La base textuelle du *Dictionnaire électronique de Chrétien de Troyes.* Établissement et fonctionnalités

Pierre Kunstmann

Le Dictionnaire Électronique de Chrétien de Troyes (désormais DÉCT) constitue à la fois un lexique de cet écrivain du XII^e^ siècle et une base textuelle qui permet de lire ou d'interroger les transcriptions de ses cinq romans (*Érec*, *Cligès*, *Lancelot* ou le *Chevalier à la Charrette*, *Yvain* ou le *Chevalier au Lion*, *Perceval* ou le *Conte du Graal*). Le DÉCT résulte de la collaboration d'un groupe de chercheurs : Pierre Kunstmann, du LFA (Laboratoire de Français Ancien, Université d'Ottawa), qui en assume la direction et rédige les articles; Hiltrud Gerner, de l'ATILF (Analyse et Traitement Automatique de la Langue Française, CNRS Nancy-Université), chargée de la présentation et de la révision des articles ainsi que de la rédaction de certaines lettres; Gilles Souvay, de l'ATILF, pour les développements informatiques; A. Stein de l'Institut für Linguistik / Romanistik, Université de Stuttgart, pour l'analyse sémantique.

Le DÉCT s'adresse à un large public : il peut intéresser le lycéen ou l'amateur éclairé aussi bien que le spécialiste de l'ancienne langue. L'ouvrage, placé sur le serveur de l'ATILF est consultable en accès libre à l'adresse suivante : www.atilf.fr/dect/. Le dictionnaire s'effectuera en deux étapes : DÉCT1 et DÉCT2. Le premier se limite à la définition des mots lexicaux (adjectifs qualificatifs, adverbes se terminant en *-ment*, substantifs, verbes) dans la copie du scribe Guiot (ms. BNF fr. 794), un des meilleurs manuscrits et surtout le seul qui présente le texte de tous les romans. Le DÉCT2 se caractérisera par un enrichissement considérable par rapport à la première version; il s'agit essentiellement de l'ajout des mots grammaticaux (quelque 553 lemmes, si l'on inclut les locutions, d'après les listes établies par M.-L. Ollier), des variantes des autres manuscrits (variations lexicales; variations également dans la collocation des

Florilegium, vol. 24 (2007): 157-61.

termes – ces variations sont nombreuses, mais impossibles à chiffrer pour l'instant), du classement sémantique des mots relevés, des axes de synonymie et d'antonymie, peut-être aussi d'holonymie et de méronymie.

La publication se fera également en plusieurs phases. La première version du DÉCT1 a été lancée officiellement à Nancy le 31 mai 2007. Elle comprend, pour le volet lexique, l'intégralité de la lettre A et, pour la base textuelle, la transcription des romans d'après le ms. BNF fr. 794, accompagnée de la lemmatisation de tous les mots occurrents et de leur indexation grammaticale. La base de texte est complète et utilisable dans sa totalité; seuls de légers changements y seront plus tard apportés, de l'ordre de la correction d'erreurs ou de la désambiguïsation (quand dans un même vers apparaissent deux formes homographes mais relevant de catégories différentes). Par souci d'homogénéité, on a limité la base aux cinq romans susmentionnés et écarté le reste de l'oeuvre de Chrétien, à savoir deux chansons courtoises d'attribution certaine (la première figure dans deux manuscrits, la seconde dans douze); puis un poème d'attribution contestée, *Philomena*, qu'on peut lire dans une vingtaine de manuscrits de l'*Ovide Moralisé*; enfin un roman d'attribution fort douteuse, *Guillaume d'Angleterre*, conservé dans deux manuscrits. Il est à remarquer que ces oeuvres exclues ne se trouvent dans aucun des manuscrits des cinq romans.

Les transcriptions ont été effectuées au LFA par le responsable du projet. On peut les considérer comme semi-diplomatiques dans la mesure où la segmentation des mots a été modernisée et la ponctuation ajoutée, ce qui constitue déjà une première interprétation critique du document. Une seconde intervention critique a consisté à signaler par un astérisque ce qui résulte d'une bévue du copiste : quand le signe en forme d'étoile suit immédiatement les lettres d'un mot, la faute porte sur ce mot; précédé et suivi d'une espace, l'astérisque signale toute faute d'un autre type, le plus souvent une omission; placé à la fin d'un vers, il indique que l'erreur concerne plusieurs éléments de cette ligne. L'usage du tréma reste discret; il s'agit essentiellement de lever une ambiguïté : par exemple entre *oie* (subjonctif présent 3 de *oïr* – 26 occurrences) et *oïe* (substantif féminin *oïe*, 2 occurrences, ou participe passé de *oïr*, 18 occurrences).

