simultanément sous des représentations diverses (texte courant, concordance, arbre). Ces annotations multiples feront l'objet de la section 2. Le panneau de droite contient un éditeur de terminologie, utilisé pour sélectionner ou modifier les jeux d'étiquettes mobilisés lors de l'annotation. Les questions de conceptualisation et de construction des jeux d'étiquettes seront abordées dans la section 3. Au préalable, nous dirons quelques mots concernant la nécessité de développer un tel logiciel.

Section 1 : Pratiques d'annotations et besoin d'un outil

Nous présenterons tout d'abord, sans nous étendre outre mesure, l'intérêt des annotations linguistiques, ainsi que la nécessité de développer un nouvel outil permettant l'annotation manuelle.

Corpus annotés en linguistique

À la différence des corpus annotés en sciences sociales, les corpus annotés dans le cadre de recherches linguistiques sont souvent enrichis de manière relativement exhaustive, car les questions posées nécessitent une vue d'ensemble; l'étude quantitative des propriétés morphosyntaxiques, description des structures syntaxiques ou des courbes intonatives n'ont de pertinence que dans la mesure où toutes les unités sont traitées.

L'exploitation des *métadonnées* associées par l'analyste au corpus nu nécessite que ce surplus d'information soit enregistré d'une façon ou d'une autre; on parle d'*annotation*. L'annotation ajoute ainsi une couche d'abstraction aux données, regroupant les différentes occurrences de mots ou de structures dans des ensembles plus ou moins homogènes. L'objectif de pareilles annotations est de permettre l'extraction de données en fonction d'une question spécifique. Ainsi, pour ne prendre qu'un exemple, un étiquetage complet en « parties du discours » permet de sélectionner, par exemple, toutes les phrases

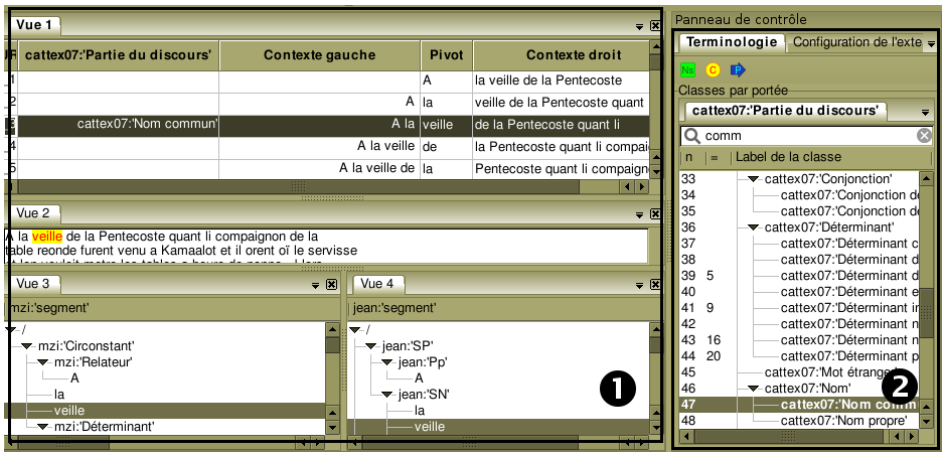

Figure 1. Environnement de travail

qui comportent un nom directement suivi d'un adverbe, ou un verbe directement précédé d'un pronom, etc. Ce type de sélection peut servir à porter au regard du chercheur des faits singuliers, qui méritent un commentaire ou une étude approfondie. D'autre part, une fois couplées à des critères extralinguistiques, comme la date de composition, le nom de l'auteur ou le type de texte, les annotations linguistiques permettent des études lexicométriques (relevant des statistiques textuelles) utiles pour étudier, par exemple, l'attribution des œuvres ou le style d'un auteur.

Besoin d'un outil

Les principales entreprises de construction de corpus de français médiéval, rassemblées en un groupe nommé *Consortium* pour les corpus de Français médiéval (http://ccfm.ens-lsh.fr/), ont choisi de partir des recommandations de la *Text Encoding Initiative* (Sperberg-McQueen, Burnard, Bauman, DeRose & Rahtz, 2007), c'est-à-dire, entre autres, d'encoder les documents en XML.

