Designing a bilingual speech corpus for French and German language learners

C. Fauth[1], A. Bonneau[1], F. Zimmerer[2], J. Trouvain[2], B. Andreeva[2], V. Colotte[1], D. Fohr[1],
D. Jouvet[1], J. Jügler[2], Y. Laprie[1], O.Mella[1], B. Möbius[2]

[1]LORIA (CNRS/INRIA/UdL), Nancy, FRANCE
[2]SAARLAND University Germany

CORPUS AIMS AND CONTEXT

- Corpus: a bilingual speech corpus recorded by French **learners** of German and German **learners** of French in their native and second languages
  - four sub-corpora: French/German, German/French, French/French and German/German
  - size: 100 speakers (50 French, 50 German), and 120 sentences (60 F, 60 G) → 6 000 non-native and 6 000 native sentence realizations
  - Beginners and advanced speakers
- Existing learner corpora: mainly for written language and mainly for English as target language; only a few parallel corpora for language pairs
- Aims:
  - data and analysis for phonetic and phonological research
  - training and test material for automated feedback system
  - make the corpus available to scientific community (audio files annotated at segmental and prosodic levels)
- Project: *Individualised feedback in computer-assisted spoken language learning (IFCASL)*, supported by ANR and DFG (Deutsche Forschungsgemeinschaft)

SELECTED SPEECH PHENOMENA

Speech phenomena of interest for the French/German pair (non exhaustive list), covering segmental and prosodic levels as well as spelling problems

- Example: French sentence "Le bateau est arrivé au port" (the boat arrived at the port) See Figure1

