Syntactic Reference Corpus of Medieval French
SRCMF [smœrf]
Un projet ANR/DFG

4e rencontre du CCFM
ATILF, Nancy, 6-7 octobre 2008

Sophie Prévost & Achim Stein

LaTTiCe

Universität Stuttgart

Lettres et sciences humaines

Syntactic Reference Corpus of Medieval French

- *Syntactic Reference Corpus of Medieval French* (SRCMF)
  - Responsables: Sophie Prévost, Achim Stein
  - Durée 3 ans, Agence Nationale de Recherche (ANR) et Deutsche Forschungsgemeinschaft (DFG)
- Coopération
  - Paris: UMR 8094-LaTTiCe (CNRS/ENS): Sophie Prévost
  - Lyon: ENS/LSH
    - Céline Guillot, Serge Heiden, Alexei Lavrentiev, Christiane Marchello-Nizia (professeur émérite)
  - Stuttgart: Institut für Linguistik/Romanistik (ILR)
    - Achim Stein, Beatrice Bischof, Nicolas Mazziotta
  - Montréal (UQAM): Fernande Dupuis
  - Experts: Richard Ingham (Birmingham), Bernard Victorri (LaTTiCe)

Syntactic Reference Corpus of Medieval French

- Justification / Motif
  - Existence de ressources syntaxiquement annotées importantes pour l'anglais (UPenn)
    - Old English (YCOE), Middle English (PPCME2)
  - Préparer une base comparable pour le français médiéval
  - Créer des corpus et outils réutilisables
    - Outils d'annotation
    - Format de représentation
    - Évaluation ou création d'environnements d'exploitation

Syntactic Reference Corpus of Medieval French

- Objectifs
  - Améliorer les deux plus grandes bases textuelles pour le français médiéval (environ 3 millions de mots chacune)
    - BFM: Base de Français médiéval (ENS-LSH Lyon)
    - NCA: Nouveau Corpus d'Amsterdam (ILR, Stuttgart)
  - Attribuer une couche d'annotation commune à ces deux corpus
    - L'annotation syntaxique sera *a priori* indépendante des couches d'annotation existantes (p.ex. morphologique)

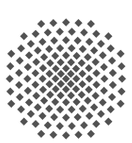
Syntactic Reference Corpus of Medieval French

Nouveau Corpus d'Amsterdam (NCA)

- Kunstmann, Gleßgen, Stein (2006)
- Le corpus
  - 296 textes (entiers ou extraits)
  - couvrant les variétés dialectale (Dees 1987)
  - 3.183.226 mots, étiquetés, lemmatisés
  - format XML, information méta-textuelle (bibliographie, descripteurs)
- Licence de recherche (gratuite):
  - Version 2 (2008): accessible en ligne
  - Outil de recherche: TWIC
- Ressources lexicales:
  - Lexique des formes (250.000 graphies), avec catégorie et lemme
- Outils
  - *TreeTagger* pour l'ancien français

Banque de Français Médiéval (BFM)

- ENS-LSH Lyon (C. Guillot)
- Le corpus
  - environ 3 million de mots
  - Ancien et Moyen français
  - Textes entiers
- Annotation morphosyntaxique partielle
  - Annotation manuelle de 5 textes
  - Développement d'un jeu d'étiquettes (CATTEX)
- Publication en ligne
- Outils de recherche
  - Weblex, IMS Corpus Workbench
  - Utilisation et développement d'outils d'étiquetage:
    - SATO, Brill



Syntactic Reference Corpus of Medieval French

- Modèle de grammaire: situé entre deux pôles
  - Grammaire à constituants immédiats (UPenn)
  - Grammaire dépendancielle (Prague Dependency Treebanks)
  - Nous annoterons la forme de certains syntagmes (constituants)
    - p.ex. groupe prépositionnel, nominal, adjectival
    - Les fonctions grammaticales seront encodés
      - ou directement (p.ex. attribut-valeur: function=objet_direct)
      - ou par leur dépendance (COD = GN qui dépend du verbe)
- Annotation la plus neutre possible par rapport aux théories syntaxiques:
  - Nous n'annoterons ni catégories vides ni traces
  - Certains phénomènes seront annotés dans le balisage (p.ex. phrases à sujet nul)

