Automatic classification of lexical stress errors for German CAPT

Anjana Sofia Vakil

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

anjanav@coli.uni-saarland.de

Abstract

[TODO Abstract (200 word limit)]

This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words. This very useful filler sentence has exactly ten different words.

Index Terms: computer-assisted pronunciation training, CAPT, word prosody, German **[TODO are these OK?]**

1. Introduction

For adult learners of a second language (L2), the phonological system of the L2 can pose a variety of difficulties. For certain L2s, such as German or English, one important difficulty involves the accurate prosodic realization of lexical stress, i.e. the accentuation of certain syllable(s) in a given word, with the placement of stress within a word varying freely and carrying a contrastive function in such languages [1]. Lexical stress is an important part of German word prosody, and has been found to have an impact on the intelligibility of non-native German speech [2]. Coping with this phenomenon in German is especially challenging for native (L1) French speakers, because lexical stress is realized very differently (or perhaps not at all) in the French language [3, 4].

To overcome this difficulty and improve their L2 word prosody, learners typically need to have their pronunciation errors pointed out and corrected by a language instructor; unfortunately, the lack of attention typically given to pronunciation in the foreign language classroom, along with other factors such as high student-to-teacher ratios, make this level of individualized attention not always feasible in a classroom setting [5, 6, 7]. Fortunately, advances in Computer-Assisted Pronunciation Training (CAPT) over recent decades have made it possible to automatically provide highly individualized analysis of learners' prosodic errors, as well as feedback on how to correct them, and thus to help learners achieve more intelligible pronunciation in the target language. However, while much research has gone into the creation and improvement of CAPT systems for English (see e.g. [8, 9]), relatively little work has been done on the development of CAPT systems for German, especially on those targeting errors in German prosody.

This paper describes work that advances the state of German CAPT by applying machine learning methods to the task of diagnosing lexical stress errors in non-native German speech, a necessary prerequisite for delivering individualized corrective feedback on such errors in a CAPT system. The paper is organized as follows: Section 2 provides background on the phenomenon of lexical stress as it is realized in German and French word prosody, motivates the creation of CAPT systems that address this error specifically, and summarizes some past work related to this topic. Section 3 describes the manual annotation of lexical stress errors in a small corpus of L2 German speech, carried out to create labeled training and test data for the classification experiments explained in section 4. Section 5 presents and analyzes the results of these experiments. Finally, section 6 offers some concluding remarks and outlines possible directions for future work.

2. Background and related work

Broadly speaking, lexical stress is the phenomenon of how a given syllable is accentuated within a word [1], i.e. how a syllable is given a more prominent role such that this syllable is perceived as "standing out" [10]. This perceived prominence of a syllable is a function not merely of the segmental characteristics of the uttered syllable, i.e. the speech sounds it contains, but rather of its (relative) suprasegmental properties, namely:

- duration, which equates on the perceptual level to length;
- fundamental frequency (F0), which corresponds to perceived pitch; and
- intensity (energy or amplitude), which perceptually equates to loudness.

In variable-stress languages, such as German and English, the location of lexical stress in a word is not always predictable, and therefore knowing a word requires, in part, knowing its stress pattern. This allows lexical stress to serve a contrastive function in these languages, e.g. distinguishing *UMfahren* (to drive around) from *umFAHRen* (to run over with a car) in German. Furthermore, in German, misplaced stress can disrupt understanding even in cases where there is no stress-based minimal pair [2]. However, in fixed-stress languages, stress is completely predictable, as it always falls on a certain position in the word (e.g. the final syllable), making the lexical stress pattern less crucial to the knowledge of a word than in variable-stress languages. Furthermore, in fixed-stress languages there may be a weaker distinction between stressed and unstressed syllables. While French has often been categorized as a fixed-stress language, given that word-final syllables are given prominence when a French word is pronounced in isolation, some argue that it may be more properly considered a language without lexical stress, in that speakers do not seem to accentuate any syllable within the word, with word-final lengthening effects explained by interactions with the realization of phrasal accent (lengthening of the final syllable in each prosodic group or phrase) [3, 4]. Regardless, French has no contrastive word-level stress [4, p. 89], and in this respect differs considerably from German.

This difference between the languages leads us to expect French learners of German to have difficulties with both perception and production of lexical stress prosody. Although little research has been done on the nature of lexical stress errors for this particular L1-L2 pair, Hirschfeld and Trouvain [7] report that such errors are commonly observed in German spoken by French natives. Research on French speakers' perception of Spanish, another contrastive-stress language, has revealed that these speakers seem to be "deaf" to lexical stress, i.e. seem to have significant and lasting difficulty perceiving and remembering stress contrasts [3]. With respect to production, studies of L2 Dutch have shown that French speakers, especially beginners, make systematic errors with lexical stress, exhibiting a tendency to stress the final syllable of Dutch words even when stress should be placed on the initial or medial syllable [11, 4]. Similar findings have also been reported for French learners of English [12]. The high (anticipated) frequency of lexical stress errors in the speech of this L1-L2 group is thus one motivating factor for the creation of CAPT systems to help learners identify and correct such errors.

