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

Anjana Sofia Vakil

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

Master’s Thesis Colloquium
16 April 2015
Lexical stress

Some syllable(s) in a word more accentuated/prominent\(^1\)

- **German**: variable stress placement, contrastive stress\(^1\)
  um·FAHR·en vs. UM·fahr·en
  *to run over* vs. *to drive around*

- **French**: no word-level stress, final syllable lengthening\(^2\)

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

---


Lexical stress errors in CAPT


Lexical stress errors by French learners of German
  Annotation of a learner speech corpus
  Inter-annotator agreement
  Frequency & distribution of errors

Diagnosis methods
  Word prosody analysis
  Diagnosis by comparison
  Diagnosis by classification

Feedback methods

de-stress: A prototype CAPT tool

Conclusion
Lexical stress errors by French learners of German
  Annotation of a learner speech corpus
  Inter-annotator agreement
  Frequency & distribution of errors

Diagnosis methods
  Word prosody analysis
  Diagnosis by comparison
  Diagnosis by classification

Feedback methods

de-stress: A prototype CAPT tool

Conclusion
Lexical stress errors in learner speech

- 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 speech\(^1\)

- German utterances by French and German speakers
  - Adults (\(>18\)) and children (15-16)
  - Levels\(^2\) 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 German word utterances by \(\sim\)55 French speakers


\(^2\)Common European Framework of Reference, www.coe.int/lang-CEFR
15 Annotators, varying by:

- Native language (L1):
  - 12 German
  - 2 English (US)
  - 1 Hebrew

- Phonetics/phonology expertise:
  - 2 Experts
  - 10 Intermediates
  - 3 Novices
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
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
  - 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
  - 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_nsyls]
- problem with audio [bad_audio]
Inter-annotator agreement

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: \( \frac{N \text{ agreed}}{N \text{ both annotated}} \)
  - Cohen’s Kappa\(^1\) (\( \kappa \)): accounts for chance agreement

---

Inter-annotator agreement

Overall pairwise agreement between annotators

<table>
<thead>
<tr>
<th></th>
<th>% Agreement</th>
<th>Cohen’s $\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>54.92%</td>
<td>0.23</td>
</tr>
<tr>
<td>Maximum</td>
<td>83.93%</td>
<td>0.61</td>
</tr>
<tr>
<td>Median</td>
<td>55.36%</td>
<td>0.26</td>
</tr>
<tr>
<td>Minimum</td>
<td>23.21%</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

- Rather low agreement (“fair”\(^1\) mean $\kappa$)
- Large variability among annotators, not explained by L1/expertise
- Single gold-standard label selected for each utterance

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

- Large variability across word types
- Beginners made more errors (vs. advanced)
- Children made more errors (vs. adult beginners)
Lexical stress errors by French learners of German
Annotation of a learner speech corpus
Inter-annotator agreement
Frequency & distribution of errors

Diagnosis methods
  Word prosody analysis
  Diagnosis by comparison
  Diagnosis by classification

Feedback methods

de-stress: A prototype CAPT tool

Conclusion
Word prosody analysis

Requires word, syllable, and phone segmentations

- Automatically produced via forced alignment\(^1\)
- This work uses existing IFCASL segmentations
- Syllable segmentations derived from words & phones

---

Duration (DUR)

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

---

Word prosody analysis: F0

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)

---
\(^2\) jsnoori.loria.fr
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
  - Maximum energy

Comparison to a single reference 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!

---

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
  - Selected by comparing speakers’ F0 mean and range (using all available recordings)

---

Diagnosis by classification

- 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 error-annotated 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:

- % accuracy (% agreement with gold-standard labels)
- \(\kappa\) with respect to gold standard

\(^1\)www.cs.waikato.ac.nz/ml/weka
Which features are most useful for classification?

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUR</td>
<td>Duration features</td>
</tr>
<tr>
<td>F0</td>
<td>Fundamental frequency features</td>
</tr>
<tr>
<td>INT</td>
<td>Intensity features</td>
</tr>
<tr>
<td>WD</td>
<td>Uttered word (e.g. <em>Tatort</em>)</td>
</tr>
<tr>
<td>LV</td>
<td>Speaker’s skill level (A2</td>
</tr>
<tr>
<td>AG</td>
<td>Speaker’s age/gender (Girl</td>
</tr>
</tbody>
</table>
How well can lexical stress errors be classified?

