ENS DE LYON

icar
UMR 5191 - CNRS

Le projet PRESTO : corpus et traitements

Vannina Goossens, ENS de Lyon, ICAR
Achille Falaise, ENS de Lyon, ICAR

Journées d'étude *Linguistique de corpus outillée : regards et expériences croisés*
Neuchâtel, 10-11 juin 2014

Plan de la présentation

1. Corpus
    1. Besoins et objectifs
    2. Structure du corpus et critères de constitution
    3. Corpus et échantillons constitués
2. Outils et traitements
    1. Base de données
    2. Étiquetage et lemmatisation

Corpus PRESTO

Structuration, critères, finalités

1. Besoins et objectifs

- Constitution d'un corpus dans le cadre de PRESTO :
  - Prérequis pour la réalisation du projet
  - Apport majeur du projet
- **Besoins** :
  - Représentation de toutes les périodes de l'histoire du français : 9ème s. au 20ème s.
  - Présence de différents types de textes et de différents genres discursifs
  - Enrichissement linguistique : étiquetage morpho-syntaxique et lemmatisation

1. Besoins et objectifs

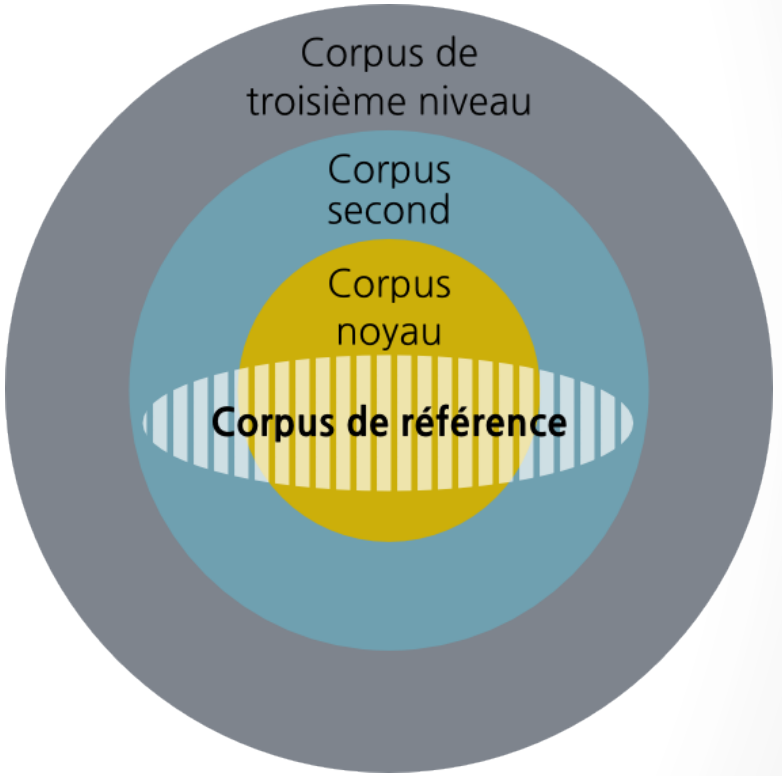
- **Objectifs** :
  - Disposer d'annotations linguistiques fiables mais également de gros volumes de textes
  - Pouvoir rendre disponibles autant que possible le corpus constitué et les outils élaborés tout en respectant les contraintes juridiques
- **Contrainte** :
  - Disposer de versions numérisées de textes de bonne qualité

→Collaboration avec diverses bases textuelles existantes :

- Frantext
- BVH
- Cologne
- ARTFL
- CEPM
- …

2. Structure du corpus

- Ces besoins et contraintes nous ont amenés à structurer le corpus PRESTO :
  - 3 niveaux hiérarchiques :
    - Corpus noyau
    - Corpus second
    - Corpus de troisième niveau
  - 1 niveau transversal :
    - Corpus de référence



2.1 Corpus noyau

- **Corpus qualitatif** :
  - Lemmatisation et étiquetage morpho-syntaxique **vérifiés et corrigés manuellement**
  - Critères stricts de sélection des textes
- **Finalités** :
  - Permettre des requêtes et des calculs utilisant des annotations les plus fiables possibles
  - Améliorer l'annotation automatique des autres textes
  - Proposer un corpus ouvert prenant en compte les besoins de certains types de public (chercheurs en sciences du langage et en lettres)

