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

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

Declaration of originality

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 original work, with all assistance acknowledged.

Anjana Sofia Vakil

Saarbrücken, 31 March, 2015

Abstract

[TODO]

Acknowledgements

[TODO]

Contents

List of Figures

List of Tables

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)[1] at Saarland University (Saarbrücken, Germany) and LORIA (Nancy, France).

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, one of the major project objectives has been the compilation of a bidirectional learner speech corpus, 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), representing the first known corpus of L2 speech in both directions of the French-Language pair.

As described by Trouvain et al. (2013) and Fauth et al. (2014), the IFCASL corpus comprises recordings of approximately 100 speakers (50 native speakers of each of the two languages) uttering carefully constructed stimuli in both languages, such that speech in both the L1 and L2 was recorded for each speaker. In recording sessions, subjects were asked to read aloud a given sentence or longer text, and for about half of the L2 stimuli, subjects were allowed to listen to a recording of the text as uttered by a native speaker of the target language before recording their utterance. An even gender distribution was maintained among the speakers

---

[1] http://www.ifcasl.org

recorded, and both children (under 18 years of age) and adults participated in the recordings (the majority being adults); while children were uniformly of beginner proficiency in their L2, adults of both beginner and advanced levels in the L2 were recorded.

While the project as a whole is also concerned with L1 German speakers learning French as L2, and the corpus collected in the IFCASL project is evenly distributed between the two languages, this thesis focuses exclusively on French L1 speakers learning German as L2. 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. As the IFCASL corpus was not annotated for lexical stress errors, one major contribution of this thesis was the annotation of such errors for a subset of the German-language utterances from L1 French speakers; the details of this annotation effort are described in Chapter 3.

Furthermore, the prototype CAPT tool developed in this thesis project has 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, especially the *Jsnoori* speech processing software developed at LORIA (see Section 2.2.2).

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 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.

[TODO Overview of system architecture to?]

In the prototype tool, implemented as a web application in the Groovy language using the Grails web framework,[2] a simple web interface presents a student with a German sentence (taken from the IFCASL corpus), one of the words of which is highlighted as the target word for that exercise. The student is prompted to submit an utterance of that sentence for assessment and feedback, with the instruction to focus on the accurate expression of the lexical stress pattern of the target word. The student's utterance is subsequently analyzed for lexical stress errors using a variety of diagnostic approaches (see **??**), and finally the student is presented with one or more types of feedback on their realization of lexical stress in the analyzed utterance (see **??**).

---

[2] `grails.org`

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.

In addition to this student-facing interface, the tool also implements an interface through which a language teacher or a researcher of L2 language acquisition can create new exercises for students to complete, where each exercise features a specific combination of the various diagnostic methods and feedback types available in the system. By allowing fine-grained control over these features, the tool enables researchers to create CAPT exercises with different features for the purposes of in-vivo studies of the effectiveness of different feedback types, and allows teachers to create exercises matching the needs of their students.

[TODO Add screenshots of both interfaces]

This prototype tool has thus been developed with both instructional and research applications in mind. Unlike with some existing tools for diagnosis 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 summarizes the contributions of this thesis project and outlines some possible directions for future work.

Background and related work | 2

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. 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). [TODO (Hahn, 2004)?]

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; Hahn, 2004; Derwing and Munro, 2005; Hirschfeld and Trouvain, 2007; Dlaska and Krekeler, 2013).

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). The importance of individualized corrective feedback is also generally acknowledged (Neri et al., 2002; Mehlhorn, 2005; Dlaska and Krekeler, 2013), 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

[TODO REORGANIZE THIS SECTION]

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).

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).

2.2.1 Automatic processing of learner speech

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);

2.2.2 The *Snorri* suite and *Jsnoori*

Another important contribution of the LORIA group is the *Snorri/Snoori* suite of software, which includes the now-outdated original *Snorri* developed in 1987 (Fohr and Laprie, 1989) and the Unix adaptation thereof developed shortly thereafter, a later Windows port,

---

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

WinSnorri (Laprie, 1999), and most recently a partial Java port of WinSnorri, *Jsnoori*[3] (Project-Team PAROLE, 2013), which is still under active development, and is the version of the *Snorri* suite most relevant to this work.

[TODO summary of speech processing capabilities of the Snooris (Fohr and Laprie, 1989; Laprie, 1999)]

Though the *Snorri* programs were originally developed primarily as research tools for speech scientists (Fohr and Laprie, 1989; Laprie, 1999), the utility of such software for pronunciation teaching has been explored by the LORIA team (Bonneau et al., 2004; Henry et al., 2007; Bonneau and Colotte, 2011), who have used this software to assess and deliver feedback on lexical stress in L1 French speakers' pronunciation of English words. In its role as a CAPT tool, Jsnoori takes 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 signal and spectrogram of each. Moreover, auditory feedback can be delivered by 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. **[TODO more details about those papers]**

[TODO Jsnoori screenshot]

As described in Chapters 4 and 5, the Jsnoori software is vital to this thesis project; the prototype CAPT tool developed in this work utilizes the signal processing capabilities of Jsnoori for speech analysis and error diagnosis (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).

2.2.3 The *Fluency* pronunciation trainer

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. 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[TM], has been shown to help real-world users significantly improve their pronunciation skills.

---

[3] http://jsnoori.loria.fr

2.2.4 The Project LISTEN Reading Tutor

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 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 more details]**

2.2.5 German and language-independent CAPT

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 **??**). 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 for teaching prosody, 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.

In the broadest terms, lexical stress is the phenomenon of how syllables are accentuated within a word (Cutler, 2005). To say that a given syllable in a word is stressed is, generally speaking, to say that that syllable is somehow accorded a more prominent role in the word than other syllables, i.e. that this syllable is perceived as somehow "standing out" (Dogil and Williams, 1999). The perceived prominence of a syllable in a word is a function not merely of the segmental characteristics of the uttered syllable, i.e. the speech sounds it contains, but rather of its (relative) suprasegmental properties, namely:

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

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 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, pointing to the fact that word-final syllables are always most prominent when a word is pronounced in isolation, 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, expressed as a lengthening of the final syllable in each

prosodic group or phrase (Dupoux et al., 2008; Michaux and Caspers, 2013). Regardless, stress at the word level does not serve any contrastive function in French (Michaux and Caspers, 2013, p. 89), which constitutes a significant difference between this language and a language with variable, contrastive lexical stress such as German or English.

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.

In addition to these difficulties with lexical stress perception, French learners of variable-stress languages such as English, German and Dutch have also been shown to have difficulties in producing stress patterns correctly. Research by Michaux et al. 2012; 2013 revealed that, as might be expected given the tendency for word- and/or phrase-final syllable prominence in French just discussed, French learners of Dutch showed a tendency to stress the final syllables of Dutch words, even when not called for by the canonical stress pattern.

While little research exists specifically investigating the errors of French, it can reasonably be expected that French learners of German will face challenges with both the perception and production of lexical stress, and that the (lack of) lexical stress system in their native language will have an influence on their realization of lexical stress patterns in German words.

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. 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 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 often been found to have a larger impact on intelligibility than segmental errors (Derwing and Munro, 2005; Witt, 2012), 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; Hahn, 2004; 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; Michaux, 2012). 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: 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

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. 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.

Table 3.1: Number of speakers in the IFCASL sub-corpus annotated for lexical stress, in terms of speakers' age, gender, and proficiency level

Age/gender	Proficiency level				Totals
	A2	B1	B2	C1	
Boy	11	0	0	0	**11**
Girl	1	1	0	0	**2**
Man	7	4	3	7	**21**
Woman	5	5	3	9	**22**
Totals	**24**	**10**	**6**	**16**	**56**

- Are lexical stress errors observed frequently in the IFCASL data? (Section 3.5.1)

- Are lexical stress errors observed more frequently with certain word types than with others? (Section 3.5.2)

- 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? (Sections 3.5.3 and 3.5.4)

[TODO preview of how the corpus will be used for supervised training]

3.1 Data

[TODO describe IFCASL corpus] [TODO Reference table of speakers in corpus by age/gender/level - table 3.1]

The IFCASL sub-corpus annotated for lexical stress errors consists of utterances of twelve word types (see table 3.2), 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). **[TODO Didn't control for phrase position - could be hypothesized that phrase position would have an effect given the way prosody works in French, but Michaux and Caspers (2013) found that wasn't the case for French learners of German]**

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 mentioned in Section 1.1, 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 3.2 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 a challenging area of active research (see e.g. Bonneau et al., 2012; Orosanu et al., 2012), 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 3.2: 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 , each of whom is listed in table 3.3 (by an arbitrary identifier, to preserve anonymity). As table 3.3 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 enrolled in an experimental phonology course, 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 3.3, 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 3.3: 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), Tatort (56), E-mail (56)
M	German	intermediate	Ringen (54), 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 3.3 lists the word types assigned to

Table 3.4: 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

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 3.3 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 3.4 lists the number of annotators for each word type.

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.
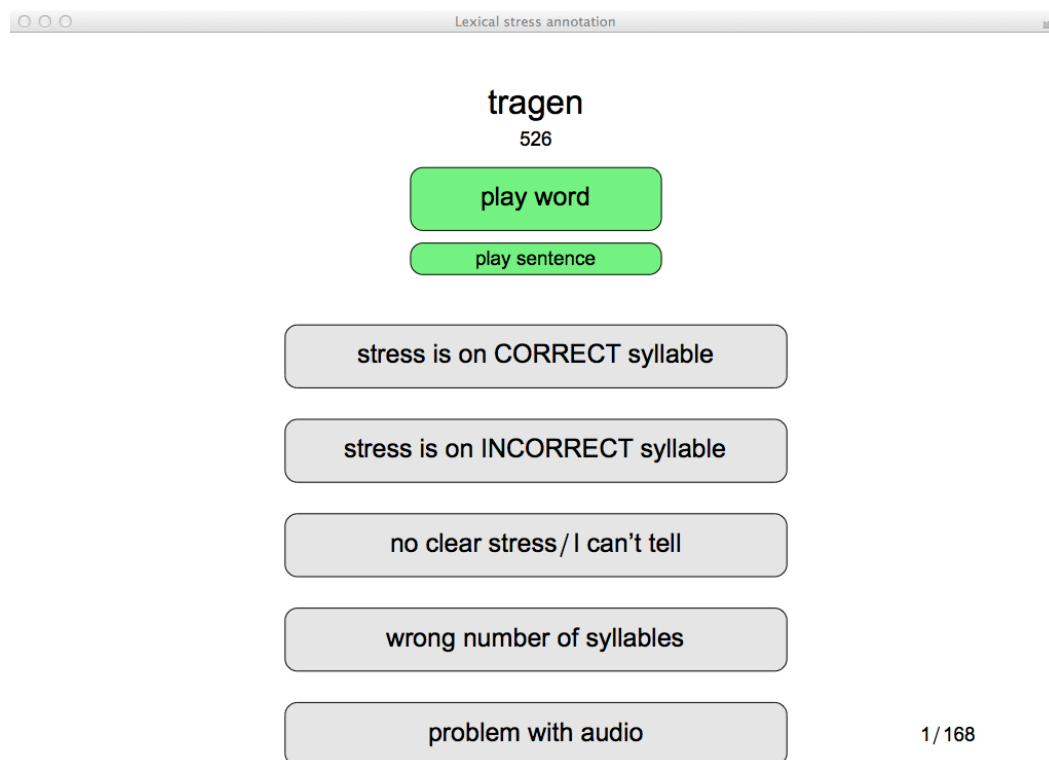
Figure 3.1: A screenshot of the graphical annotation tool scripted in Praat. Green buttons allow the annotator to listen to the word and sentence utterances. Gray buttons allow the annotator to record their judgment of stress accuracy; from top to bottom, the buttons correspond to the labels [correct], [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.

