UNIVERSITÉ DE GENÈVE
FACULTÉ DES LETTRES
Département de linguistique

Un corpus annoté du suisse allemand
Construction et utilisation

FNSNF
FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE
Fondation Ernest Boninchi
FONDS GÉNÉRAL DE L'UNIVERSITÉ DE GENÈVE

Eric Haeberli et Manuela Schönenberger
Département de linguistique

Introduction

La recherche en linguistique peut se baser sur trois types de données empiriques:

(i) Les intuitions des locuteurs.
(ii) Les données élicitées dans un cadre expérimental.
(iii) Toute autre production langagière écrite ou orale.

Les données de type (iii) ont considérablement gagné en importance depuis que la production écrite peut être stockée en format numérique et que des outils informatiques facilitent la collecte des données. Des **corpus numériques** de différents types existent aujourd'hui pour un grand nombre de langues et leur histoire.

Mais pour certains domaines de recherche, les ressources numériques restent limitées.

- Pour la plupart des **dialectes**, typiquement non-écrits, il n'existe pas de corpus qui permettraient aux linguistes d'étudier leurs propriétés souvent très intéressantes.
- Les corpus ne tiennent souvent pas compte de facteurs tels que l'origine, l'âge, le sexe ou le statut social des locuteurs, facteurs qui sont cruciaux pour l'étude de la **variation linguistique**.
- Les corpus ne permettent généralement pas d'analyses **phonétiques ou phonologiques**.

Le but de ce projet est de contribuer à combler cette lacune avec la construction d'un **corpus annoté d'un dialecte suisse allemand** basé sur la production orale spontanée d'une sélection socialement équilibrée de locuteurs.

Le corpus pourra être utilisé pour des recherches dans tout domaine linguistique. Mais notre but principal est d'explorer des aspects d'un domaine qui est encore très peu étudié, la variation syntaxique parmi les locuteurs et dans l'usage d'un même locuteur (**'inter- and intra-speaker variation'**) et son interaction avec le **changement syntaxique**.

Construction du corpus

Le corpus est basé sur des productions orales spontanées enregistrées lors d'entretiens libres avec des locuteurs d'un dialecte spécifique du suisse allemand, celui parlé à Wil (SG).

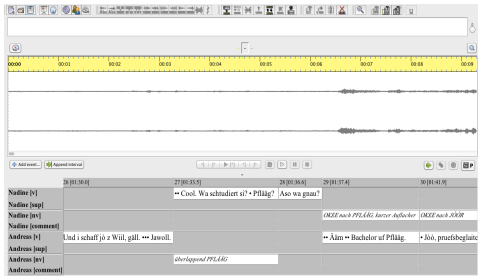
Etapes de la construction du corpus:

Interviews: 55 entretiens d'une durée moyenne de 90 minutes (plus de 80h d'enregistrements sonores).

- 59 locuteurs qui ont grandi à Wil et qui, dans la plupart des cas, y vivent encore.
- Tranches d'âge: 20-30, 45-55, 70-90; parité hommes/femmes; locuteurs provenant de différents milieux sociaux.

Transcription: Transcription des enregistrements sonores à l'aide du logiciel EXMARaLDA.

- Basé sur le système de transcription du suisse allemand de Dieth (1938/1986).
- Actuellement environ 850'000 mots.
- Le texte est lié au document sonore.



Graphique 1: Copie d'écran d'un passage transcrit

'Tokenization': Séparation des données en unités principales d'analyse syntaxique (phrases ou fragments de phrases).

- Ajout d'informations concernant: Disfluences ('faux-départs', répétitions), erreurs, catégories vides.

Etiquetage morpho-syntaxique ('POS tagging'): Ajout d'informations morpho-syntaxiques à chaque mot.

- Environ 80 'POS tags'.
- Liste basée sur un système d'annotation développé pour les *Penn Historical Corpora* (https://www.ling.upenn.edu/hist-corpora/).
- Etiquetage manuel pour un corpus d'entraînement d'environ 50'000 mots.
- Ensuite étiquetage automatique (BTagger).