2.1 Corpus noyau

- **Critères de constitution** :
  - Statut juridique ouvert permettant à terme une diffusion libre sous licence Creative Commons
  - Éditions originales ou les moins interventionnistes : importance du respect de l'orthographe d'origine
  - Répartition chronologique homogène : nombre de textes équilibrés par demi-siècles
  - Équilibrage en genres et domaines : une réflexion sur la catégorisation des textes est en cours
  - Équilibrage des formes : vers et prose

→Ensemble restreint de textes, échantillonnés au-delà de 50 000 mots

2.2 Corpus second

- **Finalités** :
  - Disposer d'un corpus structuré :
    - D'un volume plus important
    - Avec des genres textuels plus diversifiés
- **Caractéristiques** :
  - Critères de sélection des textes plus souples :
    - Statut juridique possiblement moins ouvert
    - Les éditions originales tout en étant privilégiées ne sont plus obligatoires
    - Critères de répartition chronologique et d'équilibrage en genres, domaines et formes respectés autant que possible
  - Ensemble de textes plus vaste :
    - Textes du noyau échantillonnés en version intégrale
    - Œuvres complètes de certains auteurs
    - Vaste corpus de presse (Le Figaro 19ème)
  - Lemmatisation et étiquetage morpho-syntaxique automatique

2.3 Corpus de troisième niveau

- **Caractéristiques** :
  - Corpus « de rencontre » :
    - Intégration de textes acquis au gré de rencontres mais qui ne trouvent pas leur place dans les deux premiers niveaux de corpus
  - Critères de constitution moins stricts, si ce n'est la qualité des textes
  - Lemmatisation et étiquetage automatiques

→Vivier supplémentaire plus vaste pour construire un échantillon personnalisé

2.4 Corpus de référence PRESTO

- Corpus constitué pour permettre des études longitudinales des prépositions dans le projet PRESTO
  - → Nécessité de disposer d'un corpus structuré plus volumineux que le corpus noyau
- **Caractéristiques** :
  - Réunion de textes des niveaux 1 et 2
  - Mêmes critères de sélection des textes que pour le corpus noyau : corpus très structuré
  - Pas de verrou juridique

3. Corpus constitués

- Étapes réalisées :
  - Corpus noyau :
    - Version 1 pour la période 16ème – 20ème
    - Version 2 pour la période 16ème – 18ème
  - Échantillon d'apprentissage 16ème – 18ème
- Corpus noyau 16ème – 18ème :
  - 1 M mots :
    - 5 à 7 textes par demi-siècle
    - 100 à 200 000 mots par demi-siècle
  - Genres représentés :
    - Roman, traité/essai, théâtre, poésie : pour tous les demi-siècles
    - Correspondance : excepté 16ème siècle et 1ère moitié du 18ème
    - Récit de voyage : 2ème moitié du 16ème uniquement