Les membres de cette communauté font relativement peu usage de logiciels pour mener à bien leurs études, qu'elles soient qualitatives ou quantitatives. Les relevés sont généralement faits manuellement et les annotations sont, le cas échéant, encodées directement dans le fichier sans interface spécifique, si bien que les corpus d'ancien français sont généralement peu étiquetés. La situation n'a que peu évolué par rapport à celle que décrit Habert, Nazarenko et Salem (1997, pp. 125-134).

Certains logiciels sont parfois employés pour aider à l'annotation des documents; par exemple, le *Système d'annotation de texte par ordinateur* SATO (http://www.ling.uqam.ca/sato/satoman-fr.html), qui permet l'annotation au niveau du mot. Malheureusement, le modèle de données de SATO ainsi que son implémentation sous la forme de service web rendent parfois ce travail extrêmement laborieux, en particulier quand il est question d'étiqueter *tous* les mots d'un texte. Il n'est par ailleurs pas adapté à l'annotation des relations syntaxiques. D'autres logiciels pourraient être utilisés dans ce cadre – par exemple, *WordFreak* (http://wordfreak.sourceforge.net) ou encore le *Linguistic Tree Constructor* (http://ltc.sourceforge.net/index.html) –, mais cela impliquerait de nombreuses difficultés d'interfaçage entre les différents formats de fichier annotés.

La situation peut se résumer comme suit :

- il n'existe aucun environnement d'*annotation manuelle* intégrant les fonctionnalités généralement requises;
- les formats employés pour définir les jeux d'étiquettes et la manière dont les documents sont annotés sont à chaque fois idiosyncratiques.

Partant, nous avons décidé de mettre au point NotaBene, qui se définit, déjà au stade du prototype que nous présentons ici, comme un environnement d'annotation manuelle de documents textuels XML, fondé sur les standards RDF et OWL. Enfin, bien que nous ne traiterons pas cet aspect, NotaBene a été conçu pour être extensible.

Section 2 : Annotations multiples

Une des premières conséquences du champ d'étude dans lequel NotaBene a été conçu est la multiplicité des manières d'analyser un même matériau. Si les modèles linguistiques foisonnent, les sous-branches interconnectées sont également pléthoriques. Ces dernières, bien que relevant de domaines distincts, sont reliées dans une dialectique complexe.

Ainsi, la syntaxe de la phrase, la morphologie verbale et la manière dont les verbes déterminent la forme de leurs compléments doivent pouvoir être mises en relation. Par exemple, dans la phrase présentée ci-dessus, l'emploi du circonstant *demain* est en relation avec le futur du verbe; par ailleurs, la forme et la présence du complément *Paris* est déterminée par le choix du verbe *aller*. Dans le cadre d'une analyse fine, les différents niveaux s'interconnectent « naturellement » pour le linguiste.

Principe général

Ces informations sont souvent encodées séparément avant d'être exploitées. L'annotateur n'a généralement pas besoin de connaître la fonction syntaxique

de *demain* ou d'*ira* pour indiquer qu'il s'agit respectivement d'un adverbe et d'un verbe; avoir accès au reste de la phrase n'a pas forcément d'intérêt. Néanmoins, le recours à d'autres niveaux d'analyse est sporadiquement requis, notamment pour lever les ambiguïtés syntaxiques. Une vue du texte sous forme de liste ordonnée de mots permet d'en faire une lemmatisation ou une catégorisation grammaticale rapide, mais il faut que la structure syntaxique soit également visualisable dans certains cas, sachant que ce dernier type de vue est trop complexe pour permettre une annotation rapide des lemmes.

On peut voir un exemple d'affichage simultané d'analyses différentes dans NotaBene dans la Figure 2.

À gauche se trouve une analyse syntaxique arborescente en cours d'élaboration. À droite se trouve la même phrase, affichée sous la forme de concordance (où chaque mot est présenté avec les mots qui le précédent ou le suivent directement, c'est-à-dire ses contextes de gauche et de droite), et dont les mots ont été classés suivant leur classe morphosyntaxique.

