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

UNIVERSITÄT
DES
SAARLANDES

Some syllable(s) in a word more accentuated/prominent[1]
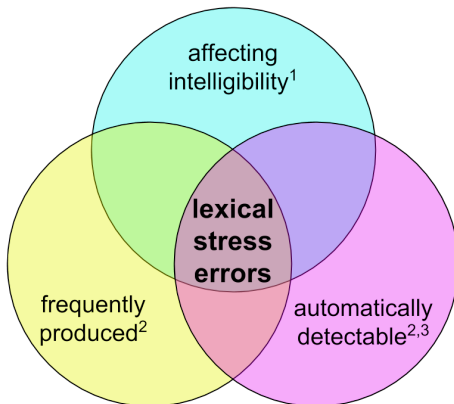
> um·FAHR·en   vs.   UM·fahr·en
> *to run over*   *to drive around*

- German: variable stress placement, contrastive stress[1]
- French: no word-level stress, final syllable lengthening[2]

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

---

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

[2] M.-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.

Lexical stress errors in CAPT

Venn diagram with three overlapping circles labeled "affecting intelligibility[1]", "frequently produced[2]", and "automatically detectable[2,3]", intersecting at the center labeled "lexical stress errors".

---

[1] U. Hirschfeld. *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Vol. 57. Forum Phoneticum. 1994

[2] A. Bonneau and V. Colotte. "Automatic Feedback for L2 Prosody Learning". In: *Speech and Language Technologies*. Ed. by I. Ipsic. InTech, 2011

[3] Y.-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?*

[TODO Explain/remove abbreviations]

Data: IFCASL corpus of French-German L1/L2 speech[1]

- 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

---

[1]C. 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

UNIVERSITÄT
DES
SAARLANDES

15 Annotators, varying by:
  ► Native language (L1):
    • 12 German
    • 2 English (US)
    • 1 Hebrew
  ► Phonetics/phonology
    expertise:
    • 2 Experts
    • 10 Intermediates
    • 3 Novices

Task: label utterances of 3
word types

Error annotation

15 Annotators, varying by:

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

Task: label utterances of 3 word types

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

Error annotation

15 Annotators, varying by:

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

Task: label utterances of 3 word types

Praat annotation tool:

tragen
526

| play word |
| play sentence |

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

How reliably can human annotators identify errors in learner utterances?

- ▶ Agreement calculated for each pair of annotators who labeled the same utterances
- ▶ Quantified by:
  - Percentage agreement: N agreed/N both annotated
  - Cohen's Kappa[1] (κ): accounts for chance agreement

---

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

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

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

UNIVERSITÄT
DES
SAARLANDES

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 among annotators, not explained by L1/expertise
- ▶ Single gold-standard label selected for each utterance

---

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

How frequently are errors actually produced by French learners of German?

How frequently are errors actually produced by French learners of German?
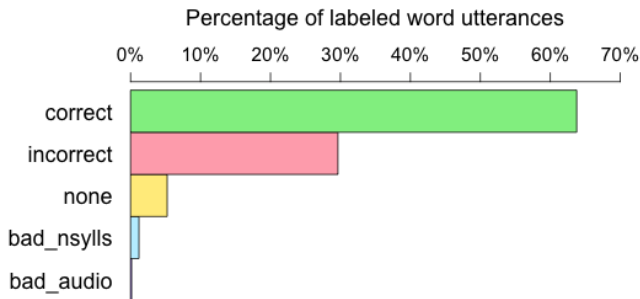
Percentage of labeled word utterances

How frequently are errors actually produced by French learners of German?

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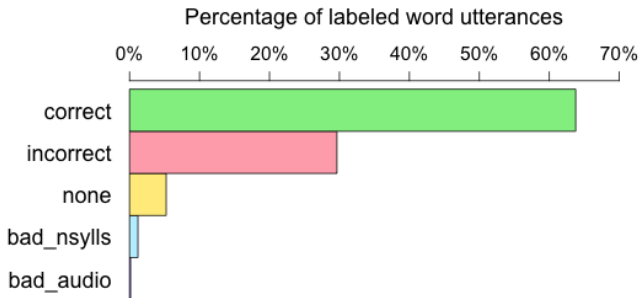
- ▶ Large variability across word types
- ▶ Beginners made more errors (vs. advanced)
- ▶ Children made more errors (vs. adult beginners)

Requires word, syllable, and phone segmentations

- ▶ Automatically produced via forced alignment[1]
- ▶ This work uses existing IFCASL segmentations
- ▶ Syllable segmentations derived from words & phones

[TODO segmentation screenshot]

---

[1]L. Mesbahi et al. "Reliability of non-native speech automatic segmentation for prosodic feedback." In: *SLaTE*. 2011.

