« L'évolution de Frantext : quelles modifications et quels usages pour Frantext 2 ? »

Proposition de poster, soumise dans la section 16, « **Projets en cours ; ressources et outils nouveaux** », du **XXVIIe Congrès International de Linguistique et de Philologie Romanes**, Nancy 2013

par Véronique Montémont et Lucia Manea

Le poster sera structuré autour de la base de données textuelles en français, Frantext, développée et maintenue à l'ATILF. Nous présenterons les transformations envisagées de la base, qui visent à la création d'un outil plus souple et plus performant. Tout d'abord, nous exposerons l'historique de Frantext, sa conception et son fonctionnement, pour pointer ensuite les objectifs et les usages attendus de Frantext 2, qui seront exploitables en linguistique comme dans d'autres domaines.

Frantext actuel est un corpus d'à peu près 4400 textes français, du Moyen Âge au XXIe siècle, associé à un moteur de recherche, Stella, qui permet d'interroger un corpus défini préalablement selon certains critères. Il contient à peu près 90 % d'œuvres littéraires et 10 % d'ouvrages techniques ; la moitié couvre la période 1900-2012. Initialement conçu pour constituer une base d'exemples qui allaient alimenter le *Trésor de la langue française*, il est devenu une base de données accessible au public à partir de 1988. À l'époque de sa création, dans les années 70, Frantext était pionnier : l'idée de construire un dictionnaire à partir d'un ensemble de textes ayant subi un traitement automatisé représentait une révolution dans le paysage lexicographique. Il existe une version catégorisée (à peu près 1940 textes) dans laquelle un étiquetage grammatical a été effectué, ce qui permet de faire des recherches syntaxiques. Une version libre de droits (500 œuvres), disponible sur le site du CNRTL, autorise l'accès aux textes complets. L'intérêt permanent que présente Frantext consiste en la combinaison d'un grand corpus et d'un moteur de recherche.

La nécessité d'améliorer cette base, déjà singulière si on la compare à Gallica ou d'autres corpus francophones (Bibliothèque de Lisieux), est apparue. Les modifications proposées tiennent aux usages et aux objectifs qui se sont précisés, notamment aux corpus et aux procédés d'interrogation. En comparaison des grands corpus qui couvrent des périodes, des collections et des bibliothèques entières, Frantext peut être vu comme une petite bibliothèque dont le contenu est sélectionné soigneusement pour servir à des projets de recherche portant sur des corpus d'auteurs précis ou sur des problématiques théoriques. Dans Frantext 2, un supplément de renseignements sur les auteurs et les œuvres sera fourni. Dans ce sens, les textes seront équipés de « métadonnées » référencées et précises. Ces métadonnées permettront de retrouver des renseignements sur les éditions utilisées, sur l'édition originale et l'édition définitive, les genres littéraires, les thématiques abordées par les textes. Grâce à Frantext 2, les utilisateurs pourront développer des recherches en fonction de critères multiples (simples ou combinés : dates biographiques de l'auteur, dates de parution des œuvres ; références bibliographiques du type : éditeur scientifique, éditeur commercial, collection, prix, paratexte, publications anthumes ou posthumes, date de création pour le théâtre).

Frantext 2 sera donc une base de données équipée d'un moteur de recherche perfectionné, assortie d'un portail bibliographique, et qui autorisera des enquêtes sur de grands corpus échantillonnés. Cette nouvelle version de Frantext souhaite offrir une interface de consultation la plus claire possible, à tous les niveaux : constitution des corpus de travail, affichage, ergonomie des fonctions de recherche. Permettant de faire des interrogations sur les états anciens de langue (ajout d'une flexion médiévale et d'une flexion XVIIe), Frantext 2 sera utile dans l'établissement d'éditions critiques, philologiques, et le chercheur pourra citer en s'appuyant sur Frantext, vérifier le lexique, les cooccurrences… Se présentera aussi la possibilité

de faire des interrogations sur les états de langue d'un même texte, en comparant plusieurs versions, éditées à des époques différentes. Les questions d'étude des diverses variétés du français seront abordables dans des corpus qui toucheront des communautés linguistiques extérieures à la France : Suisse, Belgique, Maghreb, Afrique subsaharienne, Québec. L'enrichissement des corpus d'écrits personnels conduira à des études diachronique et synchronique de l'usage du français dans certains milieux ou encore à l'étude de l'oralité dans les textes écrits. D'autres objectifs tournent autour des possibilités d'améliorer les études quantitatives et statistiques sur le lexique général, certaines isotopies, ou encore des lexiques d'auteur, par l'implémentation d'outils statistiques.

