Exploring the LangAge Corpora (with LaBB-CAT)

Erick Velázquez Godínez, Julie Kairet, Annette Gerstenberg, Valerie Hekkel

University of Potsdam (Germany)

Universität Potsdam

About

LaBB-CAT is a software published with GNU's General Public License that stores Transcriber transcripts with its corresponding Praat TextGrids and sound files. LaBB-CAT is a linguistic research tool that allows for web-based management of corpora (Fromont & Hay 2012).

LaBB-CAT functionalities are:

- Administration of user access
- Management of different corpora
- Text mining via concorder and regular expressions
- PRAAT Interface
- Statistical reports on the corpus composition: number of tokens, transcripts, turns, utterances, etc.
- Availability of annotation layers, in progress: linguistic variables, tokenisation, POS-tagging

Fromont, R., & Hay, J. (2012). *LaBB-CAT: An annotation store.* Australian Language Technology Association Workshop. http://labbcat.sourceforge.net/

Orléans

a004a.trs - **a004:**
1. ☑ le **langage** courant
2. ☑ leur **langage** est

2005 first series 56 French participants	**2012** second series 34 follow-ups	**2015** third series 22 follow-ups + 22 new

French 70–105 y. Speakers

a005a.trs - **a999:**
7. ☑ le **langage**
a019a.trs - **a999:**
8. ☑ le **langage** ça
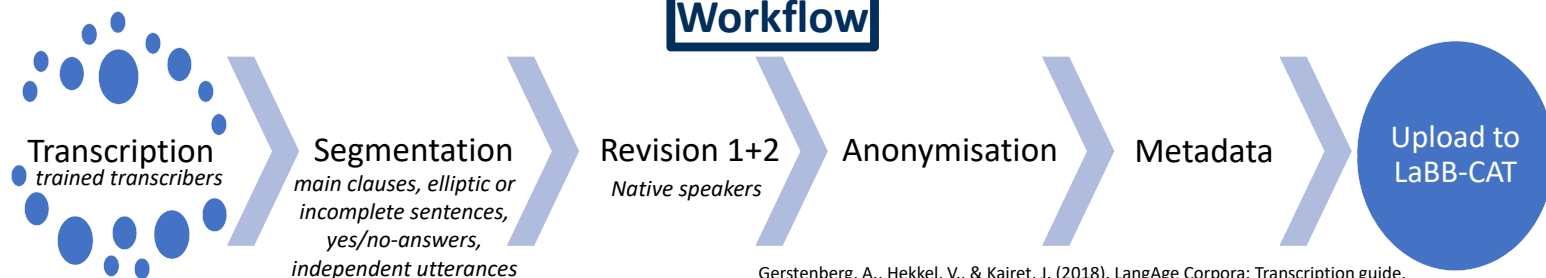☑ [select all 8 results]

CSV Export [options]

Extract Audio

Praat TextGrid ▼ Convert

2005: Biographical interviews, creative writing sessions
2012: Biographical interviews, storytelling
2015: Interviews: biography, language attitudes, later life, creative writing sessions, personality questionnaire

Workflow

Transcription *trained transcribers* › **Segmentation** *main clauses, elliptic or incomplete sentences, yes/no-answers, independent utterances* › **Revision 1+2** *Native speakers* › **Anonymisation** › **Metadata** › **Upload to LaBB-CAT**

Gerstenberg, A., Hekkel, V., & Kairet, J. (2018). LangAge Corpora: Transcription guide.
Anonymisation: Baude, O. (2006). *Corpus Oraux: Guide des bonnes pratiques*. Orléans: Presses Universitaires d'Orléans; CNRS Éditions.

Research

Julie Kairet: A longitudinal study of time management and pauses: case studies (PhD project)

Speaker	Age in 2005	Profession
046	71	Employee / Housewife
016	72	Teacher
049	74	Employee
015	78	Employee

Two samples (5 min) per speaker in 2005 and 2015: <u>40 min</u> of annotated speech

Establishment of prosodic profiles based on several variables:

- Speech rate (syll/sec)
- Articulation rate (syll/sec)
- Number and length of silent pauses, filled pauses and vowel lengthenings

Moving beyond the statement that, as compared to younger speakers, older speakers' speech rate slows down through time (Linville 2001, Bolly and al. 2016), the thesis focuses on showing the interconnection between different temporal variables to highlight how the speakers manage their time and how this time management evolves in later life.

Bolly, C., Christodoulides, G., & Simon, A. C. (2016). Disfluences et vieillissement langagier. De la base de données VALIBEL aux corpus outillés en français parlé. *Corpus [En ligne]*, 15.
Linville, S. E. (2001). *Vocal Aging*. San Diego: Singular Thomson Learning.

LangAge subsamples

- Language development in later life: prosody, sociolinguistic variables, lexical richness
- Couples over the lifespan
- Read aloud vs. spontaneous speech
- Retold stories

www.langage-corpora.org

Diachronic corpus construction using ESLO data

- Linguistic features and registers (**Erick Velázquez Godínez**)
- Sociolinguistics and language change: the causal construction *parce que* (**Valerie Hekkel**)

ESLO = Laboratoire Ligérien de Linguistique (LLL). (1968–1974; 2008–).
Enquêtes SocioLinguistiques à Orléans: [Corpus linguistique]. Université d'Orléans.
Gerstenberg, A. (2011). *Generation und Sprachprofile im höheren Lebensalter: Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews* (Analecta Romanica 76). Frankfurt am Main: Klostermann.