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 3.5. 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. **[TODO Describe these in terms of (Landis and Koch, 1977) - "fair" agreement?] [TODO Compare Kappas to those in (Michaux and Caspers, 2013)? They got substantial agreement but only had 3 expert annotators]**

[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

Table 3.5: Overall pairwise agreement between annotators **[TODO transpose this table so it's consistent with table 3.6 and table 3.7]**

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

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 table of percent/kappa min/avg/max by annotator, since graphs are difficult to read precisely?]**
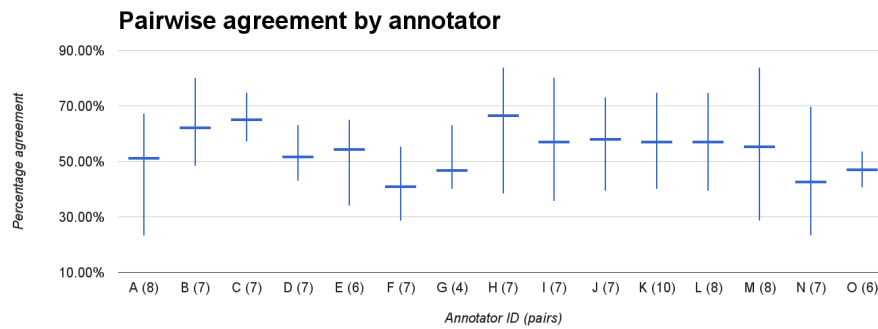
(a) Percent agreement
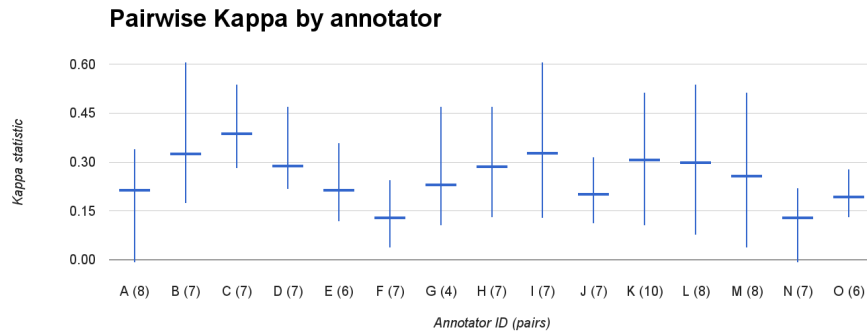
(b) Cohen's Kappa statistic

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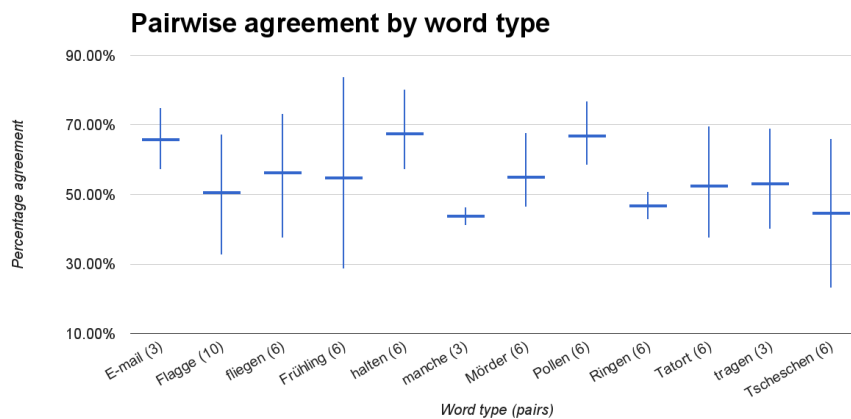
(a) Percent agreement
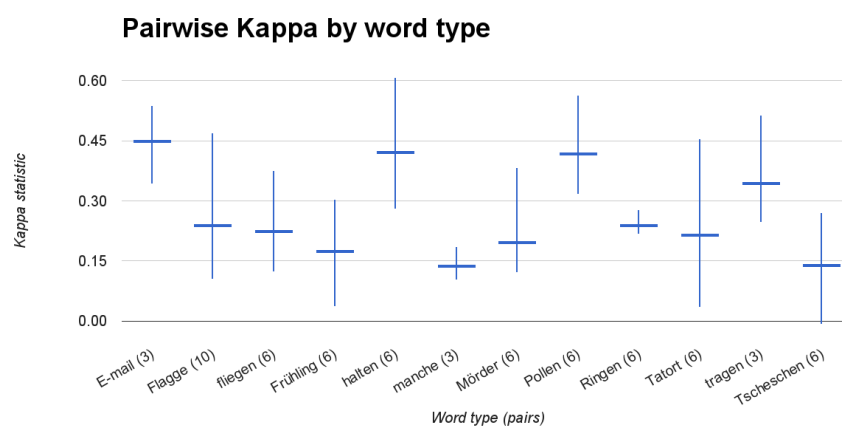
(b) Cohen's Kappa statistic

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.

[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]

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?

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 3.6, 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 3.6: 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
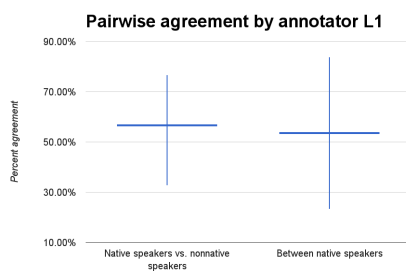
(a) Percent agreement
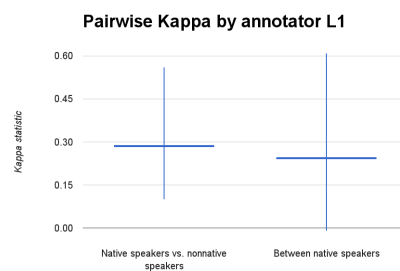
(b) Cohen's Kappa statistic

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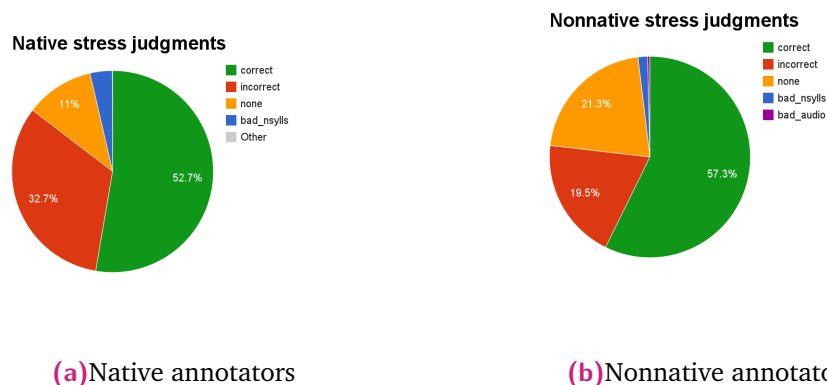
Figure 3.5: Stress judgments made by native and nonnative German speakers

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 3.7 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 3.7: 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

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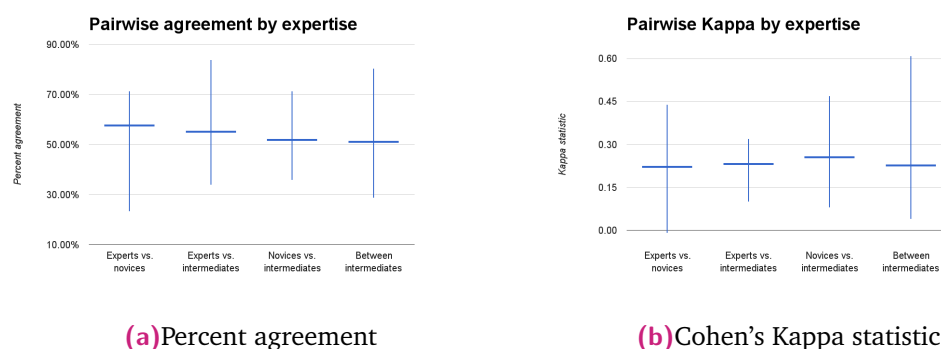
Figure 3.6: Pairwise agreement between annotators based on their level of expertise (expert, intermediate, or novice). The bottom of each vertical bar represents the minimum pairwise value, the top the maximum. Horizontal bars indicate average pairwise values.

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.

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 One could]** speculate that experts' familiarity with nonnative speech and knowledge of possible interspeaker 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 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 Conclusion]

3.4.4 Choosing gold-standard labels

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

(a) Expert

(b) Intermediate

(c) Novice
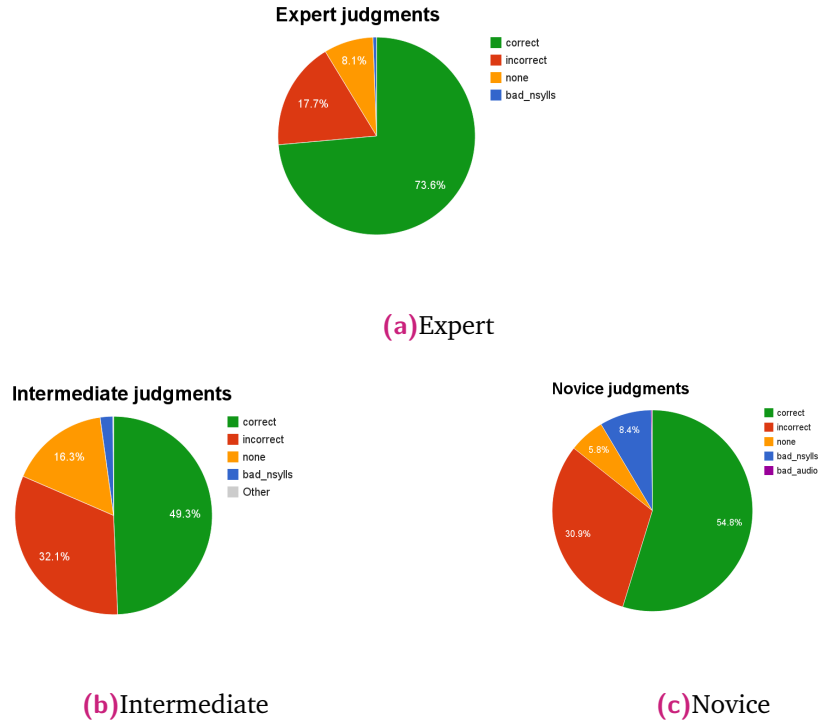
Figure 3.7: Stress judgments by annotator expertise

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-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 $s_{\mathrm{gold}}(t)$, where $s \in S = \{[\text{correct}],[\text{incorrect}],[\text{none}],[\text{bad\_nyslls}],[\text{bad\_audio}]\}$ 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: $S_{\mathbf{max}} \subseteq S_t$]** 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 3.8. 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 $N(C_{1...i})$ 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 3.8: 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 $s_{\text{certain}} \in \{[\text{correct}],[\text{incorrect}]\}$	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 3.8 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 $s_{gold}(t)$, 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 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 and 3.5.4 and **??**)

- 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? (**??**)

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 3.9 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 3.10. 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 3.9: 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%
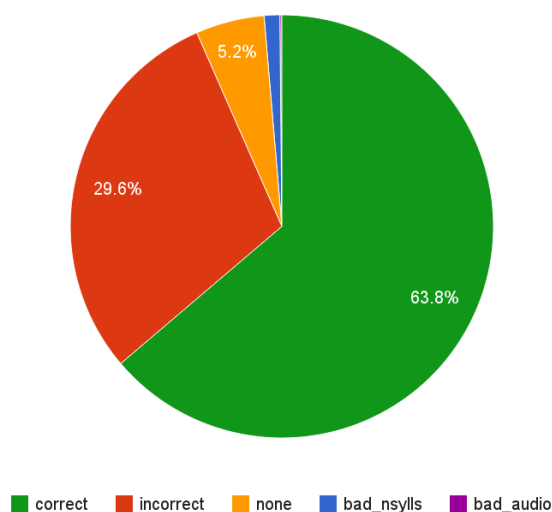
Figure 3.8: Overall distribution of lexical stress errors in the annotated corpus **[TODO retitle]**

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:

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 oː 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. **[TODO ...or it could be word frequency (Pollen less frequent), or maybe influence of phrase position or interpreted phrase position, cf. (Michaux and Caspers, 2013)?]**

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?]

