Collecte et exploitation de corpus dans le domaine berbère

Centre de Recherche Berbère, INALCO

Amina Mettouchi

Responsable du projet

- Objectif : constituer une ‘banque’ de textes oraux transcrits dans diverses variétés de berbère
- Origine : constat d ’une situation où de nombreux travaux sur le berbère à base de données textuelles sont effectués, sans qu ’il soit possible d ’avoir accès à ces données, qui restent dans les tiroirs du chercheur.
- Historique du projet
  - décembre 2000 : proposition du projet ‘Intonation et Corpus’ au CRB, par Amina Mettouchi
    - une banque de textes oraux représentant les divers parlers
    - un travail sur les intonations des variétés de berbère
  - 2001-02 : mise en œuvre du volet ‘Intonation’
  - Décembre 2002 : relance du volet ‘Corpus’, avec intégration prévue du volet ‘Intonation’ (à travers un sous-corpus d’enregistrements dédiés à l ’analyse intonative)

- Tradition et continuité

Les études berbères sont traditionnellement fondées sur des documents authentiques collectés

- Fichier de Documentation Berbère
- Etudes et Documents Berbères
- Thèses et DEA
- …

- L'innovation
  - la taille de la 'banque ' de données
  - la possibilité d'utiliser des outils de recherche automatique
  - la centralisation, à travers l'existence d 'un référent international (Centre de Recherche Berbère de l'INALCO)

- Qu'est-ce qu 'un corpus ?

  « corps de textes écrits ou transcrits qui peut servir de base pour l'analyse ou la description linguistique » (Kennedy 1998)

- Corpus ou Archive ?

  ➢ systématique et représentatif vs moins structuré, et composé de textes collectés de manière aléatoire ?

  ➢ corpus statique (composé sur une période courte) vs corpus dynamique (toujours ouvert)

- Corpus réalisés pour l'oral (anglais) : cf. http://devoted.to/corpora
  - ➢ LLC (London-Lund Corpus of Spoken English) (Svartvik 1975)
  - ➢ CSAE (Corpus of Spoken American English) (Chafe 1991)
  - ➢ ICE (International Corpus of English) (Greenbaum 1996)
  - ➢ Talkbank (International database of communicative interaction), (Mc Whinney, Bird, Liberman, Watclar 1999)
  - ➢ IViE (Intonational Variation in English) (Grabe, Post, Nolan 2000)
  - ➢ ...

Pourquoi un corpus ?

- Des données empiriques pour étayer les théorisations linguistiques
  - Dépasser la dichotomie compétence / performance, et remettre au centre des préoccupations linguistiques la langue telle qu'elle est parlée
  - Multiplier les approches, les points de vue sur la langue (intuition, introspection, élicitation et recherche sur données textuelles devraient idéalement se compléter) (Chafe 1992)
  - Voir surgir des faits de langues contredisant les idées reçues

Exemple du morphème ara en contexte négatif : l'élicitation amenait à poser un marqueur de négation discontinu *ur … ara*. L'étude d 'un échantillon de corpus (360 négations verbales) a montré que 48% de négations ne comportaient pas *ara*. On peut alors se poser la question de savoir si *ara* peut véritablement être considéré comme un marqueur négatif.

↳ avoir accès au langage en action, à des productions attestées et réalisées dans des conditions qui sont celles de l'utilisation naturelle du langage

- Une archive permettant de conserver des documents irremplaçables
  - enregistrements anciens
  - documents audio et vidéo

Exemple des zénètes du Gourara : voir l 'intervention de Rachid Bellil.

↳ participer à la constitution d'une banque de documentation orale conservant la mémoire culturelle et linguistique des berbères.

Nos objectifs

- Linguistique berbère

  ↳ un ancrage sur le terrain, des collaborations, une pluridisciplinarité (littérature et culture, anthropologie, sociologie...)
- Linguistique générale et typologique

  ↳phono-morphologie, morpho-syntaxe, sémantique, énonciation, prosodie, pragmatique…
- Traitement Automatique du Langage (à terme)

  ↳des applications industrielles possibles

⇒ constitution d'un ensemble de textes, majoritairement oraux, représentant une variété de situations de discours, locuteurs, genres, et dialectes/parlers

Concrètement …

- Un volet Intonation

(http://www.inalco.fr/pub/enseignements/langues/afrique/berbere/index.html)

