Base de français médiéval et recherches linguistiques diachroniques

Céline Guillot
Alexei Lavrentiev

Atelier Cahier Edition analytique
ENS Lyon, 30 juin – 4 juillet

(cc) BY-NC

Plan

- Historique et présentation de la BFM
- Enrichissement de la BFM
  - métadonnées
  - annotation linguistique
- Enrichissement et recherches linguistiques
  - oral représenté et variation
  - passage du latin au français

Historique

- **1989** : Début du projet (équipe ELI, resp. C. Marchello-Nizia)
  - complément au DMF (ATILF)
- **1999** : Projet Queste du Graal
  - édition du ms. Lyon BM PA 77 (texte, traduction, images ms)
- **2000** : Intégration de la Base dans Weblex
- **2002 – 2005** : Relectures et encodage XML-TEI
  - échange avec l'ATILF
- **2004** : Création du CCFM (Ottawa, Lyon, Nancy, Zurich, Stuttgart)

Historique

- **2005** : Ouverture du site http://bfm.ens-lyon.fr
- **2007- 2010** : Fermeture de la Base
  - contentieux avec la librairie Droz
- **2012** : Portail BFM **http://txm.bfm-corpus.org**
  - plateforme TXM
  - BFM 2012 (3,3 millions de mots)

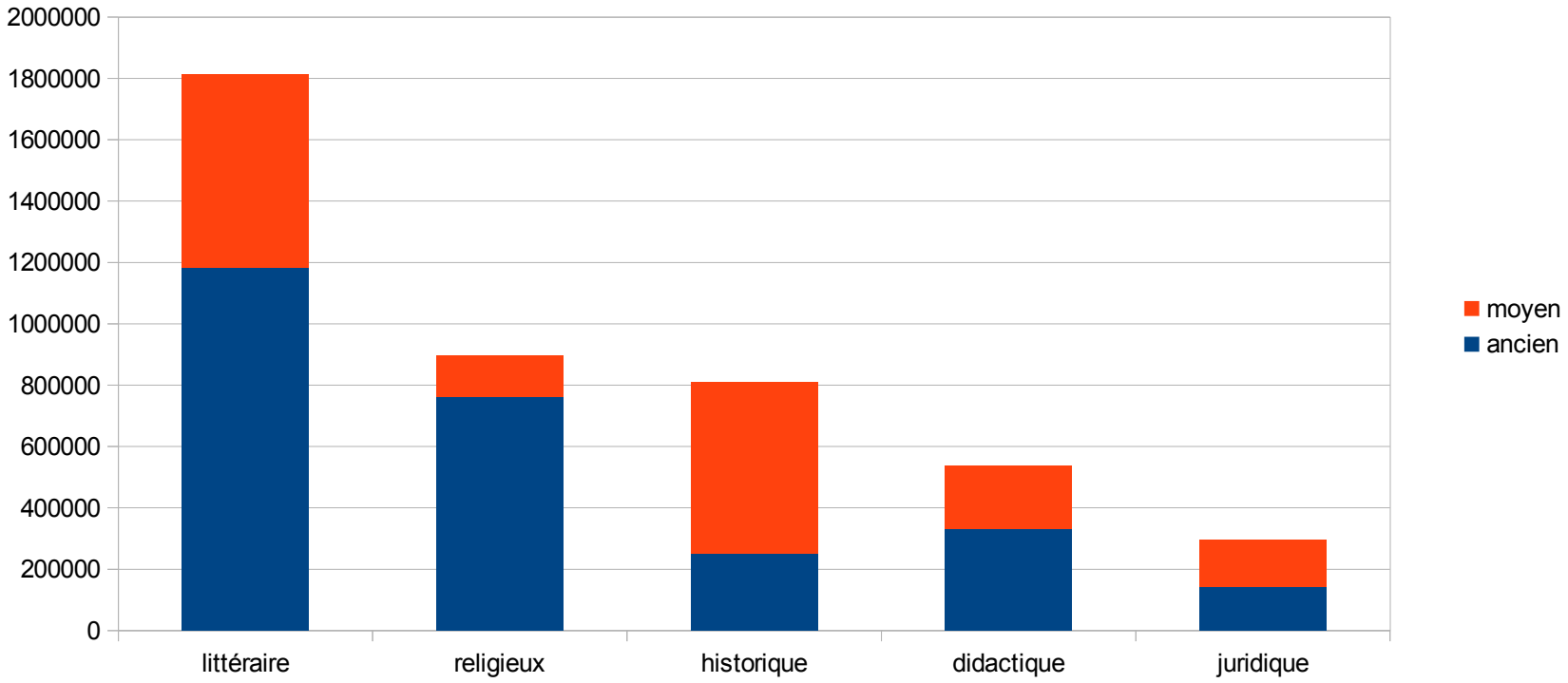
La BFM aujourd'hui (2013)