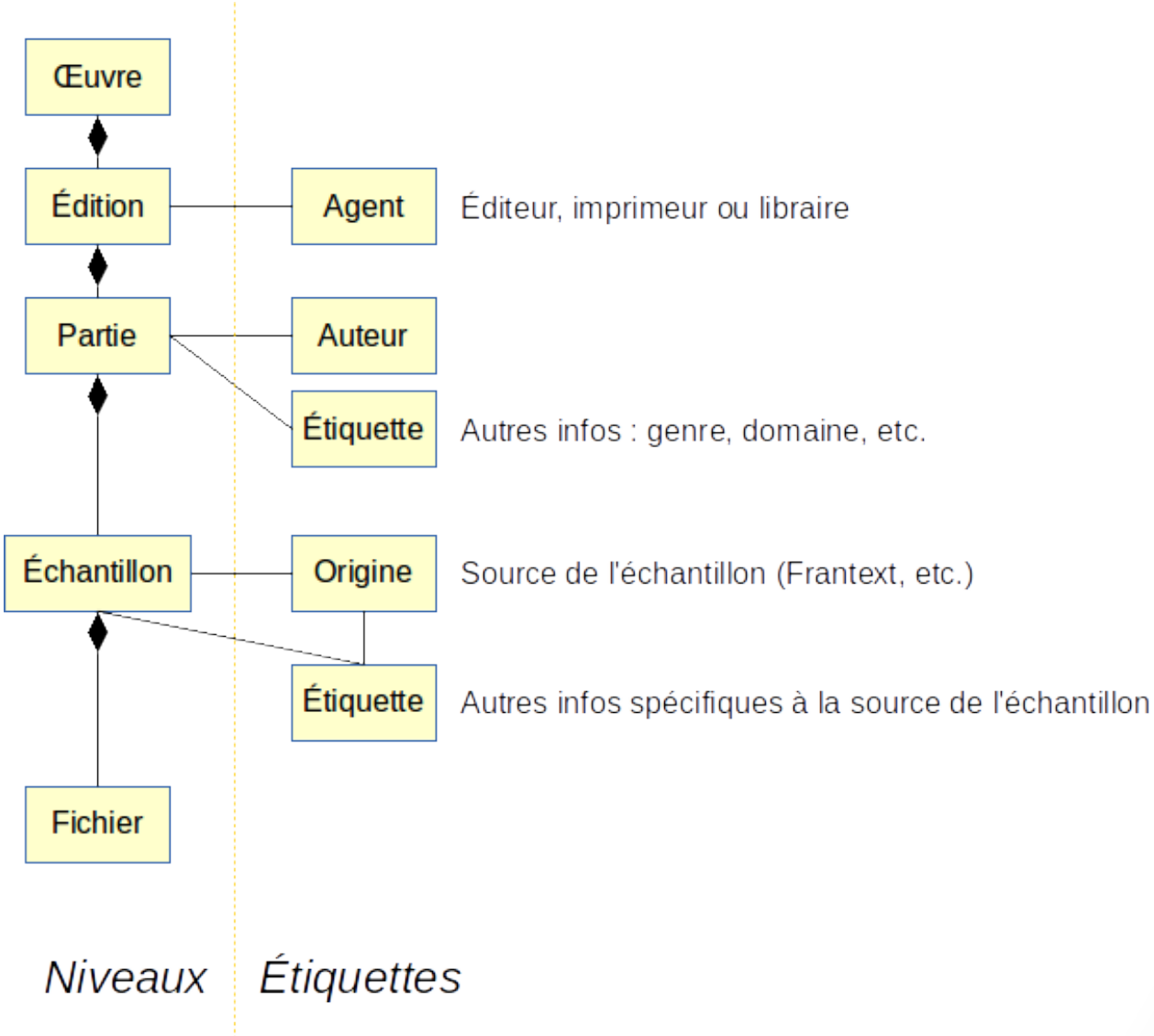
3. Corpus constitués

- Échantillon d'apprentissage 16ème – 18ème :
  - Découpage en périodes de relative stabilité orthographique, syntaxique et lexicale :
    - 1500-1530
    - 1530-1590/1610
    - 1590/1610-1660
    - 1660-1720
    - 1720-1761
    - 1761-1789
  - Prélèvement de 100 000 mots par périodes dans le corpus noyau
  - Équilibrage des genres, domaines, forme et de la répartition chronologique
  - 500 000 mots désambiguïsés et corrigés manuellement

Outils et traitements

Base de données, étiquetage et lemmatisation

1. Base de données



2. Lemmatisation & étiquetage

- Ressources
  - Collecte de textes
  - Constitution d'un jeu d'étiquettes
  - Constitution de lexiques
- Outils
  - « Projection » des lexiques/étiquettes sur les textes
  - Désambiguïsation
    - Automatique (1/2)
    - Manuelle (sur un échantillon)
    - Automatique (2/2)

2.1 Textes

- Ressources
  - Collecte de textes
  - Constitution d'un jeu d'étiquettes
  - Constitution de lexiques
- Outils
  - « Projection » des lexiques/étiquettes sur les textes
  - Désambiguïsation
    - Automatique (1/2)
    - Manuelle (sur un échantillon)
    - Automatique (2/2)