Another motivation behind this work's focus on lexical stress errors is the high impact such errors may have on the intelligibility of L2 German speech. Intelligibility, as opposed to lack of a foreign accent, is generally considered to be the most important goal of pronunciation training [13, 5, 6, 14, 9]. The exact definition of *intelligibility* is a topic of debate, but here we will follow Munro and Derwing [13, p. 289] in understanding it broadly as "the extent to which a speakers message is actually understood by a listener." Generally speaking, prosodic errors have often been found to have a larger impact on the perceived intelligibility of L2 speakers than segmental errors (Derwing and Munro, 2005; Hahn, 2004; Witt, 2012), and several studies have found lexical stress errors to have a particularly strong impact on intelligibility in free-stress languages like English and Dutch [1, 14] Though relatively little research has been done on how various pronunciation errors affect intelligibility in L2 German specifically, some studies suggest that lexical stress errors may hinder intelligibility of L2 German speech more than other types of errors [2, 7]. Stress errors may also affect perception of segmental errors in the L2 learners speech; for example, segmental errors occurring in stressed syllables may be more noticeable than those in unstressed syllables [1, 11]. It would therefore seem that there may be a strong connection between lexical stress errors and intelligibility in L2 German speech, though more research is needed to clarify the nature of this relationship. **[TODO remove that sentence?]**

Though the frequency and impact of lexical stress errors in the speech of French learners of German thus constitute strong reasons to develop CAPT tools to treat such errors, in order for such systems to be viable, the feasibility of reliable automatic detection of this type of error must be demonstrated. **[TODO Make rest of this par about how comparison-based diagnosis is usually used, then start new par with following sentence?]** To our knowledge, no work has been reported on automatic classification-based diagnosis of lexical stress errors in L2 German speech, but in recent years machine learning methods have been applied with apparent success to the classification of lexical stress patterns in English words. Kim and Beutnagel [15] experimented with various classifiers to identify stress patterns in high-quality recordings of 3- and 4-syllable English words, reporting accuracy in the 80-90% range; in pilot experiments with low-quality recordings, however, the authors report lower accuracy: 70-80% on L1 speech and 50-60% on utterances by L2 speakers. Similarly, Shahin et al. [16] trained Neural Networks to classify stress patterns in bisyllabic words uttered by L1 English children, and reported classification accuracy over 90% for some stress patterns; though this work was conducted with a view to treating childhood L1 dysprosody, its relevance to our intended application of L2 CAPT is nonetheless clear. Building on these related investigations, this paper seeks to further explore the viability of automatic classification-based detection of lexical stress errors, with a particular focus on those made by French speakers of German. To this end, a small corpus of learner speech was manually annotated for lexical stress errors, as described in section 3. Using the resulting labeled L2 data, in addition to data from L1 German speakers, a series of supervised machine learners were trained using a variety of representations of the prosodic and other features of each word utterance (see section 4.2, and these classifiers were evaluated with reference to the manually-produced labels of held-out test data (see section 4.3). Section 5 presents and analyzes the results of these evaluations.

3. Data

Error-annotated speech data from German learners is a prerequisite for the supervised training and evaluation of classifiers for lexical stress realizations in L2 German speech, yet to our knowledge no corpus of learner German with such annotation is publicly available. To fill this need, as well as to shed light on the perception of lexical stress errors in L2 German speech, a small corpus of speech by L1 French learners of German was manually annotated for such errors by native and non-native German speakers with varying levels of phonetics/phonology expertise. This section describes the data selected for annotation (section 3.1) and the method by which lexical stress realizations in this data were annotated (section 3.2), and presents an analysis of the observed inter-annotator agreement (section 3.3) and distribution of errors (section 3.4) in the annotated dataset.

3.1. The IFCASL corpus of learner speech

The learner speech data used in this work has been excerpted from the IFCASL corpus [17, 18], a collection of phonetically diverse utterances in French and German spoken by both native speakers and non-native speakers with the other language as L1. This is the first known corpus of L2 speech in both directions of the French-German language pair, and is thus an invaluable resource for research on pronunciation errors **[TODO between these languages]**.

The IFCASL corpus contains recordings of approximately 50 L1 speakers of each language reading carefully constructed sentences (and a short text) in both languages, such that both L1 and L2 speech was recorded for each speaker. Each L1 speaker group has an even gender distribution, and contains approximately 10 children (adolescents of 15-16 years of age) and 40

Table 1: Number of speakers in the portion of the IFCASL-FG corpus annotated for lexical stress, in terms of speakers' age, gender, and proficiency level [18].

Age/gender	Proficiency level				Totals
	A2	B1	B2	C1	
Boy (male, 15-16 yrs.)	11	0	0	0	**11**
Girl (female, 15-16)	1	1	0	0	**2**
Man (male, 18-30)	7	4	3	7	**21**
Woman (female, 18-30)	5	5	3	9	**22**
Totals	**24**	**10**	**6**	**16**	**56**

Table 2: Word types selected from the IFCASL corpus for lexical stress error annotation. Canonical pronunciations for each word type are given in IPA notation. The rightmost column lists the number of tokens (utterances) of each word type in the annotated dataset.

