DÉCT (*Dictionnaire Électronique de Chrétien de Troyes*): Model for Today's Lexicography?

Gilles Souvay
Centre National de la Recherche Scientifique et Université de Nancy

Pierre Kunstmann
Université d'Ottawa

The DECT is an example of today's lexicographic practice. Its realization is completely computerized from the input to the on-lining. It calls on modern concepts of data encoding (XML) and diffusion-free access on the Web. The DECT is not just a dictionary searchable from the entries. It is in fact a real lexicographic tool made up of an annotated textual base-lemma and part of speech-with the manuscript's image, and the lexicon resulting from the texts analysis. It can be consulted in a traditional way-display of a page, of a verse, of an article...—or through specialized search forms, for instance, it is possible to look for co—occurring words in the texts-lemma aimer before an adverb, or to make a multi-criteria query in the lexicon-search for a word in a verb's definition. Moreover, it is always possible for the user to go from the lexicon to the texts and vice versa. The on-line base can be accessed at http://www.atilf.fr/dect *(French and English). The DECT's computerized component is built on a platform developed at the ATILF for historical linguistics projects. The same tools allow the consultation of other lexicographic projects, about ten instancings. The DECT contributed, for a large part, to the platform development and constitutes, for it, the most successful instancing.*

1. Le corpus

1.1. *Aspects historiques*

Ce projet s'inscrit dans la ligne des travaux électroniques menés par divers groupes de chercheurs (Montréal, Ottawa, Princeton, Poitiers) sur l'œuvre de Chrétien de Troyes. Il s'agit d'établir et de mettre sur le web un dictionnaire électronique des cinq romans de Chrétien de Troyes (*Érec* 1170, *Cligès* 1176, *Lancelot* 1177, *Yvain* 1177, *Perceval* 1182). L'auteur est considéré comme un classique de la littérature française médiévale. Les textes sont des transcriptions du manuscrit BNF fr. 794, la célèbre copie du scribe Guiot.

1.2. *Étiquetage des textes*

Les textes ont été traités automatiquement à partir d'une première lemmatisation effectuée manuellement en utilisant notamment les listes établies par M.-L. Ollier (1986), chaque mot s'est vu attribuer un lemme (entrée de la version électronique de l'*Altfranzösisches Wörterbuch* d'A. Tobler et E. Lommatzsch) et un code grammatical (partie du discours). L'outil utilisé pour l'étiquetage est un programme développé à l'Université de Stuttgart. Les résultats ont été vérifiés et corrigés, si besoin était; le processus de révision continue, d'ailleurs, au fur et à mesure de la rédaction des différentes lettres du dictionnaire. À la fin du projet les textes seront téléchargeables librement au format TEI.

Exemple 1: annotation du premier vers de Yvain

1. Artus_<Artu-sp>, li_<le-art> boens_<bon-adj> rois_<roi2-sm> de_<de-prép> Bretaingne_<Bretagne-sp>,

1.3. *Rédaction des articles*

Les articles du DÉCT s'appuient sur la structure des articles du DMF (*Dictionnaire du Moyen Français*) enrichie d'informations statistiques et d'informations sur la flexion:

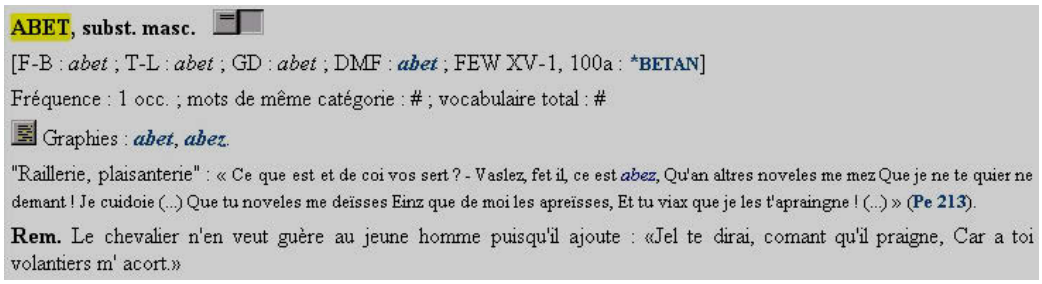
- mot vedette, ou entrée, suivi de son code grammatical,
- suite de références à d'autres dictionnaires (papier ou électroniques),
- fréquence de l'entrée dans le corpus,
- graphies: formes fléchies rencontrées dans le corpus,
- article proprement dit, hiérarchisé et découpé en paragraphes (chaque paragraphe pouvant contenir des indications métalinguistiques, une définition et être illustré d'un ou plusieurs exemples.

Les informations sont codées en XML et les articles rédigés à l'aide d'un éditeur de texte balisé.