La question de la transférabilité de l'outil peut se poser. Frantext 2 pourrait constituer un modèle pour des bases de textes littéraires similaires dans d'autres langues romanes. Dans ce sens, nous pourrions faire une comparaison avec Hyperbase latin, base consacrée au latin et aux textes de l'Antiquité classique.

L'intérêt principal que peut représenter Frantext 2 est tout d'abord d'ordre linguistique et littéraire. Il permettra ainsi la conservation, la valorisation, l'étude et la connaissance d'un état de la langue ; l'étude des genres du discours (via les mots-clés : oralité, discours familier, langage poétique…) ; l'étude des genres littéraires (en interrogeant des types particuliers de genres littéraires : poèmes en prose, tragi-comédie, drame…) ou encore l'interaction entre plusieurs genres, les genres hybrides qui naissent au contact de plusieurs genres. Si la base est intrinsèquement conçue pour servir de réservoir d'informations aux approches littéraire, lexicale, linguistique, stylistique, sémiotique, génétique, elle pourra ouvrir vers d'autres approches d'ordre sociologique, historique, bibliophile, et même juridique ou médicale.

Bibliographie

BÉHAR, Henri, *La Littérature et son golem*, Paris, Champion, 1996.

BRUNET, Étienne, *Méthodes quantitatives et informatiques dans l'étude des textes : colloque international, Université de Nice, 5-8 juin 1985, en hommage à Charles Muller*, Genève, Slatkine ; Paris, Champion, 1986.

LEMERCIER, Claire et Claire ZALC, *Méthodes quantitatives pour l'historien*, Paris, La Découverte, coll. Repères, 2008.

MONTÉMONT, Véronique, « Comment bâtir une ressource lexicale : l'exemple de Frantext », *in* Ioan Negrutiu, Jean-Paul Bravard et *al.* (éd.), *Les Ressources* (Journées Scientifiques de l'Institut Universitaire de France, Lyon, 30-31 mai 2011), Presses Universitaires de Saint-Étienne, 2012.

MONTÉMONT, Véronique, « How To Explore A Digitalized Autobiographical Corpus : The Case Of Frantext », « 3rd Global Conference "Digital Memories" », coordonnée par Daniel Riha, Prague, 16-18 mars 2011, *à paraître* (http : //www.inter-disciplinary.net/wp-contnt/uploads/2011/02/montemontdmpaper.pdf).

MONTÉMONT, Véronique, « Discovering Frantext », *in* Jan Auracher & Willie van Peer (dir.), *New Beginnings in Literary Studies*, Newcastle, Cambridge Scholars Publishing, 2008, p. 89-107.

MONTÉMONT, Véronique et Pascale BERNARD, « Voyage au cœur du langage : le *Trésor de la langue française* et *Frantext* », *Culture et Recherche*, n°124, hiver 2010-2011, p. 34-35.

NAUMANN, Bastian, *Der Textcorpus Frantext*, Studien arbeit, GRIN Verlag, 2004.

RASTIER, François, *Arts et science du texte*, Paris, PUF, 2001.

PIERREL, Jean-Marie, « Un ensemble de ressources de référence pour l'étude du français : TLFi, Frantext et le logiciel Stella », *Revue québécoise de linguistique*, t. 32, 2003, p. 155-176.

PIERREL, Jean-Marie et Éva BUCHI, « Research and Resource Enhancement in French Lexicography : the ATILF Laboratory Computerised Resources », *in* Silvia Bruti, Roberta Cella et Marina Foschi Albert (éd.), *Perspectives on Lexicography in Italy and Europe*, Newcastle-upon-Tyne, Cambridge Scholars Publishing, 2009, p. 79-117.