2.1 Textes

<head>PRéFACE DE LA SECONDE éDITION</head>
<pb n="I"/>
<p><hi rend="I"> quelques remarques sur les groupements</hi><lb/>
<hi rend="I"> professionnels : </hi><lb/>
en rééditant cet ouvrage, nous nous sommes interdit<lb/>
d'en modifier l'économie première. Un livre a une<lb/>
individualité qu'il doit garder. Il convient de lui<lb/>
laisser la physionomie sous laquelle il s'est fait<lb/>
connaître.<lb/>

2.2 Lexiques/étiquettes

- Ressources
  - Collecte de textes
  - Constitution d'un jeu d'étiquettes
  - Constitution de lexiques
- Outils
  - « Projection » des lexiques/étiquettes sur les textes
  - Désambiguïsation
    - Automatique (1/2)
    - Manuelle (sur un échantillon)
    - Automatique (2/2)

2.2 Lexiques/étiquettes

aimas;AIMER;VMIS2S0

aimasmes;AIMER;VMIS1P0

aimasse;AIMER;VMSI1S0

aimassent;AIMER;VMSI3P0

aimasses;AIMER;VMSI2S0

aimassies;AIMER;VMSI2P0

aimassiez;AIMER;VMSI2P0

aimassions;AIMER;VMSI1P0

aimassiés;AIMER;VMSI2P0

aimast;AIMER;VMSI3S0

aimastes;AIMER;VMIS2P0

2.3 Projection lexicale

- Ressources
  - Collecte de textes
  - Constitution d'un jeu d'étiquettes
  - Constitution de lexiques
- Outils
  - « Projection » des lexiques/étiquettes sur les textes
  - Désambiguïsation
    - Automatique (1/2)
    - Manuelle (sur un échantillon)
    - Automatique (2/2)

2.3 Projection lexicale

<head><wpresto dico="_UNK_">PRéF</wpresto> <wpresto dico="CE:DD0MS0|CE:PP0CNN00">CE</wpresto> <wpresto dico="DE:DF0CN0|DE:SPS00|DÉ:NCMS000">DE</wpresto> <wpresto dico="LA:NCMN000|LE:DA0FS0|LE:PP3FSA00|LÀ:RG">LA</wpresto> <wpresto dico="SECOND:NCFS000|SECONDE:NCFS000|SECONDER:VMIP1S0|SECONDER:VMIP3S0|SECONDER:VMM02S0|SECONDER:VMP00SM|SECONDER:VMSP1S0|SECONDER:VMSP3S0">SECONDE</wpresto> <wpresto dico="ÉDITION:NCFS000">éDITION</wpresto>

</head>

<pb n="I"/>

<p><hi rend="I"><wpresto dico="QUELQUE:AQ0CP0|QUELQUE:DI0CP0">quelques</wpresto> <wpresto dico="REMARQUE:NCFP000|REMARQUER:VMIP2S0|REMARQUER:VMP00PM|REMARQUER:VMSP2S0">remarques</wpresto> <wpresto dico="SUR:AQ0CS0|SUR:SPS00|SÛR:AQ0MS0|SÛR:RG">sur</wpresto> <wpresto dico="LE:DA0CP0|LES:PP3CPA00|LÈS:SPS00|LÉ:NCMP000">les</wpresto> <wpresto dico="GROUPEMENT:NCMP000">groupements</wpresto>

</hi><lb/>

2.4 Désambiguïsation

- Ressources
  - Collecte de textes
  - Constitution d'un jeu d'étiquettes
  - Constitution de lexiques
- Outils
  - « Projection » des lexiques/étiquettes sur les textes
  - Désambiguïsation
    - Automatique (1/2)
    - Manuelle (sur un échantillon)
    - Automatique (2/2)

ENS DE LYON

icar
UMR 5191 - CNRS

Le projet PRESTO : corpus et traitements

Vannina Goossens, ENS de Lyon, ICAR
Achille Falaise, ENS de Lyon, ICAR

Journées d'étude *Linguistique de corpus outillée : regards et expériences croisés*
Neuchâtel, 10-11 juin 2014