Duration (DUR)

- ▶ Perceptual correlate: length/timing
- ▶ Best indicator of German stress[1]
- ▶ Simple to extract from segmentations
- ▶ Features: Relative syllable & nucleus (vowel) lengths

[1] G. Dogil and B. Williams. "The phonetic manifestation of word stress". In: *Word Prosodic Systems in the Languages of Europe*. Ed. by H. van der Hulst. Berlin: Walter de Gruyter, 1999. Chap. 5, pp. 273–334.

Fundamental frequency (F0)

- ▶ Perceptual correlate: pitch
- ▶ 2nd best indicator of stress after duration[1]
- ▶ Pitch contours computed using JSnoori[2,3]
- ▶ Features: relative syllable & nucleus:
  - • Mean F0 (in voiced segments)
  - • Maximum F0
  - • Minimum F0
  - • F0 range (max−min)

---

[1] G. Dogil and B. Williams. "The phonetic manifestation of word stress". In: *Word Prosodic Systems in the Languages of Europe*. Ed. by H. van der Hulst. Berlin: Walter de Gruyter, 1999. Chap. 5, pp. 273–334.

[2] jsnoori.loria.fr

[3] J. Di Martino and Y. Laprie. "An efficient F0 determination algorithm based on the implicit calculation of the autocorrelation of the temporal excitation signal". In: *EUROSPEECH*. Budapest, Hungary, 1999, p. 4.

Intensity (INT)

- ▶ Perceptual correlate: loudness
- ▶ Worse predictor than DUR or F0, but still may have effect on stress perception[1]
- ▶ Energy contours computed using Jsnoori
- ▶ Features: relative syllable & nucleus:
  - Mean energy (over 60dB "silence threshold")
  - Maximum energy

---

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

**[TODO Slide previewing comparison vs. classification?]**

Diagnosis by comparison

Comparison to a single reference utterance

| Learner (L2) utterance | ⟷ | Reference (L1) utterance |
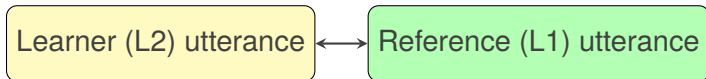
- ▶ Simplest approach, common in CAPT
- ▶ JSnoori (and predecessors) use this method[1]
  - • Assigns 3 scores (DUR, F0, INT)
    - ▶ Same syllable stressed?
    - ▶ Difference between stressed/unstressed syllables similar enough?
  - • Overall score = weighted average of 3 scores
- ▶ Problem: extremely utterance-dependent!

---

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

Diagnosis by comparison

Comparison to multiple reference utterances
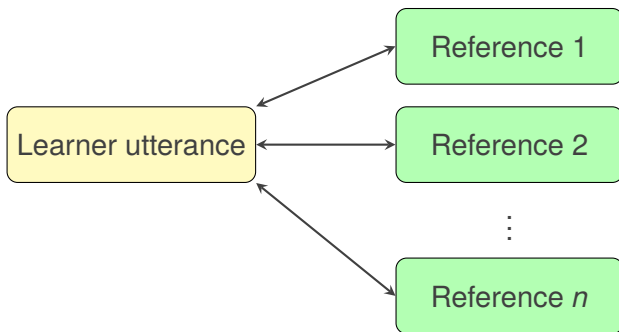
- ▶ Less common in CAPT systems
- ▶ Less utterance-dependent than single comparison
- ▶ Overall score = average of one-on-one scores

Options for selecting reference speaker(s)

- ▶ Manually
  - Learner's choice
  - Teacher/researcher's choice
- ▶ Automatically
  - May be more effective to choose reference speaker most closely resembling the learner[1]
  - Selected by comparing speakers' F0 mean and range (using all available recordings)

---

[1]K. Probst et al. "Enhancing foreign language tutors - In search of the golden speaker". In: *Speech Communication* 37.3-4 (July 2002), pp. 161–173.

- More abstract representation of L1 pronunciation
- Not yet explored for German CAPT

Research questions:

- *How well can lexical stress errors be classified?*
- *How does that compare with human agreement?*
- *Which features are most useful for classification?*

Experiments:

- ► Trained CART classifiers using WEKA toolkit[1]
- ► Used annotated L2 dataset for training/test data (gold-standard labels)
- ► Used L1 utterances of the same words as training data (all automatically labeled [correct])

Evaluated in terms of:

- ► Agreement (%, κ) with gold-standard labels
- ► Precision, Recall, F_1 and F_2 for [correct] class **[TODO explain and/or put on handout]**

---

[1] www.cs.waikato.ac.nz/ml/weka

UNIVERSITÄT
DES
SAARLANDES

Which features are most useful for classification?