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 3.11 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)

---

[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 3.10: 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%
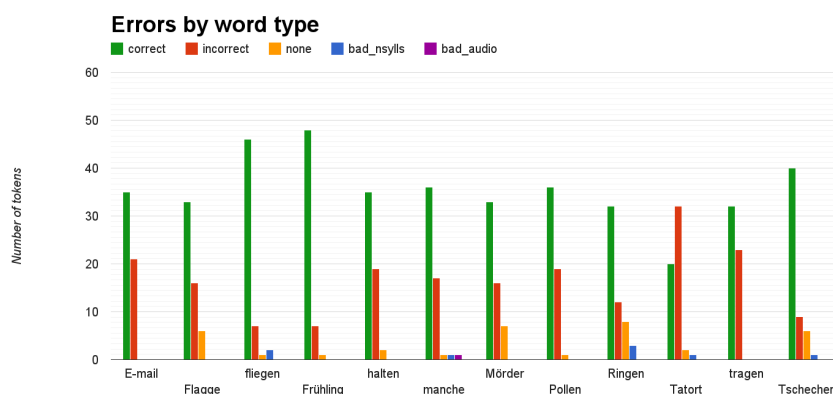
Figure 3.9: Errors by word type

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 3.11 and illustrated in figs. 3.12 and 3.13.

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

Table 3.11: 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%

such errors than learners of higher proficiency. **[TODO This conforms with findings of Michaux (2012) that beginners make more errors than advanced learners.]**

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 3.12 and illustrated in figs. 3.15 and 3.16.

Table 3.12: 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%

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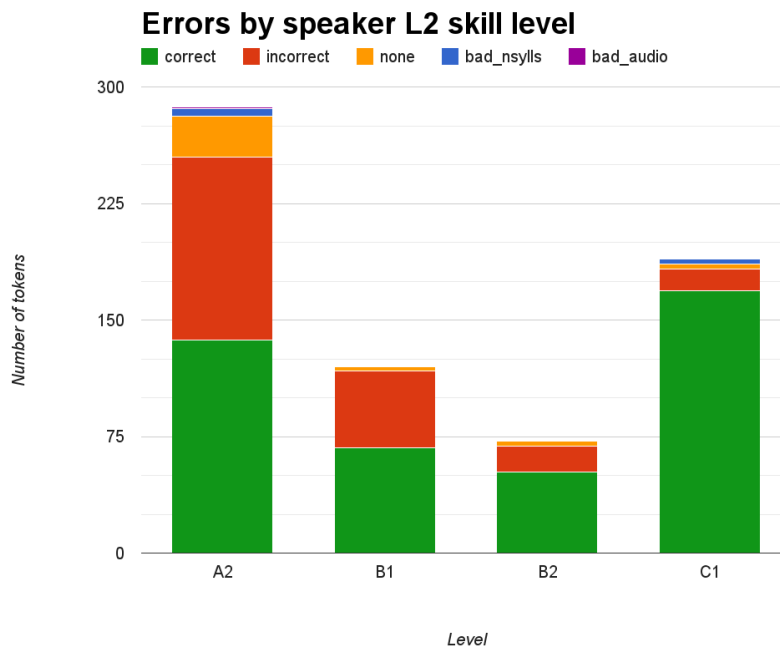
Figure 3.10: Stress judgments by speaker skill level **[TODO Exclude?]**

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 1 girl 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 3.12 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/instead? (distribution is quite similar to A2/B1)]**. Comparing the distribution of 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?]**

(a) A2

(b) B1

(c) B2

(d) C1
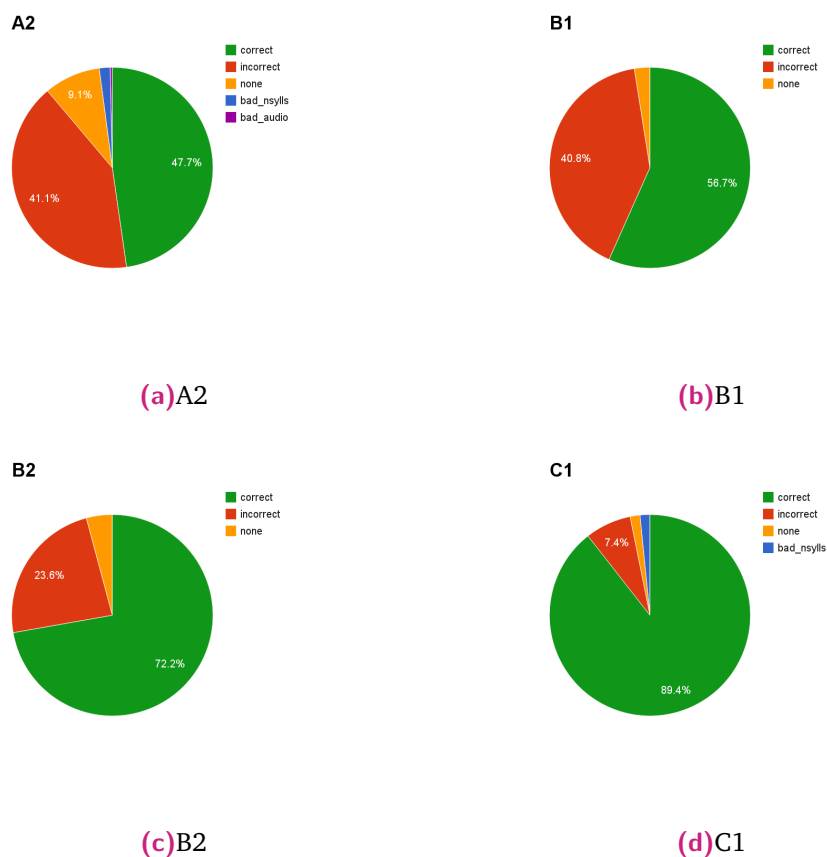
Figure 3.11: Error distribution by speaker skill level

Coming now to the question of whether there is any difference in error distribution between speakers of different genders, a brief glance at fig. 3.16 reveals that there does not seem to be a drastic difference in the distribution of errors between the two genders. Males seem to make slightly more errors in lexical stress realization than females, with 60.7% correct productions for males and 68.1% for women, though this might be explained by the fact that as noted in Section 3.1 above (see table 3.1), the group of male speakers has a higher proportion of elementary (A2) learners (18 out of 32 males, or 56.25%) than the group of female speakers (6 A2 speakers out of 24 females, or 25%). **[TODO do we need an apples-to-apples comparison based on level, as we did in the age discussion?]** Therefore, it would seem that the error distribution observed in the annotated sub-corpus provides no indication of a meaningful difference in the way speakers of different genders realize lexical stress in their L2 German.

3.6 Summary

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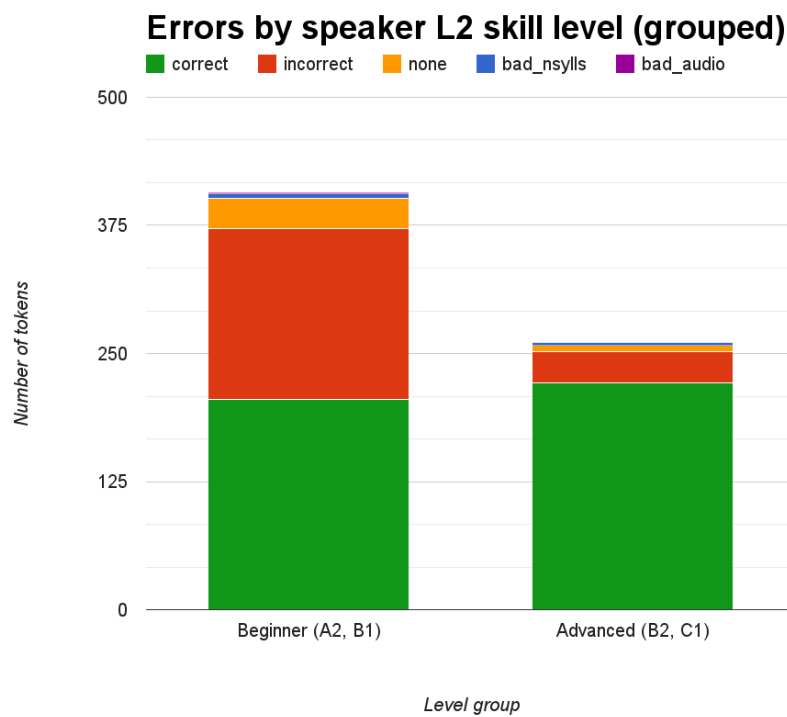
Figure 3.12: Stress judgments by speaker skill level (grouped)

(a)   (b)
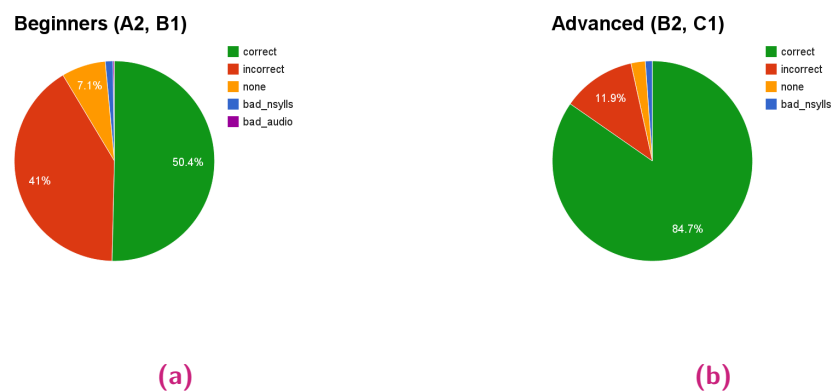
Figure 3.13: Error distribution by speaker skill level (grouped)
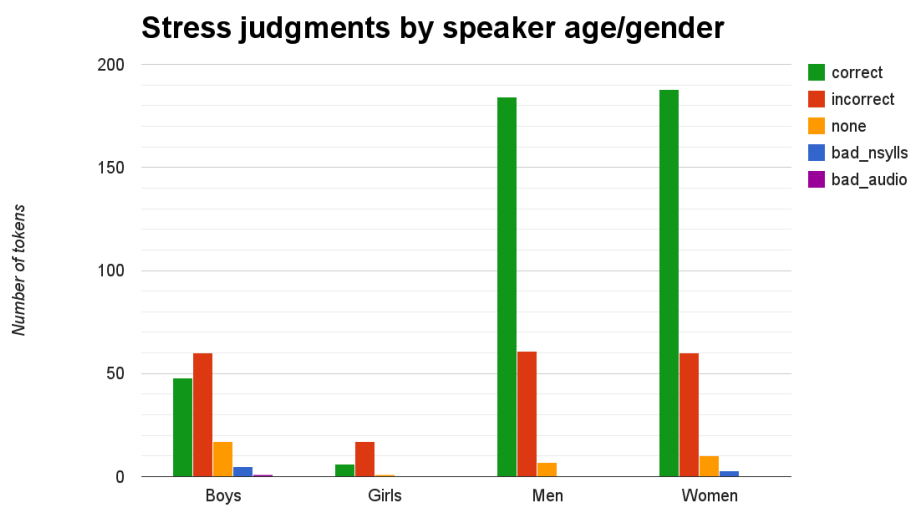
Figure 3.14: Stress judgments by speaker age/gender
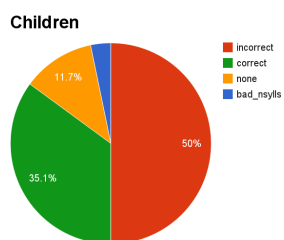
(a)Children
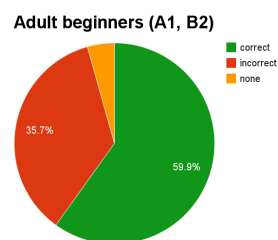
(b)Adult beginners

Figure 3.15: Error distribution by speaker age
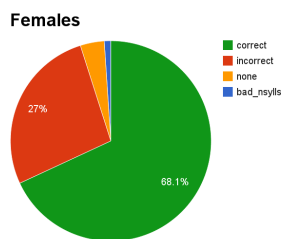
(a)Females
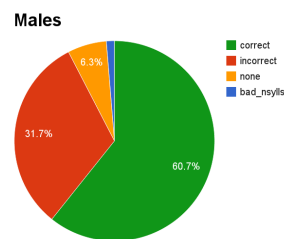
(b)Males

Figure 3.16: Error distribution by speaker gender

Chapter 3 Lexical stress errors by French learners of German **[TODO retitle?]**

Diagnosis of lexical stress errors 4

In order to provide learners with useful feedback on their lexical stress errors in the L2, the prototype CAPT tool developed in this thesis project must first be able to automatically 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) An analysis of 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 comparison of the learner's prosody to this representation.