➢ Projet « intonations du berbère » initié en 2000, en cours de réalisation (une publication en 2002, deux sous presse pour 2003)

➢ Rattachement de fichiers-sons fiables à la transcription

➢ Prise en compte, à terme, de la gestualité (vidéo)

➢ Projet similaire : IViE . Contact a été pris avec la responsable du projet, Esther Grabe. http://www.phon.ox.ac.uk/~esther/ivyweb/

- Une attention particulière portée au discours

⇒ Transcription intégrant la signalisation de la dynamique conversationnelle :

- unités intonatives,
- pauses,
- tours de parole,
- chevauchements,
- répétitions, faux-départs,
- chuchotements (rires, etc.),
- bruits (et commentaires contextuels)

Projet similaire : Talkbank (http://www.talkbank.org). Contact pris avec Brian Mac Whinney.

- Une prise en compte des genres oraux

➢ Adapter les typologies de l'anglais …

- dialogue vs monologue
- privé vs public (conversations etc. vs cours, interviews etc.)

➢ … aux réalités berbères

- en pays berbérophone / en émigration
- femmes et hommes
- monolingues et bilingues
- variation dialectale
- etc.

Représentativité

- Fonction de l'usage (proportionnalité)

⇒ La conversation : part prépondérante

(entre personnes de même sexe, conversations mixtes, conversations entre monolingues, ou monolingues et bilingues, entre bilingues (code-switching), entre personnes de la même génération, entre personnes de générations différentes (y compris adultes-enfants)…)

⇒ Autres interactions : assemblée de village, cérémonies, interviews (médias) …

⇒ Monologues privés
- Recettes
- Savoir-faire traditionnel
- Récits personnels
- Mythes, récits familiaux (généalogies…)

⇒ Monologues publics
- Contes
- Poésie orale
- Discours politiques

• Ecrit (pour certains parlers)
- Romans
- Presse
- Documents officiels

⇒ Dialogues privés

- Conversations
- Interactions médecin-patient
- …

⇒ Dialogues publics

- Débats
- Interviews (médias)
- ...

Collecte

- Collecte non centralisée (initiative laissée aux institutions et aux individus), soutenue par un site Internet offrant méthodes et outils.
- Fiche de collecte

Mise à disposition sur le site du projet, elle sert de document de référence pour l 'archivage

- Matériel de collecte

Enregistreurs mini-disques

ex. Sharp MD-MT99H(S)

Micros-cravates stéréo

ex. Sony ECM-717

Caméscope

analogique ou numérique

DAT (lorsque l'on peut disposer de ce matériel coûteux mais très fiable)

- Problèmes légaux

Respect du droit des personnes (vie privée…)

Il faut penser à donner accès à l 'enregistrement (copie) aux locuteurs, et à leur demander s'ils souhaitent désavouer certains passages, ou refuser que certains passages soient rendus publics

Rappeler que les noms n'apparaissent pas dans la transcription, et que les passages à teneur politique ne sont pas rendus publics

Elaboration d'une charte pour l'accessibilité du corpus

Problème du Copyright

Transcription

- Temps

↳environ 10 heures pour une heure de conversation (en codant les chevauchements, etc.)

- Unification

↳polices : elles seront mises sur le site du projet

↳notation usuelle ou phonétique, phonologique ?

Selon les parlers et l'usage prévu pour l'enregistrement. Voir le site.

- Encodage

↳Traductions alignées

↳ Lien fichiers-textes / fichiers-sons / fichiers-images

↳ SGML (Standard Generalized Markup Language) ?

↳ Normes de la TEI (Text Encoding Initiative) ?

Archivage

- Taille des fichiers
  - ↳ 15-20 mégabytes pour un million de mots
- Sauvegarde
  - ↳ Sur un autre site (risques)
- Gestion
  - ↳ compilation (personnel)
  - ↳ accès

Accessibilité

- Critères
  - participation au corpus
- CDRom ou Internet
  - financement
  - aspects techniques
- Lien avec d'autres bases de données
  - 'propriété' intellectuelle

Conseils et avis

- Rachid Bellil (Alger)
  - ↳terrain, dimensions culturelles, variation
- Michel Jacobson et Alexis Michaud (LACITO)
  - ↳archivage langues orales
- Maria Candea et Mary-Annick Morel (Paris III)
  - ↳traitement intonatif