Feature set	Description
DUR	Duration features
F0	Fundamental frequency features
INT	Intensity features
WD	Uttered word (e.g. *Tatort*)
LV	Speaker's L2 German skill level (A2|B1|B2|C1)
AG	Speaker's age/gender (Girl|Boy|Woman|Man)

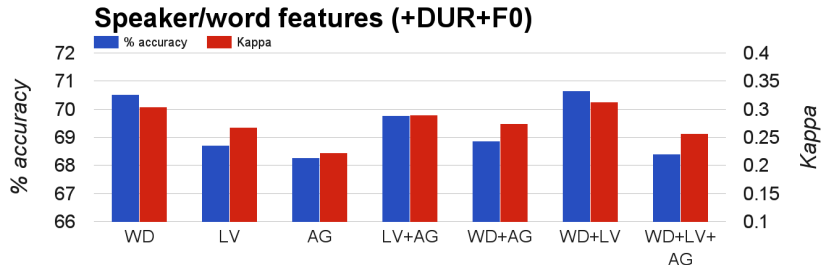*How well can lexical stress errors be classified?*

Prosodic features
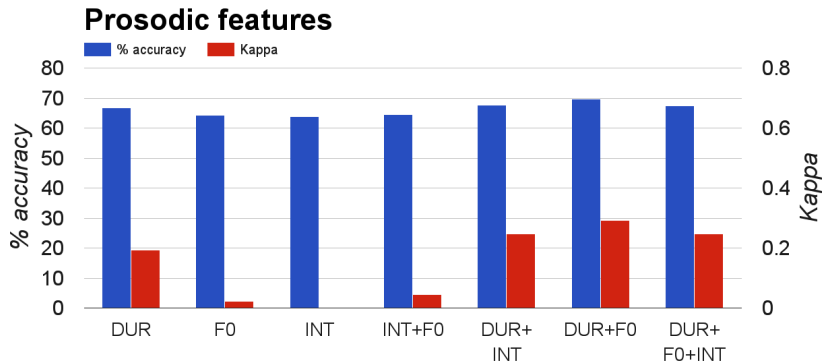
How well can lexical stress errors be classified?
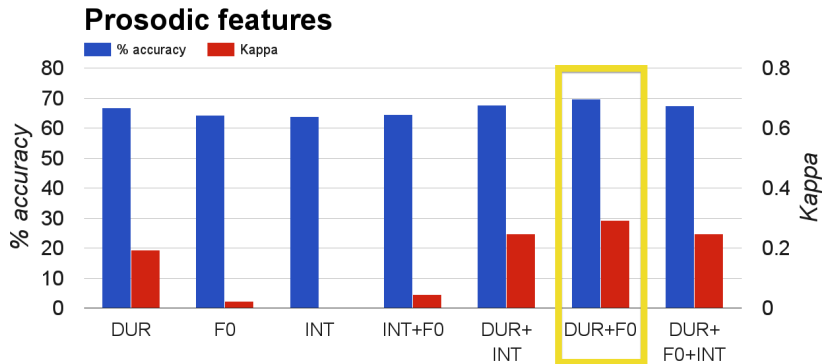
Prosodic features

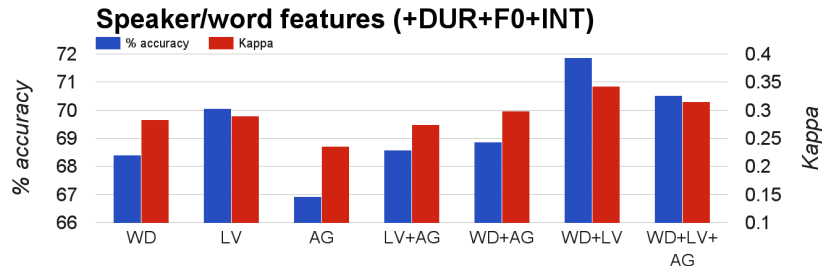Best performance using only prosodic features: DUR+F0

- % Accuracy: 69.77%
- κ: 0.29

How well can lexical stress errors be classified?

Speaker/word features (+DUR+F0)

How well can lexical stress errors be classified?

Speaker/word features (+DUR+F0+INT)

How well can lexical stress errors be classified?

