Automatic diagnosis and feedback for lexical stress errors in non-native speech

Towards a CAPT system for French learners of German

Anjana Sofia Vakil

A thesis submitted toward the degree of

Master of Science

in Language Science and Technology

Prepared under the supervision of

Prof. Dr. Bernd Möbius

Dr. Jürgen Trouvain

UNIVERSITÄT DES SAARLANDES

Saarland University

Department of Computational Linguistics & Phonetics

31 March, 2015

Anjana Sofia Vakil
anjanav@coli.uni-saarland.de
Automatic diagnosis and feedback for lexical stress errors in non-native speech
31 March, 2015
Supervisors: Prof. Dr. Bernd Möbius and Dr. Jürgen Trouvain

Saarland University
Department of Computational Linguistics & Phonetics
Fachrichtung 4.7 Allgemeine Linguistik
Postfach 15 11 50
66041 and Saarbrücken

Typeset using LaTeX 2_ε. Style adapted from the *Clean Thesis* template developed by Ricardo Langner (http://cleanthesis.der-ric.de/).

Declaration

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, 31 March, 2015

Anjana Sofia Vakil

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Abstract (different language)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgement

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

List of Figures

List of Tables

1 Introduction

For students with French as their first language (L1) who are learning German as a second language (L2), the sound system of the L2 can pose a variety of difficulties, one of the most important and interesting of which is the way in which certain syllables in German words are accentuated more than others, a phenomenon referred to as lexical stress. Learning to navigate German lexical stress is especially challenging for L1 French speakers, because this phenomenon is realized very differently in the French language.

Computer-Assisted Pronunciation Training (CAPT) systems have the potential to automatically provide highly individualized analysis of such learner errors, as well as feedback on how to correct them, and thus to help learners achieve more intelligible pronunciation in the target language (Witt, 2012). The thesis project described here aims to advance German CAPT by creating a tool which will diagnose and offer feedback on lexical stress errors in the L2 German speech of L1 French speakers, in the hopes of ultimately helping these learners become more intelligible when speaking German.

1.1 Context: The IFCASL project

This work has been conducted in the context of the ongoing research project "Individualized Feedback in Computer-Assisted Spoken Language Learning (IFCASL)" at Saarland University (Saarbrücken, Germany) and LORIA (Nancy, France). **[TODO more project info (ANR/DFG project no.? citation?)]**

The goal of the IFCASL project is to take initial steps toward the development of a CAPT system targeting native (L1) French speakers learning German as a foreign language (L2), as well as L1 German speakers learning French as their L2. To this end, a bidirectional learner speech corpus has been recorded, comprising phonetically diverse utterances in French and German spoken by both native speakers and non-native speakers with the other language as L1 (Fauth et al., 2014; Trouvain et al., 2013). While the project as a whole is thus also concerned with L1 German speakers learning French as L2, this thesis focuses exclusively on French L1 speakers learning German as L2.

[TODO Details about corpus (summary of corpus articles)]

[TODO Remove the following paragraph?] The German-language subset of the IFCASL corpus has been instrumental in training and testing the automatic diagnosis and feedback systems developed in this work. Furthermore, those systems have been designed with a view to contributing to the overall set of software developed in the context of the IFCASL project,

such that they have been as compatible as possible with the other tools developed and used by the IFCASL team.

1.2 Objectives

The main objective of this work is to investigate the automatic treatment of lexical stress errors in the context of a CAPT system for French learners of German. This includes, on the one hand, an examination of the ways in which lexical stress errors of the type made by French L1 speakers when speaking German as L2 can be reliably detected and measured automatically, and on the other, an exploration of the types of multimodal feedback on such errors that can be automatically delivered based on the aforementioned error detection. The

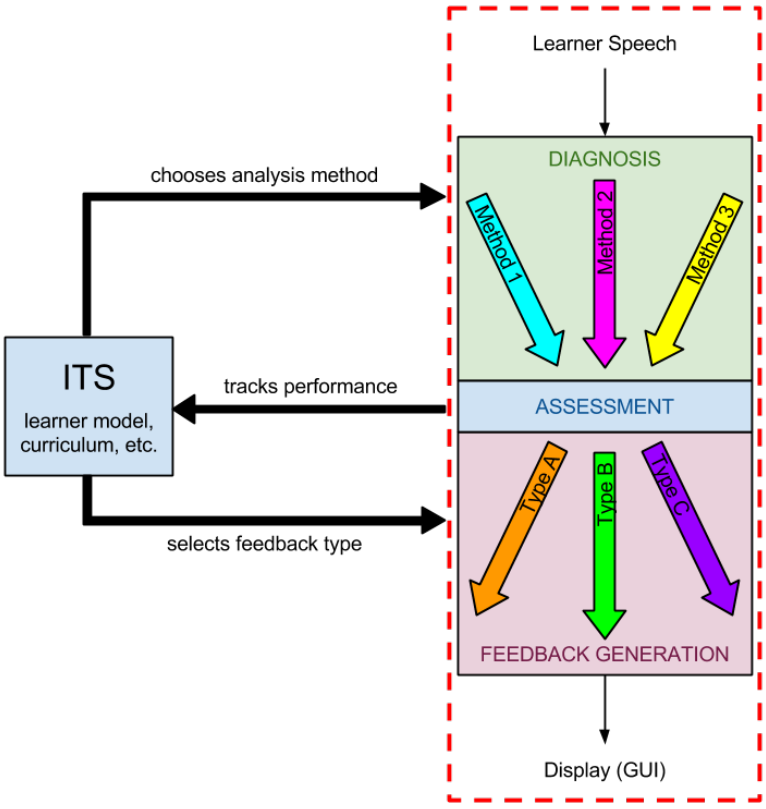


Figure 1.1: Conceptual diagram of the prototype lexical stress CAPT tool (demarcated by dotted line) and its possible function in the context of a more comprehensive Intelligent Tutoring System.

outcome of these investigations is a prototype CAPT tool, illustrated in fig. 1.1, which can diagnose lexical stress errors in different ways and present learners with different types of feedback on these errors.

This prototype tool has been developed with both instructional and research applications in mind. Unlike with some existing tools for diagnosis pf and feedback on pronunciation errors, learners can interact with the tool and interpret its feedback independently, i.e. without the assistance of a human instructor at their side. At the same time, researchers can use this modular system to study the impact of various assessment and feedback types on learner

outcomes, user engagement, and other factors impacting the success of a CAPT system. Once more is known about which diagnosis/feedback types should be delivered to which learners in which situations, this tool could become a useful component to a fully-fledged CAPT system, in which learner models and other intelligent components automatically decide which modules of the tool to activate.

1.3 Thesis overview

[TODO Add chapter names?]

Chapter 2 introduces Computer-Assisted Pronunciation Training (CAPT) in the contexts of pronunciation teaching in foreign-language education and computer-based and intelligent tutoring systems, describing some relevant past work on CAPT systems. This chapter also briefly introduces the phenomenon of lexical stress as it pertains to L1 French learners of German as L2, and outlines the motivation for focusing on lexical stress errors in this work.

Chapter 3 describes original work on the annotation and analysis of lexical stress errors in the IFCASL sub-corpus of nonnative German speech produced by native French speakers.

Chapter 4 details how the prototype CAPT tool diagnoses lexical stress errors in learner speech. It describes the methods used to automatically segment the learner's utterance, analyze the prosody of this utterance in terms of the relative pitch, duration, and intensity of the relevant syllables, and compare this analysis to one or more models of native pronunciation to produce a diagnosis.

Chapter 5 describes the multimodal feedback options that the system can deliver, and how these feedback types are generated based on the analysis of the learner's speech described in the previous chapter.

Chapter 6 briefly introduces the tools that have been developed, and the technology used to build them. **[TODO skip this chapter?]**

Chapter 7 summarizes the contributions of this work and outlines some possible directions for future work.

2 Background and related work

2.1 Pronunciation in foreign language education

In the foreign language classroom, less focus has traditionally been placed on pronunciation than other aspects of language education, such as grammar and vocabulary. However, even when pronunciation is taught in the classroom, a number of factors may limit the effectiveness of that training (Neri et al., 2002; Derwing and Munro, 2005). First of all, partly thanks to a historical lack of communication between the fields of speech science and foreign language education, many teachers lack the training in phonetics and phonology to provide helpful feedback to students and correct their articulation. Secondly, high student-to-teacher ratios may prevent teachers from giving adequate attention and feedback to individual students, and limit the amount of time each student can practice speaking. Furthermore, anxiety about speaking the L2 in front of their peers may make students less willing to practice speaking, and less able to absorb corrective feedback. **[TODO individual/specific citations for each point above?]**

Although much work still needs to be done to improve our understanding of how best to teach pronunciation, existing research reveals a few general considerations that must be kept in mind. First of all, it is important to note that intelligibility, and not lack of a "foreign accent", is generally considered to be the most important goal of pronunciation training (Neri et al., 2002; Derwing and Munro, 2005; Witt, 2012).

Research on the impact of various types of pronunciation errors on intelligibility tends to indicate that errors on the prosodic (suprasegmental) level hinder intelligibility more than segmental errors (Anderson-Hsieh et al., 1992; Derwing and Munro, 2005; Hirschfeld and Trouvain, 2007; Dlaska and Krekeler, 2013). **[TODO expand]**

For reducing these and other types of errors, perception training has been found to be very important (Derwing and Munro, 2005; Hirschfeld and Trouvain, 2007), though some researchers stress that combining this with corrective feedback on pronunciation errors leads to bigger performance gains (Dlaska and Krekeler, 2013). **[TODO expand]**

The importance of individualized corrective feedback is also generally acknowledged (Neri et al., 2002; Mehlhorn, 2005; Dlaska and Krekeler, 2013), **[TODO expand]** though there is much to be learned about exactly when and how feedback can be most effective. This is the motivation behind the feedback generation module of the proposed tool (see Chapter 5), which is intended to facilitate research on CAPT feedback.

2.2 Computer-Assisted Pronunciation Training

Computer-Assisted Pronunciation Training[1] (CAPT) stands to help make pronunciation training more accessible by overcoming some of these difficulties. With CAPT, student-to-teacher ratio is not an issue, as the learner always has the full attention of the digital tutor, and provided an effective curriculum design, a CAPT system can offer learners practically limitless practice opportunities. Interacting with a computer program may also be perceived by the learner as a lower-stakes, more comfortable environment than the classroom, where they may feel too intimidated to practice speaking in the L2. But perhaps most compelling is the potential for CAPT to deliver the type of individualized instruction which many learners may not otherwise have access to in the L2 classroom, for reasons such as those mentioned above. Indeed, in recent decades, the educational value of speech technologies has been well demonstrated (Eskenazi, 2009), with CAPT emerging as one important educational application for foreign-language education (FLE) (Neri et al., 2002; Delmonte, 2011; Witt, 2012).

2.2.1 Prosody in existing CAPT systems

The viability of CAPT has been demonstrated by a variety of systems and tools that have been developed in both academic and commercial contexts. Some focus on overall assessment of pronunciation or fluency, and others on the detection and correction of individual pronunciation errors (Eskenazi, 2009); the tool developed in this work falls into the latter category. In error-focused systems, a distinction has typically been drawn between phonemic errors, e.g. the substitution, insertion, or deletion of a segmental speech sound, and prosodic errors, such as those related to stress/accent, intonation, or rhythm (Witt, 2012). As discussed in the previous section, word-prosodic errors have a larger impact on intelligibility than segmental errors, and are therefore the focus of this work (see Section 2.4 below). With this in mind, a few prosody-aware CAPT systems relevant to this thesis are discussed below; comprehensive overviews and comparisons of these and many other systems are given by Neri et al. (2002), Eskenazi (2009), Delmonte (2011), and Witt (2012).

Both the diagnosis and feedback modules of the CAPT tool developed in this work build to a great extent on previous work by researchers in the speech group at LORIA[2] in Nancy, France, many of whom are also involved in the IFCASL project (see Section 1.1). Their work has, on the one hand, investigated the task of automatically recognizing and segmenting learners' speech, and determining how this possibly incorrect automatic segmentation can be effectively utilized in the context of pronunciation tutoring, particularly at the prosodic level (Mesbahi et al., 2011; Orosanu et al., 2012); see Chapter 4 for a discussion of how this thesis will build upon that work. Additionally, the group has developed the *Snoori* suite of software, including the PC-based WinSnoori and its partial Java port, Jsnoori (Project-Team PAROLE, 2013). These programs take as input a learner utterance, a native reference utterance, and segmentations of each, perform an acoustic comparison of the two utterances, and deliver feedback on the learner's speech in the form of e.g. annotated displays of the speech

[1] Also known as Computer-Assisted Pronunciation Teaching or Tutoring
[2] `http://www.loria.fr/`

signal and spectrogram of each. Moreover, auditory feedback can be delivered thanks to the capability of resynthesizing the learner's utterance to match the pitch contour and timing of the reference, without modifying the voice quality of the utterance, such that the learner can hear the "correct" pronunciation in their own voice. The utility of such software, and especially this resynthesized feedback, for pronunciation teaching has been explored by Bonneau and Colotte (2011), who used it to assess and deliver feedback on lexical stress in L1 French speakers' pronunciation of English words. **[TODO more about the Bonneau paper?]** As described later in this paper, the prototype CAPT tool developed in this thesis project builds on the error detection and diagnosis functionality of Jsnoori (see Chapter 4), and leverages its feedback generation capabilities to deliver a more diverse, and potentially more effective, range of feedback types (see Chapter 5).

This work also draws from research on two systems developed at Carnegie Mellon University. The first of these, the Fluency pronunciation trainer (Eskenazi and Hansma, 1998; Eskenazi et al., 2000), is a CAPT system placing particular emphasis on user-adaptivity, corrective articulatory feedback, and the integration of perceptual training (e.g. listening exercises). As with the work at LORIA described above, the Fluency system evaluates learners' speech via comparison with that of a native reference speaker, and Probst et al. (2002) found that selecting a "golden speaker" whose voice closely matched the learner's improved learning gains. **[TODO more about golden speaker?]** Fluency also implements an error-catching step to reject utterances which do not match the expected text (Eskenazi et al., 2000), in the same vein as that of Mesbahi et al. (2011) and Orosanu et al. (2012). Eskenazi et al. (2007) report that Fluency's commercial spin-off, NativeAccent™, has been shown to help real-world users significantly improve their pronunciation skills.

A second CMU system, the Project LISTEN Reading Tutor (Mostow, 2012) may not strictly be a CAPT tool, as it is designed to help children develop reading fluency in their native language. However, as it analyzes the prosody of children's read speech to measure reading fluency, and offers feedback on this prosody. it is nevertheless very relevant to CAPT and thus this thesis. Indeed, the potential for such a tool, and its underlying technologies, to enhance foreign-language education has already been demonstrated by **[TODO Weber and Bali (2010)]**, who deployed the Reading Tutor in English as a second language classes in India with encouraging initial results. In the Reading Tutor, the child's read speech is automatically segmented and compared either to a reference utterance by an adult reader, analogous to the native speaker reference in many CAPT systems, or to a generalized model of adult prosody; Duong et al. (2011) report better performance using the generalized model. Analysis of the pitch and intensity contours of the utterance(s), as well as the duration of words/syllables and the pauses between them, results in an assessment of the child's overall fluency as well as identification of words which have been pronounced (in)correctly, and feedback is delivered visually in real time by revealing the text of each word as it is spoken, with properties such as the position, color, and font size of each word reflecting various aspects of the reader's prosody (Sitaram et al., 2011). Ideas and techniques from the Reading Tutor have influenced both the diagnosis (see Chapter 4) and feedback (see Chapter 5) modules of the proposed CAPT tool. **[TODO which ideas, influenced how?]**

2.2.2 German and language-independent CAPT

[TODO flesh out this paragraph - rephrase subsection heading?] The vast majority of CAPT systems which analyze learners' speech at the prosodic level have been developed with English as the target L2, and relatively little work has been done on German. In a notable exception particularly relevant to this thesis, Bissiri et al. (2006; 2009) found that L1 Italian speakers' realizations of lexical stress in German improved when they were allowed to listen to prosodically-modified recordings of their own speech and that of native speakers (see Section 5.2). Jilka and Möhler's (1998) use of F0 contour manipulation in studying L1 English speakers' production of German represents another exploration of speech technology applications for German instruction. **[TODO details]** Language-independent tools have also been developed, such as WinPitch LTL (Martin, 2004), which enables speech signal visualization of prosodic features such as pitch contours as well as manipulation of prosody and comparison to reference utterances, with the intent that a human instructor will guide the learner in using the software and interpreting the visualizations.

