Le *Syntactic Reference Corpus of Medieval French* : Structure, outils et exploitation

Nicolas Mazziotta

Université de Stuttgart – D-70174 – Allemagne

Abstract

We introduce the project *Syntactic Reference Corpus of Medieval French*, the aim of which is to provide the two main corpora of Old French texts with a dependency annotation of syntactic relations. First, we explain the main principles that lead the analyses and how the annotation is structured. Then, we describe the annotation procedures and tool, which suit our choice of a manual annotation. We show how our syntactic description can be used by cross-comparing it with morphological and discursive annotations issued by earlier projects by studying the form and the behaviour of subjects in the *Queste del saint Graal* (13th C.).

Résumé

Dans un premier temps, nous présentons le projet *Syntactic Reference Corpus of Medieval French*, dont le but est d'enrichir les deux principaux corpus de textes en ancien français à l'aide d'une annotation syntaxique de type dépendanciel. Après avoir exposé les principes généraux d'analyse syntaxique et la structure de l'enrichissement, nous abordons brièvement la méthodologie d'annotation, cette dernière ayant été réalisée manuellement. Nous montrons ensuite comment les annotations syntaxiques peuvent être exploitées en croisant les informations fournies par ces dernières à d'autres couches indépendantes d'annotation morphosyntaxique et énonciative. Nous illustrons cela au travers de la question de la forme des sujets dans le texte de la *Queste del saint Graal* (13^{e} s.).

Mots-clés : syntaxe, corpus enrichi, annotation, ancien français, linguistique historique.

1. Introduction

Dans cette contribution, nous présentons le projet *Syntactic Reference Corpus of Medieval French* (désormais « SRCMF »), initiative subventionnée conjointement par l'*Agence Nationale de la Recherche* (France) et la *Deutsche Forschungsgemeinschaft* (Allemagne) jusqu'au 29 février 2012. Il s'agit ici de donner un aperçu général de ce projet d'annotation syntaxique des plus importants corpus disponibles pour l'étude de l'ancien français (désormais « afr. »). Notre objectif est avant tout méthodologique : nous voulons partager notre position par rapport aux matériaux et à leur enrichissement. Il ne sera pas question d'entrer dans les détails linguistiques et techniques les plus pointus, car cela nuirait à la visée générale de notre texte.

Nous aborderons successivement : les matériaux et les objectifs (section 2) ; le modèle d'analyse syntaxique employé (section 3) ; les outils et la démarche d'annotation (section 4), ainsi que les outils d'exploitation, au travers d'un exemple (section 5). Notre exposé montrera le potentiel

ouvert par les résultats, mais il sera également émaillé d'autocritiques. Enfin, nous conclurons notre article en synthétisant ces deux aspects.

2. Matériaux et objectif

Tel que défini en 2008[1], l'objectif du projet était d'enrichir les deux plus importants corpus d'afr. à l'aide d'annotations syntaxiques correspondant à un modèle commun, tout en rendant possible l'examen des interactions entre la syntaxe et les autres domaines de la grammaire qui ont bénéficié d'une annotation antérieure.

2.1. Corpus de base et enrichissements antérieurs

L'enrichissement a ainsi porté sur les deux corpus d'afr. les plus importants (ca. trois millions d'occurrences-mots chacun) :

- la *Base de Français Médiéval* (désormais « BFM », Lyon, dir. Céline Guillot, voir Guillot *et al.* 2007) ;
- le *Nouveau Corpus d'Amsterdam* (désormais « NCA », Stuttgart, dir. Achim Stein, voir Kunstmann et Stein 2007b).

Il ne s'agit pas de corpus nus. Tout d'abord, la BFM et le NCA comprennent l'un comme l'autre un ensemble de descripteurs bibliographiques associés à chaque édition numérisée : titre, auteur, éditeur, etc. Parmi ces descripteurs, on trouve aussi et surtout une importante collection de métadonnées spécifiques aux textes médiévaux et qui permettent l'analyse variationnelle de phénomènes linguistiques : date de composition, date du manuscrit, lieu de rédaction, catégorisation de la langue employée, genre textuel, etc. En fonction du corpus, la nature des métadonnées est différente et il s'en faudrait de beaucoup pour qu'elles se recouvrent complètement ; pour la BFM, voir le *Manuel de description des textes* [2]et, pour le NCA, voir la bibliographie associée au corpus (Gleßgen et Vachon 2011). À ces métadonnées, qui concernent l'ensemble de l'unité textuelle, s'ajoutent des annotations qui portent sur les occurrences-mots, sous la forme d'étiquettes morphosyntaxiques. À nouveau, les jeux d'étiquettes des deux corpus ne correspondent pas exactement, bien que les conventions de nommage des étiquettes du NCA aient été adaptées (Kunstmann et Stein 2007b : § 2.1) à celles du jeu « Cattex », développé pour la BFM[3].