This chapter describes how (a) is achieved using forced-alignment segmentation of a learner's read-speech utterance with the corresponding text, (Section 4.1); how the lexical stress analysis of (b), which is also crucial to (c), is produced by measuring the fundamental frequency, duration, and energy of relevant sections of the speech signal (Section 4.2); and the various approaches to (c) and (d) that are implemented in the prototype tool (Sections 4.3 and 4.4). Finally, it describes how the system's modular architecture allows researchers and teachers control over which of these approaches are used **[TODO (Section 4.5)]**.

4.1 Automatic segmentation of nonnative speech

[TODO Should this become a subsection of Section 4.2?]

Segmentation, or labeling, of a recorded utterance is the task of annotating the speech signal with boundaries that demarcate individual phones, syllables, words, sentences, and/or other units of speech; see fig. 4.1 for an example of a multi-level segmentation of a German utterance.

A reasonably accurate segmentation of an L2 learner's utterance is indispensable for an analysis of the accuracy of their pronunciation; **[TODO mention that there are methods which don't need segmentation, but they're more limited?]** as it allows comparison between relevant units of the learner's utterance – e.g. words, syllables, and phones – and corresponding units in native speech. The most accurate segmentation would of course be one produced by hand by a trained phonetician. However, hand-labeling is not feasible in most scenarios because of its high cost in terms of time and wages; moreover, because the ultimate goal of this work is the development of a CAPT tool which can give L2 learners helpful automatic feedback on their pronunciation, any analysis of the learner's speech

CorpusSample_German_L1German

#	d	d_	u:	h	a	s	t	t_	aI	n	@	S	2:	n	@	h	Y	t	t_	@	#
#	du		hast			eine		schöne				hütte									
du hast aI n@ S 2: n@ h Y t@																					
Du hast eine schöne Hütte.																					

Figure 4.1: **[TODO update with new example w/ SyllableTier]** 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.

signal, including the preliminary step of segmentation, must proceed fully automatically. Therefore, a means of automatically segmenting a given utterance is required.

When the content (text) of a given utterance is already known, the goal of automatic segmentation becomes aligning the boundaries of each phone in the expected sentence with the appropriate points in the recorded signal. Boundaries for larger units such as syllables and words can be inferred from the phone boundaries. An effective and widely-used technique for this is forced alignment (Fohr et al., 1996; Mesbahi et al., 2011; Fohr and Mella, 2012; Fauth et al., 2014). This technique, **[TODO *remove?* a type of speech recognition,]** requires:

- the expected text (word sequence) of the given utterance,
- a pronunciation lexicon containing the sequence of phones expected for each word, and
- an acoustic model for the target language.

The first of these requirements, the text of the utterance, is trivial when the speaker has been asked to read a given sentence aloud, which is the case in **[TODO this context *is that unclear?*]**. A lexicon of canonical word pronunciations, i.e. the pronunciations that might be given in a standard dictionary, is also relatively easy to obtain for a well-researched language such as German, for which many digital linguistic resources exist. To account for differences in how different speakers (especially non-native speakers) may pronounce

the given sentence, the lexicon should contain not only the canonical pronunciation for each word, but also any alternate or non-standard pronunciation variants (native or non-native) that might be encountered. Research **[TODO at LORIA]** has found the inclusion of non-native pronunciation variants to lead to improvements in the accuracy of automatic segmentation of non-native speech (Jouvet et al., 2011; Mesbahi et al., 2011; Bonneau et al., 2012; Orosanu et al., 2012), and one of the intended outcomes of the IFCASL project is the extraction of a non-native variant lexicon from the L2 speech in the corpus **[TODO ref?]**.

The final requirement for segmentation via forced alignment is an acoustic model, i.e. a statistical model which captures the correspondence between acoustic features extracted from the speech signal and phones in the target language. To accurately capture this correspondence, the model must be trained on a large amount of speech data in the target language; the acoustic model used to align the German IFCASL data was trained on native German speech from the Kiel corpus **[TODO ref, verify]**. However, research by Bouselmi et al. (2005; 2012) has shown that even more accurate segmentation of learners' utterances an be obtained by using acoustic models adapted to non-native speech in the target language and/or speech in the learner's L1; refining the automatic segmentation functionality using such adapted models would therefore be a logical extension of this work (see Section 6.2) **[TODO does that clause work?]**.

[TODO Details about how forced Viterbi alignment works?]

Given these resources, the Jsnoori software, **[TODO *awkward:*]** which the lexical stress CAPT tool developed in this work uses for speech processing, is capable of automatically segmenting a learner's utterance almost instantly. Unfortunately, a disadvantage of forced alignment is that it requires the entire utterance (e.g. sentence), so real-time segmentation is not possible **[TODO true? should this go in a footnote?]**. However, acoustic models and pronunciation lexicons for German have yet to be integrated into Jsnoori, which currently only has the resources to segment speech in English and French. Therefore, the prototype CAPT tool developed in this thesis project presupposes the existence of a segmentation for a given utterance, taking a "Wizard-of-Oz" approach to the automatic segmentation step by demonstrating its error diagnosis and feedback capabilities using learner (L2) and reference (L1) read-speech utterances from the German-language subset of the IFCASL corpus (Fauth et al., 2014; Trouvain et al., 2013), all of which have been segmented at the phone and word levels using the forced alignment technique described above. Once the requisite German-language resources are available in Jsnoori, the tool can easily be extended to perform on-the-fly segmentation of learner utterances (see Section 6.2).

[TODO *awkward clause:*] Although the IFCASL corpus also contains manually-corrected versions of the majority of the forced-alignment segmentations, the CAPT tool only makes use of the automatically-determined determined boundaries, even though these are potentially less accurate; this is to more accurately simulate the conditions of a fully-automatic CAPT system, which would need to perform segmentation on the fly without recourse to manual verification. Indeed, forced alignment is not a perfect method; **[TODO *unclear?:* 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. Therefore, inaccuracies in the phone boundaries determined using this technique should be expected, especially

when the alignment is performed on non-native utterances using an acoustic model trained on native speech. **[TODO Reliability of non-native speech automatic segmentation for prosodic feedback. (Mesbahi et al., 2011)]** **[TODO *work this sentence in somewhere:*]** A fully-fledged CAPT system extending this thesis project would have to cope with any problems that may result from using imperfect automatic segmentations as a starting point for analysis.

As mentioned above, the utterances in the IFCASL corpus have segmentations at the phone and word levels; however, the corpus does not contain syllable-level segmentations. As the syllable is arguably the most important unit for analysis of lexical stress realization **[TODO reference/justification]**, syllable-level segmentations had to be created for each utterance. This was accomplished by manually determining the locations of syllable boundaries in the phone sequence for each word, automatically extracting the temporal locations of these boundaries from the phone-level segmentation, and automatically combining the word-internal syllable boundaries with the boundaries in the word-level segmentation to create the syllable-level segmentation.

4.2 Analysis of word prosody

[TODO Technically these features aren't all relevant to comparison-based diagnosis, only classification - so I guess this whole section needs to be rewritten :/]

The automatically-determined word, syllable, and phone boundaries obtained as described in the previous section enable the CAPT tool to locate and analyze segments of the speech signal relevant to the realization of lexical stress. This section describes 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 **[TODO acoustic properties (and by extension their perceptual correlates)]** described in Section 2.3, namely duration (timing), fundamental frequency (pitch), and intensity (loudness). The relative utility of these features in automatically diagnosing lexical stress errors is discussed further in Section 4.4.

Throughout this section, the features discussed are illustrated with their values for a word from **[TODO *three?* 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.

Once again it should be stressed that the features described below are computed from the automatically generated segmentation of a given utterance, 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, as just discussed in Section 4.1. **[TODO *Worth mentioning? Seems a bit out of place*]** Also worth noting is the fact that we are here dealing exclusively with read, and not sponanous, 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
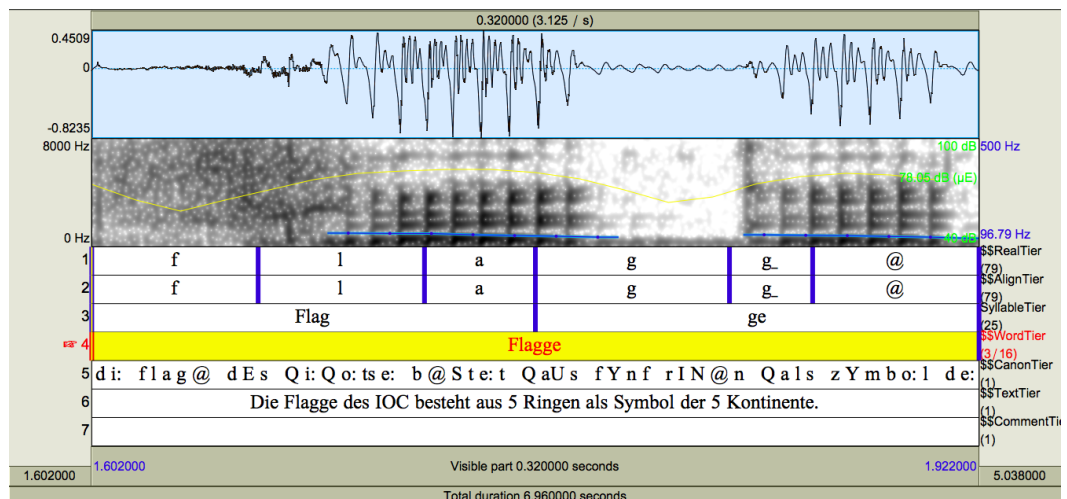
(a) L1 French speaker (F)

[TODO Add incorrect FG example?]
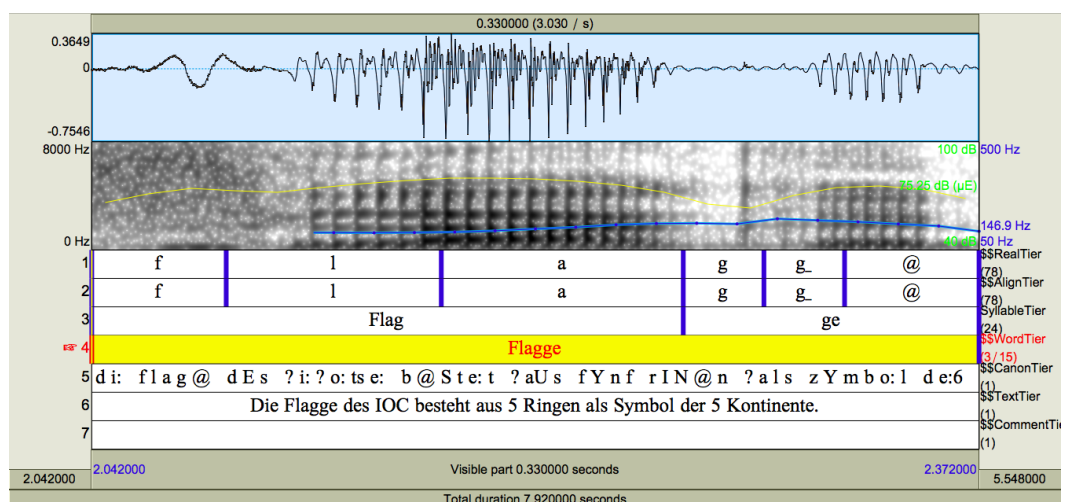
(b) L1 German speaker (G)

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]**

4.2   Analysis of word prosody   43

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, some research indicates that syllable duration may be the most important, if not the only acoustic correlate of lexical stress in German (e.g. Dogil and Williams, 1999). Duration analysis therefore figures prominently in the analysis and assessment of learners' lexical stress in this work.

Given the word-, syllable- and phone-level segmentations of an utterance (see Section 4.1), the extraction of duration features for that utterance is trivial, as it consists simply of noting the duration of each relevant segment. Following Bonneau and Colotte (2011), **[TODO we]** take into account the duration of each syllable in the word to be analyzed, and of the vowels at the nucleus of each syllable. **[TODO Mention that to find vowels I had to add German phonetic inventory to Jsnoori?]** To account for inter-speaker variability, e.g. the fact that some speakers may have an overall slower or faster speech rate than others, relative rather than absolute durations are used. The list of duration features computed for each word utterance is given in table 4.2, along with the values computed for each feature from the sample utterances shown in fig. 4.2.

