

Automatic diagnosis and feedback for lexical stress errors in non-native speech: Towards a CAPT system for French learners of German

Anjana Sofia Vakil



UNIVERSITÄT
DES
SAARLANDES

Department of Computational Linguistics and Phonetics
University of Saarland, Saarbrücken, Germany

Master's Thesis Colloquium
16 April 2015

Some syllable(s) in a word more accentuated/prominent¹

um·FAHR·en	vs.	UM·fahr·en
<i>to run over</i>		<i>to drive around</i>

- ▶ German: variable stress placement, contrastive stress¹
- ▶ French: no word-level stress, final syllable lengthening²

Goal: Computer-Assisted Pronunciation Training (CAPT) for lexical stress errors for French learners of German

¹A. Cutler. "Lexical Stress". In: *The Handbook of Speech Perception*. Ed. by D. B. Pisoni and R. E. Remez. 2005, pp. 264–289.

²M.-C. Michaux and J. Caspers. "The production of Dutch word stress by Francophone learners". In: *Proc. of the Prosody-Discourse Interface Conference (IDP)*. 2013, pp. 89–94.

Motivation

Lexical stress errors by French learners of German

- Annotation of a learner speech corpus

- Inter-annotator agreement

- Frequency & distribution of errors

Error diagnosis

- Word prosody analysis

- Diagnosis by comparison

- Diagnosis by classification

Feedback

- Implicit

- Explicit

- Self-assessment

The de-stress CAPT tool

Figure: Criteria for selecting errors to target in a CAPT system.



Lexical stress errors seem to be:

- ▶ Frequently produced by French learners of variable-stress languages^{1,2}
- ▶ More important for intelligibility in L2 German than other types of errors³
- ▶ Possible to identify automatically by comparison¹ or classification⁴

¹A. Bonneau and V. Colotte. “Automatic Feedback for L2 Prosody Learning”. In: *Speech and Language Technologies*. Ed. by I. Ipsic. InTech, 2011.

²M.-C. Michaux. “Exploring the production and perception of word stress by French-speaking learners of Dutch”. In: *Workshop on Crosslinguistic Influence in Non-Native Language Acquisition*. 2012.

³U. Hirschfeld. *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Vol. 57. Forum Phonetikum. 1994.

⁴Y.-J. Kim and M. C. Beutnagel. “Automatic assessment of American English lexical stress using machine learning algorithms”. In: *SLaTE*. 2011, pp. 93–96.

- ▶ How reliably can human annotators identify errors in learner utterances?
- ▶ How frequently are errors actually produced by French learners of German?

Data: IFCASL corpus of French-German L1/L2 speech¹

- ▶ German utterances by French and German speakers
 - Adults (>18) and children (15-16)
 - Levels A2, B1, B2, C1 (children all A2/B1)
- ▶ Word- and phone-level segmentations
(syllable level added automatically)
- ▶ Selected 12 word types (bisyllabic, initial stress)

Dataset for annotation:

668 word utterances by 55-56 L1 French speakers

¹C. Fauth et al. “Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process”. In: *9th Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland, 2014, pp. 1477–1482.

15 Annotators, varying by:

- ▶ Native language (L1):
 - 12 German
 - 2 English (US)
 - 1 Hebrew
- ▶ Phonetics/phonology expertise:
 - 2 Experts
 - 10 Intermediates
 - 3 Novices

Each annotated 3 word types in one ~15 min. session
(1 annotator did 6 word types in 2 sessions)

Figure: Praat annotation tool

tragen
526

play word

play sentence

stress is on CORRECT syllable

stress is on INCORRECT syllable

no clear stress / I can't tell

wrong number of syllables

problem with audio

Figure: Praat annotation tool

tragen
526

play word

play sentence

[correct]	stress is on CORRECT syllable
[incorrect]	stress is on INCORRECT syllable
[none]	no clear stress / I can't tell
[bad_nsylls]	wrong number of syllables
[bad_audio]	problem with audio

How reliably can human annotators identify errors in learner utterances?

- ▶ Agreement calculated for each overlapping pair
- ▶ Quantified by:
 - Percentage agreement: $N_{\text{agreed}}/N_{\text{both annotated}}$
 - Cohen's Kappa¹ (κ): accounts for chance agreement
- ▶ Overall agreement represented by mean, minimum, median, and maximum of all pairwise values

¹J. Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (Apr. 1960), pp. 37–46.

Table: Overall pairwise agreement between annotators

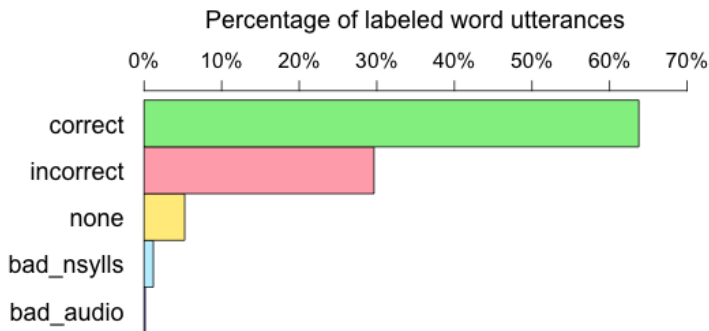
	% Agreement	Cohen's κ
Mean	54.92%	0.23
Maximum	83.93%	0.61
Median	55.36%	0.26
Minimum	23.21%	-0.01

- ▶ Rather low agreement (“fair”¹ mean κ)
- ▶ Large variability between annotators
- ▶ Not explained by L1/expertise groups

¹J. R. Landis and G. G. Koch. “The measurement of observer agreement for categorical data.” In: *Biometrics* 33.1 (1977), pp. 159–174.

Need a single label for each utterance to analyze error frequency & evaluate automatic diagnosis

- ▶ 268 utterances: no disagreement
- ▶ 265 utterances: majority vote
- ▶ remaining 135 utterances decided by rules, e.g.:
 - favor Expert judgments
 - favor certainty ([correct],[incorrect]) over [none]
 - be generous to learners if [correct] vs. [incorrect]



- ▶ Large variability across word types
- ▶ Beginners made more errors (vs. advanced)
- ▶ Children made more errors (vs. adult beginners)

[TODO]

[TODO]

[TODO]

[TODO]

[TODO]

[TODO]

[TODO]



UNIVERSITÄT
DES
SAARLANDES

[TODO]