ArchiMob: Un corpus annoté d'histoire orale en suisse allemand

Yves Scherrer, LATL, CUI/Linguistique, UNIGE

En collaboration avec:
Tanja Samardžić, CorpusLab, UZH
Prof. Elvira Glaser, Deutsches Seminar, UZH

Le projet ArchiMob (www.archimob.ch)

Le projet ArchiMob (www.archimob.ch)

« **Archimob** (archives de la mobilisation) est une association pour la collecte et l'archivage des témoignages sur la période de la Deuxième Guerre mondiale en Suisse. Elle a été fondée en 1998 sur l'initiative du cinéaste Frédéric Gonseth. Elle comprend plus de quarante historiens et cinéastes indépendants issus de toute la Suisse [. . .] »

Le projet ArchiMob (www.archimob.ch)

« **Archimob** (archives de la mobilisation) est une association pour la collecte et l'archivage des témoignages sur la période de la Deuxième Guerre mondiale en Suisse. Elle a été fondée en 1998 sur l'initiative du cinéaste Frédéric Gonseth. Elle comprend plus de quarante historiens et cinéastes indépendants issus de toute la Suisse [. . .] »

- 555 interviews avec des témoins de la période 1939–1945.
- Interviews enregistrées sur vidéo, entre 1h et 2h par témoin.
- Les témoins proviennent des différentes régions linguistiques, de différents contextes sociaux et courants politiques.

Le corpus ArchiMob

Le département d'allemand de l'Université de Zurich a pu obtenir ces enregistrements en vue de leur étude linguistique et dialectologique.

Plusieurs étapes d'annotation sont en cours :

- Sélection d'enregistrements
- Transcription
- Alignement des transcriptions avec la bande sonore
- Normalisation
- Annotation grammaticale (parties du discours)

Applications envisageables avec les données annotées

Soutien technologique pour accélérer le processus d'annotation

Le corpus ArchiMob

Le département d'allemand de l'Université de Zurich a pu obtenir ces enregistrements en vue de leur étude linguistique et dialectologique.

Plusieurs étapes d'annotation sont en cours :

- Sélection d'enregistrements
- Transcription
- Alignement des transcriptions avec la bande sonore
- Normalisation
- Annotation grammaticale (parties du discours)

Applications envisageables avec les données annotées

Soutien technologique pour accélérer le processus d'annotation

Sélection et transcription

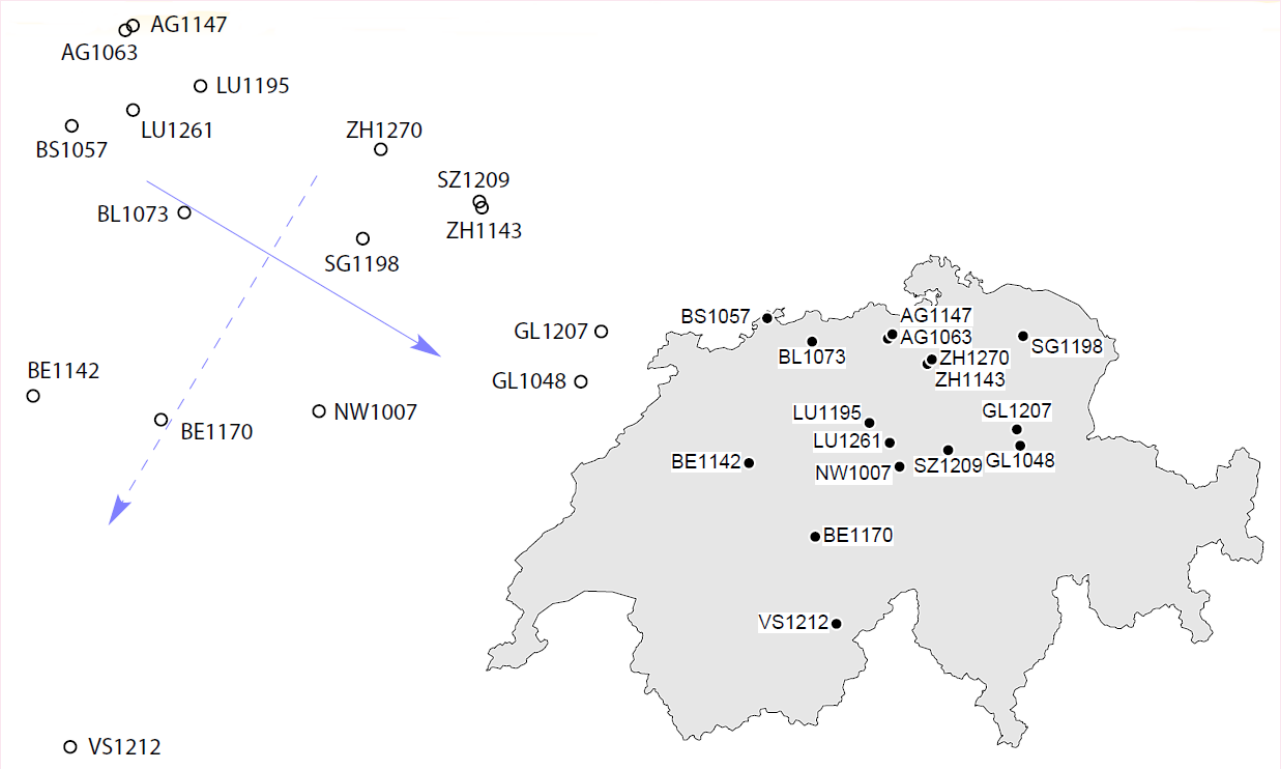
- Sélection de 45 enregistrements en différents dialectes alémaniques
- Transcription phonétique (Dieth) effectuée manuellement