**Prosodic features**

- **% accuracy**
- **Kappa**

<table>
<thead>
<tr>
<th>Prosodic Feature</th>
<th>% Accuracy</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUR</td>
<td>67%</td>
<td>0.05</td>
</tr>
<tr>
<td>F0</td>
<td>63%</td>
<td>0</td>
</tr>
<tr>
<td>INT</td>
<td>65%</td>
<td>0.05</td>
</tr>
<tr>
<td>INT+F0</td>
<td>67%</td>
<td>0.1</td>
</tr>
<tr>
<td>DUR+INT</td>
<td>68%</td>
<td>0.05</td>
</tr>
<tr>
<td>DUR+F0</td>
<td>69.77%</td>
<td>0.29</td>
</tr>
<tr>
<td>DUR+F0+INT</td>
<td>68%</td>
<td>0.15</td>
</tr>
</tbody>
</table>
How well can lexical stress errors be classified?

Best performance using only prosodic features: DUR+F0
- % Accuracy: 69.77%
- $\kappa$: 0.29
How well can lexical stress errors be classified?

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

- **WD**: % accuracy 67, Kappa 0.15
- **LV**: % accuracy 68, Kappa 0.2
- **AG**: % accuracy 69, Kappa 0.25
- **LV+AG**: % accuracy 70, Kappa 0.3
- **WD+AG**: % accuracy 71, Kappa 0.35
- **WD+LV**: % accuracy 72, Kappa 0.4
Diagnosis by classification

How well can lexical stress errors be classified?

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

% accuracy

Kappa

WD  LV  AG  LV+AG  WD+AG  WD+LV  WD+LV+AG

% accuracy: 71.87%
Kappa: 0.34
Diagnosis by classification

How well can lexical stress errors be classified?

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

- WD: % Accuracy: 66, Kappa: 0.15
- LV: % Accuracy: 71, Kappa: 0.15
- AG: % Accuracy: 68, Kappa: 0.15
- LV+AG: % Accuracy: 69, Kappa: 0.15
- WD+AG: % Accuracy: 72, Kappa: 0.15
- WD+LV: % Accuracy: 71, Kappa: 0.34
- WD+LV+AG: % Accuracy: 72, Kappa: 0.4

Best performance overall: WD+LV+DUR+F0+INT

- % Accuracy: 71.87%
- $\kappa$: 0.34
How does classification accuracy compare with human agreement?

<table>
<thead>
<tr>
<th></th>
<th>% agreement</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best classifier vs. gold standard</td>
<td>71.87%</td>
<td>0.34</td>
</tr>
<tr>
<td>Mean human vs. human</td>
<td>54.92%</td>
<td>0.23</td>
</tr>
</tbody>
</table>

- Results are encouraging in this context
- Still want better performance for real-world use
Lexical stress errors by French learners of German
Annotation of a learner speech corpus
Inter-annotator agreement
Frequency & distribution of errors

Diagnosis methods
Word prosody analysis
Diagnosis by comparison
Diagnosis by classification

Feedback methods

de-stress: A prototype CAPT tool

Conclusion
Implicit feedback

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

Im Frühling fliegen Pollen durch die Luft.

Your utterance:

Früh ling

Reference utterance 1:

Früh ling

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

Download
Explicit feedback

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

Your scores

<table>
<thead>
<tr>
<th>Metric</th>
<th>Score</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>3/10</td>
<td>I think you pronounced an incorrect number of phones in at least one of the word’s syllables.</td>
</tr>
<tr>
<td>Pitch</td>
<td>10/10</td>
<td>Your pitch was pitch-perfect, great job!</td>
</tr>
<tr>
<td>Loudness</td>
<td>6/10</td>
<td>The correct syllable is loudest, good job! But it should be even louder compared to the unstressed syllable.</td>
</tr>
<tr>
<td>Overall</td>
<td>5/10</td>
<td>Your overall score is the weighted average of your Duration (60%), Pitch (30%), and Loudness (10%) scores.</td>
</tr>
</tbody>
</table>
Self-assessment as feedback

May be linked to progress and motivation

---

**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?**

---

Lexical stress errors by French learners of German
Annotation of a learner speech corpus
Inter-annotator agreement
Frequency & distribution of errors

Diagnosis methods
Word prosody analysis
Diagnosis by comparison
Diagnosis by classification