[TODO need an outro for this (sub)section]

2.3 Lexical stress

When there is a typological difference between some segmental or prosodic feature(s) of a language learner's L1 compared to the target L2, there is a particular need for pronunciation training to bridge this gap. In the case of the French-German language pair, the prosodic realization of lexical stress is one feature which marks a striking difference between the languages.

Lexical stress is the phenomenon of how syllables are accentuated within a word (Cutler, 2005). **[TODO Elaborate]** This relates not to the segmental characteristics of an uttered syllable, i.e. the speech sounds it contains, but rather to its (relative) suprasegmental properties, namely:

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

2.3.1 German

As Cutler (2005) points out, different languages make use of this suprasegmental information in different ways. In what are termed free- or variable-stress languages, such as German, Spanish, and English, it is not always possible to predict which syllable in a word will carry the stress, and therefore knowing a word requires, in part, knowing its stress pattern. This allows lexical stress to serve a contrastive function in these languages, such that two words may share exactly the same sequence of phones and nevertheless be distinguished exclusively by their stress pattern, as is the case with *UMfahren* (to drive around) and *umFAHRen* (to run over with a car) in German. Because stress carries meaning thus, native speakers of such languages are sensitive to stress patterns, and readily able to perceive differences in

stress. Furthermore, in German, misplaced stress has been shown to disrupt understanding of a word or utterance even in cases where there is no stress-based minimal pair (Hirschfeld, 1994), supporting the theory that speakers of free-stress languages rely to a large extent on stress information in the recognition of spoken words (Cutler, 2005).

2.3.2 French

However, in the so-called fixed-stress languages, stress is completely predictable, as it always falls on a certain position in the word; in French, for example, stress is fixed on the word-final syllable, while in Czech and Hungarian, stress always falls on the initial syllable. Lexical stress may not be as crucial to the knowledge of a word in these languages as in the free-stress languages. Furthermore, although lexical stress is realized in these languages, the distinction between stressed and unstressed syllables may be weaker than in free-stress languages. While many theorists place French into this category of fixed-stress languages, others argue that it may be more properly considered a language without lexical stress, insofar as there is no systematic way in which speakers distinguish a certain syllable from others in the word, aside from the fact that French exhibits phrasal accent, i.e. lengthening of the final syllable in each prosodic group or phrase (Dupoux et al., 2008).

2.3.3 Expected pronunciation errors

As a result of this difference in the sound systems of the two languages, native speakers of French may generally be expected to lack the sensitivity to stress patterns possessed by native speakers of German. Indeed, this has been borne out by research by Dupoux et al. (2008), who found that native French speakers are "deaf" to differences in stress patterns, such that they have great difficulty discriminating between Spanish words which contrast only at the level of stress. This difficulty should also exist for French speakers when they are presented with German words in which the stress pattern is crucial to the word's meaning, as in the minimal pair above. **[TODO Furthermore, we can expect errors in their production...]**

2.4 Targeting lexical stress errors in CAPT

Learners of a foreign language typically make a wide variety of pronunciation errors, at both the segmental level (e.g. errors in producing certain vowels or consonants of the target language) and the prosodic level (e.g. errors in the speaker's intonation contour or the duration of certain syllables or words). As it is not feasible to address all of these in a prototype CAPT tool, one of the first aims of this work is to identify a single type of error which is well suited to being addressed via CAPT for L1 French/L2 German.

To guide this selection, we may consider a set of three criteria that such an error must meet; similar criteria are proposed by Cucchiarini et al. (2009). First, the error must be *produced relatively frequently* by French L1 speakers in their production of L2 German, as it would be a misuse of resources to design a system addressing an error seldom made

by learners (Neri et al., 2002). Second, the error must have a significant *impact on the perceived intelligibility* of the learner's speech; as the ultimate goal of the system is to help learners communicate more effectively in the L2, an error which is commonly made but nevertheless does not impede understanding of the learner's L2 speech, and thus does not hinder communication in the L2, is not an ideal target. **[TODO refer to intelligibility vs. accentedness as discussed above]** Third, in order for the CAPT system to provide any meaningful diagnosis and feedback, the error must lend itself to reasonably accurate and reliable *detection through automatic processing*. As illustrated in fig. 2.1, the best error to target with the CAPT system will fulfill all of these criteria, rather than only one or two of the three. For example, vowel quality errors (e.g. an L1 French speaker producing a German /ə/ as [œ]) may occur frequently in the L2 speech and may be relatively easy to detect automatically, but may not have a great impact on the intelligibility of the L2 German speech. On the other hand, equally frequent vowel quantity errors (e.g. the L1 French speaker producing a German long /eː/ as [e]) may have a greater impact on intelligibility in some cases, but may be more difficult to reliably identify automatically.

Figure 2.1: Criteria for selecting errors to target in a CAPT system.

Lexical stress errors **[TODO e.g.]** fulfill all three of these criteria, and this error type has therefore been chosen as the target of the proposed CAPT tool; the remainder of this section justifies that choice.

2.4.1 Impact on intelligibility

First, as mentioned in Section 2.1 above, errors related to prosody have generally been found to have a larger impact on intelligibility than segmental errors, and several studies have found lexical stress to be particularly important for comprehension in free-stress languages like English, Dutch, and our target language, German (Hirschfeld, 1994; Cutler, 2005). Indeed, studies on perception of German L2 speech have found that among a variety of pronunciation error types, lexical stress errors have one of the most drastic impacts on intelligibility (Hirschfeld, 1994). Furthermore, lexical stress not only impacts intelligibility on the prosodic level, but may also affect perception of segmental errors in the L2 learner's

speech; for example, segmental errors occurring in stressed syllables are more noticeable than those in unstressed syllables (Cutler, 2005). Additionally, some research indicates that prosodic errors such as lexical stress errors may have more of an impact on perceived foreign accent than segmental errors (Witt, 2012); though it must again be stressed that intelligibility is a more important goal than lack of a foreign accent, insofar as perceived accent may contribute to difficulties being understood by native speakers, this relationship between prosody and accentedness also deserves mentioning.

2.4.2 Frequency of production

Secondly, we saw in Section 2.3 that perceiving contrasts in lexical stress is notoriously difficult for native French speakers (Cutler, 2005; Dupoux et al., 2008), and given the strong link between perception and production, this is a good indication that L1 French speakers will regularly make lexical stress errors in an L2 with free, contrastive stress, such as German. Bonneau and Colotte (2011) report that in a pilot study of L1 French speakers pronouncing English words, lexical stress was frequently misplaced by beginners; given the similarities of the lexical stress systems of English and German compared to that of French, this is another sign that we can expect such errors to be produced frequently. Indeed, an analysis of lexical stress errors in the IFCASL corpus of non-native (L1 French) German speech conducted as part of this thesis project supports the expectation of frequent lexical stress errors in this particular L1/L2 pair: **[TODO verify/reword if necessary: errors were observed at all skill levels, though beginners made many more errors than advanced learners]**. See Chapter 3 for a detailed discussion of these findings.

2.4.3 Feasibility of automatic detection

Finally, although much research still needs to be done on automatic detection and diagnosis of lexical stress errors (one of the main motivations behind this work; see Chapter 4), recent work on this problem has shown encouraging results. As mentioned above, several existing CAPT tools incorporate treatment of lexical stress errors (e.g. Wik et al., 2009; Bonneau and Colotte, 2011), and Shahin et al. (2012) and Kim and Beutnagel (2011) have reported success in applying machine learning methods to the classification of lexical stress patterns in English words.

As lexical stress errors thus fulfill the aforementioned criteria for targeting with CAPT, such errors are the focus of the proposed CAPT system **[TODO reword that?]**. The following sections describe how this thesis project explores automatic diagnosis (Chapter 4) and feedback generation (Chapter 5) for this type of error.

2.5 Summary

3

Lexical stress errors by French learners of German [TODO retitle?]

[TODO "sub-corpus" gets pretty annoying in this chapter - think of better term?]

[TODO Recap of IFCASL corpus Section 1.1]

To investigate to what extent the expected lexical stress errors by French speakers of German are actually produced, a subset of the non-native German-language IFCASL corpus was annotated for such errors. **[TODO more about why this is necessary/how the corpus will be used for supervised training]** The first sections of this chapter describe the selection of material for this sub-corpus (Section 3.1), the annotators who labeled lexical stress errors in that data (Section 3.2), and the method by which annotation was performed (Section 3.3).

Once error judgments had been collected from each annotator, different annotators' judgments of the same utterances were compared to determine the reliability of the annotation, i.e. the agreement between annotators. Section 3.4 describes this analysis of inter-annotator agreement, which aims to shed light on the following questions:

- How reliably can lexical stress errors be identified by annotators, i.e. to what extent do the judgments of different annotators agree? (Section 3.4.1)
- Are there differences in how native and non-native German speakers identify errors? (Section 3.4.2)
- Are there differences in how expert and novice annotators (those without annotation experience or any training in phonetics/phonology) identify lexical stress errors? (Section 3.4.3)

As Section 3.4 will show, annotators did not always agree as to whether a given utterance exhibited a lexical stress error or not. Nevertheless, a "gold-standard" label for each utterance had to be determined; Section 3.4.4 describes how this was accomplished in cases of disagreement.

Finally, given the gold-standard labels for each utterance, the distribution of lexical stress errors in the sub-corpus was analyzed; the following questions guided this analysis, which is detailed in Section 3.5.

- Are lexical stress errors observed frequently in the IFCASL data? (Section 3.5.1)
- Is there a difference in the frequency of these errors among different groups of speakers (i.e. in terms of skill level, age, or gender) or in different contexts (e.g. after hearing a native speaker produce the word)? (Sections 3.5.3–3.5.5)
- Are lexical stress errors observed more frequently with certain word types than with others? (Section 3.5.2)
- How frequently do technical problems interfere with determining whether an error was made? (Section 3.5.6)

3.1 Data

The IFCASL sub-corpus annotated for lexical stress errors consists of utterances of twelve word types (see table 3.1), each of which is bisyllabic and canonically has its primary stress on the initial syllable. These characteristics were chosen deliberately: the selected words are bisyllabic because this simplifies comparison between stressed and unstressed syllables, and they are initial-stress because this is the stress pattern which native (L1) French speakers are expected to have the most difficulty producing in German, given the fixed final-position stress and final lengthening in French (see Section 2.3.3).

In the IFCASL corpus recordings, sentences containing these words were read aloud by L1 and L2 (L1 French) speakers. Here, only the L2 utterances were annotated; it is assumed that the L1 German speakers always realize lexical stress correctly. **[TODO justify that assumption?]**

As described in Section 1.1 **[TODO verify that this reference is appropriate]**, the IFCASL recordings were performed under two conditions: the "Sentence Read" (SR) condition, in which the L2 speaker is simply presented with the text of the sentence and asked to record themselves reading it aloud, and the "Sentence Heard" (SH) condition, in which the L2 speaker is asked to listen to an utterance of the sentence by an L1 German speaker before recording their own utterance. The sub-corpus for annotation includes recordings from both conditions, though the majority are from the SR condition **[TODO does mentioning that help or hurt?]**.

To compile the sub-corpus for annotation, utterances (tokens) of each word as produced by over 50 L2 speakers were extracted from the recordings automatically with Praat (Boersma and Weenink, 2014), using extraction times (start and end points of word utterances) taken from the word-level segmentation of each sentence utterance automatically obtained by forced alignment (see Section 4.1). Table 3.1 lists the exact number of tokens available for each word type. In total, 668 word tokens were annotated for lexical stress errors. Five tokens had to be excluded from the data, as disfluencies in the sentence utterance (e.g. false starts or repetitions of the target word) prevented the automatic extraction of the word utterance from the sentence as a whole. In a fully-fledged student-facing CAPT system, such disfluencies would need to be dealt with accordingly, e.g. by means of a pre-processing step which analyzes the student's utterance for possible disfluencies and compensates for any that are detected by, for example, prompting the student to re-record their utterance. However,

detecting disfluencies in speech, especially non-native speech, is an area of active research (see e.g. **[TODO refs]**), and the development of a disfluency-aware system is outside the scope of this thesis project; therefore, this work presupposes that no disfluencies exist in the student's utterance, and the handful of disfluent tokens have been excluded from the error-annotated sub-corpus described here.

Table 3.1: The twelve bisyllabic initial-stress words types selected from the IFCASL corpus for stress error annotation **[TODO column details]**

Orthography	Canonical pronunciation	Part of speech	English meaning	Recording condition	Number of tokens **[TODO check that these tally to 668]**
E-mail	**[TODO prons]**	noun	e-mail	SR	56
Flagge		noun	flag	SH	55
fliegen		verb	to fly	SR	56
Frühling		noun	spring (season)	SR	56
halten		verb	to hold	SR	56
manche		pronoun	some	SR	56
Mörder		noun	murderer	SR	56
Pollen		noun	pollen	SR	55
Ringen		noun	rings	SH	55
Tatort		noun	crime scene	SR	56
tragen		verb	to wear	SH	55
Tschechen		noun	Czechs	SR	56

3.2 Annotators

A total of 15 annotators participated in the annotation of this IFCASL sub-corpus **[TODO *remove?:* over the course of 2 months]**, each of whom is listed in table 3.2 (by an arbitrary identifier, to preserve anonymity). As table 3.2 shows, the annotators varied with respect to their native language, as well as with respect to their level of expertise in phonetics/phonology/linguistic annotation.

Of the 15 annotators, the majority (12) were native German speakers, two were native speakers of American English, and one annotator's first language was Hebrew. The nonnative speakers all have **[TODO *more specific?:* some knowledge]** of German as L2. In terms of expertise, the annotators can broadly be categorized into three groups:

- *expert* annotators are professional researchers with a thorough understanding of phonetics/phonology and extensive experience in annotating speech data
- *intermediate* annotators are university students **[TODO enrolled in an experimental phonology course *is that true of Frankfurt students too?*]**, and have some training in phonetics/phonology and/or experience annotating speech data
- *novice* annotators have negligible training in phonetics/phonology and lack experience annotating speech data

As shown in table 3.2, the majority of annotators (10 out of 15) fall into the *intermediate* group; two annotators can be considered *expert*, and there are three *novice* annotators.

Table 3.2: Annotators **[TODO caption]**

ID	Native language	Expertise	Word types annotated (number of tokens) **[TODO alphabetize]**
A	German	expert	Flagge (55), Ringen (55), Tschechen (56)
B	German	intermediate	halten (56), Mörder (56), Tatort (56)
C	German	novice	halten (56), Pollen (55), E-mail (56)
D	German	intermediate	Pollen (53), Flagge (49), Ringen (49)
E	English (US)	intermediate	Tschechen (56), halten (56), Mörder (56)
F	German	intermediate	Tatort (56), Frühling (56), fliegen (56)
G	Hebrew	intermediate	fliegen (0), Pollen (0), Flagge (20)
H	German	expert	Frühling (56), fliegen (56), Pollen (55)
I	German	intermediate	Ringen (55), Tschechen (56), halten (56)
J	German	intermediate	Mörder (56), Tatort (56), Frühling (56)
K	English (US)	intermediate	manche (56), E-mail (56), tragen (55), fliegen (56), Pollen (55), Flagge (55)
L	German	novice	Flagge (54), **[TODO mention?]** Tatort (56), E-mail (56)
M	German	intermediate	Ringen (54), **[TODO mention?]** Frühling (56), tragen (55)
N	German	novice	Tschechen (56), fliegen (56), manche (56)
O	German	intermediate	Mörder (56), manche (56), tragen (55)

Each annotator was assigned three word types to annotate in a single session, with the exception of one annotator who was assigned six word types over two sessions (see Section 3.3 for a description of an annotation session). Table 3.2 lists the word types assigned to each annotator, along with the number of tokens labeled for each type. Some judgments by annotators **[TODO D and G]** had to be excluded from the analysis due to technical problems; the token counts for each annotator in table 3.2 reflect only their usable judgments. **[TODO move following to agreement section in results?]** Word types were assigned such that each word type was annotated by at least two native German speakers, and to maximize the amount of overlap between annotators in order to obtain as many pairwise measures of annotator agreement as possible (see Section 3.4 for a discussion of inter-annotator agreement); table 3.3 lists the number of annotators for each word type.