Exemples de calibrage de visualisation

Prenons quelques exemples simples. Imaginons que le texte nommé *La queste del saint Graal* (texte daté de ca1230; édition en cours par Christiane Marchello-Nizia), soit en train d'être annoté par quatre équipes différentes :

- une première équipe en ferait l'annotation morphosyntaxique;
- une seconde équipe annoterait les mots désignant les personnages;
- une troisième équipe en ferait l'annotation syntaxique;
- une quatrième équipe en ferait également l'annotation syntaxique, mais suivant un modèle théorique différent.

Soit quatre ensembles d'annotation différents, qui peuvent ou non être mobilisés simultanément par les utilisateurs. Voyons ce qu'il en est au travers des trois exemples suivants.

Exemple 1. Pour commencer, l'annotation morphosyntaxique baptisée *CATTEX*, appliquée aux textes de la Base de Français Médiéval (http://bfm.ens-lsh.fr), considère chaque mot selon la classe morphosyntaxique à laquelle il appartient (on parle de « parties du discours » en grammaire traditionnelle). Pour mener à bien sa tâche, l'annotateur n'a besoin que de visualiser chacun des mots et son contexte immédiat sous la forme d'une concordance. Il n'est pas nécessaire qu'il ait accès aux trois autres couches d'annotation. Ainsi, une visualisation simple comme la suivante suffit amplement (Figure 3).

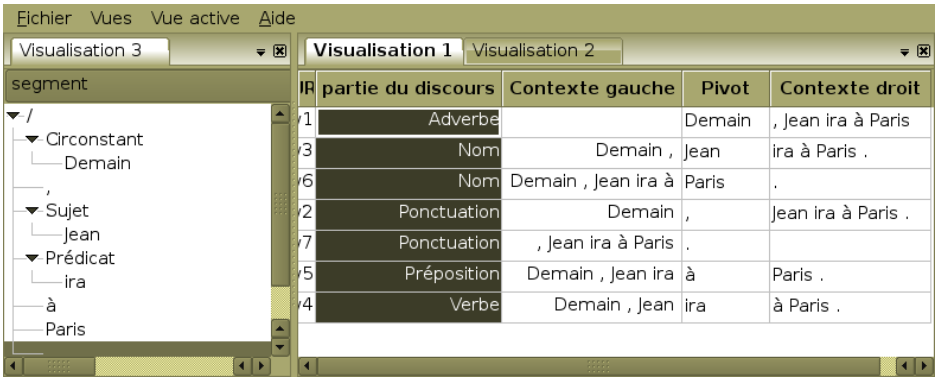

Figure 2. Deux approches différentes d'une même donnée

Figure 3. Concordance (annotation morphosyntaxique)

On voit que la deuxième colonne de la concordance (la première colonne contient la référence correspondant au mot) contient un ensemble d'annotations sous la forme cattex : '*Préposition*', cattex : '*Nom*', etc. Ces « parties du discours » sont associées aux mots de la colonne *Pivot* de la ligne

correspondante. Comme on le voit, un contexte très limité suffit à sélectionner la classe adéquate. Par ailleurs, ce type de vue permet des tris aisés sur les différentes colonnes, en vue d'une annotation plus rapide ou de corrections.

Exemple 2. En guise de deuxième exemple, voyons ce que nécessite l'annotation des mots désignant les personnages du texte – projet de thèse de Stéphanie Obry, ENS-LSH Lyon. Les noms et surnoms des personnages, mais également les unités linguistiques telles que les pronoms ou les adjectifs employés comme vocatifs sont autant d'unités susceptibles d'être annotées. Il est évident que les mots pronominaux, du fait de leur potentiel anaphorique, ne portent pas les marques qui permettent de repérer à coup sûr à quel personnage ils font référence. Par ailleurs, l'homonymie existant entre les pronoms de la troisième personne et les déterminants articles définis (*le*, *la*) oblige à prendre en compte le contexte immédiat. Ainsi, pour annoter les occurrences du mot *le*, l'utilisateur aura besoin de deux vues : une première vue permettant de discerner rapidement les pronoms des déterminants (par exemple, une concordance) et une seconde vue permettant d'associer les pronoms à leur antécédent (par exemple, le texte sous une forme courante).