Application : Mesures de similarités dialectales

Sélection et transcription

- Sélection de 45 enregistrements en différents dialectes alémaniques
- Transcription phonétique (Dieth) effectuée manuellement

Application : Mesures de similarités dialectales



Alignement des transcriptions avec la bande sonore

- Pris en charge directement par le programme de transcription

Application : Reconnaissance automatique de la parole

Développement d'un outil de reconnaissance de parole pour le suisse allemand en collaboration avec SPITCH, Zurich.

Alignement des transcriptions avec la bande sonore

- Pris en charge directement par le programme de transcription

Application : Reconnaissance automatique de la parole

Développement d'un outil de reconnaissance de parole pour le suisse allemand en collaboration avec SPITCH, Zurich.

SPITCH

Languages

We now support US-English, Swiss-German and Russian.

New languages are being added in line with market trends.

Normalisation

- Établissement d'une notation commune pour toutes les formes qui représentent « le même mot » à travers les textes :
  - Variation dialectale — hät, hed, hèt, hèd → hat
  - Variation intra-locuteur — mi, mii, miin, min → mein
  - Code-switching — main, määin, mäin → mein
- Normalisation manuelle des premiers documents
- Ensuite, validation manuelle de candidats proposés automatiquement

Soutien technologique

Utilisation de techniques de **traduction automatique** pour la proposition automatique de candidats de normalisation.

- Formes connues ambigües : modèles de langage
- Formes inconnues : traduction par caractères

Normalisation

- Établissement d'une notation commune pour toutes les formes qui représentent « le même mot » à travers les textes :
  - Variation dialectale — hät, hed, hèt, hèd → hat
  - Variation intra-locuteur — mi, mii, miin, min → mein
  - Code-switching — main, määin, mäin → mein
- Normalisation manuelle des premiers documents
- Ensuite, validation manuelle de candidats proposés automatiquement

Soutien technologique

Utilisation de techniques de **traduction automatique** pour la proposition automatique de candidats de normalisation.

- Formes connues ambigües : modèles de langage
- Formes inconnues : traduction par caractères

Normalisation

- Établissement d'une notation commune pour toutes les formes qui représentent « le même mot » à travers les textes :
  - Variation dialectale — hät, hed, hèt, hèd → hat
  - Variation intra-locuteur — mi, mii, miin, min → mein
  - Code-switching — main, määin, mäin → mein
- Normalisation manuelle des premiers documents
- Ensuite, validation manuelle de candidats proposés automatiquement

Soutien technologique

Utilisation de techniques de **traduction automatique** pour la proposition automatique de candidats de normalisation.

- Formes connues ambigües : modèles de langage
- Formes inconnues : traduction par caractères

Normalisation

- Établissement d'une notation commune pour toutes les formes qui représentent « le même mot » à travers les textes :
  - Variation dialectale — hät, hed, hèt, hèd → hat
  - Variation intra-locuteur — mi, mii, miin, min → mein
  - Code-switching — main, määin, mäin → mein
- Normalisation manuelle des premiers documents
- Ensuite, validation manuelle de candidats proposés automatiquement

Applications

- Dialectologie : Recherche de phénomènes lexicaux ou syntaxiques sans tenir compte de la variation phonétique
- Histoire : Recherche par mots-clés sans tenir compte de la variation dialectale

Annotation syntaxique (Parties du discours)

- Proposition automatique d'étiquettes, validation manuelle

Soutien technologique

Deux possibilités, selon les données disponibles actuellement :

- Utilisation d'un étiqueteur de l'allemand standard parlé (application aux formes normalisées)
- Utilisation d'un étiqueteur du suisse allemand écrit (application aux formes originales)

Applications

- Dialectologie : Recherche de phénomènes syntaxiques facilitée

Annotation syntaxique (Parties du discours)

- Proposition automatique d'étiquettes, validation manuelle

Soutien technologique

Deux possibilités, selon les données disponibles actuellement :

- Utilisation d'un étiqueteur de l'allemand standard parlé (application aux formes normalisées)
- Utilisation d'un étiqueteur du suisse allemand écrit (application aux formes originales)

Applications

- Dialectologie : Recherche de phénomènes syntaxiques facilitée