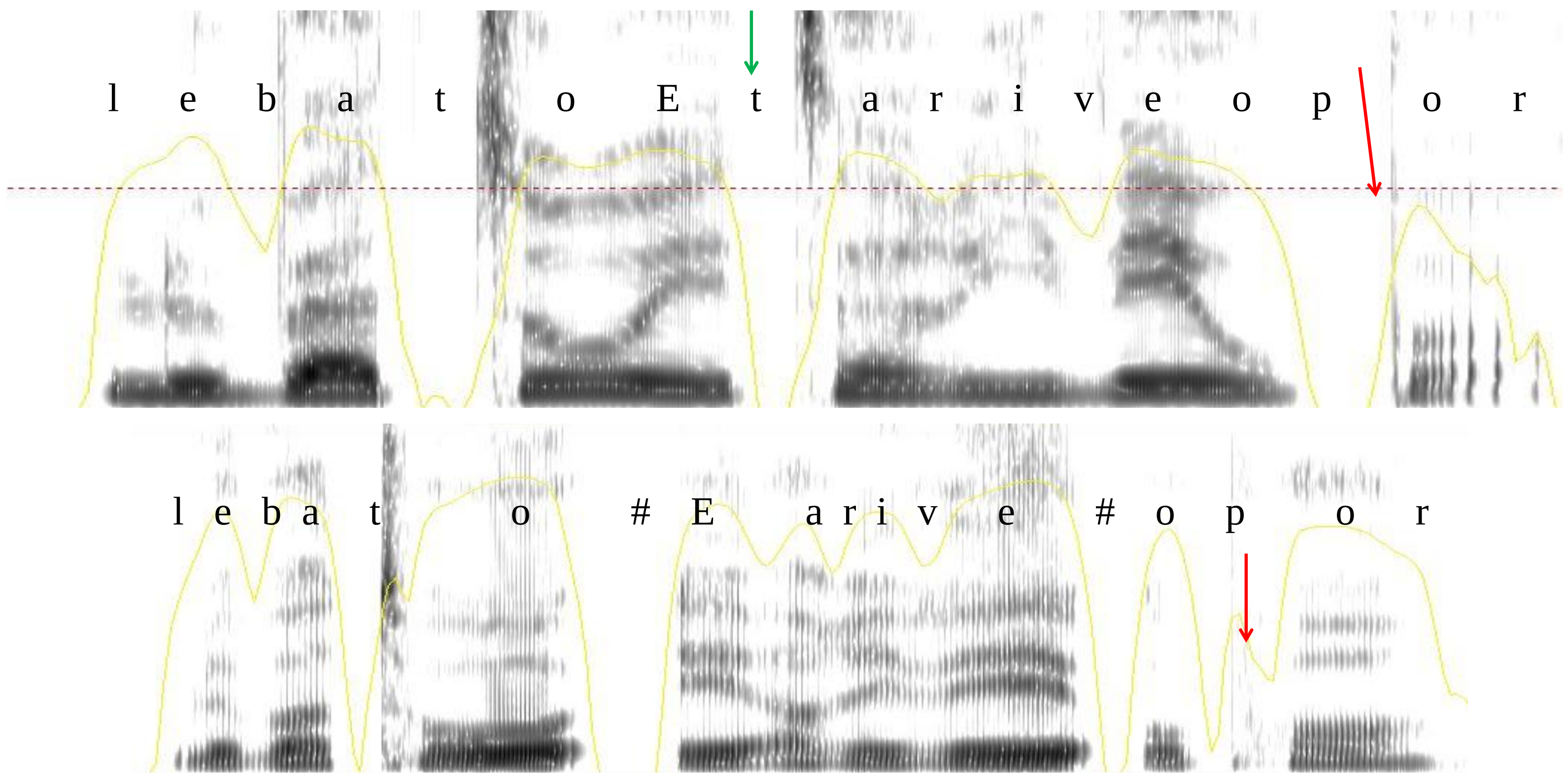

Figure 1. French sentence uttered by an advanced German speaker (top) and a beginner (bottom). The red arrows indicate the presence of aspiration on [p] for the beginner and its absence for the advanced speaker (there is no aspiration in French). The advanced speaker realizes the mandatory liaison (see the green arrow) between « est » and « arrivé ». The symbol # indicates pauses (only realized by the beginner here).

- *Speech sounds either not present in French or German, or differently realized, special phonological processes, different mapping between acoustic cues and prosodic units:*
  - German glottal fricative and glottal stop [h, ʔ], ich & ach-sound [ç, x]
  - French nasal vowels [ɛ̃ ã, ɔ̃], /r/ as consonant [ʁ, ʀ] vs. vowel [ɐ]
  - Schwa: level of rounding and confusion with full vowel
  - Final devoicing of plosives and fricatives in German
  - Aspiration of unvoiced plosives [p, t, k] in German
  - Vowel length (long *vs.* short vowels in German)
  - Liaisons in French
  - Rythm, contrastive accent and lexical stress accents

- *Spelling and cognates:*
  - Spelling-to-sound relationships: French "loup" [lu] as [lup]
  - Cognates: French "énergie" read as German [enɛɐ'giː] in L2=FR
  - Internationalisms: German "Berlin" read as French [bɛrlɛ̃] in L2=DE

LINGUISTIC MATERIAL

- Four speaking conditions:

(1) Sentence reading
(2) Sentence repetition: to exclude spelling-induced errors
(3) Focus sentences : to elicit variable locations of sentence accents
(4) Text reading : read aloud a small written text ("the three little pigs")

- Sentences (conditions 1-2) designed to contain:
  - all target speech phenomena
    - liaisons and sentence traps only in sentence reading condition
  - all speech sounds of a given language
  - quasi-minimal pairs in order to observe speech contrasts (e.g. opposition between short and long vowels) in similar contexts
  - set of sentences covering a given phenomenon for a series of sounds (e.g. voicing for stops) in variable contexts

RECORDING and LABELLING

- Recordings: High-quality recordings, using the software JCorpusRecorder (see raweb.inria.fr). Headset microphone (AKG C520) and Audiobox (M Audio Fast track).
- Six labelling tiers and a comment tier (Praat software) see Figure 2:
  - Phone level in machine-readable phonetic notation (SAMPA):
    - realised form as detected by (LORIA) automatic alignment: AlignTier
    - realised form (manual correction of AlignTier): RealTier
    - canonical form (what is expected): CanonTier
    - word, sentence (TextTier), prosodic Tier (absent in Fig.2) and comments
- Annotations:
  - insertions, deletions, substitutions,
  - special phonetic phenomena (fine phonetic transcription): assimilation of voicing, glottalisation....