Outre ces enrichissements qui, quoique complémentaires, restent du même ordre pour les deux corpus, chacun d'eux offre des ressources qui lui sont propres. Ainsi, le NCA comprend une lemmatisation générée automatiquement à l'aide de TreeTagger (Kunstmann et Stein 2007b : § 2.1), alors que l'équipe de la BFM élabore en ce moment un balisage permettant de repérer les portions textuelles correspondant au discours direct (communication personnelle).

2.2. Objectif

La couche d'annotation syntaxique s'ajoute à cette collection d'informations déjà présentes. L'idée originale était de produire une annotation similaire pour les deux corpus. Cette

1 Voir le descriptif : http:www.lattice.cnrs.fr/IMG/pdf/SRCMF-anr-dfg.pdf.

2 Voir http://ccfm.ens-lsh.fr/IMG/pdf/Manuel_Descripteurs_BFM.pdf

3 Voir http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_Manuel.pdf.

démarche unifiée rapproche les deux corpus, dont l'enrichissement a jusqu'à présent souffert d'idiosyncrasie. Toutefois, lorsque les textes figuraient dans les deux corpus, il s'est révélé impossible de projeter automatiquement les annotations syntaxiques effectuées d'un corpus à l'autre, en raison de choix méthodologiques différents en matière de segmentation des mots, de correction et de sélection des éditions de référence[4].

Fin 2011, les 18 textes et extraits ayant reçu une annotation syntaxique comptabilisent un total de 351 778 occurrences-mots pour une période courant du premier texte en afr. (*Serments de Strasbourg*, 842) à la fin du 13e siècle.

3. Modèle syntaxique et structure de l'annotation

Nous commencerons cette section par un bref aperçu des spécificités de l'afr. (3.1). Nous expliquerons ensuite les grandes lignes du modèle d'analyse que nous utilisons, modèle que l'on peut qualifier de *dépendanciel* (3.2). Pour clôturer cette section, nous présenterons brièvement notre modèle de données, ainsi que les formats que nous avons employés (3.3).

3.1. Spécificités de l'ancien français

Pour présenter les caractéristiques de l'afr. qui contrastent le plus avec celles du français moderne, examinons l'exemple suivant :

Des lors	*te*	toli	li anemis	la veue
Dès lors	toi (objet indirect)	prit	le Diable (sujet)	la vue (objet)

Dès cet instant, le Diable t'a dépouillé de ta vue. (Queste 190a, 1)

En règle générale, le verbe (*toli*) apparaît directement à la suite du complément topicalisé, qui peut être le sujet (ici, il s'agit d'un complément circonstanciel : *des lors*), et les clitiques (*te*). Si le sujet est exprimé (il peut ne pas l'être) et n'est pas topicalisé (*li anemis*), il suit le verbe, mais sa présence n'est pas obligatoire. Les autres compléments (*la veue*) suivent également le verbe. Le cas « sujet » (sorte de nominatif) est marqué par des morphèmes spécifiques (*li*, *-s*), mais de manière irrégulière. La déclinaison finit de disparaître du système du nom, bien que les pronoms conservent une opposition à trois cas qui subsiste encore de nos jours (sujet/objet/objet indirect). En d'autres termes :1/ l'ordre des mots dans la proposition exprime des informations d'ordre énonciatif plutôt que la structure syntaxique (voir Buridant 2000 : § 631) ; 2/ la morphologie nominale est pauvre et peu fiable (voir Moignet 1988 : 87).

En outre, l'étude linguistique des variétés anciennes du français doit composer avec les mêmes difficultés que présentent les langues éteintes : privée de la compétence des locuteurs natifs, elle doit impérativement recourir à un corpus. Par ailleurs, comme il est probable que de nombreuses constructions régulières et fréquentes en afr. ne sont jamais attestées dans le corpus étudié (Marchello-Nizia 1985), il est nécessaire de prendre en compte la sémantique et l'énonciation pour comprendre que afr. *La vache avra ma dame* signifie « ma femme aura la vache » et non le contraire.

4 Le NCA se révèle beaucoup plus interventionniste que la BFM en ce qui concerne l'homogénéisation des formes, notamment.

3.2. Modèle syntaxique dépendanciel

Ces difficultés complexifient la description syntaxique et il est donc important que le cadre du modèle d'analyse soit suffisamment souple. Nous en présentons ici les grandes lignes. Nous avons élaboré un modèle *dépendanciel*, proche des modèles de Tesnière 1965 et de Mel'čuk 2009. Les phrases sont décrites comme des hiérarchies de mots connectés les uns aux autres par des relations nommées. Ce modèle contraste avec la majorité des formalisations, qui reposent sur une structure composée d'une imbrication de constituants immédiats, imposant une structure plus rigide et plus complexe (pour une bonne présentation des avantages de l'analyse dépendancielle, voir Osborne 2006 : 53-58)[5].