Exemple 2: le balisage de l'article ABET

<ART><VED>ABET</VED><CODE><CodeSubstantifMasculin/></CODE><DICT><FB><LEMME>abet</LEMME></FB><TL><LEMME>abet</LEMME></TL><GD><LEMME>abet</LEMME></GD><DMF><LEMME>abet</LEMME></DMF><FEW.CONNU><VOLUME>XV-1</VOLUME><PAGE>100a</PAGE><ETYM>*betan</ETYM></FEW.CONNU></DICT><Frequences><FOcc>1</FOcc><FMMC><FVT/></Frequences><Graphies><Graphie>abet</Graphie><Graphie>abez</Graphie></Graphies><FormesGraphiques><ACompleter/></FormesGraphiques><P><DISC><DEF>Raillerie, plaisanterie@Mocking, joking</DEF></DISC><EXE><TEXTE>« Ce que est et de coi vos sert ? - Vaslez, fet il, ce est <OCC>abez</OCC>, Qu'an altres noveles me mez Que je ne te quier ne demant ! Je cuidoie (...) Que tu noveles me deïsses Einz que de moi les apreïsses, Et tu viax que je les t'apraingne !(...) »</TEXTE><REF><RefTitre>Pe</RefTitre><RefOu>213</RefOu></REF></EXE></P></ART>

Exemple 3: l'article ABET dans la base en ligne



2. La plate-forme de développement

2.1. *Présentation*

Le DÉCT pour ses aspects informatiques, s'appuie fortement sur les travaux en lexicographie réalisés à l'ATILF. En particulier il utilise les outils développés pour l'équipe Linguistique Historique Française et Romane. Il a, en retour, contribué à leur enrichissement (amélioration des interactions, ajouts de nouvelles fonctionnalités…).

La plate-forme de développement repose sur une architecture client serveur développée sur le web à l'aide de scripts selon le protocole CGI (Common Gateway Interface). Elle comporte quatre composantes: les outils, la composante textuelle, la composante lexicale et le lemmatiseur. Selon la nature du projet, certaines de ces composantes peuvent être activées ou non.

Lorsque les composantes textuelles et lexicales sont activées conjointement, il est alors possible de passer simplement des textes vers le lexique et du lexique vers les textes.

2.2. *Les outils*

Il s'agit des outils utiles aux rédacteurs pour les aider à rédiger leurs articles:

- l'outil de saisie des données balisées XML est un logiciel du commerce que l'on configure à l'aide d'une DTD et d'une feuille de style;
- un outil de contrôle accessible permet de vérifier la cohérence des données: vérifier par exemple la ponctuation, l'existence d'une entrée d'un dictionnaire que l'on cite…;

- un outil de conversion permet de transformer le texte XML en document RTF ;
- un outil d'administration permet la mise à jour interactive de la base par le responsable du projet.

2.3. *La composante textuelle*

Elle s'appuie sur une structuration des données en mots. Un mot est un triplet: forme, lemme et code grammatical. Les données dont codées en XML, la DTD étant imposée.

Exemple 4: extrait du codage XML du texte pour l'exploitation par la plate-forme (on n'est pas, à ce niveau, en codage TEI)

<phrase>
<idp>1</idp>
<page>79d</page><cp><lb>_</lb><nv>1</nv><mot><id>1</id><forme>Artus</forme><x>_</x><lemme>Artu</lemme><x>_</x><code>sp</code></mot>
, <e>_</e><mot><id>2</id><forme>li</forme><x>_</x><lemme>le</lemme><x>_</x><code>art</code></mot>
<e>_</e><mot><id>3</id><forme>boens</forme><x>_</x><lemme>bon</lemme><x>_</x><code>adj</code></mot>
<e>_</e><mot><id>4</id><forme>rois</forme><x>_</x><lemme>roi2</lemme><x>_</x><code>sm</code></mot>
<e>_</e><mot><id>5</id><forme>de</forme><x>_</x><lemme>de</lemme><x>_</x><code>prép</code></mot>
<e>_</e><mot><id>6</id><forme>Bretaingne</forme><x>_</x><lemme>Bretagne</lemme><x>_</x><code>sp</code></mot></phrase>

Le moteur de recherche qui exploite les textes est STELLa, développé à l'ATILF et qui a fait ses preuves pour le TLFi (*Trésor de la Langue Française informatisé*) et la base textuelle FRANTEXT.

La composante textuelle fournit des formulaires prédéfinis qui permettent d'accéder au texte en continu, de se constituer un corpus de travail, de rechercher des mots, de faire des cooccurrences de deux mots.