On accède à la base de deux façons : par une lecture des textes en continu ou par une recherche directe dans les textes. L'utilisateur qui choisit de lire les textes d'une manière continue sélectionne le texte désiré en cochant sur le bouton correspondant; il peut alors accéder au début du texte ou commencer sa lecture à un folio ou à un vers particulier. Un clic sur l'icone appropriée permet d'obtenir pour chaque vers les lemmes et les catégories grammaticales des mots qui le composent; cliquant alors sur le lemme qui l'intéresse, l'utilisateur voit apparaître l'article correspondant du lexique. Le principe

fondamental du DÉCT est, en effet, la navigation libre, complète et systématique entre lexique et base textuelle.

Parallèlement à la lecture du roman en mode texte, on peut avoir accès, pour *Lancelot* et *Yvain*, à l'image du folio du manuscrit. Ce qui permet de vérifier, dans certains cas douteux, l'exactitude de la transcription et offre la possibilité aux débutants ou aux amateurs de s'initier ainsi à la paléographie. L'accès aux folios se fait par liens URL avec le site du Projet Charrette à Princeton et avec la page du Projet Chevalier au Lion sur le site du LFA. Les folios des trois autres romans sont en voie de numérisation et devraient être disponibles pour la prochaine phase de publication.

L'utilisateur qui, plutôt que de faire défiler le texte d'un des romans, préfère interroger la base en formulant des requêtes précises, doit sélectionner la fonction « Recherche directe dans les textes ». Le programme lui demande alors de définir le corpus de travail (un ou plusieurs romans, ou bien l'ensemble des textes) et lui offre une option entre quatre types de recherche : sur les lemmes, sur les formes graphiques occurrentes, sur les catégories grammaticales et sur des cooccurrences de deux mots.

Pour le premier type de requête, il convient de préciser que les lemmes retenus dans le DÉCT sont ceux de l'*Altfranzösisches Wörterbuch* de Tobler-Lommatzsch (désormais TL) et, plus spécialement, de sa version électronique parue sous forme de DVD en 2002, à laquelle sont empruntés les indices numériques de désambiguïsation (*tor1*, *tor2*, *tor5*, par exemple). Il arrive exceptionnellement que le DÉCT s'écarte de ce répertoire de lemmes; ainsi, comme le *Dictionnaire du Moyen Français*, il distingue entre *afoler 1* (étymon : *fullare*) et *afoler2* (étymon : *follis*), ou entre *assomer1* (étymon : *somnus*) et *assomer2* (étymon : *summa*), alors que TL les confond. Pour des raisons de commodité, les adverbes de manière se terminant par le suffixe *–ment*, qui sont traités, dans TL, sous l'entrée du mot dont ils sont dérivés, sont, au contraire, extraits dans le DÉCT, qui leur confère le plein statut de lemme. Mais c'est essentiellement avec les lemmes grammaticaux que ce dictionnaire prend quelque distance par rapport au *Tobler-Lommatzsch*; le terrain est piégé : le choix du lemme est parfois discutable, l'analyse grammaticale souvent contestable. C'est en particulier le cas de certains déterminants du substantif, pour lesquels des corrections ont été nécessaires.

Interrogeant sur les lemmes, l'utilisateur se voit proposer une liste de lemmes par tranches alphabétiques ainsi qu'un filtre sur les lemmes. Ce dernier lui permet, entre autres choses, de chercher, en cliquant sur le bouton « expression régulière », toute suite de caractères; par exemple, en tapant « .*eor$ », il obtient les 36 lemmes à suffixe d'agent, d'*amëor* à *venëor*. Une fois le lemme sélectionné, il peut cliquer sur le bouton « Rechercher dans les textes »; apparaissent alors toutes les occurrences de ce lemme

dans le corpus défini, chacune dans un contexte qui correspond au cadre de la phrase. Mais l'utilisateur peut aussi cliquer sur le bouton « Structure de l'article » et accéder, par ce biais, aux détails de l'article du lexique. Il peut également cliquer sur « Afficher les formes » et les faire ainsi apparaître; pour le lemme *aler*, le résultat sera un total de 56 formes graphiques différentes et de 1102 occurrences de celles-ci.