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. **[TODO remove absolutes?]**

(a) Absolute features

Feature name	Description	Value (seconds) (a) F	(b) G
WORD-DUR	Duration of entire word **[TODO remove?]**	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

(b) Relative features

Feature name	Description	Value (a) F	(b) G
REL-SYLL-DUR	SYLL1-DUR/SYLL0-DUR	1.00	2.00
REL-V-DUR	V1-DUR/V0-DUR	0.67	1.80

4.2.2 Fundamental frequency

[TODO consistency in terminology: F0 vs. pitch?]

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 indication of how lexical stress is realized in that utterance, and F0 features are another crucial component of the system's prosodic analysis.

The F0 contour of a given utterance is determined using the pitch detection functionality of Jsnoori. At the heart of Jsnoori's approach to pitch detection lies the algorithm developed by Martin (1982), a frequency-domain method of pitch detection which uses a comb function with teeth of decreasing magnitude to identify harmonics of F0 in the spectrum (spectra are extracted by Fast Fourier Transform using 32-millisecond Hamming windows set 8 milliseconds apart **[TODO double-check that]**). Jsnoori also implements several improvements to pitch detection beyond Martin's algorithm (Di Martino and Laprie, 1999), including **[TODO Yves' parabolic interpolation - citation?]**, the voicing decision optimization of Secrest and Doddington (1983), and the dynamic programming technique proposed by Ney (1981) for identifying and removing incorrect points in the contour. Further improvements to Jsnoori's pitch detection capabilities, including the implementation of additional detection algorithms, are currently in development (see Section 6.2).

Thanks to these techniques, Jsnoori is capable of efficient, generally accurate detection of F0 **[TODO *keep?* within the range of 65-800 Hertz]**. Each pitch point in the contour is subsequently converted from Hertz to semitones. Using this contour, each of the relevant segments of the utterance – i.e. the word of interest, each of its syllables, and their nuclei – is analyzed in terms of its F0 mean, maximum, minimum, and range; the full list of features computed is presented in **??**. Features only take into account the non-zero points in the contour, i.e. points corresponding to voiced sections of the utterance. The F0 mean is calculated as the average of all non-zero points within the start and end boundaries of the given segment; the maximum F0 is the highest value at any of these points, and the minimum the lowest non-zero value; and F0 range is computed as the difference between the maximum and minimum values.

[TODO Jsnoori only looks at mean and max? or something? by way of transition to this next paragraph]

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 than surrounding syllables) or negative (lower pitch) (Cutler, 2005, p. 267). Therefore, the computed features capture not only the F0 maximum of each syllable, but also the minimum and range (difference between maximum and minimum) **[TODO more/separate discussion of why range might be important?]**.

4.2.3 Intensity

[TODO intensity and energy are used interchangeably - fix or explain]

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

(a)Absolute features

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
V0-F0-MEAN	Avg. F0, 1st syllable nucleus	TD	TD
SYLL0-F0-MAX	Max. F0, 1st syllable	10.73	18.25
V0-F0-MAX	Max. F0, 1st syllable nucleus	TD	TD
SYLL0-F0-MIN	Min. F0, 1st syllable	TD	TD
V0-F0-MIN	Min. F0, 1st syllable nucleus	TD	TD
SYLL0-F0-RANGE	SYLL0-F0-MAX−SYLL0-F0-MIN	1.45	4.60
V0-F0-RANGE	V0-F0-MAX−V0-F0-MIN	TD	TD
SYLL1-F0-MEAN	Avg. F0, 2nd syllable	8.24	17.51
V1-F0-MEAN	Avg. F0, 2nd syllable nucleus	TD	TD
SYLL1-F0-MAX	Max. F0, 2nd syllable	9.93	20.08
V1-F0-MAX	Max. F0, 2nd syllable nucleus	TD	TD
SYLL1-F0-MIN	Min. F0, 2nd syllable	TD	TD
V1-F0-MIN	Min. F0, 2nd syllable nucleus	TD	TD
SYLL1-F0-RANGE	SYLL1-F0-MAX−SYLL1-F0-MIN	3.66	5.86
V1-F0-RANGE	V1-F0-MAX−V1-F0-MIN	TD	TD

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 was the case with the prosodic diagnosis functionality of Jsnoori **[TODO at the start of this thesis project *reword that*]**. 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, in addition to duration and fundamental frequency, the intensity of relevant portions of an utterance are taken into account when performing prosodic analysis.

The intensity contour of a given segment (word, syllable, or syllable nucleus) in the utterance is computed in Jsnoori, which calculates the total amount of energy at frequencies from 0 to 8000 Hertz in spectra extracted from the signal by Fast Fourier Transform, using Hamming windows 20 milliseconds long spaced 4 milliseconds apart **[TODO explain why I didn't use 32ms windows?]**. Energies below a "silence threshold" of 60 decibels are not counted

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

(b) Relative features

Feature name	Description	Value (a) F	(b) G
REL-SYLL-F0-MEAN	SYLL0-F0-MEAN/SYLL1-F0-MEAN	1.13	0.90
REL-V-F0-MEAN	V1-F0-MEAN/V0-F0-MEAN	TD	TD
REL-SYLL-F0-MAX	SYLL1-F0-MAX/SYLL0-F0-MAX	1.08	0.91
REL-V-F0-MAX	V1-F0-MAX/V0-F0-MAX	TD	TD
REL-SYLL-F0-MIN	SYLL1-F0-MIN/SYLL0-F0-MIN	TD	TD
REL-V-F0-MIN	V1-F0-MIN/V0-F0-MIN	TD	TD
REL-SYLL-F0-RANGE	SYLL1-F0-RANGE/SYLL0-F0-RANGE	0.40	0.78
REL-V-F0-RANGE	V1-F0-RANGE/V0-F0-RANGE	TD	TD
F0-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
F0-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
F0-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

toward the total, as these are assumed to correspond to ambient or non-speech noise. This intensity contour is then used to calculate the mean and maximum energy in the relevant segments of the speech signal; the list of features extracted is given in table 4.3.

Using the prosodic features thus computed, the prototype CAPT tool analyzes a given learner utterance and diagnoses their lexical stress error(s) (or lack thereof) by comparing this non-native speech to that of L1 German speakers using one of several possible methods. The following section describes the various diagnostic methods explored in this thesis project, and how they make use of (subsets of) the duration, F0, and intensity features described above.

4.3 Diagnosis by direct comparison

[TODO intro]

A typical approach to assessing L2 prosody involves comparing a learner's utterance to the utterance(s) of the same word or sentence as produced by one or more native speaker of the target language; this is the approach commonly taken in CAPT systems and research

Table 4.3: Features computed for intensity analysis, and their values for the sample utterances of "Flagge" in fig. 4.2. Values are given in dB over 60dB

(a) Absolute features

Feature name	Description	Value (dB>60)	
		(a) F	(b) G
WORD-ENERGY-MEAN	TD	TD	TD
WORD-ENERGY-MAX	TD	TD	TD
SYLL0-ENERGY-MEAN	TD	TD	TD
SYLL0-ENERGY-MAX	TD	TD	TD
SYLL1-ENERGY-MEAN	TD	TD	TD
SYLL1-ENERGY-MAX	TD	TD	TD
V0-ENERGY-MEAN	TD	TD	TD
V0-ENERGY-MAX	TD	TD	TD
V1-ENERGY-MEAN	TD	TD	TD
V1-ENERGY-MAX	TD	TD	TD

(b) Relative features

Feature name	Description	Value	
		(a) F	(b) G
REL-SYLL-ENERGY-MEAN	TD	TD	TD
REL-SYLL-ENERGY-MAX	TD	TD	TD
REL-VOWEL-ENERGY-MEAN	TD	TD	TD
REL-VOWEL-ENERGY-MAX	TD	TD	TD
ENERGY-MAX-INDEX	TD	TD	TD

(e.g. **[TODO refs]**), In this comparison-based approach, the L1 utterance serves as a direct representation of how the word/sentence would be realized by a native speaker; errors are diagnosed when there are stark enough differences between the L2 learner's utterance and that of the native speaker with respect to the relevant features **[TODO *is that sentence awkward?*]**.

This section describes the various *approaches* to such diagnosis by comparison taken in the prototype CAPT tool developed in this thesis project. **[TODO is more intro/summary of upcoming sections needed here?]**

4.3.1 Using a single reference speaker

The simplest type of diagnosis by direct comparison involves comparing a single learner utterance to a single reference (native-speaker) utterance; as mentioned in **??**, this is the approach used to evaluate learner speech in Jsnoori and its predecessor WinSnoori (Bonneau et al., 2004; Henry et al., 2007; Bonneau and Colotte, 2011). In the prototype CAPT tool, this method of comparison is implemented as a type of baseline, using the pre-existing capabilities of Jsnoori for processing and comparing the learner and reference utterances, which are described in the following section.

In Jsnoori, comparison between a student's utterance and that of the reference speaker is effected through analysis of the duration (referred to as "timing" in Jsnoori), F0 ("pitch"), and intensity ("energy") of relevant segments of each utterance. **[TODO At the beginning of this thesis project, intensity was not taken into account for the analysis of learner utterances; part of this project therefore involved adding intensity analysis to Jsnoori's learner-assessment module(s).]** For each of these three feature types, a score between 0 and 1 is assigned to the learner utterance based on its correspondence with the reference, and an overall score is computed as the evenly-weighted average of the duration and F0 scores (intensity does not count towards the overall score).

[TODO *unnecessary?* As the diagnostic capabilities of Jsnoori were designed with isolated-word utterances in mind, and the data evaluated here consisted of utterances of entire sentences, the utterances could not be passed directly to Jsnoori for analysis; instead, the word under analysis in the utterance was first isolated in the segmentation by re-labeling all segments before and after that word as silence.]

The following paragraphs describe how each of these three scores is calculated, with reference to the features listed in Section 4.2. It must be noted here that the range of possible values for each of the three scores is not continuous, but discrete, and that these values fall on an ordinal rather than an interval scale. Furthermore, the same values for different scores do not necessarily correspond; for example, it is possible to achieve a timing score of 0.3 or 0.5, but possible values for pitch scores jump from 0.1 directly to 0.8. Therfore, Jsnoori's representation of scores with floating-point numbers between 0 and 1 is perhaps misleading, and the use of categorical labels might be more appropriate.

Timing (duration) score If the number of syllables in the segmentation of the learner's word utterance matches that of the reference, Jsnoori assigns an arbitrary low timing score of 0.1; similarly, if there is a match in the number of syllables but a mismatch in the number of phones within one or more of those syllables, the assigned score is 0.3. If both the number of syllables and the number of phones in each syllable match, a true analysis of the learner's timing is undertaken as follows.

First, the location of stress placement in the word utterance is determined by comparing the lengths of each syllable's nucleus, usually a vowel, i.e. by comparing the features V0-DUR and V1-DUR. The syllable with the longest vowel is taken as the stressed syllable. If the stressed syllable differs between the learner and reference utterances, the learner is assigned a timing score of 0.5. If the stressed syllable is the same, Jsnoori then computes the difference

between the length of the stressed vowel in the learner's utterance (normalized by dividing the vowel's duration by the sum of the durations of all vowels in the word) and that of the stressed vowel in the reference utterance; if that difference falls below a certain threshold **[TODO (-0.2)]**, the stressed syllable is deemed not to have been stressed clearly enough, resulting in a score of 0.8. If the difference in relative vowel lengths exceeds the threshold, the learner is assigned a perfect score of 1.0.

Pitch (F0) score   To assign a pitch score, Jsnoori identifies the stressed syllable in an utterance by comparing the F0 maxima in each syllable (i.e. SYLL0-F0-MAX and SYLL1-F0-MAX), assuming that the syllable with the highest F0 peak is the syllable that has been stressed. **[TODO this is not actually true for German - Jsnoori was developed for English]** Having identified the syllables stressed by the learner and reference speaker, the two are compared; if the syllable is the same in both utterances, stress is judged to have been placed on the correct syllable; otherwise, the learner is assessed as having placed stress on the wrong syllable and receives a score of 0.1. If the learner has stressed the correct syllable, Jsnoori assesses whether they have realized that stress clearly enough by comparing the difference between the mean pitch of the stressed and unstressed syllables in the learner's utterance (e.g. SYLL0-F0-MEAN − SYLL1-F0-MEAN) with the analogous difference in the reference utterance; if the difference between these differences is greater than a threshold **[TODO (3)]**, the learner is considered not to have expressed stress strongly enough, receiving a score of 0.8. If the difference exceeds the threshold, stress realization in terms of F0 is considered acceptable, and the learner's score is a perfect 1.0.