Table 3.3: Number of annotators by word type **[TODO add expertise levels]** **[TODO move to agreement section?]**

Word type **[TODO (Tokens)]**	Native	Nonnative	Total
E-mail	2	1	3
Flagge	3	2	5
fliegen	3	1	4
Frühling	4	0	4
halten	3	1	4
manche	2	1	3
Mörder	3	1	4
Pollen	3	1	4
Ringen	4	0	4
Tatort	4	0	4
tragen	2	1	3
Tschechen	3	1	4

3.3 Annotation method

The annotation task consisted of assigning one of the following labels to each token of the selected word types, i.e. each utterance of each word by each L1 French speaker in the corpus:

[TODO decide on format for labels ([this]?)]

- **correct**: the speaker audibly stressed the lexically stressed (initial) syllable
- **incorrect**:the speaker audibly stressed the lexically unstressed (final) syllable
- **none**: the speaker did not clearly stress either syllable, i.e. did not audibly differentiate stressed and unstressed syllables, or the annotator was unable to determine which syllable was stressed
- **bad_nsylls**: the speaker pronounced the word with an incorrect number of syllables (i.e. by inserting or deleting a syllable), rendering it impossible to judge whether stress was realized correctly or not
- **bad_audio**: a problem with the audio file (e.g. noise in the signal or very inaccurate segmentation) interfered with the annotator's ability to judge the stress realization

Annotation proceeded by means of a graphical tool scripted in Praat (Boersma and Weenink, 2014), the main interface of which is shown in fig. 3.1. At the top, a word's text is displayed, along with the IFCASL corpus ID number of the speaker whose utterance of that word will be annotated (this number is only relevant for the annotator insofar as changes in its value inform the annotator that the speaker is changing from utterance to utterance). The recording of the word is played once automatically; the annotator may then choose to click one of the green buttons to play the word again, or play the recording of the entire sentence, as many times as they wish. Once the annotator has judged the accuracy of the lexical stress realization in this utterance, they log that judgment by clicking one of the gray buttons. The

annotator is then automatically advanced to the next utterance, with the counts in the lower right corner tracking their progress towards the total number of tokens to be annotated.

A single annotation session consisted of annotating all tokens of three word types, and lasted approximately 15 minutes. As mentioned in Section 3.2 above, each annotator participated in one session, with the exception of annotator L who participated in two sessions (separated by several days) and annotated a total of six word types.

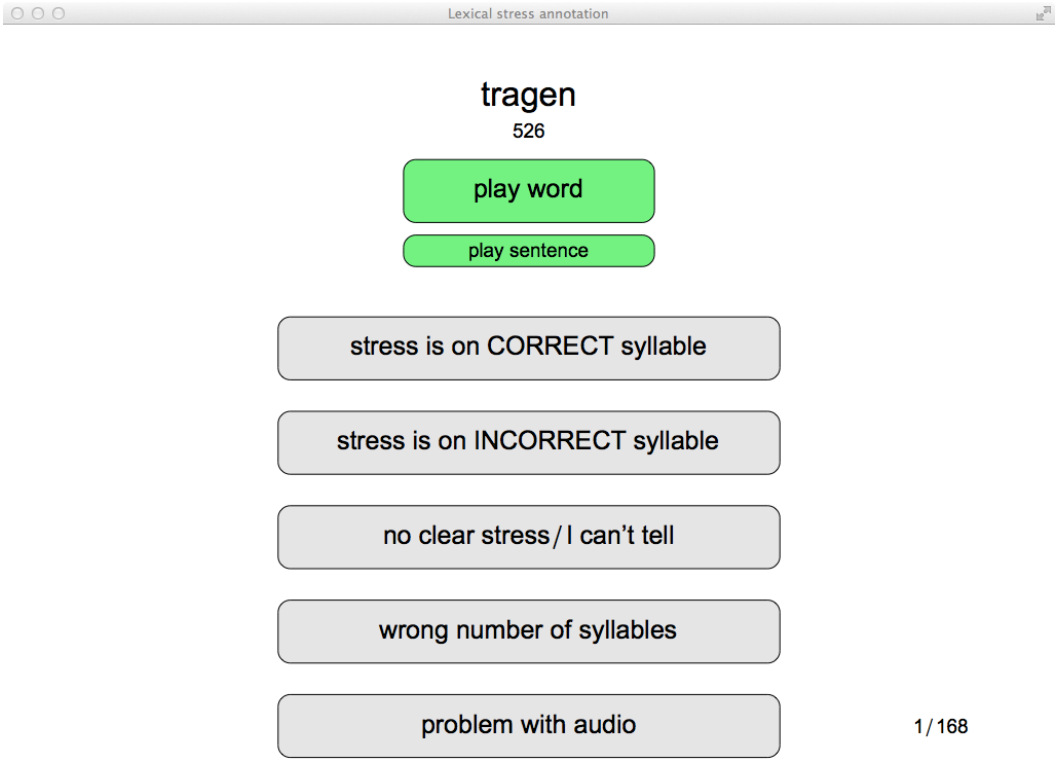


Figure 3.1: A screenshot of the graphical annotation tool scripted in Praat. Green buttons allow the annotator to listen to the word and sentence utterances. Gray buttons allow the annotator to record their judgment of stress accuracy; from top to bottom, the buttons correspond to the labels [correct], [incorrect], [none], [bad_nsylls], and [bad_audio]. **[TODO border around graphic]**

[TODO some kind of section wrap-up?]

3.4 Inter-annotator agreement

To create a useful CAPT system for lexical stress errors in nonnative German, i.e. to automatically detect whether a student has made a lexical stress error in a given utterance, it is helpful to have an understanding of the difficulty of the error-detection task, not only for machines but for humans. It is therefore useful to analyze the collected stress accuracy judgments in terms of inter-annotator agreement, in order to gain insight into the nature of the challenge this task presents. If it is uncommon for human annotators to agree about whether a given lexical stress realization is correct or incorrect, this may indicate that identifying lexical stress errors is a challenging task, and one which an automatic system

should also be expected to have difficulty with. If, on the other hand, human annotators are generally in strong agreement, this may reflect a lower level of difficulty, and give reason to judge the performance of an automatic system by a higher standard.

As stated in the previous section, lexical stress realizations in a total of 668 word utterances were each assigned to one of five classes by multiple annotators, based on whether the annotator judged the production to have correctly placed stress, incorrectly placed stress, no clear stress placement, or other problems which prevented the annotator from making a judgment about the lexical stress accuracy. The agreement between these judgments was calculated for each pair of annotators who overlapped, i.e. labeled any of the same tokens. **[TODO matrix of pairwise tokens in common (or just x/o to show which annotators overlapped?]** Two metrics were used to quantify agreement between a pair of annotators: the simple percentage of observed agreement, and Cohen's Kappa statistic (κ).

For a given pair of annotators, percentage agreement is calculated as the number of tokens to which both annotators assigned the same label, divided by the total number of tokens labeled by both annotators **[TODO as formula?]**. Possible values for percentage agreement range from 0%, representing complete disagreement between annotators, to 100%, representing complete agreement. This simple metric ignores the probability of annotators agreeing by chance, and therefore may give a somewhat optimistic picture of inter-annotator agreement, but nevertheless serves as a basic, easy-to-interpret preliminary indication of the reliability of the collected judgments.

To account for chance agreements not captured by the simple percentage of agreement, a second, more robust measure of inter-annotator agreement, Cohen's κ (Cohen, 1960), was also calculated for each pair of annotators. For a given pair of annotators who have labeled the same tokens, κ is computed as

$$\kappa = \frac{p_a - p_c}{1 - p_c}$$

where p_a is the proportion of tokens assigned the same label by both annotators (i.e. the simple percentage agreement just described) and p_c is the proportion of tokens which can be expected to receive the same label from both annotators purely by chance. The latter thus represents the probability of the two annotators agreeing by chance, and is calculated for a pair of annotators A and B as

$$p_c = \sum_{s \in S} p_A(s) \times p_B(s)$$

where s is one of the stress judgments in the set of possible labels S:

$$S = \{\text{[correct]}, \text{[incorrect]}, \text{[none]}, \text{[bad\_nsylls]}, \text{[bad\_audio]}\}$$

and $p_A(s)$ is the proportion of tokens assigned the label s by annotator A, calculated as the number of tokens assigned label s by annotator A divided by the total number of tokens labeled by annotator A; $p_B(s)$ is calculated in the same way for annotator B. As κ thus accounts for the probability of two annotators assigning a token the same label purely by chance, it provides a more conservative representation of inter-annotator agreement. A κ value of 0 indicates that the annotators do not agree any more than would be expected by chance. If agreement between annotators is less than chance, κ will take a value below

0. The maximum possible value of κ is 1.00, which indicates perfect agreement between annotators.

In the following sections, both measures are provided in the hopes of presenting a more comprehensive picture of inter-annotator agreement than either metric can convey alone.

[TODO Anything else to say here?]

3.4.1 Overall agreement

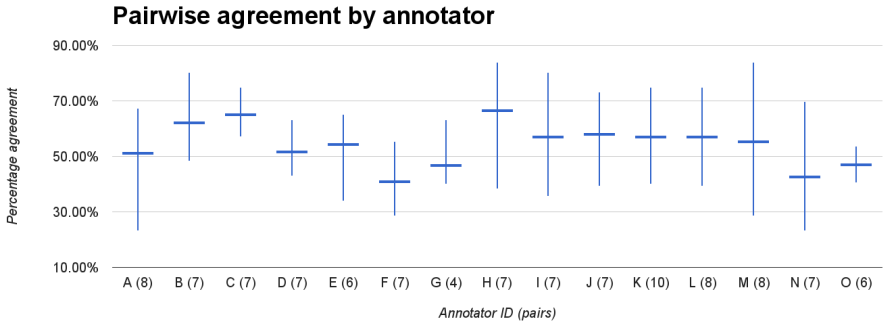
To obtain an overall measure of inter-annotator agreement for this lexical stress assessment task, the agreement between each pair of overlapping annotators was quantified by the metrics discussed in the previous section, and the minimum, median, mean, and maximum values over all pairwise comparisons were computed; these values are given in table 3.4. Though this provides a rather coarse-grained picture of the overall agreement, this simple analysis already points to a few interesting observations. First of all, we observe that the mean and median percentage agreement are near 55%, indicating that, roughly speaking, annotators agree just slightly more often than they disagree; **[TODO *fix:* this is not necessarily an encouraging ratio]**. Turning to the κ values, if we consider that $\kappa = 0$ represents agreement purely by chance while $k = 1$ represents perfect, meaningful agreement, the fact that the mean and median κ values between annotators are somewhere near 0.25 indicates that the agreement observed between annotators is closer to what would be expected simply by chance than to agreement that would indicate high reliability **[TODO *remove?:* or some type of objective truth]**. Looking next at the minimum and maximum values, we observe that while some pairs of annotators seem to exhibit relatively high agreement, indicating **[TODO *too fuzzy?* reasonably reliable]** judgments, other pairs have very low agreement; in one case, with 23.21% agreement, the annotators seem to be closer to perfect disagreement than perfect agreement, and the corresponding κ being below zero indicates that they agreed even (slightly) less than one would expect if they were merely labeling utterances randomly.

Table 3.4: Overall pairwise agreement between annotators **[TODO transpose this table so it's consistent with table 3.5 and table 3.6]**

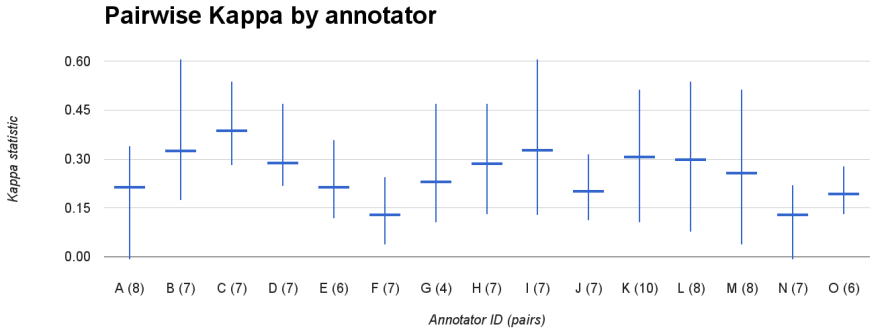
Agreement measure	Minimum	Median	Maximum	Mean
Percentage agreement	23.21%	55.36%	83.93%	54.92%
Cohen's κ	-0.01	0.26	0.61	0.23

[TODO Is this section even meaningful? Should it be left out?] It seems, then, that there may be stark differences in reliability from annotator to annotator. Analysis of the set of pairwise comparisons between a given annotator and all overlapping annotators provides more insight into that annotator's individual reliability; fig. 3.2 illustrates the pairwise agreements involving each of the 15 annotators. These figures should be interpreted with caution because they do not account for differences in the number of overlapping annotators/tokens available for each annotator **[TODO reference overlap table]**; nonetheless, it seems that there is indeed some noticeable variation from annotator to annotator **[TODO finish this**

paragraph] [TODO table of percent/kappa min/avg/max by annotator, since graphs are difficult to read precisely?]



(a) Percent agreement



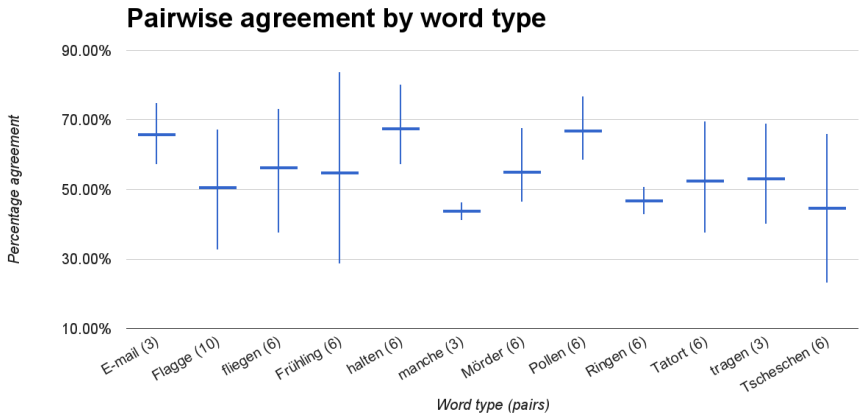
(b) Cohen's Kappa statistic

Figure 3.2: Each annotator's pairwise agreement with all other annotators with whom they overlapped. Numbers in parentheses indicate the number of pairwise comparisons involving each annotator. The bottom of each vertical bar represents the minimum pairwise value, the top the maximum. Horizontal bars indicate mean pairwise values.

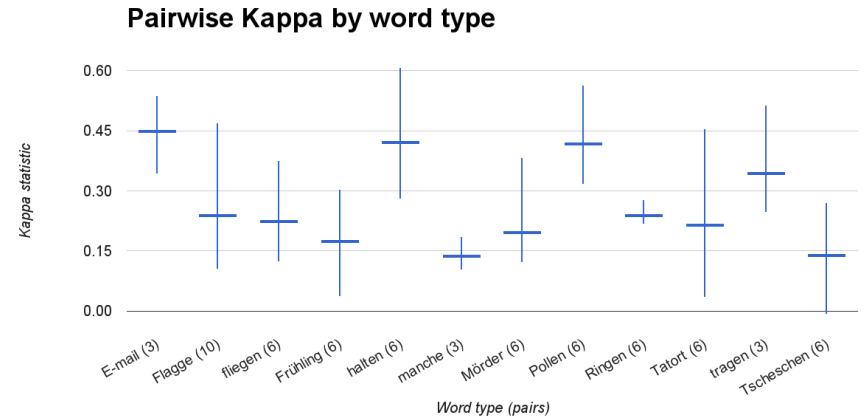
It is also of interest to analyze the overall inter-annotator agreement for each word type in the annotated sub-corpus. As fig. 3.3 illustrates, there are noticeable differences between word types, with annotators exhibiting relatively high agreement on certain words (e.g. *E-mail*, *halten*, and *Pollen*), and on other words (e.g. *manche* and *Ringen*) exhibiting agreement values closer to chance. **[TODO DISCUSSION (reference error breakdown by word in sec:results:overall]**

[TODO Move this?] On the whole, then, it seems that inter-annotator agreement in this lexical stress error annotation task is relatively low, which indicates that the task of assessing a given lexical stress realization as correct or incorrect is a relatively difficult one.