Illustration:

(1)
```
I/PRO
waiss/VBP
nöd/NEG
,/PUNC
öb/C
{AUDIO:s superfluous}/CODE
*/PRO
di/PRO
magsch/VBP
erinnere/VB
?/PUNC
```

'Parsing': Annotation (semi-automatique) de la structure syntaxique.

- Environ 50 étiquettes supplémentaires.

Illustration:

(2)
```
( (IP-MAT (CODE Thea:)
          (FS (PRO Me) (DOP tot) (CODE s<slash>))
          (CODE ää)
          (NP-SBJ (D d) (NPRS Schwizer))
          (DOP tönd)
          (NP-OB1 (PRO s))
          (NEG nöd)
          (ADVP (ADV diräkt))
          (VB säge)
          (PUNC ,)
          (DIP jò)
          (PUNC .)))
```

Correction manuelle: Comparé à d'autres corpus, la taille de ce corpus est relativement petite. Pour obtenir le plus de données pertinentes que possible lors des collectes automatiques, il est essentiel que le nombre d'erreurs d'annotation soit minimisé.

Utilisation: Une étude de cas

La **collecte des données** peut se faire avec les fichiers EXMARaLDA (recherches lexicales; accès direct au fichier sonore) ou avec les fichiers annotés (recherches syntaxiques linéaires ou hiérarchiques à l'aide du logiciel CorpusSearch).

Une étude de cas (Schönenberger 2017)

- Dans les **questions indirectes** en suisse allemand, le **complémenteur** *dass* est parfois présent (3a) et parfois absent (3b).

(3) a. I waiss nöd, [$_{wh}$ *wie alt*] **dass** si isch.
Je sais pas comment vieille que elle est
'Je ne sais pas quel âge elle a.'
b. I waiss nüme, [$_{wh}$ *wie*] _ si hässt.
Je sais plus comment elle s'appelle

- Une analyse de 1'066 exemples montre que le statut de *dass* est lié au poids phonologique de l'élément *wh* qui précède: Dans la grande majorité des cas, *dass* est présent (+DFC) quand l'élément *wh* est polysyllabique (non-mono) et absent (-DFC) quand l'élément *wh* est monosyllabique (mono).

Tab. 4: Distribution of DFCs in wh-complement clauses (with mono- and non-monosyllabic wh-constituents) in spontaneous production data (St. Galler German)

Age groups	mono +DFC	mono -DFC	non-mono +DFC	non-mono -DFC
G1: young (n = 7)	12 (7.2%)	154 (92.8%)	66 (92.9%)	5 (7.1%)
G2: middle-aged (n = 15)	7 (3.6%)	190 (96.4%)	63 (91.3%)	6 (8.7%)
Interviewers (n = 2)	4 (1.3%)	305 (98.7%)	66 (89.2%)	8 (10.8%)
G3: elderly (n = 11)	1 (0.7%)	152 (99.3%)	23 (85.2%)	4 (14.8%)
Total (n = 35)	24 (2.9%)	801 (97.1%)	218 (90.4%)	23 (9.6%)

- Les exceptions à cette généralisation sont typiquement liées à des particularités prosodiques (accent sur *wh*, disfluence etc.).
- Ces observations suggèrent que la présence/absence de *dass* favorise certaines configurations d'intonation, et que les facteurs prosodiques jouent un rôle primordial dans cette variation.

Références

Dieth, E. 1986. Schwyzertütschi Dialäktschrift. Dieth-Schreibung. In C. Schmid-Cadalbert (ed.), *Lebendige Mundart*. Band 1. Aarau/Frankfurt am Main: Verlag Sauerländer.

Schönenberger, M. 2017. Are doubly-filled comps governed by prosody in Swiss German? The chameleonic nature of *dass* 'that'. In E. Aboh et al. (eds), *Elements of Comparative Syntax: Theory and Description*. Berlin: De Gruyter Mouton. 185-220.

Logiciels

BTagger: http://clcl.unige.ch/SOFTWARE.html
EXMARaLDA: http://www.exmaralda.org/
Corpus Search 2: http://corpussearch.sourceforge.net/