Energy (intensity) score   Jsnoori's method for assessing lexical stress realization in terms of energy is analogous to that for F0, with the exception that the location of the stressed placement in the utterance is determined with reference to the maximum, not mean, energy observed in each syllable (i.e. SYLL0-ENERGY-MAX vs. SYLL1-ENERGY-MAX), such that the syllable with the higher energy peak is assumed to contain the stress. **[TODO again, this is not actually true for German]** As with pitch, if the learner has stressed the wrong syllable they are assigned a score of 0.1, whereas if they have stressed the correct syllable, their score depends on the difference in maximum intensity between the stressed and unstressed syllables, comparing their utterance to the reference: if the difference in differences exceeds a threshold, they are assessed as having not realized stress clearly enough, and receive a score of 0.8. Otherwise, they receive a perfect intensity score of 1.0.

The scores calculated thus for each of the three feature types are used in the prototype CAPT tool to generate various types of feedback for the learner, including the explicit feedback of reporting their scores directly; see Chapter 5 for a detailed discussion of the use of these diagnoses in feedback delivery.

4.3.2   Using multiple reference speakers

When using a single native-speaker utterance for reference, even if the reference speaker has been chosen carefully (see Section 4.3.3 below), analysis of the learner's pronunciation 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 is therefore 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, **[TODO *remove?* the hope being that the variability between these reference utterances will capture more general traits of native pronunciation]**.

In the lexical stress CAPT tool, this is 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. This is accomplished simply by averaging each of the duration, pitch, and energy scores assigned by Jsnoori in each one-on-one comparison; for example, the final pitch score for a learner's utterance when compared to three different reference utterances will be the average of those three one-on-one pitch scores as computed by Jsnoori. Other means for combining several single-reference scores into one multiple-reference score are certainly conceivable, and this could be an interesting direction for future work (see Section 6.2).

4.3.3 Reference speaker selection

Inspired and informed by the investigations of Probst et al. (2002), this work also examines different ways of selecting the reference speaker against which a learner's utterance will be judged, given a pool of potential references.

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 therefore enables the choice of a reference from a set of available speakers, **[TODO with that set optionally being constrained by one or more properties of the speaker (e.g. gender, age)]**. When designing an exercise, the researcher can either manually select a reference utterance for all students who will complete that exercise **[TODO FIXED]**, or enable each student to manually select their own reference **[TODO MANUAL]**

Automatically selecting a reference

A different and perhaps more interesting means of selecting a reference speaker is to automatically choose a speaker whose voice resembles that of the learner; as described in Section 2.2.3, research by Probst et al. (2002) seems to indicate that using an carefully-selected reference speaker can help learners improve their pronunciation. This requires some representation of the relevant features of each speaker's voice; the prototype CAPT tool follows Probst et al. (2002) in using F0 mean and range for that representation, where ach speaker's overall F0 mean and range are computed as the average of each of these features across all available whole-sentence utterances by that speaker. To determine the best reference speaker for a given learner, the respective absolute differences between that learner's overall F0 mean and range and those of each of the available native speakers are calculated and added together, and the native speaker with the lowest total difference

from the learner is selected as the reference. While this accomplishes the goal of providing automatic reference selection as an alternative to the more common manual selection of the reference speaker, it remains a rather simplistic way of representing each speaker's voice, and an exploration of how speakers can be compared using other representations would be a worthwhile future endeavor (see Section 6.2).

4.4 Diagnosis by classification

[TODO Finally, a different approach may be to abstract away from the reference speaker(s).] [TODO comparison approach has disadvantages (still not general enough, limits tutoring exercises to sentences for which we have reference utterances, ...)] By 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, [TODO we] may be able to overcome these shortcomings of the comparison approach, as such a model could theoretically abstract away from any remaining speaker- or utterance-dependent influence from the reference utterance(s) and enable the creation of exercises with arbitrary text, including sentences for which no reference utterance has been recorded.

[TODO This diagnostic approach, using generalized lexical stress modeling, is the one which has been least explored in CAPT research, ... although these people have done it: ....] [TODO *move the following to Chapter 2 Background?*] Shahin et al. (2012) and Kim and Beutnagel (2011) use machine learning to categorize English words based on their stress patterns. 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.

Given the relative novelty of this type of diagnosis for prosodic errors, diagnosis by classification was of particular interest in this thesis project. A series of classification experiments was conducted in an effort to determine:

- how well lexical errors can be identified by a classification-based approach, in comparison to the accuracy of human listeners in identifying such errors, [TODO combine with 2nd point?]

- which of the features discussed in Section 4.2 are most useful for diagnosis by classification, and

- whether a classification-based approach can lead to reasonably accurate diagnosis for words or speakers not seen in the training data [TODO do I need to say more about that here?].

The sections that follow describe these experiments and their findings. [TODO *rephrase this paragraph*]. The objective of these experiments

4.4.1 Data and method

In addition to the motivation of analyzing the frequency and distribution of lexical stress errors in L2 German speech by L1 French speakers, another motivation behind the annotation of these errors in a subset of the IFCASL corpus (described in Chapter 3) was the creation of labeled data for a supervised machine learning approach to diagnosing learner errors. In addition to the non-native utterances and their gold-standard labels from the annotated sub-corpus (see Section 3.4.4), the corresponding utterances of the selected word types from the L1-German portion of the IFCASL corpus were also included in training data, with each native utterance labeled as [correct] based on the assumption that native speakers always realize lexical stress correctly.

[TODO *awkward*] Using a classifier trained on (a subset of) this data, it is possible to predict a label (e.g. [correct] or [incorrect]) for a given learner utterance based on the values of (a subset of) the features described in Section 4.2, and then compare the predicted label to the label assigned to that instance in the gold-standard data to evaluate the accuracy of the prediction. This was accomplished by training and evaluating classifiers in various configurations using the WEKA machine learning toolkit (Hall et al., 2009). In the experiments reported below, the classifiers used are simple Classification And Regression Trees (CARTs) **??**. **[TODO explanation of how CARTs work and why I chose decision trees?]**. However, as WEKA implements a wide variety of other classifiers, some of which can be much more powerful than the algorithm used here, it would be interesting to compare different classification algorithms to see if other classifiers are more effective for this type of data (see Section 6.2).

For each relevant configuration (see below), a CART is trained to classify utterances as belonging to one of the five categories described in Section 3.3. However, in practice these trees classify every utterance as either [correct] or [incorrect], neglecting [none] and the other labels due to their comparatively low frequency in the data **[TODO need to say more?]**. Overall classification accuracy on the annotated sub-corpus was assessed by using held-out portions of the annotated data as test sets, and performing cross-evaluation on multiple train/test splits of the data. The features and data splits used in each experiment are described in the sections below. **[TODO *reword that?*]**

Overall accuracy of each classifier's performance on its test set was quantified in terms of the following measures:

- Percent accuracy (% acc.): The number of samples given the correct label, divided by the total number of samples in the test set
- Kappa (κ): Agreement between the labels assigned by the classifier and the true labels (see Section 3.4)

For the two most frequently observed classes in the data, [correct] and [incorrect], the following evaluation metrics were also computed:

- Precision (P): the number of instances the classifier correctly assigned to this class (i.e. the number labeled as this class that were truly of this class), divided by the total number of instances it assigned to this class

- Recall (R): the number of instances the classifier correctly assigned to this class, divided by the number of instances which truly belong to this class

- F-measure (F), also known as F-score or F_1 measure: a metric which combines Precision and Recall by taking their harmonic mean, given by the formula $F = 2PR/(P + R)$

- **[TODO F_2 measure (weights R twice as much as P)** $F_2 = (1 + 2^2) \cdot PR/(2^2 \cdot P + R) = 5PR/(4P + R)$**]**

Given the intended application of error detection in a student-facing CAPT system, the recall for the [correct] class should be accorded particular importance, since it informs us of the proportion of truly correct utterances that the system marks as having some type of error. This type of misclassification is more dangerous for a CAPT system than misclassifying incorrect pronunciations as correct, because telling a student that they have made a mistake when in fact they have not can be more damaging to their motivation and willingness to continue learning with the system than telling them that they have stressed a word correctly when in fact they have made a mistake **[TODO citation?]**. Therefore, [correct] recall should be high (close to 1.0) for a classifier that will be used for error diagnosis. However, recall of 1.0 can be trivially achieved by simply classifying everything as [correct], though this would defeat the purpose of an error diagnosis system. Therefore, a balance must be struck between high recall for the [correct] class, i.e. a low proportion of correct utterances misclassified as incorrect, and high precision, i.e. a low proportion of incorrect utterances misclassified as correct. The evenly-weighted F-measure

In each cross-validation, these evaluation statistics were averaged over all folds, i.e. over each train/test split in the data. **[TODO *fit that in better*]**

4.4.2 Feature performance **[TODO *retitle?*]**

As mentioned above, a series of experiments was carried out in an effort to determine which of the prosodic features described in Section 4.2 give the best accuracy in the task of classifying lexical stress errors. Determining the best-performing features not only enables the creation of the most accurate diagnosis-by-classification system possible, but may also have implications for the way these acoustic features of the speech signal correspond (or fail to correspond) with the perception of lexical stress in non-native speech.

A 10-fold cross-validation was performed on the entire set of available training data, i.e. the utterances of 12 German word types produced by L1 French speakers which had been annotated for lexical stress errors as described in Chapter 3, along with the utterances of these word types by native German speakers, labeled as correct. As the goal of error diagnosis by classification in the CAPT context is to classify non-native, and not native, speech, including native utterances in the test data was not appropriate; therefore, for each of the 10 folds of the cross-validation, one tenth of the non-native utterances were randomly

selected to be held out as the test data set, and the other nine-tenths were combined with the native utterances to create the training data set.

To evaluate feature performance, classifiers were trained using various subsets of the complete feature set, which are listed in table 4.4. For each of these feature combinations, classifiers were trained on each of the 10 training sets created as just described, and tested on the corresponding test set. The averages of each of the aforementioned evaluation metrics (see Section 4.4.1) across all 10 folds are reported.

Prosodic features

[TODO Hypothesis: Dur>F0>Intensity]

Table 4.5 lists the results of experiments with the prosodic features described in table 4.4a. The results obtained using features representing each of the three acoustic correlates of lexical stress, duration, F0, and intensity (energy), which are displayed in the first three rows of table 4.5, conform with findings from other research regarding the correspondence between these three **[TODO features]** and lexical stress (see Section 2.3): of the three feature sets, duration seems to be the best predictor of lexical stress errors, insofar as a classifier trained on duration features alone has higher accuracy, κ, and F-scores than one trained on F0 features alone, which in turn outperforms a classifier trained using only features related to intensity. However, it should be noted that the F0 and intensity features do seem to be at an advantage over the duration features in one respect: the latter has an average recall of only 0.91 for [correct] utterances, while the other two features exhibit perfect recall (1.0). As mentioned earlier (Section 4.4.1), lower [correct] recall means that correct pronunciations are being misclassified as incorrect, which is a dangerous type of mistake for the CAPT system to make. Therefore, it is worth bearing in mind that while F0 and intensity features may not lead to the best overall accuracy in error diagnosis, they may constitute "safer" alternatives to the duration features, in that they tend to make classifiers more conservative in labeling utterances as [incorrect]. **[TODO *rewrite that sentence?*]**

Compared to using each of the three prosodic feature sets in isolation, it is clear from the figures in the lower rows of table 4.5 that even better performance can be achieved by combining these features. Interestingly, however, better performance was observed with classifiers trained only on duration and F0 features (omitting intensity features) than with those trained on features of all three types. This duration-F0 pairing resulted in the best-overall averages for any of the prosodic feature combinations: 69.77% accuracy, $\kappa = 0.29$, and $F_1 = 0.8$ for the [correct] class.