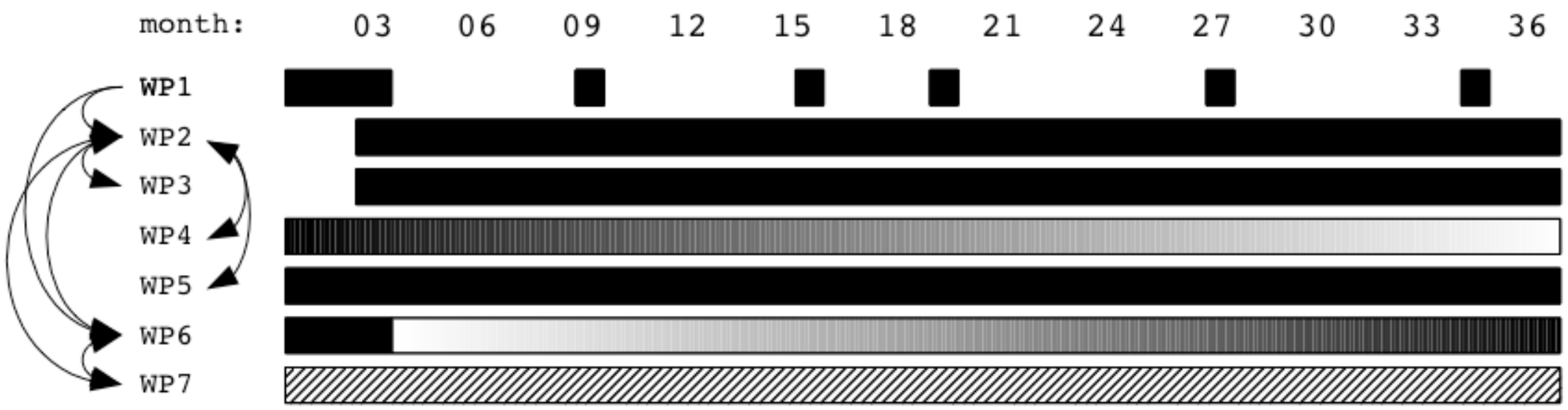
Annotation manuelle (automatique partielle?)

- Outil d'annotation manuelle
  - *NotaBene* (N. Mazziotta)
- Essais d'annotation automatique partielle (*chunking*)
  - basée sur les étiquettes de la couche morphosyntaxique
  - limitée à des syntagmes de haute fréquence et non ambigues
- Question:
  - Retire-t-on un gain de temps de l'annotation automatique (plus vérification)?

```
<s id="s3" line="3">
<word pos="ADV">tresqu</word>
<CHUNK-PP id="s3.pp1" gov="s3" func="mod">
  <word pos="PRE">en</word>
  <CHUNK-NP id="s3.np1" gov="s3" func="obj">
    <word pos="DET">la</word>
    <word pos="NOM">mer</word>
  </CHUNK-NP>
</CHUNK-PP>
<word pos="VER">cunquist</word>
<CHUNK-NP id="s3.np2" gov="s3" func="obj">
  <word pos="DET">la</word>
  <word pos="NOM">tere</word>
</CHUNK-NP>
<word pos="ADJ">altaigne</word>
</s>
```

TWIC chunker: reconnaissance des GN et des GP
tresqu en la mer cunquist la tere altaigne

Syntactic Reference Corpus of Medieval French



- Paquets de travail (selon la demande de projet ANR/DFG)
  - WP1: partial in-depth annotation: tous
  - WP2: in-breadth annotation: SP, AS, NM, BB, AL
  - WP3: Morphosyntactic Annotation and Chunking: SP, AS, FD, CMN
  - WP4: Tool for Manual Syntactic Annotation: NM, BB, AL
  - WP5: Corpus Administration, Elaboration and Quality Control: CG, AL
  - WP6: Formalisms, Query Tools, Portability and Reusability: SH, AL
  - WP7: Evaluation and Prototypical Research: tous, plus experts externes

Syntactic Reference Corpus of Medieval French

- WP 1: Annotation "en profondeur"
  - Annotation maximale (pour certains textes)
- WP 2: Annotation "en largeur"
  - Un sous-ensemble des catégories de WP 1
  - Visée maximale: appliquée à tout le corpus
    - Catégories
      - Phrases: type de phrase
      - GN, GP, pronoms personnels: fonction, dépendance
- WP 3: Annotation morposyntaxique
  - Indépendante de l'annotation syntaxique
- WP 4-6: Outils, formalismes, procédures
- WP 7: Experts, partenaires, chercheurs en syntaxe diachronique