Speaker/word features (+DUR+F0+INT)
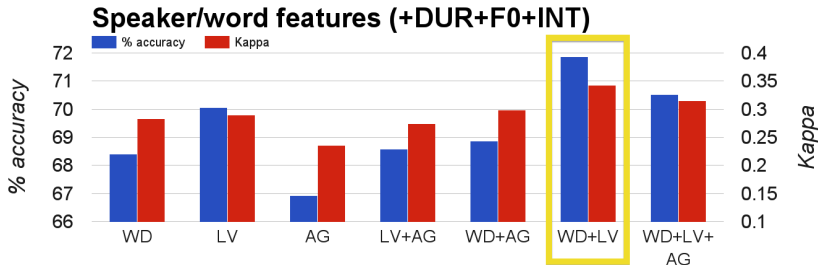
Best performance overall: WD+LV+DUR+F0+INT

- % Accuracy: 71.87%
- κ: 0.34

*How does classification accuracy compare
with human agreement?*

	% agreement	κ
Best classifier vs. gold standard	71.87%	0.34
Mean human vs. human	54.92%	0.23

- ▶ Results are encouraging in this context
- ▶ Still want better performance for real-world use

Allows learner to notice features of their utterance/reference utterance, without explicitly evaluating their pronunciation

Im **Frühling** fliegen Pollen durch die Luft.

Your utterance:

2SR23_FGWB1_536_frühling

Duration (width): 58.0% of word
Pitch (height): 100.0% of mean
Intensity (darkness): 0.54% of mean

Früh ling

Reference utterance 1:

2SR23_GGMC1_034_frühling

Download

Früh ling

Directly calls learner's attention to error(s) and/or offers corrective instruction

Your scores

Duration		3/10

I think you pronounced an incorrect number of phones in at least one of the word's syllables.

Pitch		10/10

Your pitch was pitch-perfect, great job!

Loudness		6/10

The correct syllable is loudest, good job! But it should be even louder compared to the unstressed syllable.

Overall		5/10

Your overall score is the weighted average of your Duration (60%), Pitch (30%), and Loudness (10%) scores.

May be linked to progress and motivation[1]

Self-assessment

Listen to your utterance and the reference utterance(s).

Then answer these questions:

Which syllable did you stress?

- ○ The first syllable (correct)
- ○ The second syllable (incorrect)
- ○ Neither syllable (incorrect)

Is the stress as clear in your utterance as it is in the reference utterance?

- ○ Just as clear as in reference
- ○ Not as clear as in reference
- ○ I don't know

What could you work on for next time?

[ ]

Continue

---

[1] A. Neri et al. "The pedagogy-technology interface in computer assisted pronunciation training". In: *Computer Assisted Language Learning* (2002).

UNIVERSITÄT
DES
SAARLANDES
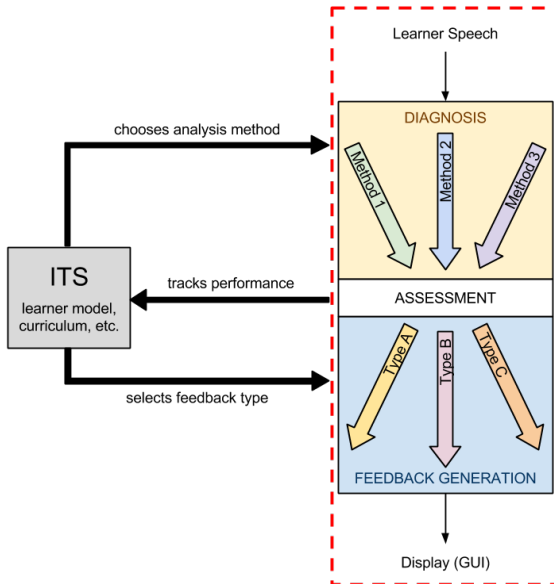
de-stress

🏠 Home   📄 Exercise List

Create Exercise

Name * `Comparison-StyleText`

Description *
```
This exercise uses a simple one-on-
one comparison method and
delivers feedback via stylized text.
Learners are asked to self-assess
before feedback is delivered.
```

Word * `fliegen ▼`

Diagnosis Method * `SimpleComparison-1refs-MANUAL ▼`

Feedback Method `TextStylization-SelfAssessed ▼`

Lessons

📋 Create

UNIVERSITÄT
DES
SAARLANDES

Create DiagnosisMethod

Name * `SimpleComparison`

Description `Single ref. comparison`

Scorer * `Comparison ▾`

Number Of References * `1`

Selection Type `MANUAL ▾`

🗐 Create

Create FeedbackMethod

Name * `TextStylization-SelfAsses`

Description ` `

Requires Scorer Type `▾`

Show Skill Bars ☐

Play Feedback Signal ☐

Display Shapes ☐

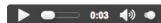Style Text ☑

Display Messages ☐

Self Assessment ☑

🗐 Create

[TODO]