[TODO transition]



(a) Percent agreement



(b) Cohen's Kappa statistic

Figure 3.3: Pairwise agreement between annotators for each word type. Numbers in parentheses indicate the number of pairwise comparisons available for each word type. The bottom of each vertical bar represents the minimum pairwise value, the top the maximum. Horizontal bars indicate average pairwise values.

3.4.2 Native vs. nonnative annotators

Going beyond the coarse-grained analysis of inter-annotator agreement described in the previous section, we come now to the second question raised at the beginning of this chapter:

Are there differences in how native and non-native German speakers identify errors?

[TODO expectations/speculations of what we will find?]

To answer this question, it is useful to look at the inter-annotator agreement between native and non-native annotators, as well as at the distribution of label types within each group.

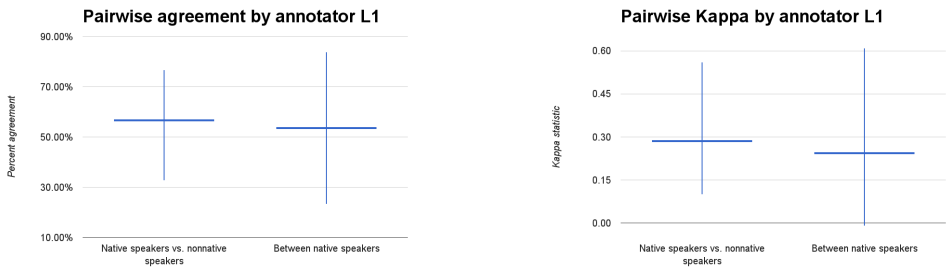
Figure 3.4 illustrates the inter-annotator agreement for all pairs in which one annotator was a native German speaker and the other a non-native speaker, as well as agreement between

pairs in which both annotators were native speakers. Due to the small size of the non-native group (3 annotators) and the aforementioned technical problems with annotator G's data (see Section 3.2), there was very little overlap between non-native annotators (only one pairwise comparison), preventing meaningful analysis of agreement within the non-native group. The precise mean, maximum, median, and minimum pairwise values for the two agreement metrics are listed in table 3.5, for both native-nonnative pairs and native-native pairs.

Looking at these statistics, we see little difference between the two types of pairs; in particular, the mean percentage agreement and κ values for native-nonnative and native-native pairs are quite close. If anything, it would appear that agreement within the native annotator group is slightly lower and more varied than agreement between the native and nonnative groups, though this may be explained by the larger number of native-native pairs compared to native-nonnative. It would therefore seem that these inter-annotator statistics do not tell us much about difference between how the two groups of annotators judge lexical stress accuracy.

Table 3.5: Inter-annotator agreement between native and non-native annotators (pairwise)

	Native vs. nonnative		Native vs. native	
	% Agreement	Cohen's κ	% Agreement	Cohen's κ
Mean	56.98%	0.29	53.87%	0.25
Maximum	76.79%	0.56	83.93%	0.61
Median	57.14%	0.25	50.91%	0.23
Minimum	32.65%	0.10	23.21%	-0.01

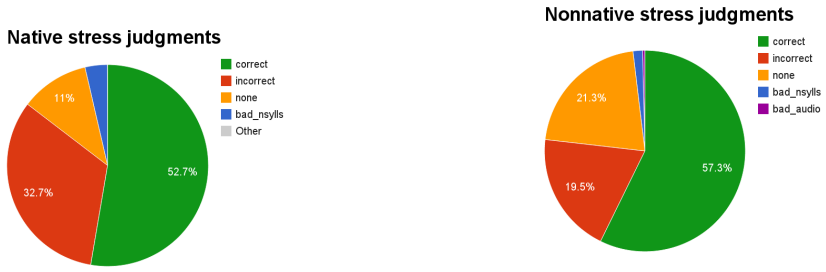


(a) Percent agreement **(b)** Cohen's Kappa statistic

Figure 3.4: Pairwise agreement between annotators based on their L1 (native or nonnative German speaker). The bottom of each vertical bar represents the minimum pairwise value, the top the maximum. Horizontal bars indicate average pairwise values.

However, in comparing the relative frequencies of the different labels assigned by annotators in these two L1 groups, a more noticeable difference between the groups begin to emerge. As illustrated in fig. 3.5, we observe that the native and nonnative speakers judge utterances as having correct lexical stress with approximately the same frequency: 52.7% of native annotators' judgments are **[correct]**, vs. 57.3% for nonnative annotators. However, non-native speakers seem to choose the **[none]** label somewhat more frequently than native speakers (21.3% vs. 11%); this could indicate that nonnative speakers are less confident

about how stress should be realized in German, resulting in less certainty about whether a given utterance is correct or not. **[TODO update/verify this paragraph]**



(a) Native annotators **(b)** Nonnative annotators

Figure 3.5: Stress judgments made by native and nonnative German speakers

Though the differences between native and nonnative annotators are interesting **[TODO from the perspective of X]**, the ultimate goal of this thesis project is to create a CAPT tool which will help L1 French speakers be more intelligible when speaking German as L2, and therefore the way in which native German speakers perceive lexical stress in non-native speech is of more relevance to this work than the way it is perceived by non-native speakers. Therefore, the remainder of this chapter is concerned exclusively with the judgments of native annotators, and judgments by non-native annotators are not included in the analyses that follow.

3.4.3 Expert vs. novice annotators

[TODO Transition] This section seeks to answer the last of the questions raised at the beginning of the chapter concerning inter-annotator agreement in the stress-annotated IFCASL sub-corpus, namely:

Are there differences in how expert and novice annotators identify lexical stress errors?

Given the general difficulty of the task of identifying lexical stress errors, evidenced by relatively low overall inter-annotator agreement as discussed in Section 3.4.1 above, it might seem reasonable to suppose that training in phonetics/phonology or experience annotating (non-native) speech might have a positive impact on an annotator's ability to reliably judge the accuracy of lexical stress realizations by non-native speakers. However, it once again bears mentioning that the ultimate goal of this work is to help L2 learners communicate intelligibly in German, and it can safely be assumed that in the vast majority of cases such learners will be communicating more often with native speakers who possess little formal knowledge of speech science than with expert phoneticians. Therefore, even if differences in reliability do exist between expert and novice annotators, it is important that the perception of non-native lexical stress errors by non-experts not be ignored in favor of perception of such errors by experts **[TODO reword that, or just scrap this sentence?]**.

[TODO Better transition needed here?]

Just as the previous section analyzed native vs. non-native annotations in terms of inter-annotator agreement and differences in label distributions between those groups, this section uses analogous data to investigate the differences between annotators of the three different expertise levels – expert (exp.), intermediate (int.), and novice (nov.) – described in Section 3.2 above.

To determine inter-annotator agreement between the three expertise groups, percentage agreement and κ were tabulated for each pairing of annotators from different groups, i.e. for each of the following three pair types:

- Expert annotator vs. novice annotator
- Expert annotator vs. intermediate annotator
- Novice annotator vs. intermediate annotator

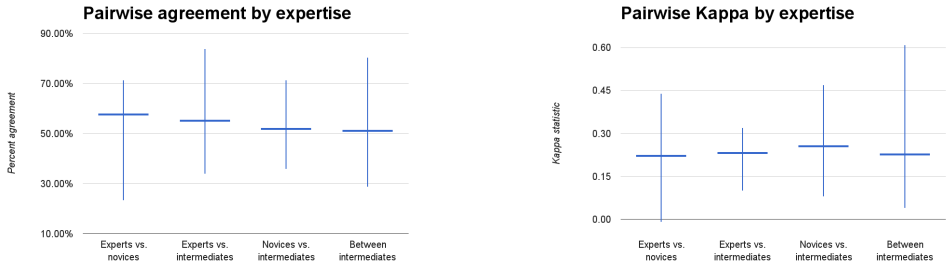
Additionally, pairwise agreement was tallied for pairings between two intermediate annotators, as a measure of inter-annotator agreement within this expertise group. Due to the small size of the expert and novice groups (two and three annotators, respectively), as well as the fact that expert annotators were deliberately not assigned overlapping tokens to label in an effort to maximize the number of tokens labeled by at least one expert **[TODO does that contradict the above statement about novice judgments being just as important as expert ones?]**, overlap within these groups was insufficient to calculate meaningful intra-group agreement statistics, so none are reported here. The small size of these two groups should also be kept in mind throughout the following analysis, as we should hesitate to draw firm conclusions from such small samples.

The agreement measures between groups and within the intermediate group are presented in table 3.6 and illustrated in fig. 3.6. As these figures show, the mean values of both percentage agreement and κ between the different expertise groups are quite close, and close to the overall means for all annotator pairs; interestingly, the highest mean percentage agreement observed in this comparison (though only by a small margin) is that of expert-novice pairings, which might be a preliminary indication that there is no relevant difference in reliability between expertise levels.

Table 3.6: Pairwise agreement between expertise groups **[TODO explain abbreviations]**

	Exp vs. Nov		Exp vs. Int		Nov vs. Int		Int vs. Int	
	% Agr.	κ	% agr.	κ	% agr.	κ	% agr.	κ
Mean	57.89%	0.23	55.30%	0.23	52.12%	0.26	51.44%	0.23
Maximum	71.43%	0.44	83.93%	0.32	71.43%	0.47	80.36%	0.61
Median	68.46%	0.24	49.95%	0.25	51.70%	0.26	47.58%	0.22
Minimum	23.21%	-0.01	33.93%	0.10	35.71%	0.08	28.57%	0.04

Figure 3.7 illustrates the relative number of each label type as assigned by annotators of the three expertise levels described in Section 3.2 above, and while any analysis of this data should bear in mind the small sample sizes of the expert and novice groups (two and three annotators, respectively), it does appear that some interesting differences may exist between the three groups.



(a) Percent agreement **(b)** Cohen's Kappa statistic

Figure 3.6: Pairwise agreement between annotators based on their level of expertise (expert, intermediate, or novice). The bottom of each vertical bar represents the minimum pairwise value, the top the maximum. Horizontal bars indicate average pairwise values.

Expert annotators seem to be far more "generous" in their labeling than intermediate or novice annotators, in that the experts assigned the **[correct]** label 73.6% of the time, in contrast with 49.3% and 54.8% for the other two groups respectively. **[TODO *person?:* One could]** speculate that experts' familiarity with nonnative speech and knowledge of possible inter-speaker variations in lexical stress realization may be the cause for this willingness to "accept" a high proportion of utterances as correct. **[TODO too many scare quotes in this paragraph?]**

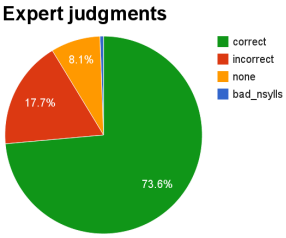
Another interesting difference can be observed between the intermediate and novice annotator groups: compared with the intermediate annotators, novices assign the **[none]** label less frequently (5.8% of the time, versus 16.3% for intermediates) and the **[bad_nsylls]** label more frequently (8.4% of the time, versus 2.1% for intermediates). Still keeping in mind the discrepancy in sample sizes when comparing 10 intermediate annotators to three novices, **[TODO *person?:* we might]** speculate that if experts' extensive experience with nonnative speech could be an explanation for their "generosity" with the correct label, novice annotators' lack of experience with nonnative speech could in a similar way make them "harsher" in judging nonnative utterances as having an incorrect number of syllables. **[TODO This paragraph sucks. Also too many scare quotes.]**

[TODO Conclusion]

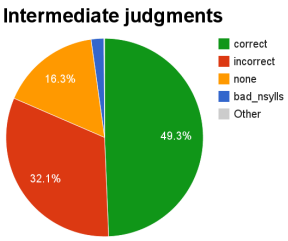
3.4.4 Choosing gold-standard labels

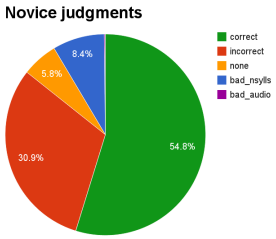
[TODO Rephrase gold-standard as ground truth or some other term?]

As the previous sections have illustrated, having multiple annotators judge the accuracy of each lexical stress production was useful insofar as it led to some interesting observations about the difficulty of reliably assessing lexical stress accuracy and differences in how judgments by annotators with different native languages and levels of expertise compare. However, if the annotations are to be used for **[TODO training an automated error-**



(a)Expert





(b)Intermediate

(c)Novice

Figure 3.7: Stress judgments by annotator expertise

diagnosis system], each token in the sub-corpus must ultimately be assigned a single "gold-standard" label from the set of possible labels described in Section 3.3.

In some cases, this assignment was trivial, while in others a decision had to be made between competing candidate labels. This section describes the procedure by which a single gold standard label was chosen for each word token (utterance) in the data set described in Section 3.1. In the remainder of this section, the gold-standard label chosen for a given word token t will be referred to as $s_{\text{gold}}(t)$, where $s \in S = \{$[correct],[incorrect],[none],[bad_nyslls],[bad_audio]$\}$ stands for one of the possible labels.

To prepare for gold-standard labeling, all available annotations for t were tallied by their label type s, resulting in a set $S_t \subseteq S$ of labels assigned to t by any of the native annotators who labeled this token (non-native annotators' judgments were omitted as mentioned in Section 3.4.2 above). For each label $s(t) \in S_t$, the number of "votes" for that label was recorded as the number of annotators who assigned this label to token t; henceforth **[TODO *need better notation for maxvotes set:* $S_{\text{max}} \subseteq S_t$]** will refer to the set of labels for t with the highest number of votes **[TODO reword?]**.

Given the observed labels and their vote counts, a rule-based procedure was followed to assign a gold-standard label $s_{\text{gold}}(t)$ to each token t in the annotated sub-corpus; this procedure is outlined in table 3.7. At each step i in the procedure, any tokens whose set S_t of observed labels fits condition C_i are assigned the gold-standard label described in column $s_{\text{gold}}(t)$; the number of tokens matching C_i is given as $N(C_i)$, and $N(C_{1\ldots i})$ represents the

total number of tokens which have been assigned a gold-standard label at the end of step i in the labeling procedure (i.e. the number of tokens matching C_i or any previous condition).

Table 3.7: Procedure for choosing a gold-standard label $s_{\text{gold}}(t)$ for a given token t. At step i, tokens matching C_i are assigned the label in column $s_{\text{gold}}(t)$. The rightmost columns $N(C_i)$ and $N(C_{1...i})$ list the number of tokens labeled in step i and the total number that have been labeled at the end of step i, respectively.

Step (i)	Condition (C_i) **[TODO reword w/ S_t&S_{max}]**	$s_{\text{gold}}(t)$	$N(C_i)$	$N(C_{1...i})$
1.	There is only one label s_{only} with any votes	s_{only}	268	268
2.	One label s_{max} has more votes than any other labels	s_{max}	265	533
3.	There is a label s_{expert} assigned by an expert	s_{expert}	51	584
4.	[bad_nsylls] is one of the competing labels, and there is only one other competing label s_{only} **[TODO reword]**	s_{only}	17	601
5.	There is a three-way tie between labels	[none]	6	607
6.	There are two competing labels: [none] and $s_{\text{certain}} \in$ {[correct],[incorrect]}	s_{certain}	21	628
7.	There are two competing labels: [correct] and [incorrect]	[correct]	40	668

[TODO Check following paragraphs for consistency of terms - shouldn't "Condition 1" be C_1, etc.?]

For 268 of the 668 tokens annotated, there was no disagreement whatsoever between annotators: for each of these 268 tokens, all annotators who labeled the token made the same judgment **[TODO How does this fit in with Section 3.4.1? Should it be mentioned there?]**, making it easy to assign this label as the gold standard for this utterance. Condition 1 in table 3.7 captures this category of tokens. For another 265 tokens, a majority of annotators assigned the same label, though one or more annotators dissented, so assigning the majority-vote label as the gold standard is logical; these are captured by Condition 2 **[TODO C_2?]** in the table. Therefore, for a total of 533 tokens (approximately 80% of the word utterances in the sub-corpus), the choice of $s_{\text{gold}}(t)$ was uncontroversial.