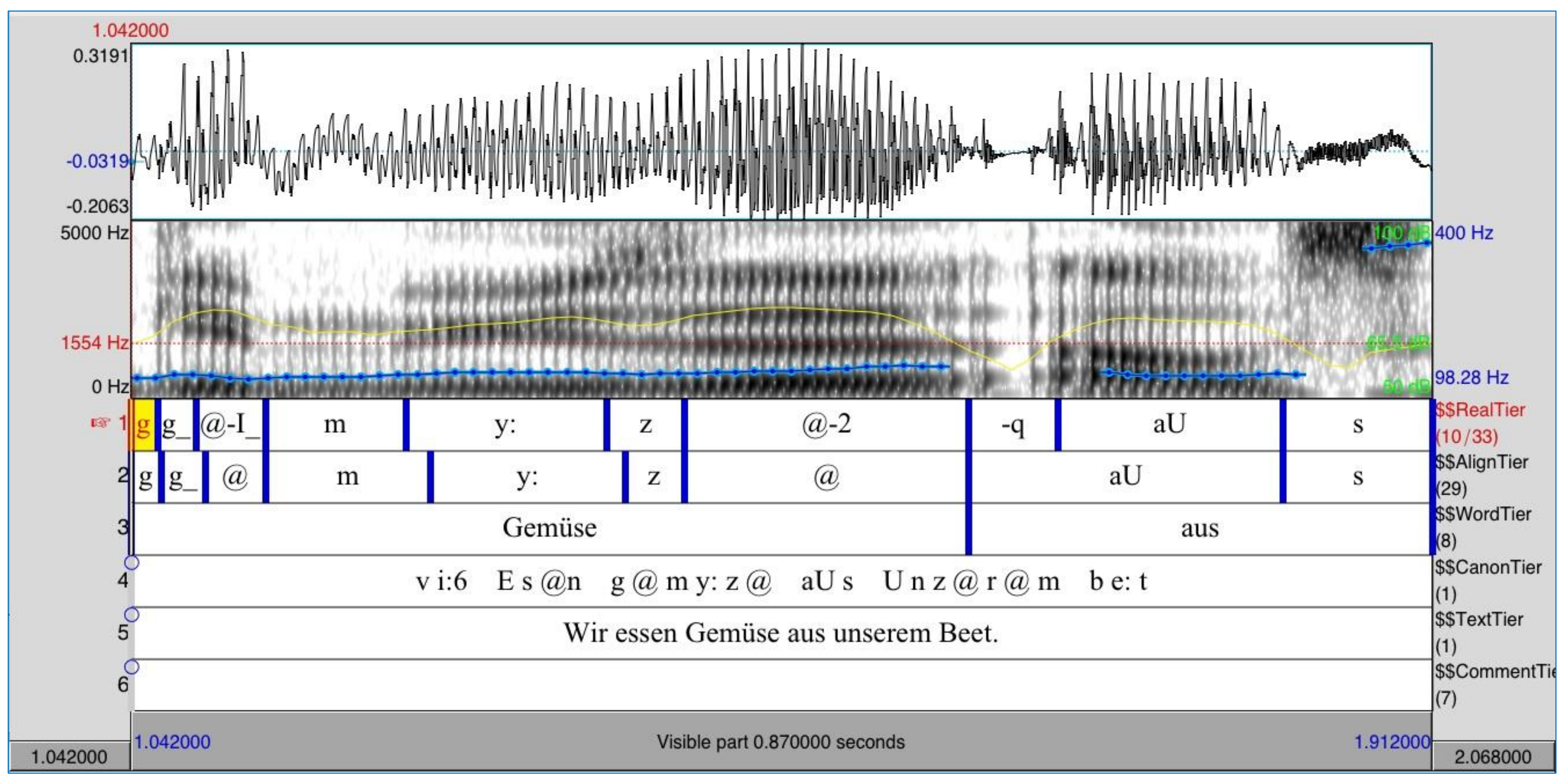

Figure 2. Automatic alignment and corrections of a German sentence produced by a French speaker learning German (FG) "Wir essen Gemüse aus unserem Beet" "We eat vegetables from our patch"

Language	Read	Rep.	Focus	Texte
French	31	29	7	1
German	31	29	7	1

Subjects	Beg. T	Beg. A.	Adv. A.	Total
French	10	20	20	50
German	10	20	20	50

Corpus size: number of sentences for each language (left) and number of subjects (left). Rep. for repeated, Beg. for beginners, Adv. for advanced and T. for teenagers

METHODOLOGY: a two step process to constitute the corpus, choose phenomena of interest and their distribution in sentence conditions

(1) a bilingual corpus with few speakers (14) including all sounds of each language and all speech phenomena of potential interest was recorded and analysed

- Its analysis revealed/confirmed:
  - the existence of special strategies due to sentence reading and sentence listening conditions
  - the importance of recording duration (the whole corpus should not last more than one hour to avoid subjects' fatigue)
  - the frequence and importance of some mispronunciations (voicing problems, erroneous presence (or absence) of /h/ for German (or French) non-native speakers, rhythm ...)

(2) Constitution of the final corpus, presented here, which puts a focus to the problems revealed by the preliminary corpus