- Corpus **large** et **standardisé**
  - 250 utilisateurs inscrits
  - 142 textes / 4 700 000 occ.
  - toute la période médiévale (9e – 15e siècle)
  - encodage XML/TEI P5 avec ODD
- Corpus **en ligne** et **ouvert** (licence Creative commons BY-NC-SA)
- Corpus **organisé** et **enrichi** (métadonnées, annotation)

Métadonnées de la BFM

- Système de **descripteurs** :
  - Données bibliographiques (auteur, titre, date, etc.)
  - Données philologiques et pragmatiques (région, etc.)
  - Typologies discursives : 6 **domaines** (religieux, littéraire, didactique, historique, juridique, sources documentaires/actes de la pratique), 50aine de **genres discursifs**
- Standardisation des descripteurs :
  - Site du CCFM
  - Entrepôt http://weboai.sourceforge.net/ (Cahier)

Domaines par période



Annotation linguistique

- Balisage du **discours direct** (quasi-total)
- **Syntaxe** :
  - annotation manuelle (15 textes / 280 000 occ.) ; modèle conçu pour le français méd.
- **Morphosyntaxe**
  - étiquetage automatique (total) / vérifié (19 textes / 850 000 occ.) ; jeu spécialisé pour le français méd. (Cattex 2009/59 étiquettes)
- **Lemmatisation** (en cours)

Enrichissement et recherches linguistiques

- Textes et métadonnées :
  - → contrastes externes
- Structures textuelles et propriétés : numéro, titre...
  - → contrastes internes
  - → cooccurrences
- Unités lexicales et propriétés :
  - → choix du niveau d'analyse (eg lemme>verbes)

Oral représenté et variation diasystématique

- Problématique et méthodologie de recherche

  *Définition de l'**oral représenté** en français ancien*

  – De l'écrit qui se donne comme de l'oral (**discours direct**)

  – Des marques explicites de balisage

  Quel statut pour l'oral représenté ?

  – Le discours direct présente-t-il une **grammaire** spécifique ? Comment l'analyser ?

  – L'axe discours direct *vs* non discours direct est-il un **paramètre de variation** à prendre en compte dans les analyses linguistiques ? Comment se combine-t-il avec les autres paramètres de variation connus ?