Si, plutôt que d'opérer une recherche sur les lemmes, l'utilisateur désire formuler une requête portant sur les formes, il cliquera sur le bouton « Recherche sur les formes ». Il obtiendra, comme dans le cas précédent, une liste de formes par tranches alphabétiques ainsi qu'un filtre sur les formes. Sélectionnant par exemple la forme *afole*, il trouvera en réponse que cette forme est citée dans le lexique comme indicatif présent 3 et subjonctif présent 3 de *afoler1*, et comme subjonctif présent 3 de *afoler2*. Un clic sur chacun de ces lemmes permet de parvenir à l'article correspondant du lexique. L'utilisateur peut également rechercher les attestations de la forme ou des deux lemmes dans les textes (8 occurrences dans le premier cas, respectivement 3 et 5 occurrences dans le second).

Pour des requêtes sur des catégories grammaticales, il faut cliquer sur le bouton « Attestation d'un mot », puis sur « Code ». Si l'on retient la catégorie « adjectif », on devrait obtenir toutes les formes étiquetées comme telles lors de l'encodage des textes; en fait, la version actuelle limite l'accès aux 200 premières réponses. Celles-ci nécessitent d'ailleurs un tri, car y sont également inclus tous les lemmes marqués « adjectif » mais possédant aussi d'autres étiquettes (adverbe, substantif, pronom) : ainsi *tant* est étiqueté « adj/adv/s » et *tot* « adj/pron/adv ». Il est possible (et souhaitable) que, dans une nouvelle version de la base, on arrive à préciser l'étiquetage en fonction du contexte et à élargir l'éventail des catégories grammaticales sur lesquelles portent les requêtes, sans nuire au bon fonctionnement de l'ensemble.

Enfin, si la requête porte sur une cooccurrence de mots, l'utilisateur dispose d'un formulaire comportant d'abord deux cellules (une par mot), avec, dans chacune, un choix entre lemme, forme et code; suit une autre cellule avec des options concernant l'ordre des mots. Par exemple, un linguiste s'intéressant à la place de l'adjectif épithète dans *Yvain* peut pour le mot1 cocher la case « lemme » et sélectionner « grant », pour le mot2 cocher « code grammatical » et sélectionner « substantif féminin », puis cocher « mot1 et mot2 contigus » et « mot1 avant mot2 », il verra en réponse les 85 passages d'*Yvain* où cet adjectif est antéposé à un nom féminin; inversément, si la requête porte sur la postposition de l'épithète, on notera que cette construction n'est pas attestée dans le roman (la seule occurrence chez Chrétien est au vers 5050 de *Lancelot* : *rote grant*, probablement à cause de la rime).

Quelle que soit la façon dont on interroge la base textuelle, il faut souligner qu'on peut toujours passer, par un ou deux clics, des lemmes aux formes, des mots aux

contextes, des contextes aux textes intégraux, et de chacun de ces éléments aux sections du lexique qui y correspondent. Le logiciel est conçu de façon à assurer la navigation la plus souple entre ces différentes composantes. Cette souplesse interne se retrouve à l'extérieur, car le DÉCT est une base ouverte, qui donne et qui reçoit; d'abord, et avant tout, dans le cadre des diverses ressources linguistiques informatisées du site ATILF : le lexique se situe dans la filiation du *Dictionnaire du Moyen Français*, c'est-à-dire dans la lignée du *Trésor de la Langue Française*; la base textuelle entretient des rapports importants avec la Base textuelle du Moyen Français et avec les divers programmes conçus par G. Souvay, sur ce site, pour l'étude de l'ancienne langue : des innovations faites d'un côté sont souvent répercutées de l'autre. La même osmose devrait se produire avec la parution prochaine sur le site ATILF du Nouveau Corpus d'Amsterdam, récemment publié et accessible sur CD-Rom.

Université d'Ottawa

Bibliographie

ATILF/Équipe « Moyen français et français préclassique », 2003-2005, *Dictionnaire du Moyen Français (DMF). Base de Lexiques de Moyen Français (DMF1)*. www.atilf.fr/blmf.

Chevalier au Lion, projet du Laboratoire de Français Ancien, www.uottawa.ca/academic/arts/lfa/activites/textes/chevalier-au-lion/chlpresduprojet.html.

Kunstmann P., H. Gerner et G. Souvay. *Dictionnaire Électronique de Chrétien de Troyes (DÉCT)*, www.atilf.fr/dect/.

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

Ollier M.-L. *Lexique et concordance de Chrétien de Troyes d'après la copie Guiot*. Montréal : Institut d'Études Médiévales, Paris : Vrin, 1986.

The Princeton Charrette Project, www.princeton.edu/~lancelot/ss/.

Tobler-Lommatzsch : Altfranzösisches Wörterbuch, édition électronique conçue et réalisée par Peter Blumenthal et A. Stein (DVD et guide). Stuttgart : Steiner, 2002.

Trésor de la Langue Française Informatisé, ATILF / CNRS, www.atilf.atilf.fr/tlf.htm.