Speaker- and word-related features

As discussed in Section 3.5, differences in the realization of lexical stress can sometimes arise due to features not specific to the utterance itself, but rather to the speaker making the utterance or to the word type uttered. Therefore, in addition to the prosodic features

Table 4.4: Feature sets used in classification experiments **[TODO replace "-" in feature sets with "+", to avoid confusion]**

(a) Prosodic features (see Section 4.2)

Set name	Features
DURATION	REL-SYLL-DUR, REL-V-DUR
F0	REL-SYLL-F0-MEAN, REL-SYLL-F0-MAX, REL-SYLL-F0-MIN, REL-SYLL-F0-RANGE, REL-VOWEL-F0-MEAN, REL-VOWEL-F0-MAX, REL-VOWEL-F0-MIN, REL-VOWEL-F0-RANGE, F0-MAX-INDEX, F0-MIN-INDEX, F0-MAXRANGE-INDEX
ENERGY	REL-SYLL-ENERGY-MEAN, REL-SYLL-ENERGY-MAX, REL-VOWEL-ENERGY-MEAN, REL-VOWEL-ENERGY-MAX, ENERGY-MAX-INDEX
DUR-F0	DURATION + F0
DUR-ENER	DURATION + ENERGY
ENER-F0	ENERGY + F0
ALL	DURATION + F0 + ENERGY

(b) Speaker/word features **[TODO move to separate table in Section 4.4.2?]**

Set name	Feature(s)
WORD	The word being uttered (e.g. *Tatort*)
LEVEL	Speaker's L2 German skill level (e.g. A2)
GENDER	Speaker's age/gender category (Girl/Boy/Woman/Man)
LVL-GEN	LEVEL, GENDER
WD-LVL	WORD, LEVEL
WD-GEN	WORD, GENDER
WD-SPKR	WORD, LEVEL, GENDER

Table 4.5: Results of experiments with prosodic features [TODO explain stats] [TODO bold best values]

Feature set	% acc.	κ	[correct] class			
			P	R	F_1	F_2
DURATION	66.78	0.19	0.69	0.91	0.79	0.86
F0	64.37	0.02	0.64	1.00	0.78	0.90
ENERGY	63.77	0.00	0.64	1.00	0.78	0.90
ENER-F0	64.52	0.04	0.65	0.98	0.78	0.89
DUR-ENER	67.68	0.25	0.71	0.89	0.79	0.85
DUR-F0	69.77	0.29	0.72	0.91	0.80	0.86
ALL	67.52	0.25	0.71	0.89	0.79	0.85

described in Section 4.2, another series of experiments included features related to the speaker and word of a given utterance (listed in table 4.4b), to ascertain whether the inclusion of such features could lead to any performance gains. Table 4.6 presents the results of those experiments. [TODO Hypothesis: all are helpful]

Comparing the performance of the two speaker-related features, the speaker's proficiency level (LEVEL) seems to be a better predictor of stress accuracy than their age/gender category (GENDER, which refers to whether the speaker is a girl, boy, woman, or man, and not exclusively to whether they are a male or a female; see table 4.4b). This is not surprising, considering the large discrepancy between skill level groups observed in the error distribution analysis (see Section 3.5.3). If the intended application of the classifier were the assessment of a learner's proficiency level, including this feature would obviously be nonsensical; however, in this context the application is not assessment but training, and in that case it makes sense to allow the system to take the learner's level into account.

Interestingly, for many of the word/speaker feature combinations tried, slightly different results were obtained when these features were combined with the full set of prosodic features (ALL) than with the best-performing subset of the prosodic features, DUR+F0 [TODO *What can we infer from that?*]. For example, when combined with the feature set ALL, the feature LEVEL slightly outperformed the word of the utterance (WORD) as a predictor of stress accuracy, yet when combined with only the duration and F0 features (DUR+F0), WORD outperformed LEVEL. In both situations, the combination of these two features, WD+LVL, seemed to give better results than any combination involving GENDER, including the combination using all three word/speaker features. In fact, the best-performing classifiers resulted from using the word of the utterance, the speaker's proficiency level, and the entire set of prosodic features (WD+LVL+ALL), yielding an average accuracy of 71.87%, κ of 0.34, and F_1 and F_2 measures of 0.81 and 0.87, respectively.

Though these statistics are the best of any of the experiments reported in this section, they are still not terribly impressive, and we would perhaps like to see better accuracy and F-scores before placing such an error-diagnosis system in front of actual students. However, these results should be interpreted in the context of the particular task at hand. Considering

Table 4.6: Results of experiments with speaker and word features **[TODO Reorder subtables?] [TODO fix width] [TODO bold best values]**

(a) In combination with ALL feature set

Feature set (+ALL)	% acc.	κ	[correct] class			
			P	R	F_1	F_2
WORD	68.41	0.28	0.72	0.88	0.79	0.84
LEVEL	70.07	0.29	0.71	0.92	0.80	0.87
GENDER	66.93	0.24	0.71	0.88	0.78	0.84
LVL-GEN	68.57	0.27	0.72	0.89	0.79	0.85
WD-GEN	68.87	0.30	0.73	0.87	0.79	0.83
WD-LVL	71.87	0.34	0.73	0.92	0.81	0.87
WD-SPKR	70.52	0.31	0.72	0.91	0.80	0.86

(b) In combination with DUR-F0 feature set

Feature set (+DUR-F0)	% acc.	κ	[correct] class			
			P	R	F_1	F_2
WORD	70.52	0.30	0.72	0.92	0.81	0.87
LEVEL	68.72	0.27	0.71	0.91	0.79	0.86
GENDER	68.26	0.22	0.69	0.94	0.80	0.88
LVL-GEN	69.77	0.29	0.72	0.91	0.80	0.86
WD-GEN	68.86	0.27	0.71	0.91	0.80	0.86
WD-LVL	70.65	0.31	0.72	0.92	0.81	0.87
WD-SPKR	68.41	0.26	0.71	0.91	0.79	0.86

the relatively low inter-annotator agreement observed when humans were asked to diagnose lexical stress errors, (detailed in Section 3.4), the classification accuracy and κ scores do not seem shockingly low. Indeed, the best average κ between the classifier's decision and the gold-standard labels (0.34) exceeds the observed average human-human κ (0.23), and the best average percentage accuracy for that classifier (71.87%) is substantially higher than the average human-human percentage agreement (54.92%). **[TODO *fix this nightmare sentence-paragraph:*]** If, as seems possible given the low inter-annotator agreement observed in the annotation described in Chapter 3, there is an element of subjectivity in the decision of whether or not lexical stress was realized correctly in an utterance of a given word, such that an indisputable diagnosis of that utterance cannot be produced, and different listeners could and would make different assessments of that utterance, it would not be realistic to expect an automatic diagnosis system to reproduce humans' assessments with perfect accuracy, and any agreement between the classifier's output and the gold-standard labels which exceeds the average agreement between humans should be interpreted in a positive light.

4.4.3 Unseen speakers and words

Table 4.7: Results of experiments with unseen speakers, averaged over all 56 held-out speakers **[TODO explain stats] [TODO fix width] [TODO bold best values]**

Feature set	% acc.	κ	[correct] class			
			P	R	F_1	F_2
DUR-F0	69.16	0.19	0.68	0.90	0.74	0.85
LEVEL+DUR-F0	69.33	0.22	0.69	0.87	0.74	0.82
WD-LVL+DUR-F0	70.22	0.24	0.68	0.90	0.75	0.84

[TODO No need to have these in the same section - split into 2 separate subsections]

In the feature experiments described in the previous section, each speaker and word type in the testing data had already been seen in the training data, i.e. the training data included utterances by that speaker or of that word type (uttered by both native and non-native speakers). However, as stated in **[TODO ??]**, one motivation for a classification-based approach to error diagnosis is that the abstraction from concrete reference utterances afforded by such a method can theoretically enable the diagnosis of errors in new word types without requiring additional recordings of native speakers pronouncing these words; to assess the feasibility of diagnosing utterances of new words, the classifier's performance on words not seen in the training data must be evaluated and compared with the performance observed on seen words, with the hypothesis that the former will be better than the latter. Furthermore, in order for a CAPT tool to be useful in practice, it must be able to diagnose the speech of a learner from their very first time using the system. In the experiments reported above, it is possible that the training data's inclusion of other utterances by the speaker in question resulted in higher performance than would be observed if such utterances were not available as training instances, so the classifier's performance on unseen speakers must be compared to the performance reported above, once again with the hypothesis that the classifier will perform more poorly on unseen test data. To test these hypotheses, another series of experiments was carried out using different configurations of testing and training data sets.

Unseen speakers

To create training and testing data for experiments with utterances from unseen speakers, the entire set of labeled non-native utterances (see Section 4.4.1) was divided into 56 subsets, each containing all utterances from one of the 56 unique non-native speakers. Using a classifier trained on the native German utterances as well as those of the other 55 speakers, performance on each of these 56 test sets was evaluated. To investigate whether adding features related to speaker or word characteristics is more helpful than prosodic features alone when diagnosing unseen speakers' errors, a comparison was made between classifiers trained using several combinations of the best features from the experiments described in Section 4.4.2 above. The results of the unseen-speaker experiments, and the different feature sets used for training, are given in table 4.7.

As expected, the best performance on unseen speakers, achieved using the word and proficiency level features in addition to the duration/F0 features, is slightly lower than the overall best performance when utterances from each speaker are available as training instances (using word, level, and all prosodic features, as described in Section 4.4.2). Specifically, **[TODO we]** observe a drop in accuracy from 71.87% to 70.22%, in κ from 0.34 to 0.24, and in F_1 and F_2 measures from 0.80/0.86 to 0.75/0.84, respectively. While these performance losses do not seem drastic, they do seem to indicate that the system has greater success evaluating the accuracy of a learner's lexical stress production when it has seen labeled utterances from that speaker before; in future work it would be interesting to explore techniques for enabling improvements in accuracy as a new learner continues to use the system (see Section 6.2).

Unseen words

To evaluate classification performance on unseen word types, the non-native utterances annotated for lexical stress (see Section 4.4.1) were divided by their word type, resulting in twelve different data sets to be used for testing. To construct the training set for each test set, utterances of the target word type were removed from the complete set of native German utterances, and the resulting set of 11 word types uttered by L1 German speakers was combined with the utterances of those 11 word types by L2 speakers. Given the improvement in performance observed with the inclusion of speaker-related features (see Section 4.4.2), along with the apparent **[TODO interaction *in the everyday, not scientific sense*]** between such features and the prosodic feature set (such that overall performance improved when all prosodic features were included, not just the best-performing prosodic features of duration and F0), classifiers were trained using a few different combinations of features, to ascertain how these features affect performance when the test data consists of unseen words. It might be the case that including information about the speaker, i.e. their age/gender and proficiency level, may be more helpful when there are no training instances available for the given word type.

The results of these experiments are presented in table 4.8. As that table shows, the best average performance on all twelve unseen word types was achieved using classifiers trained only on the best-performing prosodic features,(DUR+F0), and not taking either of the speaker-related features into consideration: this yielded an accuracy of 66.85%, κ of 0.17, and F_1 and F_2 measures of 0.77 and 0.84, respectively.

A more detailed breakdown of classification performance, using these best-performing features, for each of the twelve held-out word types, is presented in table 4.9. As this table shows, there are relatively large differences in accuracy from word to word. Accuracy ranges from 83.93% (on the word *fliegen*) to 50.91% (*Tatort*), κ from -0.10 (*Frühling*) to 0.47 (*Tschechen*), and F_1 and F_2 from 0.91/0.96 (*fliegen*) to 0.49/0.55 (*Tatort*). This recalls the large differences in human-human annotator agreement observed from word type to word type, as described in Section 3.4.1, reinforces the impression that lexical stress realizations are more difficult to assess for some word types than for others, and provides an additional motivation for future investigations into the acoustic-phonetic features responsible for this observed difference in accuracy among word types (see crefsec:conclusion:future).

