Designing a bilingual speech corpus for French and German language learners

Jürgen Trouvain[1], Yves Laprie[2], Bernd Möbius[1],
Bistra Andreeva[1], Anne Bonneau[2], Vincent Colotte[2],
Camille Fauth[2], Dominique Fohr[2], Denis Jouvet[2],
Odile Mella[2], Jeanin Jügler[1], Frank Zimmerer[1]

[1]Phonetics, Saarland University, Saarbrücken
[2]Équipe Parole, LORIA, Nancy

UNIVERSITÄT DES SAARLANDES

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Motivation [1]

- Project *Individualised feedback in computer-assisted spoken language learning (IFCASL)*
  - Supported by ANR and DFG (Deutsche Forschungsgemeinschaft)
  - For i) French learners of German and ii) German learners of French
  - What are the phonetic and phonological deviations of our learners?
    - personal and anecdotal experience
    - (theoretical) contrastive comparisons
    - not many corpus data available
- Aims of the corpus
  - Training and test material for automated feedback system
  - Data and analysis for phonological research

Motivation [2]

- Existing learner corpora
  - mainly for written language
  - mainly for English as target language
  - only a few parallel corpora for language pairs
  - available corpora:
    - HABLA (Hamburg Adult Bilingual LAnguage) corpus with bilinguals (L1: French & German) (Kupisch et al. 2012)
    - IPFC-allemand (Interphonologie du Français Contemporain) with L1: German L2: French (very advanced level) (Pustka 2012)
- Need for our corpus
  - Audio files
  - Annotations on the segmental and prosodic level
  - Availability

Features of the corpus: subjects

- Groups of subjects
  - 50 learners with L1: French and L2: German
  - 50 learners with L1: German and L2: French
- For each group
  - 10 teenagers (age 14-16 years) at beginners level (A1/A2)
  - 20 adults at beginners level (A1/A2)
  - 20 adults at advanced level (C1/C2)
- Subject acquisition
  - Subjects w/ L1: French in Nancy, w/ L1: German in Saarbrücken
  - Teenagers at various schools
  - Adults at University courses, Goethe-Institut

Legal and administrative matters

- Small remuneration for subjects
- Consent to be signed
  - Subject stays owner of the data
  - If wished data can be accessed by the owner at any time
  - Data can be used
    - in anonymous form for scientific purposes (oral and written)
    - for speech signal processing
    - for improvement of language learning software
- Access of data (audio + annotations + meta-data)
  - Data of subjects **not** for public use (except explicitally indicated)
  - on request for research purposes

Questionnaire

- Linguistic biography (in L1)
  - L1 and age (residence in first 16 years and in school time)
  - Highest educational degree
  - For each L2:
    - school time (years of instruction)
    - stay abroad
    - use (w/ partner, parents, tandem partners etc.)
    - certificates
- Self-assessment
  - Self-assessment of language skills, esp. pronunciation
  - Motivation
  - Attitude towards language learning
  - Opinion on learning languages with a computer

Recording sessions of the corpus

- Features
  - Read sentences and texts (no spontaneous speech)
  - To be read in **two** languages ("double parallel")
  - Good acoustic quality (quiet office)
  - Head-mounted close-talk microphone (nearly invisible for speaker)
  - Software "Corpus Recorder" (developed in Nancy)
    - Display of sentence to be read aloud
    - One sentence one audio signal file
  - Duration
    - Questionnaire: ~10 min.
    - Speech material: between 40 and 60 min.