Feedback methods

de-stress: A prototype CAPT tool

Conclusion
de-stress: A prototype CAPT tool
de-stress: A prototype CAPT tool

Learner Speech

DIAGNOSIS

Method 1
Method 2
Method 3

tracks performance

ASSESSMENT

Type A
Type B
Type C

selects feedback type

FEEDBACK GENERATION

chooses analysis method

Intelligent Tutoring System

Display (GUI)
Teacher/Researcher interface

de-stress

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
Teacher/Researcher interface

Create Diagnosis Method

Name * SimpleComparison
Description Single ref. comparison
Scorer * Comparison
Number Of References * 1
Selection Type MANUAL

Create

Create Feedback Method

Name * TextStylization-SelfAsses
Description
Requires Scorer Type
Show Skill Bars
Play Feedback Signal
Display Shapes
Style Text
Display Messages
Self Assessment

Create
Learner interface

de-stress

Im Frühling fliegen Pollen durch die Luft.

Your utterance: Pollen

Native speakers: Pollen

You stressed the correct syllable. Great job!
Lexical stress errors by French learners of German
  Annotation of a learner speech corpus
  Inter-annotator agreement
  Frequency & distribution of errors

Diagnosis methods
  Word prosody analysis
  Diagnosis by comparison
  Diagnosis by classification

Feedback methods

de-stress: A prototype CAPT tool

Conclusion
Conclusion

Main contributions of the thesis:

▶ Annotation & analysis of lexical stress errors in small corpus of German spoken by French speakers
  • Rather low inter-annotator agreement
  • Roughly one-third of utterances contained errors

▶ Exploration of classification for error diagnosis
  • 71.87% accuracy, $\kappa = 0.34$ wrt. gold-standard labels
  • Slightly better than mean inter-annotator agreement

▶ The de-stress CAPT tool
  • Integrates various diagnosis and feedback methods
  • Allows teachers/researchers control over methods used

Future work:

▶ In vivo studies using de-stress

▶ Improve classification performance (e.g. new algorithms)
Conclusion

Main contributions of the thesis:

▶ Annotation & analysis of lexical stress errors in small corpus of German spoken by French speakers
  • Rather low inter-annotator agreement
  • Roughly one-third of utterances contained errors

▶ Exploration of classification for error diagnosis
  • 71.87% accuracy, $\kappa = 0.34$ wrt. gold-standard labels
  • Slightly better than mean inter-annotator agreement

▶ The de-stress CAPT tool
  • Integrates various diagnosis and feedback methods
  • Allows teachers/researchers control over methods used

Future work:

▶ In vivo studies using de-stress
▶ Improve classification performance (e.g. new algorithms)
Conclusion

Main contributions of the thesis:

▶ Annotation & analysis of lexical stress errors in small corpus of German spoken by French speakers
  • Rather low inter-annotator agreement
  • Roughly one-third of utterances contained errors

▶ Exploration of classification for error diagnosis
  • 71.87% accuracy, $\kappa = 0.34$ wrt. gold-standard labels
  • Slightly better than mean inter-annotator agreement

Future work:

▶ In vivo studies using de-stress CAPT tool
  • Integrates various diagnosis and feedback methods
  • Allows teachers/researchers control over methods used

29 / 29
Main contributions of the thesis:

- Annotation & analysis of lexical stress errors in small corpus of German spoken by French speakers
  - Rather low inter-annotator agreement
  - Roughly one-third of utterances contained errors
- Exploration of classification for error diagnosis
  - 71.87% accuracy, $\kappa = 0.34$ wrt. gold-standard labels
  - Slightly better than mean inter-annotator agreement
- The de-stress CAPT tool
  - Integrates various diagnosis and feedback methods
  - Allows teachers/researchers control over methods used

Future work:

- In vivo studies using de-stress
- Improve classification performance (e.g. new algorithms)
Main contributions of the thesis:

- Annotation & analysis of lexical stress errors in small corpus of German spoken by French speakers
  - Rather low inter-annotator agreement
  - Roughly one-third of utterances contained errors

- Exploration of classification for error diagnosis
  - 71.87% accuracy, $\kappa = 0.34$ wrt. gold-standard labels
  - Slightly better than mean inter-annotator agreement

- The de-stress CAPT tool
  - Integrates various diagnosis and feedback methods
  - Allows teachers/researchers control over methods used

Future work:

- In vivo studies using de-stress
- Improve classification performance (e.g. new algorithms)
Thanks for listening!

Many thanks to:

- DFG/ANR Project IFCASL
- Bernd Möbius
- Jürgen Trouvain
- Yves Laprie
- Julie Busset
- Frank Zimmerer
- Jeanin Jügler
Selected references

- 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