For the remaining utterances, choosing gold-standard labels was a less straightforward task, and the decisions made in steps 3-7 are somewhat more controversial. If either of the two expert annotators had labeled one of the remaining tokens, the expert's judgment was taken as the gold standard; 51 tokens met this condition (C_3). Next, in step 4, if there were exactly two labels in S_{max} and one of them was [bad_nsylls], the other label was chosen as $s_{\text{gold}}(t)$. The reasoning behind this step is that since the label [bad_nsylls] was intended to be applied to utterances for which no stress judgment was possible, then if at least one annotator was able to make a judgment, the [bad_nsylls] label must not be appropriate and should be rejected. This condition (C_4) applied to 17 tokens. The following step (5) addressed tokens for which the set of competing labels S_{max} had three members, i.e. for which there was a three-way tie between labels. The fact that so many different labels

were assigned to each of these tokens was taken as an indication that the accuracy of the lexical stress realization in this utterance was quite difficult to judge, i.e. it is unclear which syllable in the uttered word has been stressed; as the label [none] is intended to capture such cases, this label was chosen as the gold standard for the 6 utterances matching this condition (C_5). The next condition (C_6) captured the 21 cases in which S_{max} contained exactly two labels competing for gold-standard status, with one of the labels being [none] and the other being one of the two labels associated with certainty about the accuracy of the lexical stress realization, i.e. [correct] or [incorrect]. In these cases, [none] was rejected in favor of the certain label, based on the assumption that if at least one annotator was able to categorically classify the given utterance as correct or incorrect in terms of lexical stress, other native-speaking listeners might be inclined to make the same judgment **[TODO fix that sentence]**. The remaining 40 utterances were captured by the seventh and final condition, C_7, in which S_{max} contained exactly two labels: [correct] and [incorrect]. In these cases, the learner's utterance was assessed generously and the [correct] label was chosen as $s_{\text{gold}}(t)$, to capture the fact that as mentioned in Section 3.4.1, assessing the accuracy of a lexical stress realization seems to be a somewhat difficult task, and if at least one of the native speakers who heard the given utterance were willing to accept its stress realization as correct, the learner should not be **[TODO "penalized"]** by an [incorrect] label.

Despite the necessarily controversial nature of some of the labeling decisions described above, in the remainder of this thesis, the gold-standard labels chosen thus are taken as the ground truth for the distribution of lexical stress errors in this annotated subset of 668 word utterances from the IFCASL corpus. These gold-standard labels are used to analyze the distribution of errors in the corpus (see the following section), and also serve as training data for the supervised machine learning approach to stress error diagnosis described in Chapter 4**[TODO more specific section reference]**.

3.5 Results

[TODO Choose consistent naming for tables/graphs in this section - now some are "Errors" and others are "Stress judgments"]

[TODO intro]

[TODO Are the subsections in this section really necessary? Can they all be rolled into one?]

Given the final stress accuracy judgments compiled as described in the previous section, it is finally possible to return to the most important questions raised at the beginning of this chapter:

[TODO put these in the right order]

- Are lexical stress errors observed frequently in the IFCASL data? (Section 3.5.1)
- Is there a difference in the frequency of these errors among different groups of speakers (i.e. in terms of skill level, age, or gender) or in different contexts (e.g. after hearing a native speaker produce the word)? (Sections 3.5.3–3.5.5)
- Are lexical stress errors observed more frequently with certain word types than with others? (Section 3.5.2)
- How frequently do technical problems interfere with determining whether an error was made? (Section 3.5.6)

In the hope of providing tentative answers to these questions, this section describes and analyzes the distribution of errors in the IFCASL sub-corpus of 668 word tokens of 12 bisyllabic initial-stress word types as pronounced by L1 French speakers learning German as L2 (see Section 3.1), given the assessment of these errors made by native German speakers as described in Sections 3.2, 3.3 and 3.4.4.

3.5.1 Overall frequency of lexical stress errors

[TODO Maybe this section should go last instead of first?]

The overall distribution of the lexical stress accuracy judgments observed in the annotated IFCASL sub-corpus is detailed in table 3.8 and illustrated in fig. 3.8. Evidently, the majority (63.77%) of learners' lexical stress productions were judged to be correct; in other words, almost two-thirds of the time, learners clearly stressed the correct (initial) syllable in the uttered word. However, incorrect productions (productions in which the learner clearly stressed the incorrect syllable) and productions in which the learner did not clearly stress either syllable (corresponding to the [none] stress judgment, as described in Section 3.3), also occurred regularly: 29.64% of the productions were judged incorrect and 5.24% were labeled [none]. If we consider both of these types of productions as types of lexical stress errors, then errors were observed in just over one-third (34.88%) of the utterances annotated.

This sizable proportion of errors seems to give an affirmative answer to the question of whether lexical stress errors are observed frequently in L2 German speech by L1 French speakers. Bearing in mind that frequency of production is one of the criteria mentioned in Section 2.4.2 above for choosing a good error to target with CAPT, this provides further justification of the choice of lexical stress errors as the error type to focus on in this thesis project.

3.5.2 Errors by word type

To take a more detailed look at the errors observed in the annotated data, error judgments were broken down by word type, with the results of this analysis presented in **??** and illustrated in table 3.9. As should be expected, the most word types exhibit a distribution of errors quite similar to the overall distribution, i.e. a ratio of correct to incorrect utterances of approximately 2:1, broadly speaking. However, for the words *fliegen*, *Frühling*, and *Tschechen*,

Table 3.8: Overall frequency of stress judgments in the annotated corpus

Label	Tokens	% of corpus
correct	426	63.77%
incorrect	198	29.64%
none	35	5.24%
bad_nsylls	8	1.20%
bad_audio	1	0.15%
Total	668	100%

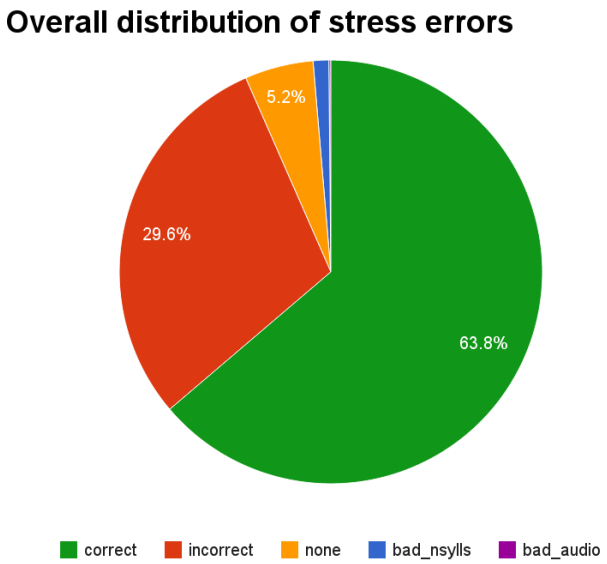


Figure 3.8: Overall distribution of lexical stress errors in the annotated corpus **[TODO retitle]**

a much higher proportion of correct stress realizations was observed, and for one word, *Tatort*, incorrect realizations actually exceeded correct productions by a noticeable margin (32 or 58.18% versus 20 or 36.36%, respectively).

Unfortunately, no clear explanations for these discrepancies between word types readily present themselves, though a few speculations will be offered here. Of the words with uncommonly high proportions of correct utterances, two of the three (*fliegen* and *Frühling*) occurred in the same sentence in the IFCASL corpus – *In Frühling fliegen Pollen durch die Luft* – along with another of the annotated word types, *Pollen*. This sentence, in part due to the occurrence of these three bisyllabic initial-stress words in immediate succession, exhibits a very regular metrical pattern **[TODO is this iambic or trochaic? explanation? reference?]**:

In	Früh-	ling	flie-	gen	Pol-	len	durch	die	Luft
-	x	-	x	-	x	-	x	-	x

As a result of this regularity, correctly realizing the prosody of each word in this sentence may present less of a challenge to L1 French speakers than a less regular sentence, and may thus explain their uncharacteristically flawless productions of the words therein. The fact that no fewer than the expected proportion of errors were observed in utterance of *Pollen* would seem to contradict this speculative explanation; however, unlike the other words, *Pollen* is doubly challenging for L1 French speakers, insofar as its first vowel is a short ɔ, as opposed to the long o: in the word *Polen* (meaning *Poland* in English), with which *Pollen* forms a minimal pair.[1] Differentiating between long and short vowels when speaking German is a notorious hurdle for French speakers **[TODO reference(s)]**. It may be the case that the short vowel in *Pollen*, and the existence of another similar-sounding word, is responsible for some of the errors observed in the data, even though *Pollen* and *Polen* share the same stress pattern (stress on the initial syllable), for one of two reasons: either the added challenge of producing a difficult vowel in *Pollen* distracts French speakers from the simple, regular prosody of the sentence, causing them to produce more prosodic errors with this word, or the native German annotators (most of whom, as discussed in Section 3.2, are not trained in phonetics or phonology) who are tasked with assessing the correctness of the word's prosody are distracted by the incorrect vowel quantity/quality in French speakers' productions of this word, and erroneously interpret the flaw(s) that they detect in the word's pronunciation as having to do with lexical stress, when in fact they are segmental errors.

As for the uncharacteristically large proportion of errors in *Tatort*, it may be the case that this word's resemblance to the common French words *ta* (*your*) and *tort* (*wrong*) interferes with French speakers' production, such that they realize the word as the plausible French word sequence *ta tort* (*your wrong(doing)* or *your mistake*), in which *tort* would unmistakably be the focus **[TODO is that inaccurate?]**, instead of realizing it correctly as a compound of the German words *Tat* (*act*) and *Ort* (*place*). However, once again it should be noted that this purely speculative explanation is not (and cannot be) verified by the data collected here.

[TODO outro?]

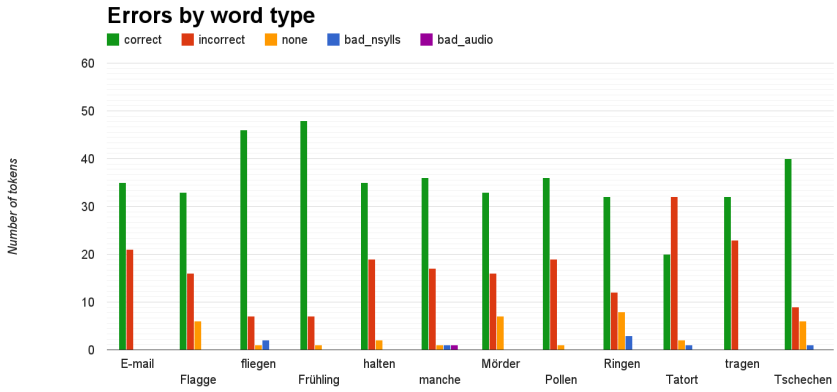


Figure 3.9: Errors by word type

[1] This minimal pair was of interest to the researchers who constructed the IFCASL corpus, and *Polen* also appears in the corpus, though it was not selected for inclusion in the sub-corpus to be annotated for lexical stress errors.

Table 3.9: Errors by word type **[TODO put %s in parens and lose 2nd row] [TODO omit 0 values?]**

Word type	correct		incorrect		none		bad_nsylls		bad_audio	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%
E-mail	35	62.50%	21	37.50%	0	0.00%	0	0.00%	0	0.00%
Flagge	33	60.00%	16	29.09%	6	10.91%	0	0.00%	0	0.00%
fliegen	46	82.14%	7	12.50%	1	1.79%	2	3.57%	0	0.00%
Frühling	48	85.71%	7	12.50%	1	1.79%	0	0.00%	0	0.00%
halten	35	62.50%	19	33.93%	2	3.57%	0	0.00%	0	0.00%
manche	36	64.29%	17	30.36%	1	1.79%	1	1.79%	1	1.79%
Mörder	33	58.93%	16	28.57%	7	12.50%	0	0.00%	0	0.00%
Pollen	36	64.29%	19	33.93%	1	1.79%	0	0.00%	0	0.00%
Ringen	32	58.18%	12	21.82%	8	14.55%	3	5.45%	0	0.00%
Tatort	20	36.36%	32	58.18%	2	3.64%	1	1.82%	0	0.00%
tragen	32	58.18%	23	41.82%	0	0.00%	0	0.00%	0	0.00%
Tschechen	40	71.43%	9	16.07%	6	10.71%	1	1.79%	0	0.00%

Table 3.10: Errors by speaker skill level

	correct	incorrect	none	bad_nsylls	bad_audio	Total	(% corpus)
A2	137	118	26	5	1	287	42.96%
B1	68	49	3	0	0	120	17.96%
B2	52	17	3	0	0	72	10.78%
C1	169	14	3	3	0	189	28.29%
Beginner (A2,B1)	205	167	29	5	1	407	60.93%
Advanced (B2,C1)	221	31	6	3	0	261	39.07%

3.5.3 Errors by L2 proficiency level

[TODO Be consistent with skill/proficiency or OK to use them interchangeably?]

As Section 3.1 stated, the L1 French speakers whose recordings comprise the annotated IFCASL sub-corpus span four levels of L2 German proficiency: A2 (elementary), B1 (intermediate), B2 (upper intermediate), and C1 (advanced). The rightmost column of table 3.10 gives the number and proportion of utterances from speakers of each level in the annotated sub-corpus, along with the number of utterances from speakers of each level that were assigned to each of the five possible stress-accuracy labels, and these figures are illustrated in figs. 3.10 and 3.11. Because the total number of utterances by speakers of each of the two intermediate (B) levels in the corpus is lower than the number by speakers of the lowest (A2) and highest (C1) levels, the judgments have also been grouped into two broader categories for easier comparison: beginners (A2 and B1) and advanced speakers (B2 and C1). The breakdown of stress errors by these groups is given in the lower portion of table 3.10 and illustrated in figs. 3.12 and 3.13.

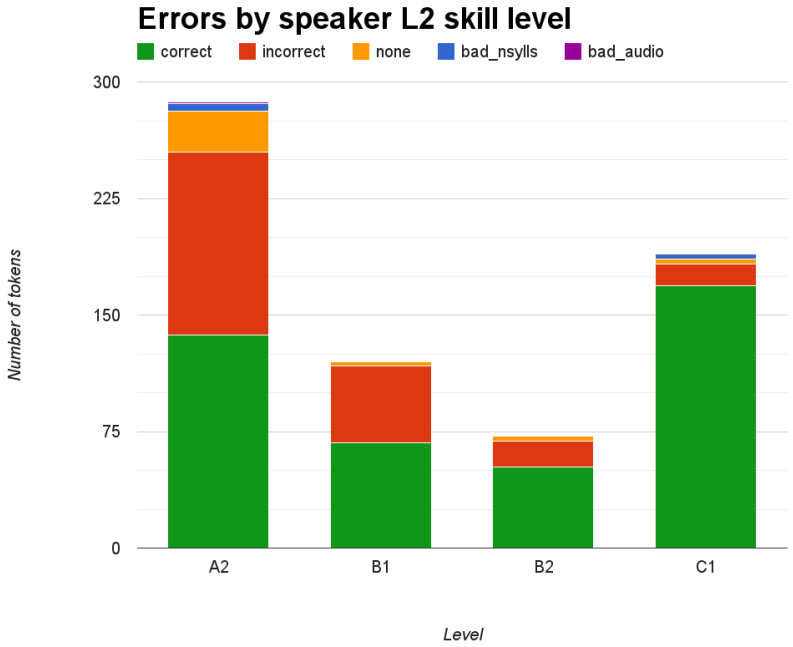


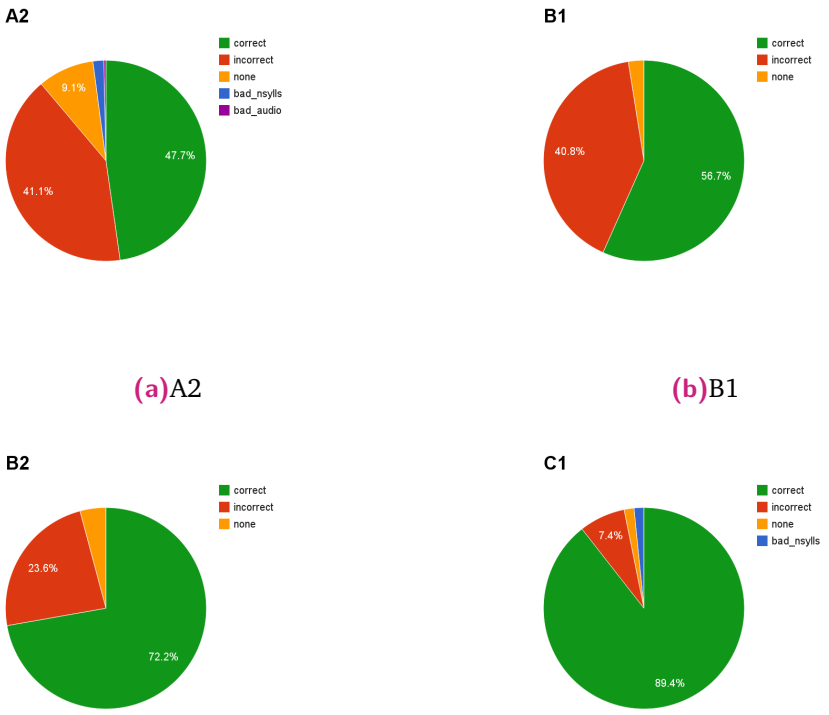
Figure 3.10: Stress judgments by speaker skill level **[TODO Exclude?]**

Unsurprisingly, these figures reveal that speakers of the higher levels (B2 and C1) seem to make a proportionally lower number of errors than speakers of the lower ones (A2 and B1), with each level exhibiting a lower proportion of errors than the level below it. Generally speaking, beginners (A2 and B1) seem to realize lexical stress correctly in about half of their utterances, whereas for upper intermediate (B2) learners the proportion of correct utterances is closer to three-fourths, and it approaches 90% for advanced (C1) learners. As previously established (see Section 2.4), a CAPT system targeting a particular type of error will only be useful if that error is produced with considerable frequency by the learners using the system; therefore, it would seem from the frequency of lexical stress errors in their speech that learners of lower proficiency levels may benefit more from a CAPT system targeting such errors than learners of higher proficiency.