Word	Pronunciation	Part of speech	English meaning	Tokens
E-mail	/ˈiː.meɪl/	noun	e-mail	56
Flagge	/ˈfla.gə/	noun	flag	55
fliegen	/ˈfliː.gn̩/	verb	to fly	56
Frhling	/ˈfryː.lɪŋ/	noun	spring (season)	56
halten	/ˈhal.tn̩/	verb	to hold	56
manche	/ˈman.çə/	pronoun	some	56
Mrder	/ˈmœʁ.dɐ/	noun	murderer	56
Pollen	/ˈpɔ.lən/	noun	pollen	55
Ringen	/ˈʁɪŋ.ən/	noun	rings	55
Tatort	/ˈtaːt.ʔɔʁt/	noun	crime scene	56
tragen	/ˈtʁaː.gn̩/	verb	to wear	55
Tschechen	/ˈtʃɛ.çn̩/	noun	Czechs	56

adults. A variety of self-reported L2 proficiency levels are also represented in the corpus: the recorded adults span levels A2 (beginner) through C1 (advanced) of the Common European Framework of Reference **[TODO footnote url]**, the children levels A2 (beginner) and B1 (low intermediate).

While L2 French speech is thus also captured in the IFCASL corpus, the annotation effort described here focuses exclusively on the German-language subset of the corpus. Only utterances from the sub-corpus of L2 German speech by L1 French speakers (henceforth IFCASL-FG) were manually annotated; native utterances from the L1 German sub-corpus (IFCASL-GG) were assumed to contain only correct lexical stress realizations. Details about the speakers included in the IFCASL-FG sub-corpus are given in table 1.

The subset of IFCASL-FG selected for manual error annotation, (henceforth simply the dataset) consists of utterances of twelve bisyllabic word types (see table 2), each of which has primary stress on the initial syllable. Only bisyllabic words were selected to simplify comparison between stressed and unstressed syllables, and only initial-stress words because this is the stress pattern which native (L1) French speakers are expected to have the most difficulty producing in German, given the phenomenon of final lengthening in French (see section 2).

3.2. Annotation method

3.3. Inter-annotator agreement

3.4. Error distribution

4. Method

4.1. Machine learning algorithm

4.2. Feature sets

4.3. Datasets for training and testing

5. Results

5.1. Feature performance

5.2. Performance on unknown words

5.3. Performance on unknown speakers

6. Conclusions and future work

7. References

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

[2] U. Hirschfeld, *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*, ser. Forum Phoneticum, 1994, vol. 57.

[3] E. Dupoux, N. Sebastián-Gallés, E. Navarette, and S. Peperkamp, "Persistent stress 'deafness': The case of French learners of Spanish," *Cognition*, vol. 106, pp. 682–706, 2008.

[4] M.-C. Michaux and J. Caspers, "The production of Dutch word stress by Francophone learners," in *Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013)*, 2013, pp. 89–94.

[5] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer Assisted Language Learning*, 2002.

[6] T. M. Derwing and M. J. Munro, "Second Language Accent and Pronunciation Teaching: A Research-Based Approach," *TESOL Quarterly*, vol. 39, no. 3, pp. pp. 379–397, 2005.

[7] U. Hirschfeld and J. Trouvain, "Teaching prosody in German as foreign language," in *Non-Native Prosody: Phonetic Description and Teaching Practice*, J. Trouvain and U. Gut, Eds. Walter de Gruyter, 2007, pp. 171–187.

[8] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, Oct. 2009.

[9] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 1–8.

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

[11] M.-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.

[12] A. Bonneau and V. Colotte, "Automatic Feedback for L2 Prosody Learning," in *Speech and Language Technologies*, I. Ipsic, Ed. InTech, 2011, no. 1977.

[13] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 49, no. Supplement s1, pp. 285–310, 1999.

[14] J. Field, "Intelligibility and the Listener: The Role of Lexical Stress," *TESOL Quarterly*, vol. 39, no. 3, p. 399, Sep. 2005.

[15] Y.-J. Kim and M. C. Beutnagel, "Automatic assessment of American English lexical stress using machine learning algorithms," in *SLaTE*, 2011, pp. 93–96.

[16] M. Shahin, B. Ahmed, and K. Ballard, "A neural network based lexical stress pattern classifier," *Qatar Foundation Annual Research Forum Proceedings*, no. 2012, p. CSP22, Oct. 2012.

[17] J. Trouvain, Y. Laprie, B. Möbius, B. Andreeva, A. Bonneau, V. Colotte, C. Fauth, D. Fohr, D. Jouvet, O. Mella, J. Jügler, and F. Zimmerer, "Designing a bilingual speech corpus for French and German language learners," in *Corpus et Outils en Linguistique, Langues et Parole: Statuts, Usages et Mésuages*, no. ii, Strasbourg, France, 2013, pp. 32–34.

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