Dans cette capture d'écran (voir Figure 4), on peut voir que la troisième ligne de la concordance (au-dessus) a comme pivot un pronom, ce qui n'est pas le cas de la première et de la quatrième ligne. En tant que pronom, le mot fait potentiellement référence à un personnage, en l'occurrence, Lancelot, ce qui est aisément dégagé d'une vue d'ensemble du passage où figure le pronom à annoter (en-dessous).

Exemple 3. Imaginons à présent que les équipes travaillant sur la base de modèles syntaxiques différents aient besoin de comparer les résultats de leurs analyses pour évaluer l'adéquation de leur modèle aux données analysées. Dans ce cas, il est nécessaire d'afficher simultanément les deux analyses (voir Figure 5).

Les deux analyses concurrentes suivent, à gauche, un modèle inspiré de Lucien Tesnière (1965), et, à droite, l'analyse en constituants immédiats (voir p. ex. Gleason 1969). Comme on le voit, le découpage des unités et la manière dont elles sont groupées sont fondamentalement différents, de même que la valeur des étiquettes qui sont placées sur les structures.

On dispose ainsi d'un moyen visuel de comparer les analyses. En recherchant et en alignant les différentes analyses gravitant autour d'un mot en particulier, on est en mesure de déterminer laquelle des deux modélisations convient le mieux aux données.

Figure 4. Concordance et texte courant (référence aux personnages)

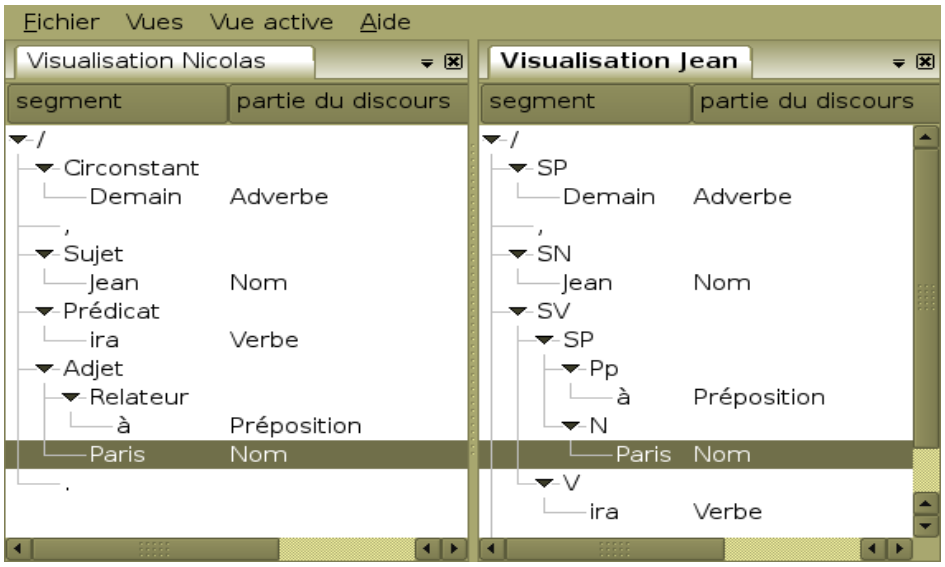

Figure 5. Analyses syntaxiques concurrentes (heuristique épistémologique)

Section 3 : Conceptualisations multiples

En sciences humaines les annotations sont le reflet d'une conceptualisation subjective de l'objet observé. Très souvent, comme on vient de le voir dans l'exemple associé à la Figure 5 ci-dessus, deux annotateurs de formation proche construisent deux ensembles différents de concepts. Un « inévitable éparpillement des étiquetages » (Habert, Nazarenko & Salem, 1997, pp. 23-24) se forme ainsi, faute d'un formalisme unificateur. Cette difficulté déteint fatalement sur les possibilités de comparaison entre les modèles. Pour pallier ce manque, nous avons décidé de choisir un langage d'expression unique de ces modèles (métamodèle), qui nous permet de représenter de manière similaire des conceptualisations différentes. Ainsi, les jeux d'étiquettes correspondant aux analyses syntaxiques concurrentes dont nous venons de parler peuvent être exprimés comme le montre la Figure 6.