[TODO anything else to say about this?]

3.5.4 Errors by speaker age and gender

Given that the IFCASL corpus, and by extension the sub-corpus annotated for lexical stress errors, contains recordings from speakers of both genders and from adult speakers (those over age 18 **[TODO check that]**) as well as children (see Section 3.1), an analysis of the errors observed in terms of the age and gender of the speakers is of interest, to determine whether any discernible differences exist between the different groups of speakers. The breakdown of errors for each of these groups is presented in table 3.11 and illustrated in figs. 3.15 and 3.16.



(a)A2 (b)B1

(c)B2 (d)C1

Figure 3.11: Error distribution by speaker skill level

With regard to the two different age groups of speakers, any interpretation of the results presented here must bear in mind the considerable difference in size between the two different groups: **[TODO reference the actual number of each type of speaker - should already have been presented in Section 3.1 or earlier]** of the 668 tokens annotated in total, 513 (over three-fourths) were from adult speakers while only 155 (less than one-fourth) were utterances by children. Furthermore, it must be highlighted that there is a strong interaction between age and proficiency level: all of the child speakers recorded in the IFCASL corpus are beginners (the majority at the A2 level with only 2 girls at B1), while the adults span all four levels. Given the discrepancies between L2 proficiency levels discussed in the previous section, then, it is not surprising to see that over half of children's utterances are judged to have lexical stress errors, with correct stress productions making up only 35.1% of utterances (54 utterances) by this age group. Adults, on the other hand, seem to realize lexical stress correctly in the majority of their utterances, with only 23.6% (121) incorrect productions and 3.3% percent (17) utterances with no clear lexical stress realization ([none]). However, this is not an entirely just comparison, given that the group of child speakers only includes beginners; therefore, instead of comparing the children's error distribution to that of all adults, it is helpful to restrict the comparison to adults of the lower proficiency levels. Table 3.11 lists the statistics for **[TODO *remove?:* adults at the A2 proficiency level only as well as for]** adults of both beginner levels (A2 and B1), and fig. 3.15b illustrates the error distribution for the latter group **[TODO include adult A2 chart also? (distribution is quite similar to A2/B1)]**. Comparing the distribution of

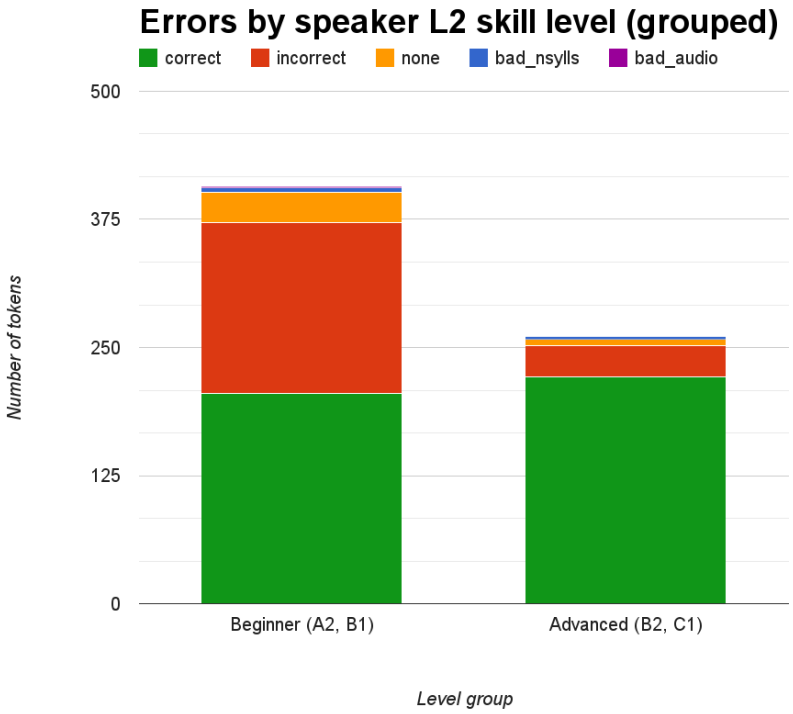


Figure 3.12: Stress judgments by speaker skill level (grouped)

children's errors to that of adult beginners, the difference is less drastic but still noticeable, as adult beginners realize lexical stress correctly in the majority (approximately 60%) of their utterances. Considering the comparatively high proportion of lexical stress errors in children's speech, therefore, it seems that just as **[TODO we]** concluded in the previous section that beginners may benefit more from a CAPT system targeting lexical stress errors than advanced learners would, so also may children stand to gain more from such a system than adult beginners. **[TODO anything else to say about age?]**

Coming now to the question of whether there is any difference in error distribution between speakers of different genders,

3.5.5 Errors by recording condition

[TODO]

3.5.6 Impact of technical problems **[TODO remove?]**

[TODO remove?]

(a) (b)

Figure 3.13: Error distribution by speaker skill level (grouped)

Table 3.11: Errors by speaker age and gender **[TODO add %ages]**

Group	correct	incorrect	none	bad_ nsylls	bad_ audio	Total	% of corpus
Boys	48	60	17	5	1	131	19.61%
Girls	6	17	1	0	0	24	3.59%
Men	184	61	7	0	0	252	37.72%
Women	188	60	10	3	0	261	39.07%
Children	54	77	18	5	1	155	23.20%
Adults	372	121	17	3	0	513	76.80%
Adults (A2 only)	86	49	9	0	0	144	21.56%
Adults (A2, B1)	151	90	11	0	0	252	37.72%
Females	194	77	11	3	0	285	42.66%
Males	232	121	24	5	1	383	57.34%

3.6 Summary

[TODO]

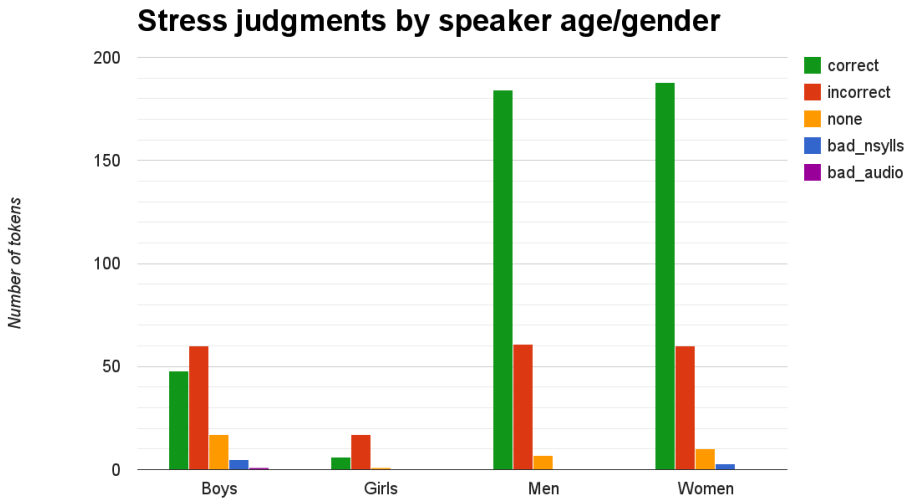


Figure 3.14: Stress judgments by speaker age/gender

(a) Children **(b)**Adult beginners

Figure 3.15: Error distribution by speaker age

(a) Females **(b)**Males

Figure 3.16: Error distribution by speaker gender

Table 3.12: Stress judgments by recording condition

Judgment	SH tokens	% of SH	SR tokens	% of SR
correct	97	58.79%	329	65.41%
incorrect	51	30.91%	147	29.22%
none	14	8.48%	21	4.17%
bad_nsylls	3	1.82%	5	0.99%
bad_audio	0	0.00%	1	0.20%
Total	165		503	
% of corpus	24.70%		75.30%	

(a)Sentence Read (b)Sentence Heard

Figure 3.17: Error distribution by recording condition

4 Diagnosis of lexical stress errors

In order to provide learners with useful feedback on their lexical stress errors in the L2, the CAPT system must first be able to detect and diagnose such errors in a learner's utterance. This requires at least:

(a) Reasonably accurate word-, syllable- and phone-level segmentation of the learner's L2 utterance;

(b) A means of analyzing how lexical stress is realized in the given utterance;

(c) A representation of how native speakers of the target language (would) realize lexical stress in the given sentence; and

(d) A way of comparing the learner's prosody to this representation.

In this section, we will examine how (a) will be achieved using forced alignment, and how problems in accuracy of the resulting segmentation can be overcome (Section 4.1); how the lexical stress analysis in (b) can be performed by measuring the fundamental frequency (F0), duration, and energy of the relevant parts of the speech signal (Section 4.2); and finally a variety of approaches to (c) and (d) (Section 4.3).

4.1 Automatic segmentation of nonnative speech

4.1.1 Segmentation via forced alignment

The native and non-native read speech recordings comprising the IFCASL corpus (Fauth et al., 2014; Trouvain et al., 2013) have been automatically segmented via forced alignment (Fohr et al., 1996; Mesbahi et al., 2011). This technique requires the expected text of the given utterance, acoustic models of the target language, and a pronunciation lexicon that describes the sequence of phones expected for each word. To account for non-native pronunciations, the lexicon is supplemented with a lexicon of non-native variants that might be encountered for each word.

The IFCASL recordings have already been segmented at the phone and word levels, and a subset of these automatic segmentations has been manually verified. However, segmentation at the syllable level still needs to be performed. This may be accomplished based on the word- and phone-level annotations by automatically or manually determining the sounds between which syllable boundaries are expected in each sentence from the text and phonetic lexicon, automatically extracting the locations of these boundaries from the phone-level segmentation, and automatically combining those boundaries with the word-level boundaries to create a new annotation level.

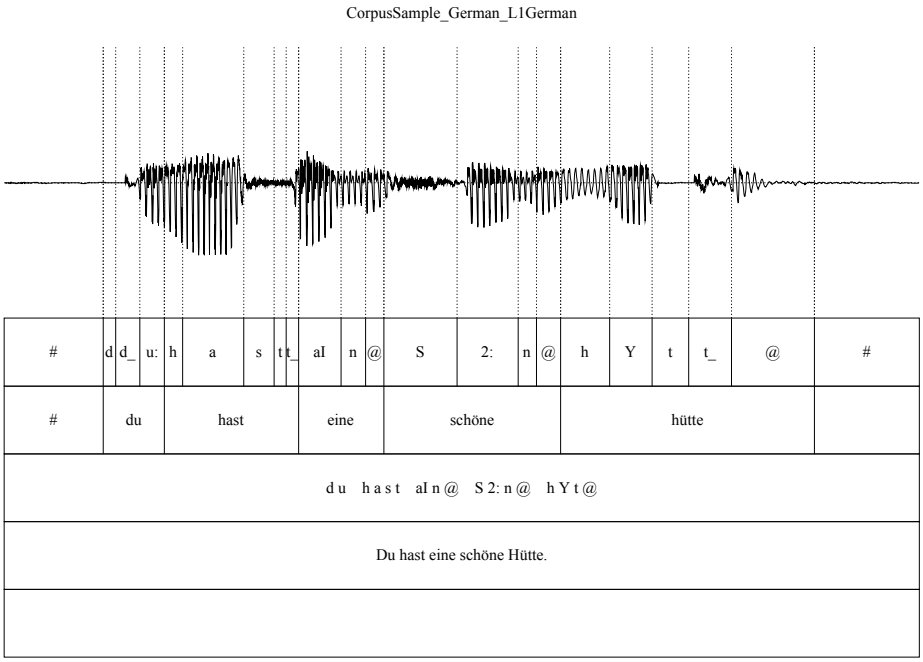


Figure 4.1: An example of a German utterance that has been segmented at the phone level (first row) and word level (second row). The third row contains the canonical (expected) native pronunciation of each word in the sentence, while the fourth row contains the written sentence of which the utterance is a reading.

4.1.2 Evaluation of segmentation accuracy

The accuracy of the forced-alignment segmentation can be assessed by computing inter-annotator agreement between the automatically produced segmentation and one or more manually-verified segmentations. The team at LORIA in Nancy has already completed this evaluation for the French IFCASL sub-corpus using the CoALT tool (Fohr and Mella, 2012). In cooperation with that team, the German sub-corpus (or a subset thereof) will be evaluated in the same way. A similar evaluation will be carried out for the syllable-level segmentations, a subset of which will be manually verified.

4.1.3 Coping with segmentation errors

Forced alignment is not a perfect method; because of the constraints put on the recognition system, the aligner will always find a match between the given text and audio, even if they do not correspond. Incorrect segmentation can lead to mistakes in diagnosis, so CAPT systems must have a means of reducing, or at least monitoring, the amount of error introduced by inaccurate segmentation (Eskenazi, 2009). In the proposed CAPT tool, this function may be served by the development of a simple sentence- and/or word-level confidence measure. While it is very difficult to compute such a measure directly from the decoding scores of the forced aligner, it may be possible to determine from the aforementioned accuracy evaluation which types of boundaries (e.g. between a sonorant and a vowel) the aligner typically has trouble detecting accurately, and then to calculate, for a given utterance, the proportion of

error-prone boundaries. While a very simplistic measure, this could nevertheless provide some indication of when (not) to trust the automatic alignment, thus impacting decisions on how and whether to attempt error diagnosis (or feedback). Other error-management strategies may also be explored, such as the type of error-filtering methods described by Mesbahi et al. (2011), Bonneau et al. (2012), and Orosanu et al. (2012), in which utterances which do not correspond to the expected text are detected and rejected before alignment is attempted.

4.2 Analysis of word prosody

This section will describe the features by which the system analyzes the lexical stress prosody of an utterance, be it the utterance of a learner or of a native speaker. These features relate to the three properties described in Section 2.3, namely duration (timing), fundamental frequency or F0 (pitch), and intensity (loudness). The features computed for each property are described in the corresponding sections below. Where possible, the diagnosis module of the CAPT tool will provide researchers control over the features used; for example, there may be an option to include all F0 and duration features but ignore intensity features.