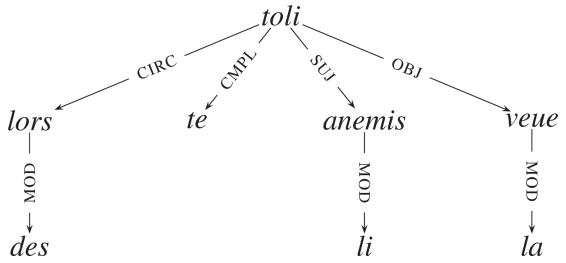

Figure 1 : Représentation dépendancielle

Nous ne ferons pas ici la liste de tous les types de structures que l'on peut rencontrer : nous nous bornerons à une très brève introduction. Schématiquement, les mots sont représentés par des nœuds hiérarchisés. On dit qu'un nœud lié par une relation à un nœud supérieur est le *dépendant* de ce *gouverneur*, dont il est dit qu'il *dépend*. Comme on le voit sur la fig. 1 (analyse de l'ex. 1), le verbe principal n'a pas de gouverneur et surplombe (*gouverne*) ses compléments, y compris le sujet et les compléments circonstanciels. Par ailleurs, l'ordre linéaire, considéré comme une marque et non une relation (Mel'čuk 2009 : 24), n'est pas illustré dans le schéma.

3.2.1. Classes de dépendants

Chaque relation est étiquetée du nom de la fonction représentée. Les principales sont : 1/ au niveau des dépendants du verbe, sujet (abrégé *Suj*), objet *(Obj),* complément régi autre que l'objet (*Cmpl*)*,* attribut du sujet (*AtSj*), circonstant (*Circ*) ; 2/ à tous les autres niveaux, modifieur (*Mod*). Cette petite liste appelle déjà deux commentaires.

Premièrement, la distinction entre les différents types de fonctions est beaucoup plus fine au niveau des dépendants du verbe. Cette différence est due à la structure de la langue étudiée. Le verbe est la « partie du discours » qui impose le plus de contraintes formelles et sémantiques à ses dépendants. Ces derniers peuvent être classés en tenant compte de la disparité des contraintes (correspondance avec la forme d'un pronom, utilisation d'une préposition, présence du sème locatif, etc.). Les noms, les adjectifs et les adverbes sélectionnent leur régime de manière beaucoup plus souple, ce qui se manifeste par une indigence des marques qui empêche un

5 Dans l'histoire du projet, la perspective dépendancielle a progressivement effacé l'approche en constituants, bien que les deux modèles eussent été employés simultanément au début de l'entreprise (les analyses liant aussi bien des occurrences-mots que des constituants).

classement aisé. En conséquence, une seule classe de dépendants du « non-verbe » existe, là où nous distinguons douze classes de dépendants du verbe à proprement parler[6].

Deuxièmement, les termes employés sont loin d'être consensuels. La question de la terminologie s'est ici révélée fondamentale. Il a été nécessaire de définir avec précision (quoique souvent de manière assez pragmatique) des notions aussi lourdement marquées par la tradition que celles de *phrase* ou de *sujet*. Nous ne prétendons pas avoir résolu tous les problèmes, mais les choix posés ont permis à sept personnes différentes d'annoter les textes sur une base commune, synthétisée dans un *Guide de l'annotateur*[7] qui permettra à quiconque de se positionner par rapport à nos décisions tant théoriques que pratiques.

3.2.2. Relations particulières

L'entreprise qui consiste à départir les relations en classes qui correspondent à leur fonction par rapport à leur gouverneur est considérable. Toutefois, elle ne doit pas pour autant masquer les questions relatives aux relations elles-mêmes. Dans ce cadre, chaque relation entre deux occurrences-mots est caractérisée par une *orientation*, qui détermine laquelle des deux unités doit être qualifiée de gouverneur. Il est parfois problématique de déterminer l'orientation d'une relation, et les critères avancés par les théories les plus abouties (Mel'čuk 2009) sont parfois inopérants pour l'afr. Nous retiendrons ici un cas particuliers : celui des relateurs (conjonctions et prépositions).

La question de savoir si le relateur gouverne ce qu'on appelle traditionnellement son « régime » est difficile à résoudre en afr. Il est clair que les phénomènes de juxtaposition, de parataxe ou de détermination directe existent, mais certains compléments sont nécessairement introduits par un mot-outil. Si l'approche habituelle consiste à faire des relateurs des gouverneurs plutôt que des dépendants (cf., p. ex., Mel'čuk 2009), SRCMF en a pris le contrepied. Les raisons de ce choix sont simples : 1/ historiquement, les relateurs ont eu une importance fonctionnelle variable depuis le latin et celle-ci est difficile à évaluer pour l'afr. ; 2/ il est en revanche resté constant que la proposition, en tant que représentation linguistique de phénomènes, met en relation des unités lexicales, qui sont le fondement du sens.