Material [1]

- Four speaking conditions
  1. Sentence reading
     - Read aloud written sentences
  2. Sentence repetition
     - Read aloud sentences presented in written and spoken form (prerecorded with a native speaker)
     - Purpose: to exclude spelling-induced errors
  3. Focus sentences
     - Listen to a question, then read aloud the answer (also indicated by capitalised letters)
     - Purpose: to elicit variable locations of sentence accents
  4. Text reading
     - Read aloud written texts: i) informative, ii) narrative text

Material [2]

- Various blocks

	L2=FR*		L1=DE*	
task	no. of sent.	no. of words	no. of sent.	no. of words
sentence reading	25	183	51	359
sentence repetition	29	207	-	-
focus sentences**	24	291	25	144
text reading***	8 + 12	154 + 205	10 + 13	127 + 215
total	98	1040	101	845

* L1=FR, L2=DE no. of sentences/words vice versa

** versions of FR and DE very similar

*** here translations of the same text in both languages

Phenomena [1]

- Segmental level (selection)
  - Glottal stop [ʔ] and glottal fricative [h]
  - Liaison and enchaînement consonantique
  - Nasal vowels [ɛ̃, ã, õ]
  - Final devoicing of plosives and fricatives
  - Aspiration of unvoiced plosives [p^h, t^h, k^h]
  - ich- and ach-sound [ç, x]
  - Schwa: level of rounding and confusion with full vowel
  - /r/ as consonant [ʁ, ʀ] vs. vowel [ɐ]
  - Vowel length [iː-ɪ, eː-ɛ, ɛː-ɛ, aː-a, oː-ɔ, uː-ʊ, yː-ʏ, øː-œ]
  - Consonant clusters
  - Reductions

Phenomena [2]

- Prosodic
  - Word stress
  - Contrastive accent
  - Pitch range
- Unsure spelling-to-sound relationships
  - French "plus tard" as [plys], French "loup" as [lup]
- Internationalisms and cognates
  - French "énergie" read as German [enɛɐ'giː] in L2=FR
  - German "Berlin" read as French [bɛr'lɛ̃] in L2=DE
- Misreadings and influences of other L2s
  - German "Licht" [lɪçt] (Engl. "light") read as [laɪtʃ]

Examples

- Sentences contain:
  - minimal pairs, e.g. "Paar-Bar" or "pont-bon"
  - all phonemes of the given language at least once
  - "In jeder Bank gibt es eine Kasse."
    - Nasal vowel in "Bank"
    - Missing aspiration in "Kasse"
    - No vocalised /r/ in "jeder"
    - Rounded schwa in "es, eine, Kasse"
  - "Marie a rangé ses lunettes sans son étui."
    - Unclear nasal vowels in "rangé, sans, son"
    - Missing liaison in "son étui"
    - Inserted aspiration of [t] in "lunettes"
    - Inserted glottal stops in "Marie a" and "son étui"