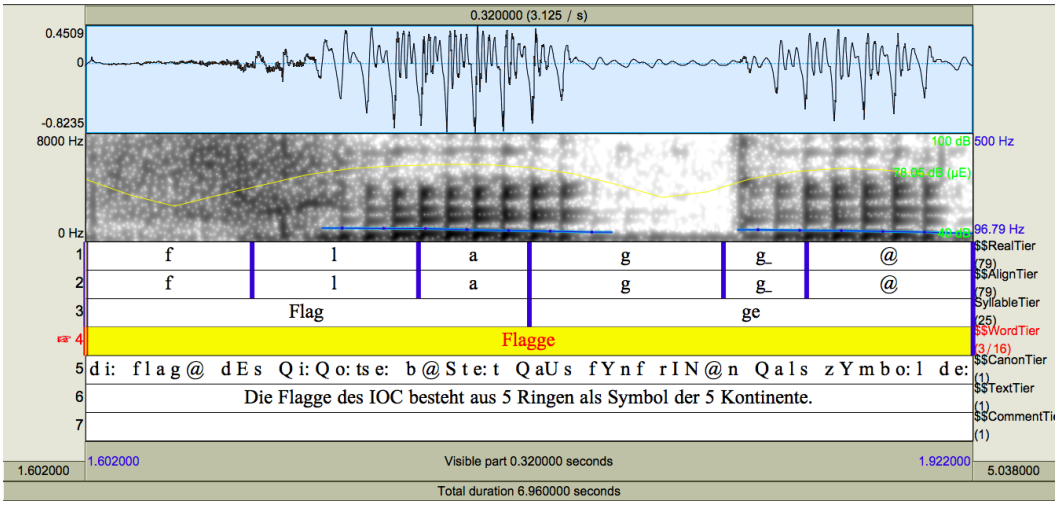
In this section, the features discussed are illustrated with their values for a word from two sample utterances of a German word selected from the IFCASL corpus; one by a L1 French speaker and the other by a L1 German speaker. The oscillogram, waveform, and annotation for these samples are shown in fig. 4.2.

In this work, the features described below have been computed from the automatically generated segmentation of an utterance **[TODO (see section X)]**, and not from a hand-corrected segmentation; as a result, the computed values may be slightly (or in some cases, significantly) inaccurate due to errors in the forced-alignment segmentation process. This reliance only on automatically-detected segment boundaries is intentional, as it simulates the conditions of an automatic, real-time tutoring system, which would need to perform segmentation on the fly and would not have recourse to human verification of segment boundary locations.

A potential complication of this analysis that should be pointed out relates to the fact that we are here dealing exclusively with read, and not sponaneous, speech. As Cutler (2005, p. 275) remarks, “acoustic differences between stressed and unstressed syllables are relatively large in spontaneous speech. With laboratory-read materials, however, such differences do not always arise”. Therefore, the task of recognizing prosodic deviations in learners’ read speech may be somewhat different than the corresponding task for spontaneous speech, and this difference should be kept in mind in the discussion that follows.

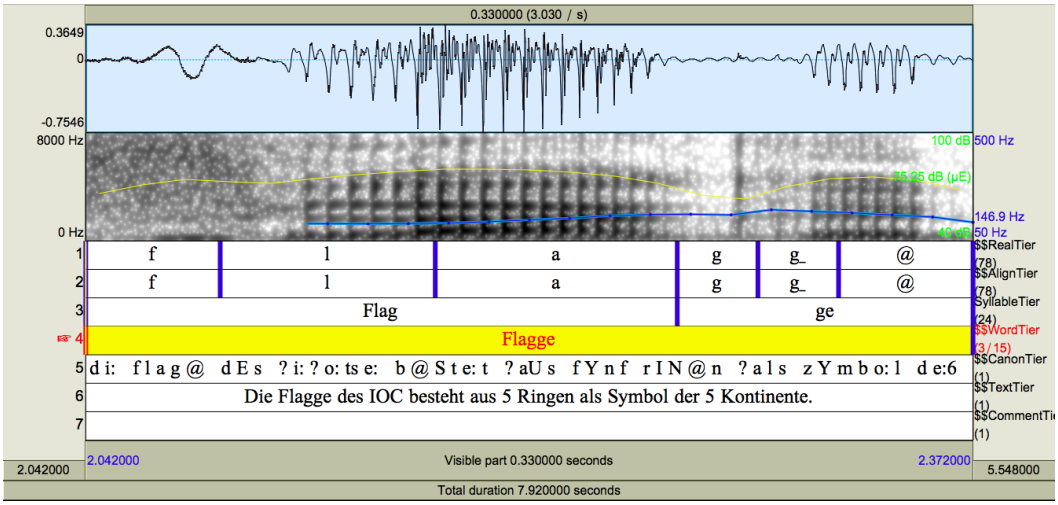
4.2.1 Duration

Analysis of duration (timing) is extremely important for detecting stress patterns; indeed, syllable duration may be the most important acoustic correlate of lexical stress in German (Dogil and Williams, 1999). Duration analysis therefore figures prominently in the analysis



(a)L1 French speaker (F)

[TODO Add incorrect FG example?]



(b)L1 German speaker (G)

Figure 4.2: Two sample utterances of the word "Flagge" from the IFCASL corpus, used to illustrate the features discussed in this section. [TODO description]

and assessment of learners' lexical stress in this work. Following Bonneau and Colotte (2011), we take into account the relative duration of each syllable of the word to be analyzed, as well as the relative duration of the vowels at the nucleus of each syllable. The complete list of features computed for each utterance is given in table 4.2, along with the values computed for a sample utterance from the IFCASL corpus **[TODO reference?]**.

Table 4.1: Features computed for duration analysis, and their values for the sample utterances of "Flagge" in fig. 4.2. Values are given in seconds.

Feature name	Description	Value (seconds)	
		(a) F	(b) G
WORD-DUR	Duration of entire word	0.32	0.33
SYLL0-DUR	Dur. of 1st syllable	0.16	0.22
SYLL1-DUR	Dur. of 2nd syllable	0.16	0.11
V0-DUR	Dur. of vowel in 1st syllable	0.04	0.09
V1-DUR	Dur. of vowel in 2nd syllable	0.06	0.05
SYLL-REL-DUR	SYLL0-DUR/SYLL1-DUR	1.00	2.00
V-REL-DUR	V0-DUR/V1-DUR	0.67	1.80

The relative utility of these features in automatically diagnosing lexical stress errors is discussed further in Section 4.4.

4.2.2 Fundamental frequency

As described in Section 2.3, the fundamental frequency (F0) of an utterance, which corresponds at the perceptual level to its pitch, also provides a strong signal of how lexical stress is realized in that utterance, and F0 features should therefore also contribute to the system's prosodic analysis. Much of the work on assessing non-native lexical stress has been conducted with English as the L2, and thus often makes the assumption that a stressed syllable should have a higher F0 than unstressed syllables (Bonneau and Colotte, 2011). In German, the F0 of a stressed syllable also tends to differ from the surrounding contour, but the difference may be positive (the stressed syllable has a higher pitch) or negative (lower pitch) (Cutler, 2005, p. 267). Therefore, features used to represent F0 may include the absolute value of the difference in average F0 between each pair of adjacent syllables in the word, or perhaps between the syllable which should carry (primary) stress and the rest of the word. To guard against unvoiced segments interfering with the F0 analysis, syllables may be represented by the vowels that form their nuclei. Relative differences between syllables may be more helpful than absolute differences. The F0 variation (range) over the entire word might also be informative of whether or not the speaker failed to stress any syllable.

4.2.3 Intensity

Research on lexical stress prosody has generally indicated that intensity is the least important of the three features, i.e. corresponds least closely to lexical stress patterns (Cutler, 2005). Indeed, existing lexical stress assessment tools may not take intensity into account, as is the case in the system described by Bonneau and Colotte (2011). However, intensity can

nonetheless have an impact on the perception of lexical stress, especially in combination with pitch or duration, or both (Cutler, 2005); Therefore, the diagnosis system should ideally take intensity into account when performing its prosodic analysis. This could be as simple as computing the total energy of the part of the signal corresponding to each syllable of the word in question, although more complex measures may be explored if time allows.

4.3 Comparison of native and nonnative speech

This thesis will explore a variety of approaches to modeling the lexical stress prosody of native speech in such a way that the learner's utterance can be automatically compared to that native model. This investigation, and the creation of a CAPT tool that allows researchers to easily switch between approaches to study their effects, will be one of the primary contributions of the thesis.

4.3.1 Using a single reference speaker

The most common approach to assessing L2 prosody involves comparing a learner's utterance to the same utterance produced by a native speaker of the target language; this approach is taken by Bonneau and Colotte (2011) and others.

Manually selecting a reference

The most basic way of selecting a reference speaker is to choose one manually. As a type of baseline, the CAPT tool will therefore enable the learner and/or the instructor/experimenter to choose a reference from a set of available speakers, with that set potentially being constrained by one or more properties of the speaker (e.g. gender).

Automatically selecting a reference

Another means of selecting a reference speaker would be to automatically choose a speaker whose voice resembles that of the learner (Probst et al., 2002). By analyzing speaker-dependent features of the speech of each reference candidate and of the learner – possibly in their L1 (French) as well as the L2 (German) – it should be possible for the system to rank reference candidates by proximity to the learner's voice. Relevant features may include F0 mean/range as well as spectral and duration-based features.

4.3.2 Using multiple reference speakers

However, when using a single native-speaker utterance for reference, even if the chosen speaker has been chosen carefully, we may be "over-fitting" to speaker- or utterance-dependent characteristics of the reference utterance that do not accurately represent the

"nativeness" of the reference speech. It would therefore be advantageous not to limit the diagnosis to comparison with a single reference speaker, but to instead compare the learner's speech with a variety of native utterances. This could be accomplished by conducting a series of one-on-one comparisons, pairing the learner utterance with a different reference utterance for each comparison, and then combining the results from all the comparisons. Factors to explore in this approach might include whether the set of reference speakers should be more or less constrained (e.g. by gender), and which metrics can be used to synthesize the one-on-one comparisons into a single diagnosis.

4.3.3 Using no reference speaker

Finally, a different approach may be to abstract away from the reference speaker(s). In their work on assessing children's reading fluency, Duong et al. (2011) found that evaluating a child's utterance in terms of a generalized prosody model, which predicts how a given text should be uttered, yielded more accurate fluency predictions than comparing it to a reference utterance of the text in question. It would be interesting to investigate whether the same principle applies in our CAPT scenario, so if time permits, this work will explore the possibility of constructing a more general model of native lexical stress realization, and comparing the learner's utterance directly to this model instead of to one or more reference utterances. This would theoretically enable the creation of exercises with arbitrary text, including sentences for which no reference utterance has been recorded. Possibilities for generalized lexical stress modeling include using word-prosody predictions from a text-to-speech synthesizer such as MARY (Schröder and Trouvain, 2003), as well as classification-based machine learning approaches such as those used by Shahin et al. (2012) and Kim and Beutnagel (2011) to categorize English words based on their stress patterns.

As this last diagnostic approach, using generalized lexical stress modeling, is the one which has been least explored in CAPT research, it will be the first priority for this thesis work after the baseline approach (manually selecting a single reference speaker) has been implemented. The next highest priority will be comparing the learner's speech to multiple reference speakers, followed by automatically selecting a reference speaker to match the learner's voice; these approaches will only be explored as time allows.

4.4 Evaluation

Lexical stress errors in the manually-annotated subset of the IFCASL corpus have not been explicitly labeled. We can assume that the utterances from L1 German speakers exhibit only correct German stress patterns, but a subset of the L1 French utterances will need to be annotated for lexical stress errors. This labeled data will be needed to assess the accuracy of the various error diagnosis methods which will be explored, and potentially to train classifiers to recognize correctly and incorrectly stressed words.

4.5 Summary

Table 4.2: Features computed for fundamental frequency (F0) analysis, and their values for the sample utterances of "Flagge" in fig. 4.2. Values are given in semitones.

Feature name	Description	Value (semitones)	
		(a) F	(b) G
WORD-F0-MEAN	Average (Avg.) F0, entire word	8.78	16.36
WORD-F0-MAX	Maximum (Max.) F0, entire word	10.73	20.08
WORD-F0-MIN	Minimum (Min.) F0, entire word	6.27	13.65
WORD-F0-RANGE	WORD-F0-MAX−WORD-F0-MIN	4.46	6.43
SYLL0-F0-MEAN	Avg. F0, 1st syllable	9.29	15.81
SYLL0-F0-MAX	Max. F0, 1st syllable	10.73	18.25
SYLL0-F0-MIN	Min. F0, 1st syllable	TD	TD
SYLL0-F0-RANGE	SYLL0-F0-MAX−SYLL0-F0-MIN	1.45	4.60
SYLL1-F0-MEAN	Avg. F0, 2nd syllable	8.24	17.51
SYLL1-F0-MAX	Max. F0, 2nd syllable	9.93	20.08
SYLL1-F0-MIN	Min. F0, 2nd syllable	TD	TD
SYLL1-F0-RANGE	SYLL1-F0-MAX−SYLL1-F0-MIN	3.66	5.86
SYLL-REL-MEAN	SYLL0-F0-MEAN/SYLL1-F0-MEAN	1.13	0.90
SYLL-REL-MAX	SYLL0-F0-MAX/SYLL1-F0-MAX	1.08	0.91
SYLL-REL-MIN	SYLL0-F0-MIN/SYLL1-F0-MIN	TD	TD
SYLL-REL-RANGE	SYLL0-F0-RANGE/SYLL1-F0-RANGE	0.40	0.78
SYLL-MAX-INDEX	$\begin{cases} 0, & \text{if SYLL0-F0-MAX} > \text{SYLL1-F0-MAX} \\ 1, & \text{if SYLL0-F0-MAX} < \text{SYLL1-F0-MAX} \end{cases}$	0	1
SYLL-MIN-INDEX	$\begin{cases} 0, & \text{if SYLL0-F0-MIN} < \text{SYLL1-F0-MIN} \\ 1, & \text{if SYLL0-F0-MIN} > \text{SYLL1-F0-MIN} \end{cases}$	1	0
SYLL-MAXRANGE-INDEX	$\begin{cases} 0, & \text{if SYLL0-F0-RANGE} > \text{SYLL1-F0-RANGE} \\ 1, & \text{if SYLL0-F0-RANGE} < \text{SYLL1-F0-RANGE} \end{cases}$	1	1

5 Feedback on lexical stress errors

Since the focus of this thesis is on pronunciation training, not pronunciation assessment (see Section 2.2.1), feedback on the errors diagnosed via the methods described in Chapter 4 will be an important component of the proposed CAPT tool. As mentioned in Section 2.1, the particular importance of corrective feedback in pronunciation training is generally acknowledged, though much remains to be learned about when and how feedback can be most effective. Therefore, one aim of this thesis is the creation of a feedback generation module for the lexical stress CAPT tool which will offer a variety of possible feedback types, and a Graphical User Interface (GUI) allowing a researcher or instructor to easily switch between feedback types. While it is outside the scope of the thesis to carry out in vivo studies with learners to determine which feedback types are most effective in which situations, the tool will hopefully facilitate such studies going forward.

