PRINCIPES DE SELECTION DES TEXTES DU CORPUS CORPTEF

Coordonnés par
Céline Guillot (Celine.Guillot@ens-lsh.fr)
Alexei Lavrentiev (Alexei.Lavrentev@ens-lsh.fr)

CNRS / Université de Lyon (ENS-LSH), UMR 5191 ICAR
Document élaboré dans le cadre du projet CoRPTeF financé par l'ANR

VERSION 1

Table de mise à jour du document

- 20 octobre 2008, rédaction de la 1ère version

Objectifs du corpus :

Rendre accessible et exploiter un ensemble de textes antérieurs au XIIIe siècle dans le cadre de recherches linguistiques. Ces recherches visent à **décrire le « très ancien français »**, à la fois en cherchant le plus haut niveau de **généralité** possible et en même temps en rendant compte de la **variation des usages**.

Les différents usages du français de cette période doivent donc être représentés dans le corpus et ils doivent être décrits de façon explicite dans la documentation para-textuelle (cf. notre document sur les descripteurs des textes).

Ce corpus est destiné à être exploité dans le cadre de recherches aussi bien quantitatives que qualitatives. Il importe donc de veiller à un **équilibre relatif des usages représentés**. On ne cherchera pas à calquer cet équilibre sur ce que nous savons de la production documentaire de cette période, mais on recherchera plutôt l'équilibre quantitatif des usages à l'intérieur du corpus lui-même.

Cet équilibre sera nécessairement relatif, compte-tenu du nombre de documents médiévaux qui nous sont parvenus pour chaque type (aléas de leur transmission, déséquilibres de départ dans la production écrite en très ancien français).

Les critères de sélection des textes ne rendent par ailleurs pas compte de tous les paramètres de variation possible. Pour certains d'entre eux, ils correspondent aux paramètres qui ont été définis comme particulièrement pertinents pour l'étude de la variation dans le cadre de ce projet. En outre, ces critères ne doivent pas correspondre à des catégories trop 'fines', pour lesquelles on ne pourrait pas trouver suffisamment de témoins ou pour lesquelles les informations dont on dispose sont beaucoup trop souvent lacunaires.

D'autre part, certains critères de sélection sont sans rapport avec la variation linguistique mais **garantissent plutôt l'exploitation future des données** : qualité/fiabilité de l'édition critique, taille du texte (relativement à l'exploitation qui en sera faite), disponibilité du texte.

Liste des critères de sélection des textes du corpus

Impor-tance	Critères liés à la variation	Critères sans lien avec la variation
[1]	• Variation diachronique : o date de composition du texte ; o date de composition du manuscrit • Variation diaphasique: o forme (vers VS prose),	• Qualité / fiabilité de l'édition critique
[2]	o genres et domaines textuels • Variation diatopique : dialecte	
[3]		• Taille du texte
[4]		• Accessibilité et Droits

Principes d'application des critères de sélection

Ces différents critères sont de niveau d'importance différent quant à la sélection des textes.

▪ On distingue les **critères d'importance 1**, qui sont discriminants dès le premier niveau de sélection (l'ordre est non pertinent, ils sont d'égale importance) :

Date de composition du texte
Nécessairement inférieure à 1200.

Date du manuscrit
On privilégie toujours les mss les plus proches de la date de composition présumée de l'œuvre.
On privilégie les mss antérieurs au XIIIe siècle.
On admet cependant des mss plus récents et/ou plus éloignés de la date de composition pour les types textes qui sont représentés par moins de témoins en général (c'est-à-dire dans la pratique, à peu près exclusivement pour les textes en prose).
Les manuscrits postérieurs à 1250 sont exclus.

Forme
On privilégie la prose en raison de sa rareté dans le total de la production écrite et de la volonté d'équilibrage du corpus.

Qualité/fiabilité de l'édition
On choisit la meilleure édition quand il y en a plusieurs, et on exclut celles qui sont vraiment reconnues comme étant « interventionnistes ».

▪ Les critères de niveau d'**importance 2** interviennent dans un second temps :

Domaine* et *genre textuels
On privilégie la diversité.
Priorité notamment aux textes qui ne relèvent pas des domaines littéraire et religieux/

Dialecte
On privilégie la diversité.
Les textes anglo-normands étant les plus nombreux, on choisit en priorité les autres dialectes.

Le critère de niveau d'**importance 3** intervient ensuite :

Taille du texte
On préfère les textes relativement courts, pour offrir un plus grand choix de textes et pour favoriser la diversité typologique. On évite en même temps les textes « trop » courts (par exemple, les gloses), à partir desquels on pourrait tirer trop peu de conclusions lors de l'exploitation des données.

Deux conditions supplémentaires, sans rapport avec l'intérêt du texte pour le corpus et la qualité de l'édition (**importance 4**), doivent néanmoins être remplies pour qu'un texte puisse être intégré au corpus :

- Nous devons être en mesure d'**accéder à l'édition source** (l'acquérir si possible ou l'emprunter dans une bibliothèque)
- Nous devons obtenir l'**autorisation de numériser** de la part des ayants droit (l'éditeur scientifique et la maison d'édition) pour les textes qui ne sont pas libres de droit

CoRPTeF et BFM

Le corpus CoRPTeF bénéficie de l'apport du corpus de la Base de Français Médiéval (BFM). Les textes de la BFM qui répondent aux critères de niveau 1 sont sélectionnés par défaut.

Pour la numérisation dans le cadre du projet CoRPTeF la priorité est donnée aux types de textes absents ou sous-représentés dans la BFM.