3.3. Modèle de données et format des ressources

La nature hiérarchique des annotations et la présence de relations d'ordres différents complexifient la question de la formalisation des données. Du point de vue de l'encodage de l'information, les index (associations clef-valeur) ou les documents XML traditionnels (modèle arborescent strictement projectif) se révèlent insuffisants. Par contre, le modèle linguistique dépendanciel est très facilement formalisable en suivant les principes du modèle classique *entité-relation* (Chen 1976). La structure de données est fondée sur une collection de *tuples* de trois éléments : deux entités liées par une relation. Pour chaque phrase analysée, cette collection forme un graphe acyclique connexe qui peut être traité efficacement par les librairies logicielles modernes. On transpose facilement la fig. 1 sous une forme linéaire. La structure des notations est ici la suivante : <entité1, entité2>:relation :

6 Nous n'entrerons pas ici dans les détails de la distinction entre les dépendants du verbe et les dépendants de la proposition, qui sont assimilés aux premiers dans le cadre de SRCMF.

7 Voir http://www.srcmf.org.

- <*toli*, Verbe>:est_un
- <*lors*, *toli*>:circonstant_de
- <*des*, *lors*>:relateur_de
- ...

La hiérarchie de terminologie elle-même peut être exprimée de la sorte :

- <Sujet, Actant>:sous-classe_de
- <Sujet personnel, Sujet>:sous-classe_de
- <Régime, Actant>:sous-classe_de
- ...

Le *Resource Description Framework* (RDF, Klyne *et al.* 2004) permet de mettre en relation les structures et les mots, ces derniers mot possédant un identifiant (par exemple, l'*Uniform Resource Locator*, ou « URL » utilisé pour localiser les pages webs). Nos ressources sont annotées de façon « débarquée » (Loiseau 2007), c'est-à-dire en séparant les données « brutes » de leur analyse (en pratique, dans deux fichiers différents). L'existence d'un standard de description ontologique (*OWL Web Ontology Language*, Bechhofer *et al.* 2004) rend aisée la formalisation d'un jeu d'étiquettes selon le même modèle de données.

En conséquence, nos ressources sont réparties dans trois modules : 1/ les observables (textes numérisés), encodés en XML-TEI ; 2/ la terminologie sous la forme d'un graphe OWL ; 3/ les annotations décrivant la structure syntaxique, liant incidemment cette dernière à la terminologie.

4. Annotation

Avant tout, nous préciserons les choix méthodologiques qui ont été posés préalablement à l'analyse et quelles sont les conséquences immédiates à en tirer (4.1). Nous donnerons ensuite un aperçu du logiciel employé pour effectuer l'annotation (4.2) ainsi que des procédures garantissant la qualité du résultat (4.3).

4.1. Philosophie générale

Dès le début du projet, deux options ont été choisies *a priori* : 1/ la couche d'annotation syntaxique serait élaborée indépendamment des analyses déjà disponibles ; 2/ l'annotation serait effectuée manuellement, c'est-à-dire par un humain, sans aucune procédure automatique ou semi-automatique.

Comme nous l'avons vu ci-dessus (2.1), nous disposons de banques lexicales et de quelques textes annotés de manière fiable en parties du discours. Toutefois, du fait de la disparité des deux corpus, ces enrichissements ne suffisent pas à une analyse *bottom-up* satisfaisante et il ne saurait être question de baser la description syntaxique sur eux de manière systématique.

En outre, un avantage non négligeable de la disjonction des couches est que ces dernières peuvent ainsi être comparées l'une à l'autre *a posteriori*, se validant mutuellement ou, au contraire, se contredisant. Par exemple, il sera possible de vérifier l'adéquation de l'étiquette Cattex correspondant à la notion de *préposition* en vérifiant si les occurrences-mots concernées

sont bien annotées comme des relateurs de structures non propositionnelles dans SRCMF (et vice-versa).

Quant à l'annotation manuelle, elle présente à la fois des avantages et des inconvénients. Il est clair qu'une annotation manuelle est nécessairement plus coûteuse en temps et en main d'œuvre, la nature de la langue étudiée impliquant en outre un niveau d'expertise élevé (tous les annotateurs du projet ont soutenu une thèse touchant de près la linguistique de l'ancien français).

Pour une langue comme l'afr., où les marques segmentales sont loin d'être déterministes et où les contraintes séquentielles ont souvent une valeur énonciative, l'annotation automatique n'est pas envisageable *a priori* : outre le choix de ne pas utiliser les couches d'annotations disponibles pour les raisons mentionnées dans la sous-section qui précède, l'orthographe de l'afr. est trop hétérogène pour servir de repère fiable.

D'un autre côté, l'annotation manuelle génère beaucoup plus d'incohérences que l'analyse automatique, mais elle présente deux avantages majeurs. Tout d'abord, les erreurs sont beaucoup moins systématiques. Une analyse statistique de phénomènes potentiellement mal analysés fait ressortir les erreurs comme des aberrations. Moins d'artefacts en résulteront. Ensuite, la procédure manuelle est inductive ; elle ne masque pas les structures exceptionnelles non connues *a priori* comme pourrait le faire une analyse déductive automatique.