2.4. *La composante lexicale*

Elle s'appuie sur une structuration des données en XML. La DTD est un paramètre de la composante, la seule contrainte étant l'existence d'un identifiant unique pour chaque article (en général l'entrée).

Comme pour la composante textuelle, elle fournit des interfaces standards pour interroger les données: liste des entrées, interrogation plein texte, interrogation du contenu d'une balise et combinaison multicritère. Lors de la visualisation, on peut choisir d'afficher l'article complet ou simplement sa structure.

Deux moteurs de recherches coexistent dans cette composante:

- le moteur STELLa destiné aux applications volumineuses qui nécessitent une optimisation des requêtes. La phase de montage des données est lourde et réalisée par un informaticien.
- un moteur simplifié qui est déployé pour des applications de "petite taille", qui permet une administration complète de la base par le responsable du projet.

2.5. *Le lemmatiseur*

L'orthographe médiévale n'est pas encore stabilisée, et lors de la consultation d'un dictionnaire de cette époque, il n'est pas toujours facile de savoir sous quelle entrée a été traité le mot recherché. Le lemmatiseur est la composante qui va permettre de gommer cette grande variation graphique et de faire des propositions à l'utilisateur du dictionnaire. Il s'agit du lemmatiseur LGeRM développé à l'ATILF pour le DMF.

3. La base en ligne

Le DÉCT est une base en accès libre sur le web (à l'adresse *http://www.atilf.fr/dect*) ouverte depuis mai 2006. La composante textuelle est terminée et complètement opérationnelle avec les cinq textes annotés. La composante lexicale ne comporte à cette date (mars 2008) que les lettres A et B.

Fonctionnalités principales du DÉCT qui seront présentées pour la démonstration:

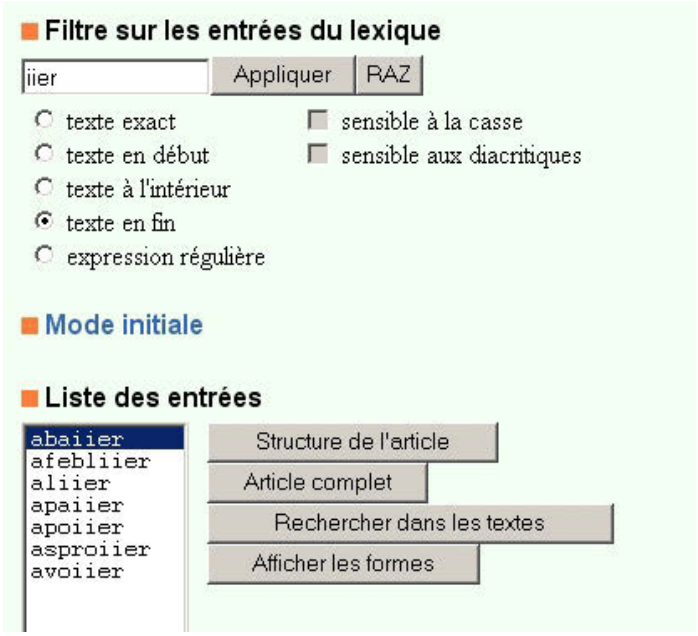
- Texte en continu: consultation folio par folio, vers par vers, des textes avec accès à l'image du manuscrit. La transcription à partir du manuscrit est de type semi-diplomatique, avec segmentation des mots et ponctuation modernes. On peut accéder aux lemmes et codes grammaticaux et par un simple clic accéder à l'article du dictionnaire correspondant au lemme.
- Recherche dans le lexique
  - Recherche sur les entrées: le mode classique d'entrée dans un dictionnaire, avec en plus possibilité de travailler sur l'initiale du mot, sa finale, aide du lemmatiseur… Une fois l'entrée sélectionnée, affichage de la structure de l'article, affichage de l'article complet, affichage de toutes les attestations de l'entrée dans le corpus de textes.
  - Recherche plein texte: recherche d'un mot sans tenir compte du balisage des informations.
  - Recherche avancée: choisir l'élément à interroger avec possibilité de combiner plusieurs critères
- Recherche dans les textes
  - Recherche sur les lemmes, les formes: afficher les vers contenant un lemme ou une forme donnée. Regarder la définition d'un mot dans le lexique.
  - Cooccurrences de deux mots: chercher des mots en cooccurrence en définissant leur ordre, leur distance. On peut travailler sur la forme, le lemme ou le code grammatical.

Exemple 5: affichage du texte et des annotations des premiers vers d'Yvain.