OR : Méthodologie de recherche

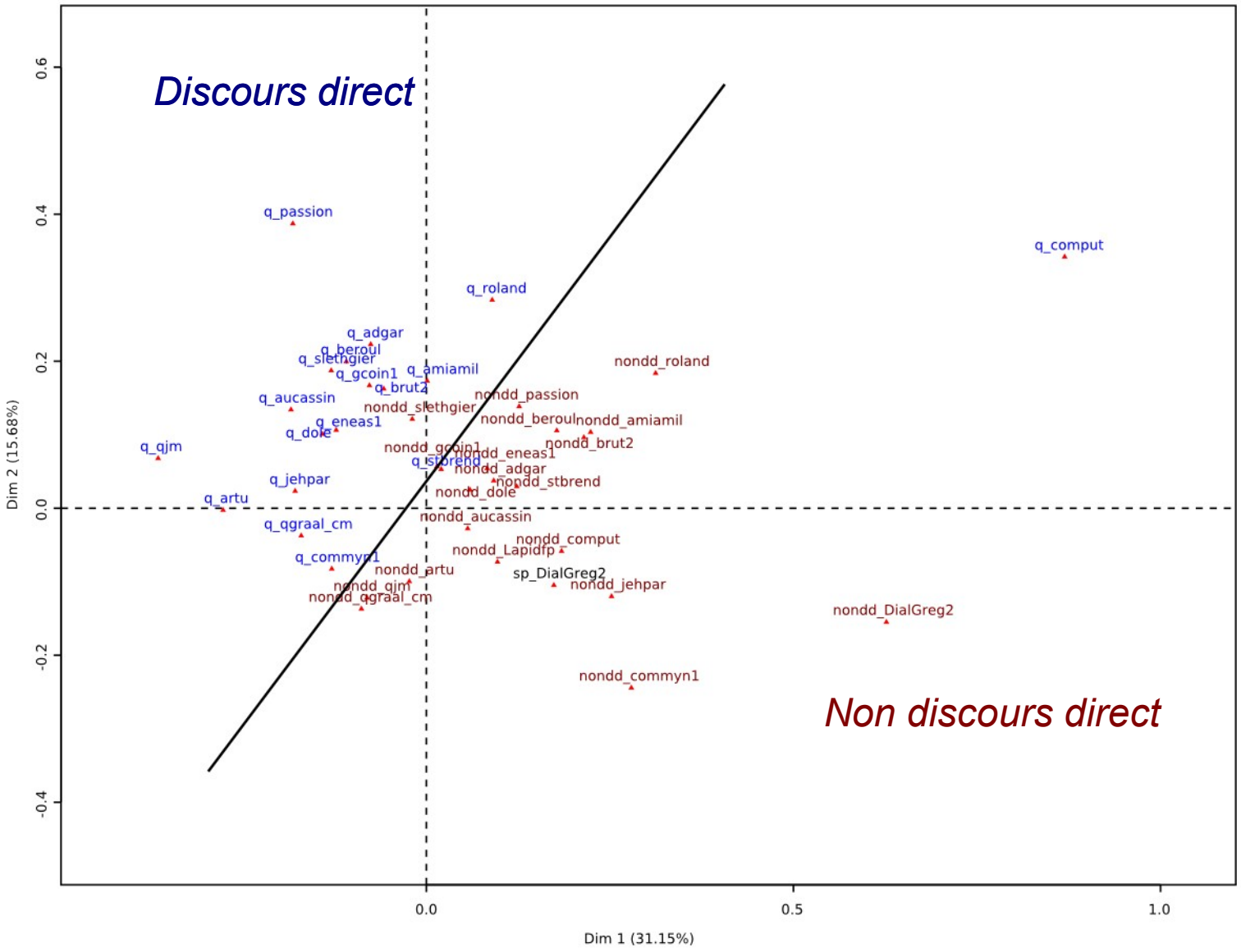
- *Constitution d'un corpus adapté*
  - Balisage du discours direct
  - Étiquetage des parties du discours
  - Sélection des paramètres de variation et création de contrastes
- *Instrumentation du corpus avec l'outil d'analyse textométrique TXM*
  - Construction de sous-corpus et calcul de fréquences
  - Calcul de scores de spécificité
  - Analyses factorielles des correspondances
- *Interprétation des résultats*
  - Sélection de vues et de données significatives
  - Retours sur la procédure d'analyse