4.2. Outil : NotaBene RDF Annotation Tool

Les choix théoriques et les objectifs définis, leur mise en œuvre est loin d'être évidente. La complexité des structures syntaxiques rend l'annotation directe dans les fichiers sources totalement irréalisable. Nous avons donc développé un logiciel capable de manipuler cette couche d'annotation de manière ergonomique. Ce logiciel, baptisé *NotaBene RDF Annotation Tool*[8] (voir notamment Mazziotta 2010), permet d'employer le modèle décrit sous 3.2 et de le stocker dans la base de connaissance décrite sous 3.3 de manière accessible.

8 Disponible à l'adresse https://sourceforge.net/projects/notabene/. Version alpha. Merci de contacter l'auteur en cas d'utilisation.

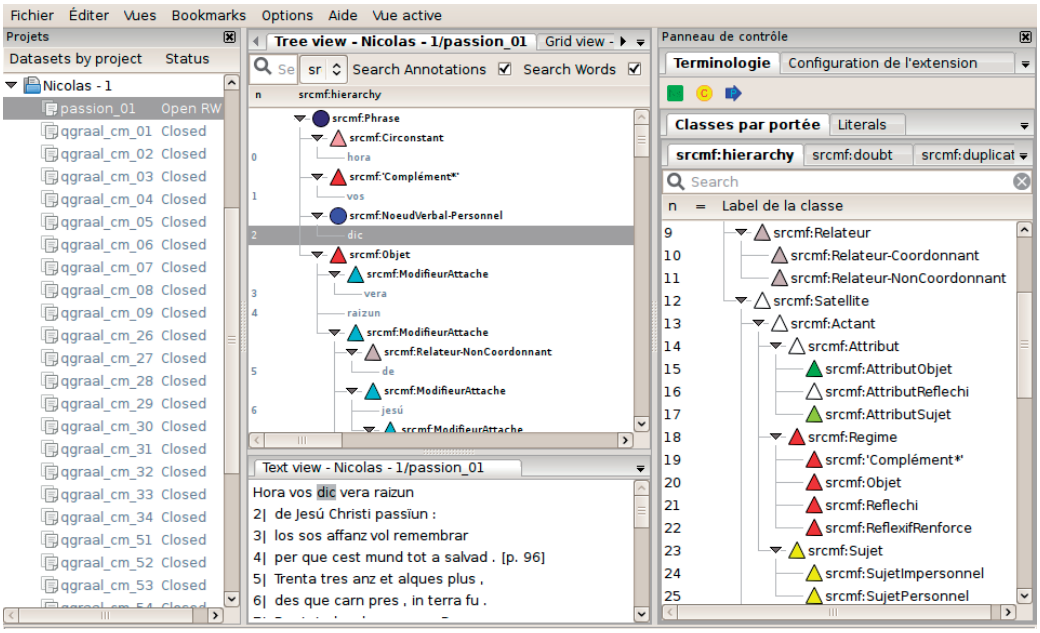

Figure 2 : Interface de NotaBene employée pour annoter la Passion de Clermont

Comme on le voit dans la fig. 2, différentes vues sont synchronisées. L'analyse syntaxique du texte (au milieu en bas) est représentée par une arborescence (au milieu en haut)[9]. À chaque niveau de l'arbre se trouve soit un symbole étiqueté, soit un terme lexical. Dans les grandes lignes, nous dirons qu'un disque représente un gouverneur (comme c'est le cas du nœud verbal personnel *dic* ci-dessus), ou une structure qui n'a pas de gouverneur (la phrase, par exemple). Un triangle indique un dépendant de ce gouverneur, étiqueté de sa fonction. La terminologie prend également la forme d'une arborescence (à droite).

4.3. Contrôle qualité

La complexité naturelle de la structure de l'analyse syntaxique hiérarchique implique une irréductible complexité de l'annotation. Cette dernière entraîne à son tour une inévitable complexité de la pratique concrète de l'encodage qui mène à des erreurs. Par ailleurs, même si l'équipe de travail a atteint un consensus suffisant et documenté concernant la terminologie employée, de nouveaux cas sont découverts dans chaque extrait annoté. En conséquence, il est impossible que deux annotateurs soient en permanence en accord – la cohérence d'un annotateur pris indépendamment est même parfois sujette à caution.

Erreurs et inconsistances sont évidemment dommageables dans une entreprise de ce type. Par conséquent, la procédure d'annotation se devait d'être stricte. Elle se résume comme suit :

1. Deux annotateurs (A et B) annotent l'extrait en parallèle sans concertation.
2. Comparaison des résultats par A et suggestions de correction à B : A corrige sa propre annotation s'il trouve que l'analyse de B est meilleure et celle de B dans le cas contraire. Au terme de cette étape, les deux ensembles d'annotations sont identiques.

9 Elle aurait pu être représentée par un graphe pour rester plus isomorphe du modèle représenté. Cependant, les test d'ergonomie que nous avons réalisés nous ont fait préférer une représentation arborescente.