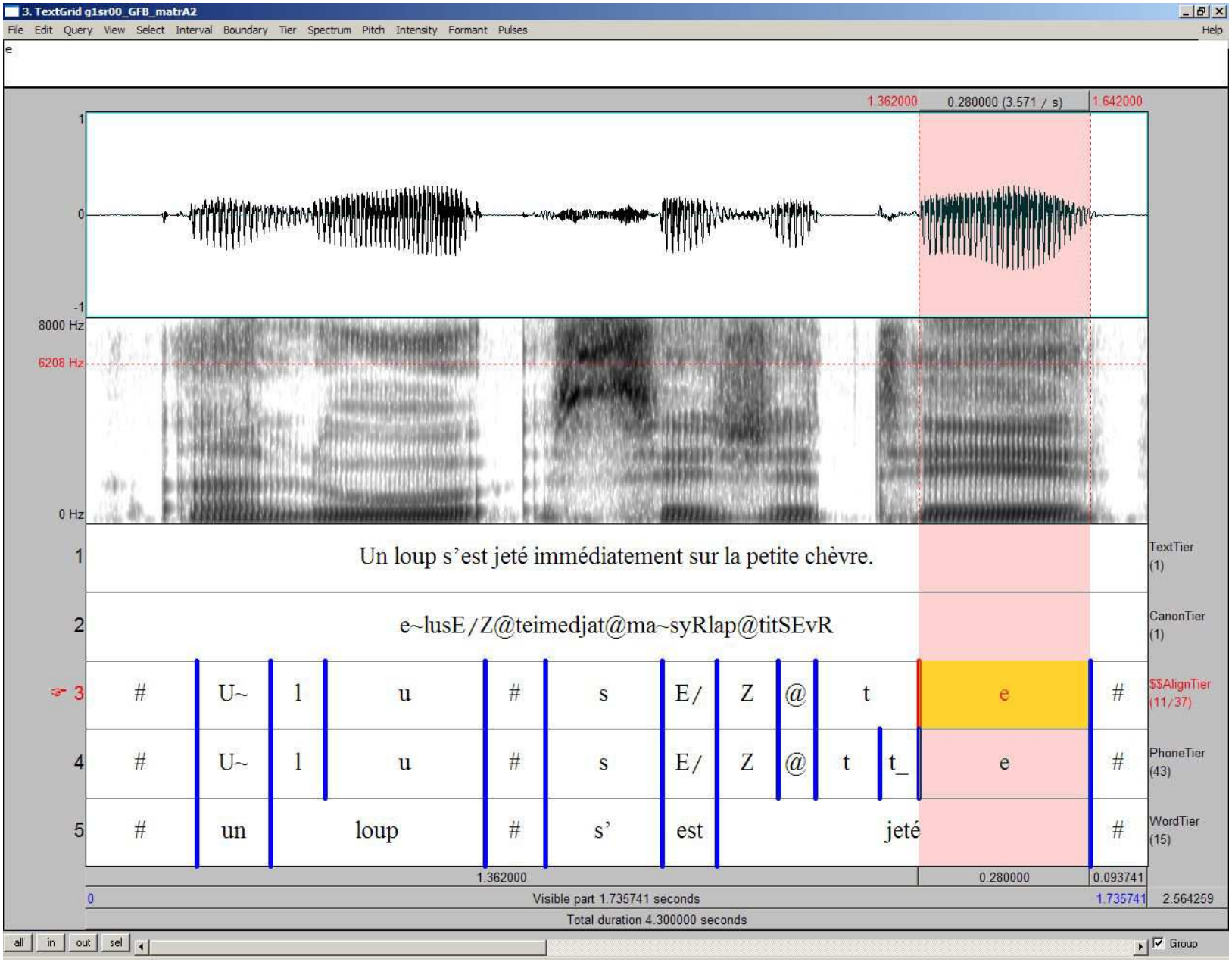
Examples [2]

- Focus sentences
  - (Der Fremde besucht Paris?) Der TOURIST besucht Paris.
  - (Der Tourist besucht München?) Der Tourist besucht PARIS.
  - (Der Tourist geht in Paris spazieren?) Der Tourist BESUCHT Paris.

Labelling [1]

- Signal file
- Annotation file
  - Sentence and word level (orthography)
  - Phone level in machine-readable phonetic notation (SAMPA):
    - Canonical (expected) form
    - Realised form
      - Substitutions
      - Insertions
      - Deletions
  - Prosodic level:
    - Deviations of word stress and sentence accent, if realised

Labelling: example



Questions

- What are **frequent** deviations?
  - Some deviations will occur more frequent than others
  - Some deviations will be pertinent also on the advanced level
- Which deviations are more **important** which are less?
  - Some deviations will lead to reduced intelligibility, others not
  - Some deviations will lead to a strong foreign accent
- Which deviations can be automatically **recognised** and **repaired**?
  - Some deviations will be easily recognised, others not
  - Some deviations can be "repaired", others not

Example application: Language teaching

- Focus on important mistakes
  - L2: German
    - Vowel length
    - Location of word stress
    - Schwa
- No focus on less important mistakes
  - L2: German
    - ich-sound [ç] vs. sch-sound [ʃ]

Example application: Automatic feedback

- L2: FR – Aspiration of unvoiced plosives
  - "… sur la petite chèvre."  

- L2: DE – Vowel length
  - "… Frühling fliegen Pollen durch die Luft."  
  - "… der schnellste Weg nach Polen ist."  

Thanks!

[mɛɐ'si bo'kʰuː]

[fiː'lœn 'dãk]