OR: Méthodologie de recherche

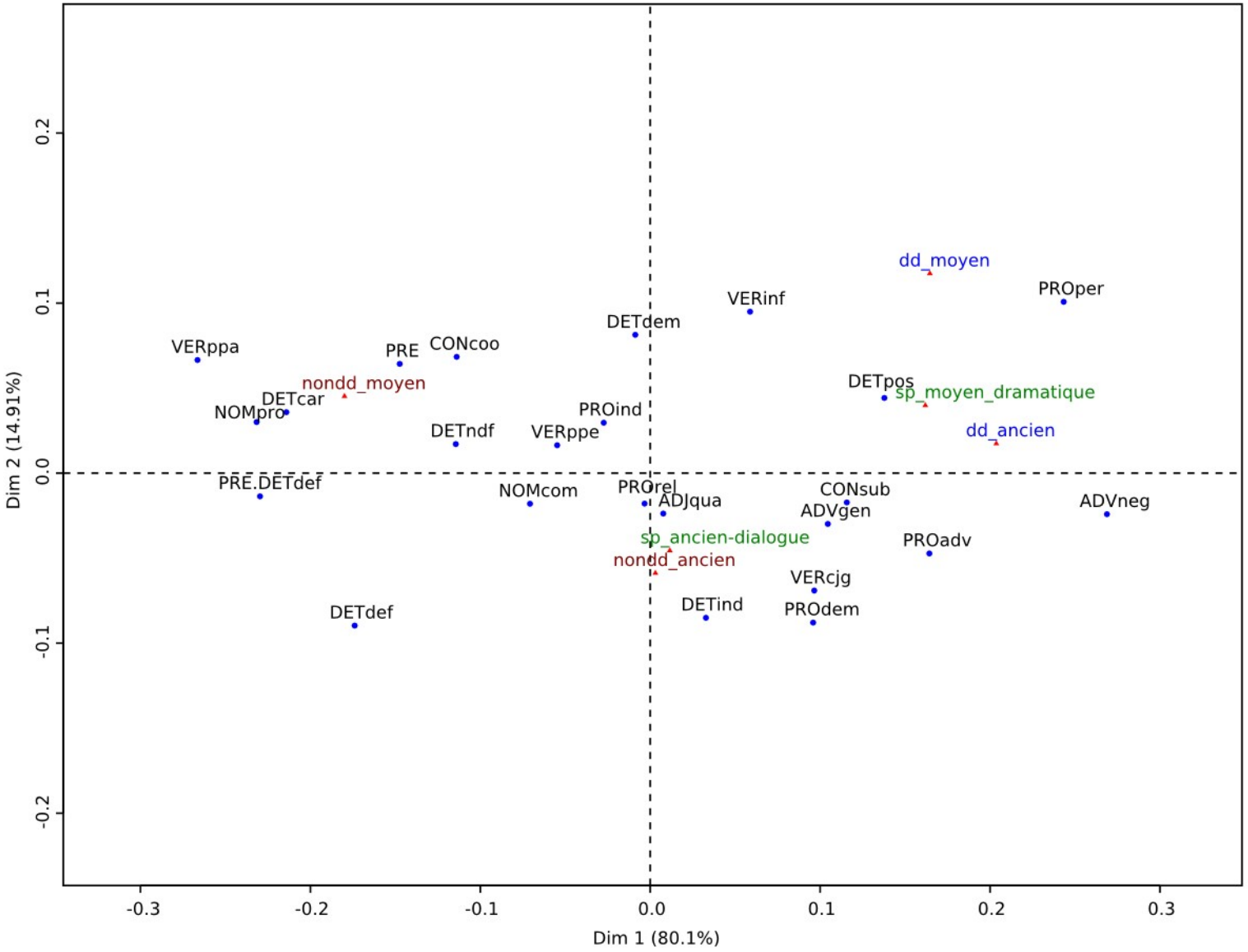
- *Axes de variation*
  - **Axe diachronique** : périodes *ancien f. / moyen f.* (9ème–fin 15ème)
  - **Axe des domaines discursifs** : *didactique / historique / juridique / littéraire / religieux*
  - Axes de la forme (vers/prose), des genres, des dialectes, etc.
- *Annotation des données*
  - Balisage du discours direct à partir des marques formelles : *DD* ou *SP* (reste = *non DD*)

    → **Axe DD / nonDD**
  - Étiquetage des **parties du discours** (58 étiquettes Cattex 2009 : *NOMcom*, *VERcjg*, *DETpos*, *PRE.DETdef*, etc.)