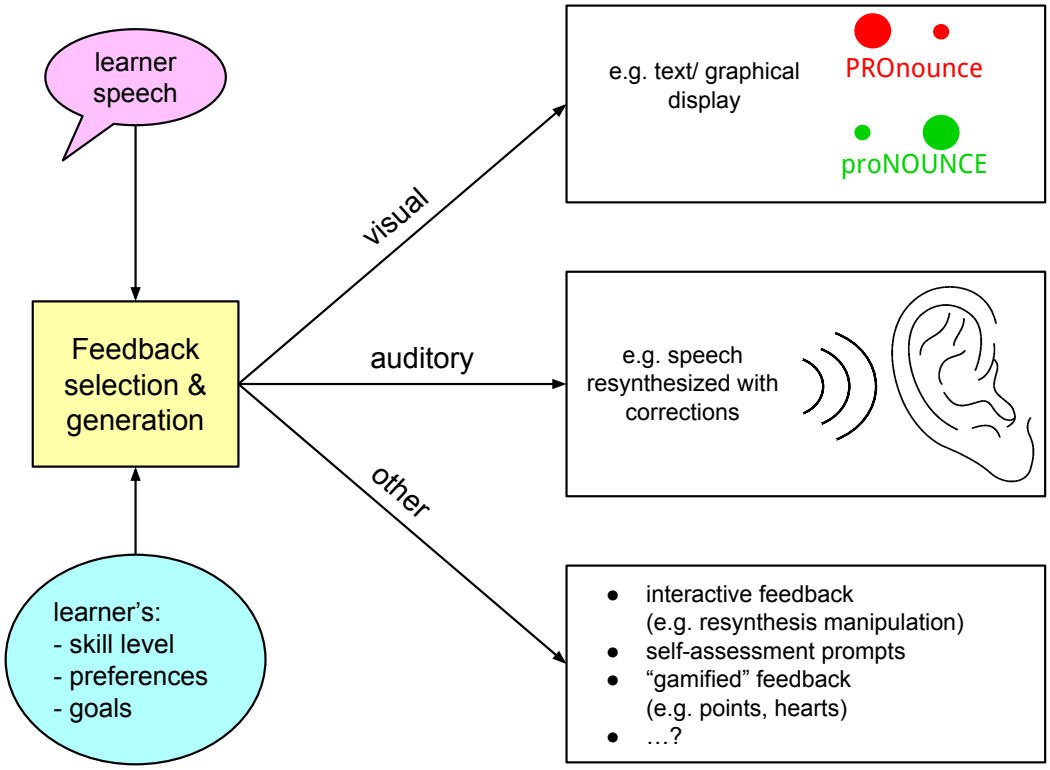


Figure 5.1: Delivery of prosody feedback in different modalities.

5.1 Visual feedback

5.1.1 Visualizations of the speech signal

In several existing CAPT tools, the learner is presented with relatively direct visualizations of the speech signal, such as its waveform (oscillogram) and spectrogram, often with overlays highlighting perceptually relevant properties such as the pitch contour and durations of various parts of the utterance. However, as Neri et al. (2002) point out, waveforms and spectrograms are signal representations designed for speech researchers, not language learners, and the latter may have difficulty understanding these visualizations without the proper training. To research whether this conjecture holds, these direct visualizations must be compared with alternatives in user studies with learners; several options for such alternative visualizations are explored in this section.

5.1.2 Graphical representations of prosody

One type of alternative would be a more abstract graphical representation of the lexical stress pattern in the native reference speaker and/or the learner's speech. Classroom materials for pronunciation instruction sometimes represent lexical stress patterns using dots or other shapes, one for each syllable, whose relative sizes indicate each syllable's prominence in the word (Hirschfeld and Reinke, 1998). This type of visualization would be relatively simple to implement, given that the reference or learner utterance can be classified into one of a set of stress patterns (Kim and Beutnagel, 2011; Shahin et al., 2012). It would also be possible to map the acoustic features of each syllable in the utterance(s) to graphical features of the representative shape, e.g. using size to represent duration, vertical position to represent F0, and darkness to represent intensity. To facilitate studies on which mappings, if any, make this feedback useful to the learner, the researcher-facing GUI should offer control over the different possible mappings.

5.1.3 Stylized text

This is essentially the approach used by Sitaram et al. (2011), though they modify the text of each word instead of a more abstract visual representation. As text stylization is also often used in pronunciation instruction materials (Behme-Gissel, 2005; Hirschfeld et al., 2007), it would be logical for the CAPT tool to offer text stylization as a feedback option. As with the shapes mentioned above, and following Sitaram et al. (2011), it would be interesting to explore the possible mappings between acoustic features and properties of the text of each syllable (e.g. size, weight, underlining/decoration, etc.), with these mappings controllable by the researcher via the GUI.

5.1.4 Other

Given some visual representation of the learner's utterance, be it textual or more abstract, visual feedback should also be given on what the learner can do to improve their lexical stress realization. Bonneau and Colotte (2011) deliver such feedback in the F0 dimension by displaying arrows which indicate whether the user should raise or lower the pitch of a given syllable to make their realization more like that of the reference speaker, and this is one option

for the CAPT tool. Another might be the use of animation to transform the visualization of the learner's (incorrectly realized) utterance into a corresponding visualization of the correct realization, e.g. by growing or shrinking the size of the dot or text for each syllable to visualize the desired change in duration, or showing it moving up or down to convey the desired change in pitch.

Implementation of at least one visual feedback type will be of high priority in this work. Stylized text and graphical representations will be explored first. If time allows, animation will be added to convey corrective feedback to the learner.

5.2 Auditory feedback

In foreign language classrooms, feedback on correct pronunciation is often given implicitly by allowing the learner to listen to a native speaker's production of the target utterance and/or a recording of their own production. However, previous work on delivering lexical stress feedback (see Section 2.2.1) has revealed that learners seem to benefit more from prosodically modified implicit feedback, either in the form of a learner utterance modified to reflect the "correct" prosody of a native reference utterance (Bonneau and Colotte, 2011), or a native utterance modified to place exaggerated emphasis on the stressed syllable (Bissiri et al., 2006; Bissiri and Pfitzinger, 2009).

At least one type of audio feedback type will be implemented in the CAPT tool, with the highest-priority option being prosodic modification of the learner's utterance to match a single, manually-selected reference utterance, following Bonneau and Colotte (2011); Jsnoori (Project-Team PAROLE, 2013) will be used to perform this modification. If a generalized lexical stress model is successfully integrated into the diagnostic module (see Section 4.3), the next highest-priority task will be performing prosodic modification of the learner's utterance based on this model. Emphasizing stressed syllables in the native reference utterance(s) will be of lowest priority.

5.3 Alternative feedback types

Other options, which will only be explored if time allows, include (in order of priority) feedback encouraging self-assessment and self-correction, metalinguistic feedback, and interactivity. Self-assessment and self-correction can be encouraged by presenting learners with targeted questionnaires before delivering diagnosis and feedback, e.g. asking learners to listen to their utterance and assess whether they have placed stress on the correct syllable, or asking how the speaker of an incorrect production could have realized stress properly ("By making the first syllable longer", etc.). Metalinguistic feedback, e.g. reminding learners of the stress rule(s) affecting the target utterance, could be delivered either visually (e.g. text displayed on the screen), auditorily (e.g. playback of an instructor's voice), or both. Interactivity could be achieved by allowing learners to interact with the resynthesis component to modify the prosody of their utterance, as is done in WinPitch LTL (Martin, 2004). By allowing researchers to easily control which of these feedback options to present

to the learner, the tool could facilitate research into the effects of alternative feedback types such as these, which have not yet been adequately studied in CAPT.

5.4 Summary

6 System overview?

This section is conceptually weak - **[TODO figure out what this section's purpose is, other than just to collect random technology facts that don't have any other home]**

6.1 Goal and architecture

6.2 Tools and technologies

6.2.1 Speech processing with Jsnoori

6.2.2 Machine learning with Weka

6.2.3 Web interface with Grails

6.3 User interface

6.3.1 For language learners

6.3.2 For teachers and CAPT researchers

7 Conclusion and outlook

Bibliography

Anderson-Hsieh, Janet, Ruth Johnson, and Kenneth Koehler (1992). "The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentais, Prosody, and Syllable Structure". In: *Language Learning* 42.4, pp. 529–555 (cit. on p. 5).

Behme-Gissel, Helma (2005). *Deutsche Wortbetonung: ein Lehr- und Übungsbuch*. Iudicium (cit. on p. 50).

Bissiri, Maria Paola and Hartmut R. Pfitzinger (2009). "Italian speakers learn lexical stress of German morphologically complex words". In: *Speech Communication* (cit. on pp. 8, 51).

Bissiri, Maria Paola, Hartmut R. Pfitzinger, and Hans G. Tillmann (2006). "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis". In: *Proceedings of the 11th Australian International Conference on Speech Science & Technology* (cit. on pp. 8, 51).

Boersma, Paul and David Weenink (2014). *Praat: doing phonetics by computer* (cit. on pp. 14, 17).

Bonneau, Anne and Vincent Colotte (2011). "Automatic Feedback for L2 Prosody Learning". In: *Speech and Language Technologies*. Ed. by Ivo Ipsic. 1977. InTech (cit. on pp. 7, 11, 45, 46, 50, 51).

Bonneau, Anne, Dominique Fohr, Irina Illina, Denis Jouvet, Odile Mella, Larbi Mesbahi, and Luiza Orosanu (2012). "Gestion d'erreurs pour la fiabilisation des retours automatiques en apprentissage de la prosodie d'une langue seconde". In: *Traitement Automatique des Langues* 53, pp. 129–154 (cit. on p. 43).

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46 (cit. on p. 19).

Cucchiarini, Catia, Ambra Neri, and Helmer Strik (2009). "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback". In: *Speech Communication* 51.10, pp. 853–863 (cit. on p. 9).

Cutler, Anne (2005). "Lexical Stress". In: *The Handbook of Speech Perception*. Ed. by David B Pisoni and Robert E Remez, pp. 264–289 (cit. on pp. 8–11, 43, 45, 46).

Delmonte, Rodolfo (2011). "Exploring Speech Technologies for Language Learning". In: *Speech and Language Technologies*. Ed. by Ivo Ipsic. InTech (cit. on p. 6).

Derwing, Tracey M and Murray J Munro (2005). "Second Language Accent and Pronunciation Teaching: A Research-Based Approach". In: *TESOL Quarterly* 39.3, pp. 379–397 (cit. on p. 5).

Dlaska, Andrea and Christian Krekeler (2013). “The short-term effects of individual corrective feedback on L2 pronunciation”. In: *System* 41.1, pp. 25–37 (cit. on p. 5).

Dogil, Grzegorz and Briony Williams (1999). “The phonetic manifestation of word stress”. In: *Word Prosodic Systems in the Languages of Europe*. Ed. by Harry van der Hulst. Berlin: Walter de Gruyter. Chap. 5, pp. 273–334 (cit. on p. 43).

Duong, Minh, Jack Mostow, and Sunayana Sitaram (2011). “Two methods for assessing oral reading prosody”. In: *ACM Transactions on Speech and Language Processing* 7.212, pp. 1–22 (cit. on pp. 7, 47).

Dupoux, Emmanuel, Núria Sebastián-Gallés, Eduardo Navarette, and Sharon Peperkamp (2008). “Persistent stress ’deafness’: The case of French learners of Spanish”. In: *Cognition* 106, pp. 682–706 (cit. on pp. 9, 11).

Eskenazi, Maxine (2009). “An overview of spoken language technology for education”. In: *Speech Communication* 51.10, pp. 832–844 (cit. on pp. 6, 42).

Eskenazi, Maxine and Scott Hansma (1998). “The Fluency pronunciation trainer”. In: *Proc. of Speech Technology in Language Learning*, pp. 77–80 (cit. on p. 7).

Eskenazi, Maxine, Yan Ke, Jordi Albornoz, and Katharina Probst (2000). “The Fluency Pronunciation Trainer: Update and user issues”. In: *Proc. of InSTIL 2000, Dundee* (cit. on p. 7).

Eskenazi, Maxine, Angela Kennedy, Carlton Ketchum, Robert Olszewski, Garrett Pelton, Forbes Ave, and Pittsburgh Pa (2007). “The NativeAccent(TM) pronunciation tutor: measuring success in the real world”. In: *SLaTE*, pp. 124–127 (cit. on p. 7).

Fauth, Camille, Anne Bonneau, and Frank Zimmerer (2014). “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, pp. 1477–1482 (cit. on pp. 1, 41).

Fohr, Dominique and Odile Mella (2012). “CoALT: A Software for Comparing Automatic Labelling Tools.” In: *LREC*, pp. 325–332 (cit. on p. 42).

Fohr, Dominique, JF Mari, and Jean Paul Haton (1996). “Utilisation de modèles de Markov pour l’étiquetage automatique et la reconnaissance de BREF80”. In: *Journées d’Etude de la Parole* (cit. on p. 41).

Hirschfeld, Ulla and Jürgen Trouvain (2007). “Teaching prosody in German as foreign language”. In: *Non-Native Prosody: Phonetic Description and Teaching Practice*. Ed. by Jürgen Trouvain and Ulrike Gut. Walter de Gruyter, pp. 171–187 (cit. on p. 5).

Hirschfeld, Ursula (1994). *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Vol. 57. Institut für Phonetik, JW Goethe-Universität (cit. on pp. 9, 10).

Hirschfeld, Ursula and Kerstin Reinke (1998). *Phonetik Simsalabim: Ein Übungskurs für Deutschlernender (Begleitbuch)*. Langenscheidt (cit. on p. 50).

Hirschfeld, Ursula, Christian Keß ler, Barbara Langhoff, Kerstin Reinke, Annemargret Sarnow, Lothar Schmidt, and Eberhard Stock (2007). *Phonothek intensiv: Aussprachetraining*. Ed. by Ursula Hirschfeld, Kerstin Reinke, and Eberhard Stock. Langenscheidt (cit. on p. 50).

Jilka, M and G Möhler (1998). “Intonational foreign accent: speech technology and foreign language teaching”. In: . . . . *ESCA Workshop on Speech Technology in* . . . (Cit. on p. 8).

Kim, Yeon-Jun and Mark C Beutnagel (2011). “Automatic assessment of american English lexical stress using machine learning algorithms.” In: *SLaTE*, pp. 93–96 (cit. on pp. 11, 47, 50).

Martin, Philippe (2004). “WinPitch LTL II, a multimodal pronunciation software”. In: *InSTIL/ICALL Symposium 2004* (cit. on pp. 8, 51).

Mehlhorn, G (2005). “Learner autonomy and pronunciation coaching”. In: *Proceedings of the Phonetics Teaching and Learning Conference, University College London* (cit. on p. 5).

Mesbahi, Larbi, Denis Jouvet, Anne Bonneau, and Dominique Fohr (2011). “Reliability of non-native speech automatic segmentation for prosodic feedback.” In: *SLaTE* (cit. on pp. 6, 7, 41, 43).

Mostow, Jack (2012). “Why and how our automated reading tutor listens”. In: *International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)* (cit. on p. 7).

Neri, A., C. Cucchiarini, H. Strik, and L. Boves (2002). “The pedagogy-technology interface in computer assisted pronunciation training”. In: *Computer Assisted Language Learning* (cit. on pp. 5, 6, 10, 50).

Orosanu, Luiza, Denis Jouvet, Dominique Fohr, Irina Illina, and Anne Bonneau (2012). “Combining criteria for the detection of incorrect entries of non-native speech in the context of foreign language learning”. In: *SLT 2012 - 4th IEEE Workshop on Spoken Language Technology* (cit. on pp. 6, 7, 43).

Probst, Katharina, Yan Ke, and Maxine Eskenazi (2002). “Enhancing foreign language tutors – In search of the golden speaker”. In: *Speech Communication* 37.3-4, pp. 161–173 (cit. on pp. 7, 46).

Project-Team PAROLE (2013). *Activity Report 2013*. Tech. rep. Nancy: LORIA (cit. on pp. 6, 51).

Schröder, Marc and Jürgen Trouvain (2003). “The German text-to-speech synthesis system MARY: A tool for research, development and teaching”. In: *International Journal of Speech Technology* 6, pp. 365–377 (cit. on p. 47).

Shahin, Mostafa Ali, Beena Ahmed, and Kirrie J. Ballard (2012). “Automatic classification of unequal lexical stress patterns using machine learning algorithms”. In: *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 388–391 (cit. on pp. 11, 47, 50).

Sitaram, S, J Mostow, Y Li, A Weinstein, D Yen, and J Valeri (2011). “What visual feedback should a reading tutor give children on their oral reading prosody?” In: *SLaTE* (cit. on pp. 7, 50).

Trouvain, Jürgen, Yves Laprie, and Bernd Möbius (2013). “Designing a bilingual speech corpus for French and German language learners”. In: *Corpus et Outils en Linguistique, Langues et Parole: Statuts, Usages et Mésuages*. ii. Strasbourg, France, pp. 32–34 (cit. on pp. 1, 41).

Weber, Frederick and Kalika Bali (2010). “Enhancing ESL education in India with a reading tutor that listens”. In: *Proceedings of the First ACM Symposium on Computing for Development - ACM DEV ’10*. New York, New York, USA: ACM Press, p. 1 (cit. on p. 7).

Wik, P, R Hincks, and JB Hirschberg (2009). “Responses to Ville: A virtual language teacher for Swedish”. In: (cit. on p. 11).

Witt, Silke M (2012). “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)*, pp. 1–8 (cit. on pp. 1, 5, 6, 11).