Valentin Vydrin

Corpus électronique des textes mandingue : un état des lieux

Suite à la discussion au 2e Colloque à St. Petersbourg (2008), l'équipe de St. Petersbourg (Vydrin, Davydov, Maslinsky, Erman) a commencé, fin de 2009, l'élaboration des logiciels nécessaires pour un corpus électronique des textes bambara. Plus tard, elle a été rejointe par Adrij Rovenchak.

Résultats de travail :

1. La base de données lexicale (sur la base du dictionnaire de Charles Bailleul) a été créée. Le travail accompli : la systématisation des gloses, des marques des parties de discours, de la présentation des mots composés, des entrées de référence. Le travail en cours : systématisation de la présentation de la polysémie.

2. Un analyseur morphologique des textes (parser), s'appuyant sur la base de données lexicale et une présentation formelle de la morphologie bambara. Une première version a été faite en janvier 2011; nous sommes en train de perfectionner ce logiciel.

3. Les logiciels pour l'introduction des métadonnées et la désambiguïsation des textes traités par l'analyseur morphologique ont été élaborés vers décembre 2010 – janvier 2011. Actuellement, nous sommes en train de les perfectionner.

4. Un correcteur automatique d'orthographe bambara pour Open Office a été élaboré (juillet 2011).

5. Les premiers essais d'exportation de textes glosés en Open Office et de la machine de recherche on été effectués.

6. On a commencé la désambiguïsation des textes bambara.

7. Une bibliographie des publications en bambara a été composée.

Les objectifs prochains du travail :

– créer un échantillon de textes glosés et désambiguïsés de 10 à 100 000 mots pour commencer les essais de désambiguïsation automatique (moyennant les logiciels statistiques);

– compléter le traitement de la polysémie dans la base lexicale;

– élaborer la machine de recherche.

Perspectives plus avancées : travailler sur les corpus électroniques maninka et dioula.

Un élargissement du groupe de travail est prévu ; un projet franco-allemand est en formation.