3. B compare le résultat des étapes 1 et 2 pour ses annotations. Il accepte ou rejette les corrections de son annotation qui ont été effectuées par A. Au terme de l'étape, les différences entre les annotations de A et celles de B correspondent à deux analyses différentes.
4. Deux correcteurs (C et D) comparent les résultats et choisissent entre l'analyse de A et celle de B en répétant les étapes 2 et 3 en se substituant aux annotateurs.
5. Les différences subsistant sont discutées et éventuellement éliminées.

5. Exploitation

Après une revue des principaux formats d'exportation envisagés par SRCMF (5.1), nous donnerons un exemple de traitement à l'aide du format d'exportation *TigerSearch* (5.2).

5.1. Exportations

À chaque exportation correspondent des conventions de notation et de représentation propres. Nous avons prévu essentiellement trois modules d'exportation :

1. un module générant des visualisations graphiques ;
2. un module générant une analyse formatée selon les conventions CoNLL[10], qui servira de base (« Gold Standard ») pour entraîner les parseurs dépendanciels en vue d'annotations automatiques ultérieures ;
3. un module générant une analyse dans le format TigerXML, adapté au logiciel d'extraction *TigerSearch* (König *et al.* 2003), pour permettre l'exploitation du corpus par des requêtes sur les analyses syntaxiques, le logiciel étant spécialisé dans ce sens[11].

Contrairement à l'analyse, les exportations intègrent non seulement les annotations SRCMF, mais également les autres ressources disponibles dans les deux corpus sur lesquels nous avons travaillé.

5.2. Exemple : les sujets dans la Queste del saint Graal

En croisant les enrichissements préalables avec les annotations, il devient possible d'user de procédures quantitatives pour répondre à des questions qui ne pouvaient être traitées sans un important travail préparatoire. Nous ne ferons pas ici une étude détaillée d'un problème précis, ce qui aurait donné lieu à un article complet. Nous résumerons en quelques lignes comment le corpus aide à trouver des éléments de réponse à un questionnement spécifique.

5.2.1. Définition de la question

Selon l'étude que nous avons menée précédemment (Glikman et Mazziotta à paraître), le discours oral représenté du texte de la *Queste del saint Graal* (texte du 13e s., voir Marchello-

10 Concernant *CoNLL-2009 Shared Task : Syntactic and Semantic Dependencies in Multiple Languages*, voir http://ufal.mff.cuni.cz/conll2009-st/.

11 Le format est actuellement en cours de révision : la prochaine version sera plus adaptée aux analyse dépendancielles que ne l'est la version actuelle. Voir http://korpling.german.hu-berlin.de/tiger2/homepage/index.html.

Nizia 2009) a plusieurs spécificités par rapport à la narration. Entre autres, il se caractériserait par un nombre plus important de sujets exprimés, sachant, comme nous l'avons mentionné ci-dessus (3.1), que l'afr. est une langue dont les phrases peuvent ne pas contenir de sujet au sens traditionnel du terme.

Cette découverte donne envie d'examiner plus avant la forme des sujets selon leur environnement. Nous définirons donc nos individus (les sujets exprimés) à l'aide de variables correspondant aux questions suivantes :

1. Quelle est sa forme grammaticale ? Cette question décrit l'aspect morphologique du sujet. En l'occurrence, nous voulons avant tout savoir s'il s'agit d'un pronom ou d'un syntagme plus étendu – variable *PRO*, avec deux modalités : *0* « non pronominal » et *1* « pronominal »).
2. Apparaît-il dans une subordonnée ou une principale ? Il s'agit ici de la dimension syntaxique de l'individu – variable *SUB*, avec deux modalités : *0* « en subordonnée » et *1* « en principale »).
3. Apparaît-il dans un énoncé correspondant à de l'oral représenté ou non ? Cet aspect discursif est hiérarchisé – variable *DD*, avec trois modalités : *0* « narration », *1* « oral représenté dans la narration », *2* « oral représenté dans l'oral représenté (récursif) » ?
4. Se trouve-t-il devant le verbe ? La question combine marques séquentielles et syntaxe à proprement parler – variable *PRV*, avec deux modalités : *1* « devant le verbe », *0* « après le verbe ».

Ces quatre questions mobilisent trois types de ressources : l'annotation SRCMF permet de collecter les individus et de répondre à la question 2. Les annotations morphosyntaxiques de la BFM (en l'occurrence, le texte n'est pas dans le NCA) répondent à la question 1. L'annotation du discours cité dont bénéficie notre texte dans la BFM répond à la question 3. Enfin, la combinaison de l'édition et des analyses SRCMF permet de répondre à la question 4.

5.2.2. Extraction des données

Nous travaillons sur l'exportation des annotations au format employé par *TigerSearch*. Ce dernier nous permet d'extraire plusieurs jeux de données, que nous combinons en un seul tableau. Nous soumettons ainsi le corpus à la requête entrée dans la fenêtre de gauche de la figure 3. Elle signifie littéralement : « 1/ extraire tous les sujets personnels (cat="SjPer") et les associer à la variable *sj* (#sj:) ; 2/ extraire le premier mot de chacun de ces sujets (>@l) et les associer à la variable *w* ; 3/ ne tenir compte que des sujets de phrases subordonnées (ligne 2) ; 4/ ne tenir compte que des sujets suivant le verbe (lignes 3 et 4). » Nous pouvons exporter les résultats en sélectionnant, pour chaque variable, les éléments de description qui correspondent à notre questionnement (fenêtre de droite de la fig. 3)[12].