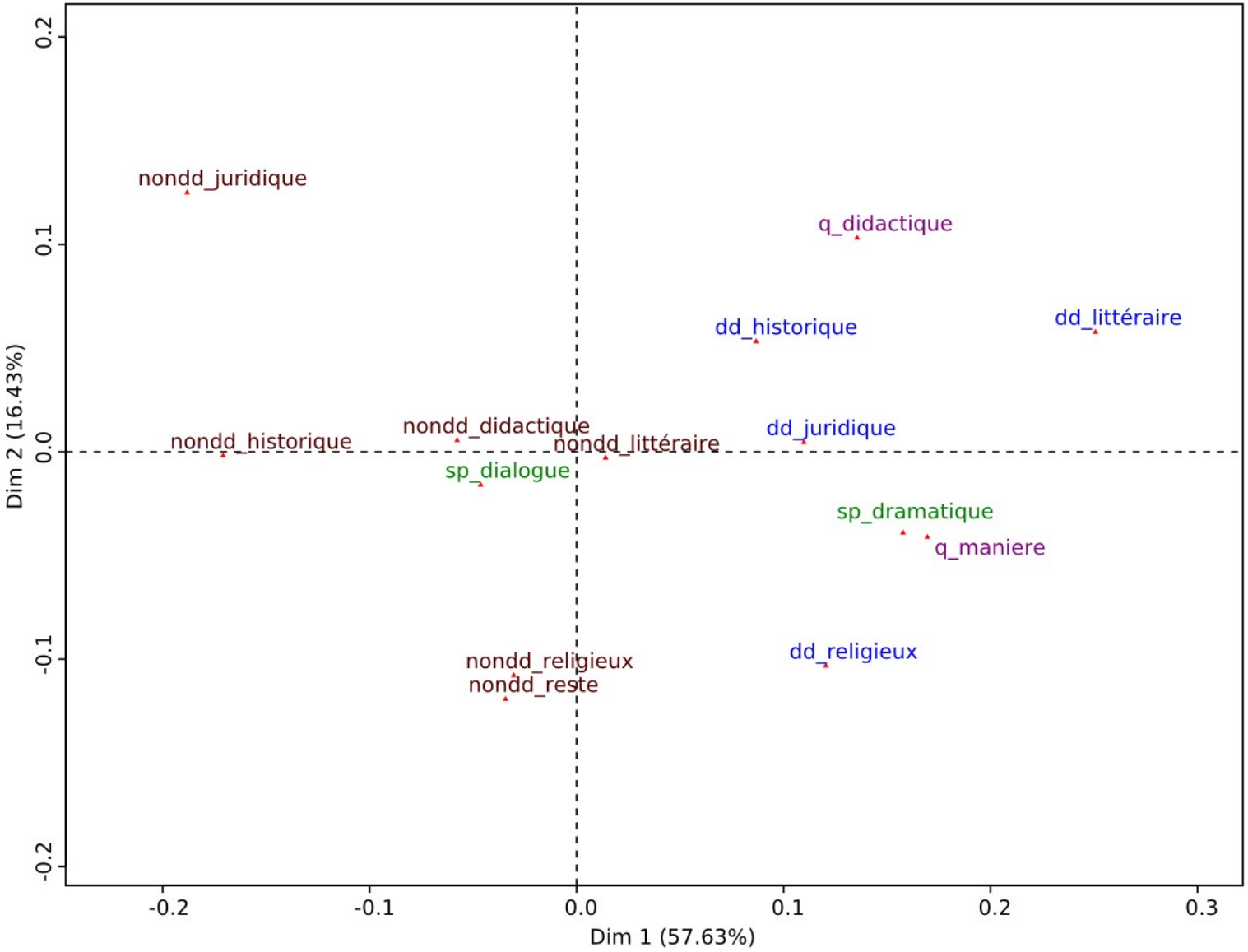
OR : AFC au niveau des textes



OR : AFC au niveau des périodes



OR : AFC au niveau des domaines



Oral représenté : bilan

- Convergence des résultats
  - sous-corpus vérifié / corpus complet
  - axe « DD / nonDD » > autres axes de variation (diachronie, domaines...)
  - stabilité de la répartition des étiquettes
- Caractéristiques morphosyntaxiques
  - DD : pronom pers. et imper., négation, infinitifs, conj. de sub.
  - nonDD : nom, déterminant, préposition, conj. de coord.
- Il est important de prendre en compte la variation « DD / nonDD » dans l'analyse de données linguistiques

Passage du latin au français

Création d'un **corpus bilingue latin tardif / français ancien** en collaboration avec le Lasla (U. de Liège) :

- corpus **global** : 36 textes latins et français indépendants
- sous-corpus **aligné** : 3 textes en relation de traduction alignés au niveau des pages

Corpus bilingue global : partie latine

23 textes - 69 801 occurrences	
date composition	350-750 (10), 800 -1160 (6)
période	**mérovingien**, carolingien
région	Gaule, **Italie**
domaine	**religieux**, littéraire
genre	dialogue, dramatique, **hagiographie**
forme	prose

Corpus bilingue global : partie française

11 textes – 230 838 occurrences	
date composition	883 - 1275
région	ouest, anglo-normand, bourguignon, normand, wallon
domaine	**religieux**, didactique
genre	dialogue, dramatique, **hagiographie**
forme	prose, vers

Corpus global : les démonstratifs

BLLAT		BLFRO	
T= 69 801, t= 3 310, soit 4,7%		T= 230 838, t= 1 822, soit 0,8%	
IS	1411	CIL	1200
HIC	686	CIST	622
ILLE	553		
IPSE	335		
IDEM	199		
ISTE	126		

Sous-corpus aligné

Trois textes en prose (VIe-Xe) traduits en français (fin XIIe-déb. XIIIe)

- *Dialogue de l'Âme* : Saint Isidore (*Synonyma*), VIIe / traduction anonyme v.1200, ms lorrain
- *Vie de saint Benoît* : saint Grégoire, fin VIe, Italie / traduction anonyme XIIe, wallonie
- *Vie de saint Eustache* : anonyme, Xe / début XIIIe, peu de traits dialectaux

Sous-corpus aligné