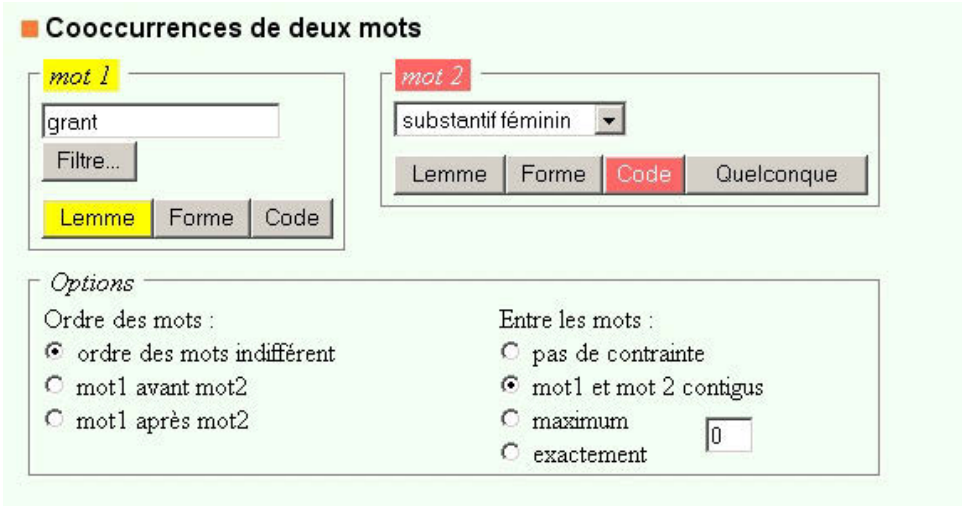
Yv Chrétien de Troyes *Yvain* ou *Le Chevalier au Lion*, 1177, 79d

1	Artus **Artu** sp	li **le** art	boens **bon** adj	rois **roi2** sm	de **de** prép	Bretaingne **Bretagne** sp
2	La **le** art	cui **que2/3** pron/adv/rel/inter	proesce **pröece** sf	nos **nos1** pronpers	enseigne **enseignier** v	

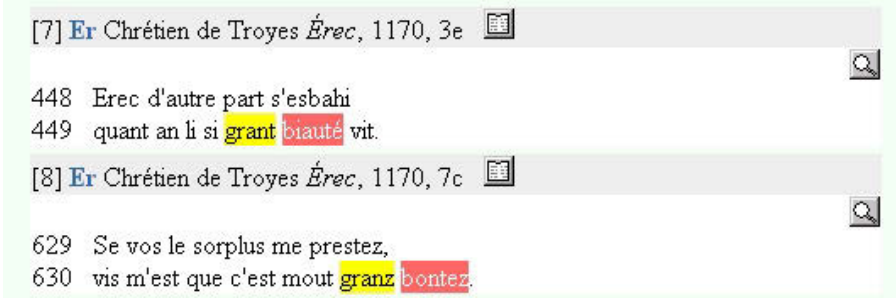
Exemple 6: recherche des entrées se terminant par *iier*.



Exemple 7a: le formulaire de recherche de cooccurrence entre le lemme *grant* et un substantif féminin sur l'ensemble des textes:



Exemple 7b: le résultat de la recherche



Nous proposerons aussi au cours de la démonstration les outils ayant servi à la rédaction du dictionnaire.

Le DÉCT reçoit le soutien financier du Conseil de Recherches en Sciences Humaines du Canada.

References

Articles et livres cités

Dendien, J. (2002). “STELLA et ses fonctionnalités”. Dans *Congrès international de Rouen, L’édition électronique en littérature et dictionnairique: évaluation et bilan*.

Kunstmann, P.; Gerner, H.; Souvay, G. (à paraître). “Dictionnaire électronique de Chrétien de Troyes: état actuel du projet”. Dans Iliescu, M. *et al*. (éds.). *Actes du XXVe Congrès International de Linguistique et de Philologie Romanes (Innsbruck, 3-8 septembre 2007)*. Tübingen: Niemeyer.

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

Souvay, G. (2007). “LGeRM: un outil de lemmatisation du moyen français”. Dans Trotter, D. (éd.). *Actes du XXIV^e Congrès International de Linguistique et de Philologie Romanes*. Vol. I. Tübingen: Niemeyer. 457-466.

Ressources électroniques citées

Blumenthal, P.; Stein, A. (2002). *Tobler-Lommatzsch: Altfranzösisches Wörterbuch* [4 CD-Roms et DVD]. Stuttgart: Steiner.

[DMF]. *Dictionnaire du Moyen Français* [en ligne]. ATILF / Nancy Université - CNRS. *http://www.atilf.fr/dmf*.

FRANTEXT [en ligne]. *http://www.frantext.fr/*.

[TEI]. *Text Encoding Initiative* [en ligne]. *http://www.tei-c.org/*.

[TLFi]. Trésor le la Langue Française informatisé [en ligne]. *http://www.frantext.fr/*.