12 Via le menu *Query* > *Statistics*. On se reportera au manuel de *TigerSearch* pour plus de détails concernant les exportations (König *et al.* : 99-107) et au manuel d'exploitation de SRCMF par *TigerSearch* pour plus de détails sur les éléments de description disponibles (http://www.srcmf.org).

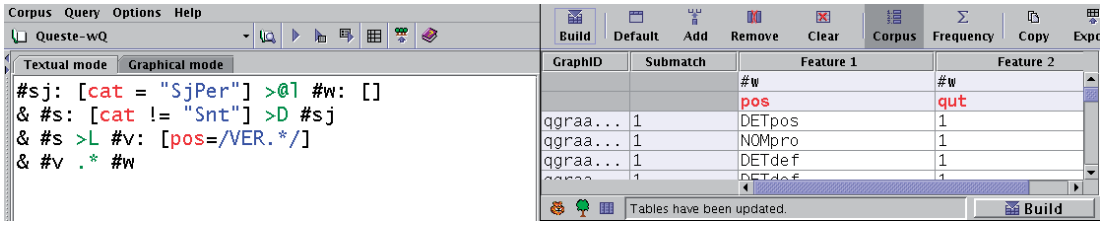

Figure 3 : Exportation des statistiques sous TigerSearch

Ensuite, nous ajoutons aux résultats obtenus ceux produits par trois autres requêtes extrayant les sujets de subordonnées situés devant le verbe, puis ceux des principales, devant et derrière le verbe. Les résultats sont prêts à être exportés dans quatre tableaux, qui pourront être manipulés par un logiciel d'analyse statistique.

5.2.3. Traitement des données et premiers éléments d'interprétation

Après quelques manipulations que nous ne détaillerons pas ici, le tableau des individus comprend 4256 lignes structurées comme celles présentées dans le tab. 1. Une analyse factorielle des correspondances multiples[13] (dont on a retiré les pronoms relatifs à cause de leur comportement trop prévisible) donne les résultats de la fig. 4. On y remarque que la structure morphologique des sujets (pronominale ou non) est davantage influencée par des facteurs syntaxiques (opposition entre proposition subordonnée ou principale) et séquentiels (position préverbale ou non) que par des facteurs discursifs (discours cité ou non) : le sujet pronominal a tendance à être préverbal et à apparaître en subordonnée. Bien entendu, l'examen de la question est par trop superficiel et nous nous garderons de conclure à ce stade, mais il serait aisé d'introduire d'autres variables et d'effectuer des analyses complémentaires.

SUB	DD	PRO	PRV
0	0	1	1
0	0	1	1
0	1	1	1

Tableau 1 : Variables correspondant à trois sujets

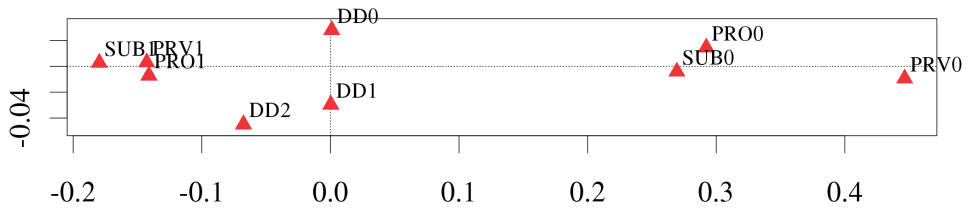

Figure 4 : ACM des modalités décrivant le sujet

13 En réalité, une *Joint Correspondance Analysis* (Greenacre/Nenadic 2006), réalisée à l'aide du package « ca » pour le logiciel R. Nous ne donnons pas ici les valeurs numériques associées à cette analyse.

6. Conclusion

Le projet SRCMF s'articule avec d'autres ressources, dont il se sert ou qui peuvent être mobilisées de manière concomitante. Les interactions entre ces différentes ressources sont résumées dans la fig. 5, où l'on remarquera prioritairement les éléments qui suivent.

1. SRCMF s'appuie sur les textes encodés en XML des corpus du NCA et de la BFM (qui utilise XML-TEI). Ces textes sont centraux. Avec les annotations et la terminologie spécifiques et construites dans le cadre du projet, ils constituent la ressource primaire produite par ce dernier.
2. L'annotation est effectuée par l'équipe sans tenir compte des annotations déjà présentes dans la BFM et le NCA, mais en recourant quand cela est nécessaire aux éditions qui fondent les bases textuelles préexistantes.
3. La ressource SRCMF est ensuite fusionnée aux ressources extérieures du NCA et de la BFM pour que les ressources exportées dans des formats facilement exploitables à l'aide d'outils existants (parseurs *CoNLL*, outil de requête *TigerSearch*, tableurs, etc.) contiennent l'ensemble des enrichissements disponibles à ce jour. Les utilisateurs peuvent ainsi pleinement profiter des corpus et ont la possibilité, le cas échéant, de contribuer à les améliorer par une évaluation critique croisée.