En effet, dans la perspective de la mise en relation des différentes analyses, la visualisation simultanée de celles-ci est un premier pas, mais ne nous paraît pas suffisante. Il y a une nécessité méthodologique d'ordre collectif de documenter les schémas d'analyse employés, la structure des classes de concepts mobilisés et la manière dont ceux-ci interagissent avec les données. En d'autres termes, il faut que le chercheur puisse décrire son vocabulaire dans un langage qui puisse être compris par d'autres. Ce vocabulaire est désigné sous le nom d'*ontologie.*

La représentation présentée dans la Figure 6 est une manière de visualiser les vocabulaires formalisés à l'aide d'un langage spécifique d'une manière adaptée à la l'interaction entre l'humain et la machine. Le choix de NotaBene au niveau du langage formel est d'employer le *Web Ontology Language* (OWL, voir Bechhofer, Van Harmelen, Hendler, Horrocks, McGuiness, Patel-Schneider & Stein, 2004) pour assurer la communication entre les entreprises d'analyse. Ce langage a de multiples qualités, dont les plus intéressantes dans le cadre ici défini sont :

- plusieurs degrés de formalisation de l'ontologie sont possibles et varient du plus lâche au plus strict (Bechhofer et al., 2004, §8);
- OWL exploite le potentiel des espaces de nommage (Bray, Hollander, Layman & Tobin, 2006), qui délimitent précisément chaque ontologie, de façon à ce que plusieurs d'entre elles puissent interagir ou être comparées sans introduire d'équivoque;

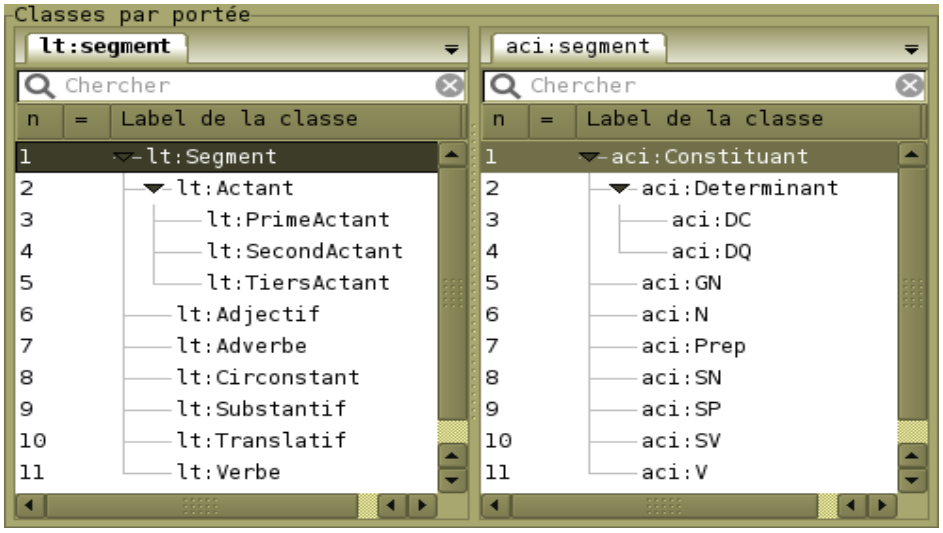

Figure 6. Jeux d'étiquettes correspondant aux analyses concurrentes

- OWL, conçu pour être exprimé en XML, peut être traduit dans n'importe quel langage de représentation de triplets (association entre trois éléments ordonnés) RDF (Klyne & Carroll, 2004, en partic. §3).

Cela dit, l'adoption du langage OWL ne signifie pas que nous adhérions à l'idée couramment véhiculée avec lui que les ontologies sont formulées *a priori* et sont nécessairement consensuelles (Laublet, Charlet & Reynaud, 2007, p. 105). Dans le cadre de la linguistique historique et, plus généralement, des sciences humaines, le travail scientifique implique un constant effort de conceptualisation nouvelle et de révision des concepts anciens. Techniquement, les ontologies OWL sont évolutives et mutualisables. Dans cette optique, il est important que l'utilisateur puisse modifier lui-même la terminologie d'une ontologie simple, comme le montre la Figure 7.

En refusant le figement, nous adoptons le point de vue selon lequel les ontologies sont dynamiques et conçues suivant une dialectique permanente (Iacovella, Bénel, Pétard & Helly, 2007, pp. 118-119). Toutefois, les ontologies ainsi élaborées peuvent être momentanément capturées pour permettre la comparaison de points de vues ou la mutualisation de la terminologie.

Figure 7. Éditeur d'ontologie intégré

La dernière version de travail de la spécification de NotaBene, qui détaille les formalismes choisis pour le logiciel, est accessible en ligne à l'adresse suivante : http://notabene.svn.sourceforge.net/viewvc/notabene/trunk/doc/specification.pdf.

Références

Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuiness, D.L., Patel-Schneider, P.F., & Stein, L.A. (Éds). (2004). *OWL Web ontology language reference. Reference. W3C recommendation 10 february 2004.* Document consulté le 26 mai 2010 de http://www.w3c.org/TR/2004/REC-owl-ref-20040210.

Bray, T., Hollander, D., Layman, A., & Tobin, R. (Éds). (2006). *Namespaces in XML 1.0* (2e éd.). *W3C Recommendation 16 August 2006.* Document consulté le 26 mai 2010 de http://www.w3c.org/TR/2006/REC-xml-names-20060816.

Gleason, H.A. (1969). *Introduction à la linguistique.* Paris : Larousse.

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus.* Paris : Armand Colin.

Iacovella, A., Bénel, A., Pétard, X., & Helly, B. (2007). Corpus scientifiques numérisés : savoirs de référence et point de vue expert. Dans R.T. Pédauque (Éd.), *La redocumentarisation du monde* (pp. 117-130). Toulouse : Cépaduès.

Klyne, G., & Carroll, J. (Éds). (2004). *Resource description framework (RDF) : Concepts and abstract syntax W3C Recommandation 10 February 2004*. Document consulté le 26 mai 2010 de http://www.w3c.org/TR/2004/REC-rdf-concepts-20040210/.

Laublet, P., Charlet, J., & Reynaud, C. (2007). Sur des aspects primordiaux du Web sémantique. Dans R.T. Pédauque (Éd.), *La redocumentarisation du monde* (pp. 99-116). Toulouse : Cépaduès.

Sperberg-McQueen, C.M., Burnard, L., Bauman, S., DeRose, S., & Rahtz, S. (2007). *TEI P5. Guidelines for electronic text encoding and interchange. Version 1.0.1*. Document consulté le 26 mai 2010 de http://www.tei-c.org/release/doc/tei-p5-doc/en/html/.

Tesnière, L. (1965). *Éléments de syntaxe structurale.* Paris : Klincksieck.

Nicolas Mazziotta *est actuellement* akademischer Mitarbeiter *pour le projet Syntactic Reference Corpus of Medieval French à l'Université de Stuttgart. Il est titulaire d'un doctorat en Langues et littératures romanes (Université de Liège). Ses centres d'intérêt sont essentiellement les systèmes graphiques médiévaux, l'édition numérique et la syntaxe du français (en partic. médiéval). Parallèlement à ses recherches, il développe les logiciels nécessaires à l'annotation de structures syntaxiques et à l'exploitation de ces annotations. Ses contrats de recherche l'ont amené à collaborer avec les principales initiatives de constitution de corpus de français médiéval (Base de français médiéval à l'ENS-Lyon et Nouveau corpus d'Amsterdam à l'Université de Stuttgart).*