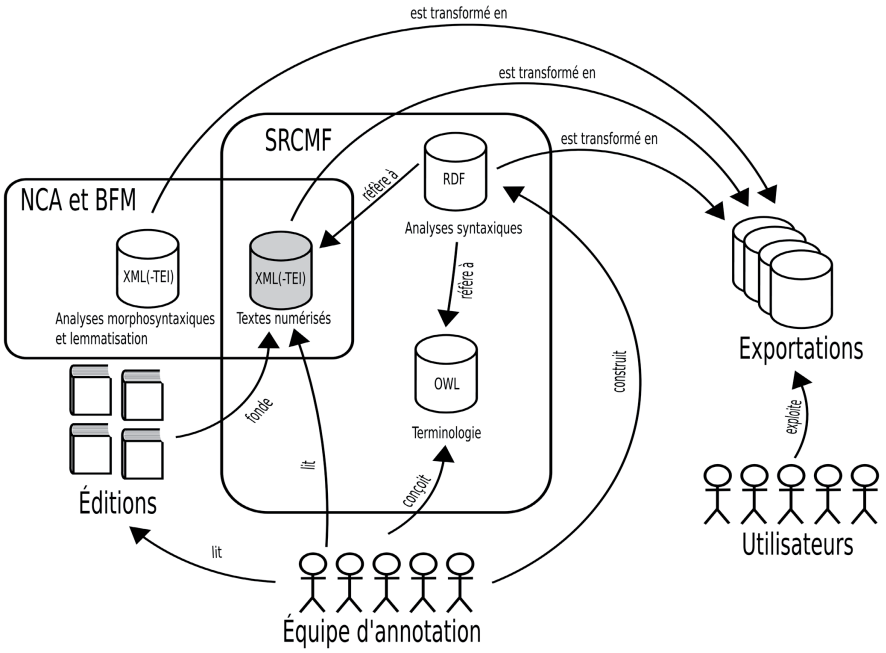

Figure 5 : Ressources et interactions principales

Références

Bechhofer S., van Harmelen F., Hendler J., Horrocks I., McGuinness D.L., Patel-Schneider P.F. et Stein L.A. (2004). *OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004* [http://www.w3.org/TR/owl-ref/].

Buridant C. (2000). *Grammaire nouvelle de l'ancien français*. Sedes.

Chen P.P.-S. (1976). The entity-relationship model – Toward a unified view of data, *ACM Transactions on Database Systems*, vol. 1 : 9-36.

Glikman J. et Mazziotta N. (à paraître). Représentation de l'oral et structures syntaxiques dans la prose de la *Queste del saint Graal (1225-1230)*. In *Actes de Representations du sens linguistique V*, Chambéry, 2011.

Greenacre M. et Nenadic O. (2010). *ca: Simple, Multiple and Joint Correspondence Analysis R package version 0.33* [http://CRAN.R-project.org/package=ca].

Guillot C., Marchello-Nizia C. et and Lavrentiev A. (2007). La base de français médiéval (bfm) : états et perspectives. In Kunstmann et Stein 2007a, pp. 143-152.

Klyne G. et Carroll J.J., éditeurs (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax W3C Recommendation 10 February 2004 [http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/].

Kunstmann P. et Stein A., éditeurs (2007a). *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*. Steiner.

Kunstmann P. et Stein A. (2007b). Le Nouveau Corpus d'Amsterdam. In Kunstmann et Stein 2007a, pp. 9-27.

Loiseau S. (2007). CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations. *Corpus*, vol. 7 : 153-186.

Marchello-Nizia C. (1985). Question de méthode. *Romania*, vol. 106 : 481–492.

Marchello-Nizia C., éditeur (2009). *Queste del Saint Graal. Édition numérique interactive. Manuscrit Lyon, Bibliothèque municipale, P.A. 77* [http://textometrie.risc.cnrs.fr/txm/texte/quete].

Mazziotta N. (2010). Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples. Recherches qualitatives. Hors-série "Les actes", 9, 83-94.

Mel'čuk I. (2009). Dependency in natural language. In Polguère A. et Mel'čuk I., éditeurs. *Dependency in linguistic description*. Benjamins, 1-110.

Moignet G. (1988). *Grammaire de l'ancien français*. Klincksieck.

Osborne T. (2006). Shared Material and Grammar : Toward A Dependency Grammar Theory of Non-Gapping Coordination for English and German. *Zeitschrift für Sprachwissenschaft*, vol. 25 : 39-93

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

König E., Lezius W. et Voormann H. (2003) *TIGERSearch 2.1 User's manual. University of Stuttgart* [http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/manual.shtml]