Table 4.8: Results of experiments with unseen words, averaged over all 12 held-out word types **[TODO explain stats] [TODO fix width] [TODO bold best values]**

Feature set	% acc.	κ	[correct] class			
			P	R	F_1	F_2
DUR-F0	66.85	0.17	0.69	0.88	0.77	0.84
LEVEL+DUR-F0	65.51	0.16	0.69	0.89	0.77	0.84
LVL-GEN+DUR-F0	65.05	0.16	0.69	0.88	0.76	0.84
ALL	65.66	0.19	0.70	0.85	0.76	0.82
LEVEL+ALL	64.16	0.11	0.67	0.88	0.75	0.83
LVL-GEN+ALL	64.31	0.12	0.68	0.90	0.77	0.85

Table 4.9: Best classification results (using feature set DUR+F0) on unseen words, by word type **[TODO explain stats] [TODO fix width] [TODO bold best values]**

Held-out word	% acc.	κ	[correct] class			
			P	R	F_1	F_2
Tatort	50.91	0.08	0.41	0.60	0.49	0.55
Mörder	62.50	0.15	0.62	0.94	0.75	0.85
manche	71.43	0.39	0.79	0.83	0.81	0.82
Pollen	64.29	0.03	0.65	0.97	0.78	0.88
Ringen	47.27	0.14	0.65	0.53	0.59	0.55
Tschechen	76.79	0.47	0.85	0.88	0.86	0.87
Flagge	60.00	0.00	0.60	1.00	0.75	0.88
Frühling	78.57	-0.10	0.85	0.92	0.88	0.90
tragen	63.64	0.15	0.62	1.00	0.76	0.89
halten	71.43	0.33	0.72	0.94	0.81	0.89
E-mail	71.43	0.30	0.69	0.97	0.81	0.90
fliegen	83.93	0.16	0.84	1.00	0.91	0.96
Average	66.85	0.17	0.69	0.88	0.77	0.83

[TODO Graph or table comparing the classifier accuracy/kappa for each word type to the human-human stats?]

In comparison with the best results obtained on seen words as reported in the previous section (**[TODO recap results for DUR+F0 and WD+LVL+ALL]**), these statistics seem to confirm the hypothesis that performance would be lower on words not seen in the training data. Perhaps most striking is the difference in κ, which constitutes a drop from fair to slight agreement with the gold-standard labels (using the Landis and Koch (1977) thresholds), which also constitutes a fall below the average inter-annotator κ of 0.23 observed between human annotations (see Section 3.4.1). However, the drop in the overall accuracy percentage, as well as in the two F-measures, seems less drastic, which seems encouraging from the perspective of extending the diagnostic capabilities of the CAPT tool to novel word types using a classification-based approach, especially considering the possibilities for improving

classification performance by other means, e.g. using additional features or more powerful machine learning algorithms (see Section 6.2).

4.5 Controlling diagnosis in the system

In the prototype CAPT tool developed in this thesis project, a researcher (or teacher) can choose between the various possible approaches for diagnosing lexical stress errors in learner's speech which have been discussed in this chapter. The choices available in the tool are illustrated in fig. 4.3. When creating a new exercise, in which a learner is asked to read one of the sentences extracted from the IFCASL corpus **[TODO (see ??]** in order to receive feedback on their realization of lexical stress in a target word in that sentence, the researcher is asked to select a DiagnosisMethod for that exercise. The Diagnosis method captures the researchers' choices among the various diagnostic options offered by the system; a screenshot showing the creation of a simple example is presented in fig. 4.4. As illustrated in **??**, this DiagnosisMethod is combined with a (compatible) FeedbackMethod when the exercise is created; see Section 5.4 for a description of FeedbackMethods and their compatibility with the various possible DiagnosisMethods.

4.6 Summary

[TODO]

Which diagnosis approach?
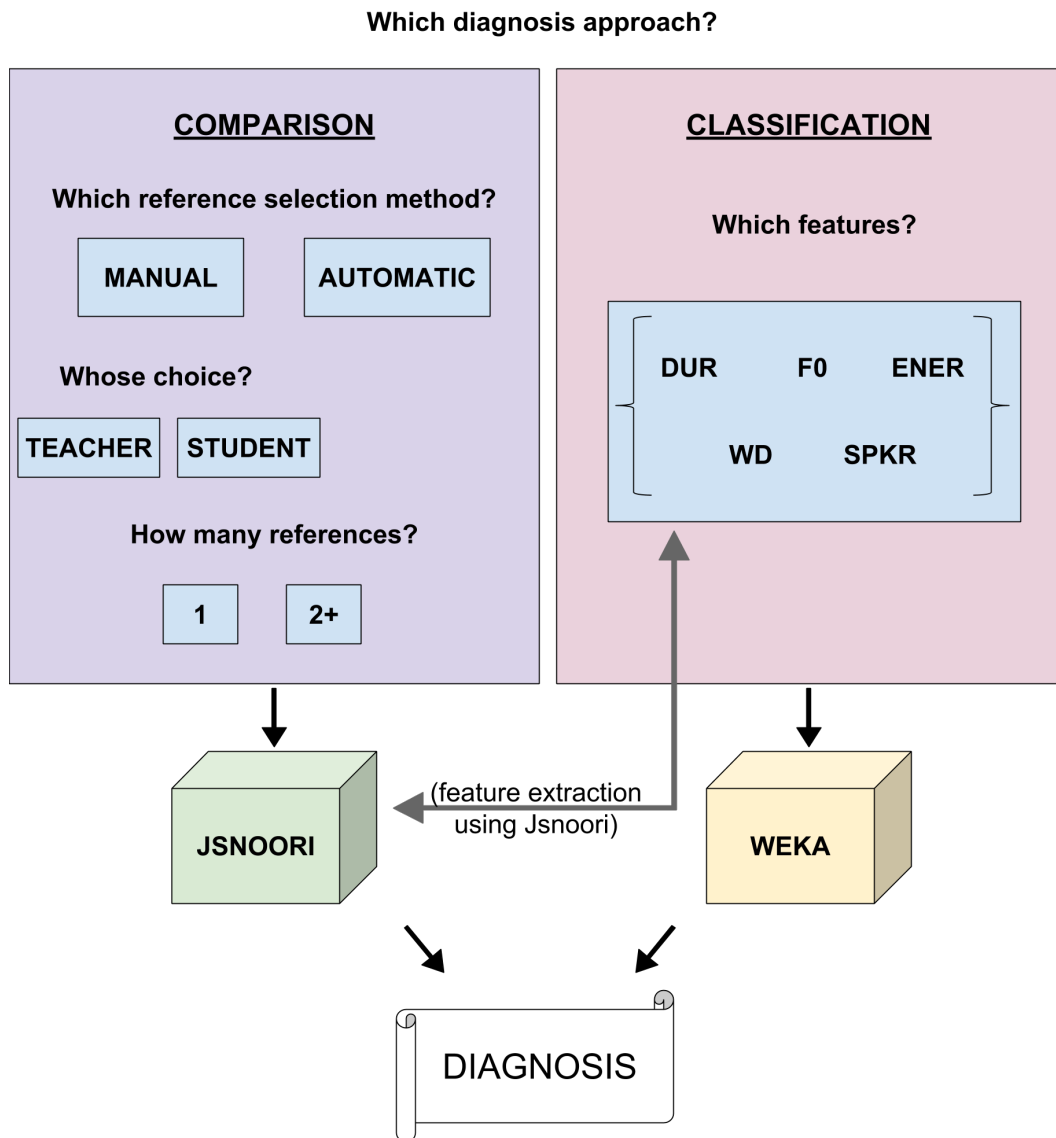
Figure 4.4: Screenshot of the researcher-facing interface to create a DiagnosisMethod

Create DiagnosisMethod

Name * SimpleComparison

Description

Scorer * Jsnoori

Number Of References * 1

Selection Type MANUAL

Create

Figure 4.5: Screenshot of the researcher-facing interface to create an Exercise

Create Exercise

Name * Comparison-TextStyle

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

Word * fliegen

Diagnosis Method * SimpleComparison-1refs-MANUAL

Feedback Method TextStylization-SelfAssessed

Lessons

Create

Feedback on lexical stress errors   5

Since the focus of this thesis is on pronunciation training, not pronunciation assessment (see Section 2.2), feedback on the errors diagnosed via the methods described in Chapter 4 is an important component of the prototype CAPT tool developed in this work. 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, an important contribution of this thesis is the creation of a feedback module for the lexical stress CAPT tool which offers a variety of possible feedback types, and a Graphical User Interface (GUI) allowing a researcher or instructor to easily switch between feedback types. The hope is that researchers can use this modular tool in in-vivo studies with language learners to compare the effects of various feedback types on the acquisition of L2 German prosody by L1 French speakers (or perhaps even speakers of other L1s); though it is outside the scope of the thesis to carry out such studies, the tool has been designed with this application in mind. Ultimately, once research has given us a better understanding of which feedback types are most effective in which situations, the modular feedback delivery system developed here could theoretically be embedded in a full-featured intelligent tutoring system, where models of the relevant learning contexts (such as the objectives of the current exercise, or the student's past achievements and personal goals and preferences) could be used to automatically select the most useful feedback type to present to the learner, as mentioned in Section 1.2 and illustrated in figs. 1.1 and 5.1.

This chapter presents the various options for the types of feedback that can be generated given a diagnosis of the learner's lexical stress realization, **[TODO *rephrase:* guided by the notion that]** to maximize its utility in future feedback research, the CAPT tool should offer as wide a variety of feedback options as possible, especially those offering types of feedback not commonly seen in existing CAPT systems.

5.1 Implicit feedback

[TODO Intro]

5.1.1 Visual

Visual delivery of feedback on learner errors (or lack thereof) is a widely used technique in CAPT. In many existing CAPT tools (e.g. Martin, 2004; Henry et al., 2007), 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. Indeed,
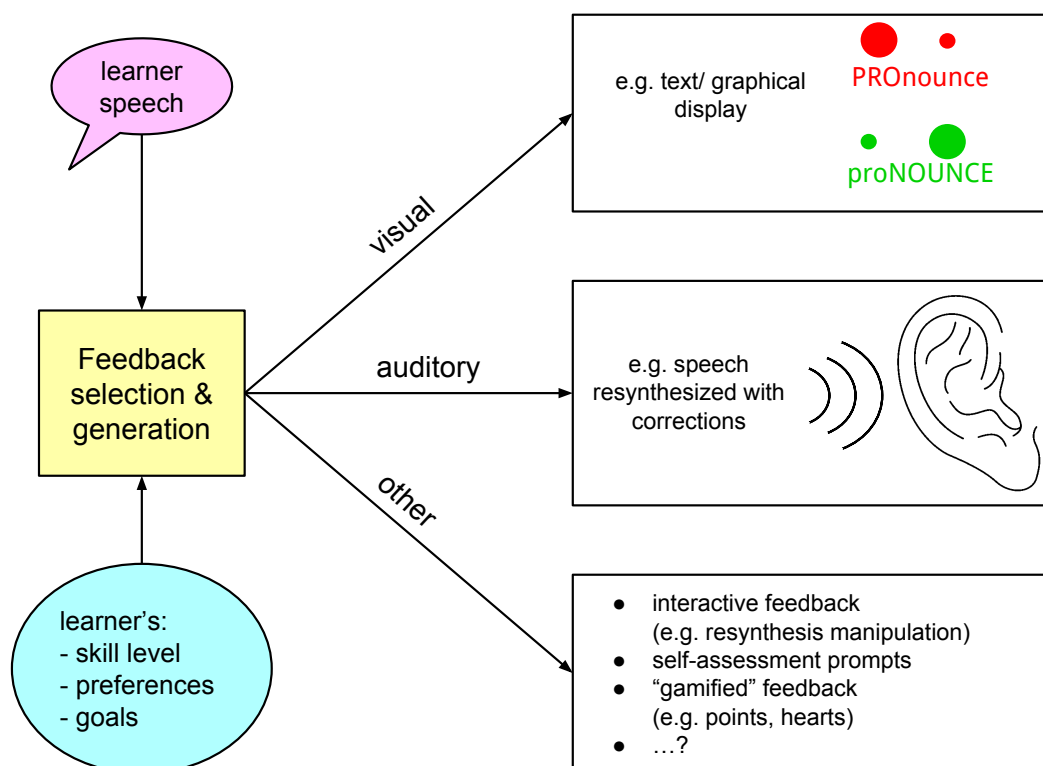
Figure 5.1: Delivery of prosody feedback in different modalities. **[TODO redo or remove]**

this is the case in Jsnoori, as seen in **[TODO Jsnoori screenshot]**. 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; to this end, visual feedback in the lexical stress CAPT tool developed in this thesis project focuses on alternatives to direct signal visualizations, as described in this section.

Graphical abstractions of prosody

One type of alternative to direct visualizations of the speech signal is 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). By mapping the acoustic features of each syllable in the utterance(s) to graphical features (e.g. height, width) of a geometrical shape it is possible to dynamically create a visual abstraction of the relevant properties of the learner's utterance as well as those of the reference utterance(s). This abstracted visual representation of prosody may be easier for the learner to interpret than the complex and possibly overwhelming signal visualizations more commonly used to give prosodic feedback in CAPT as mentioned in the previous section.

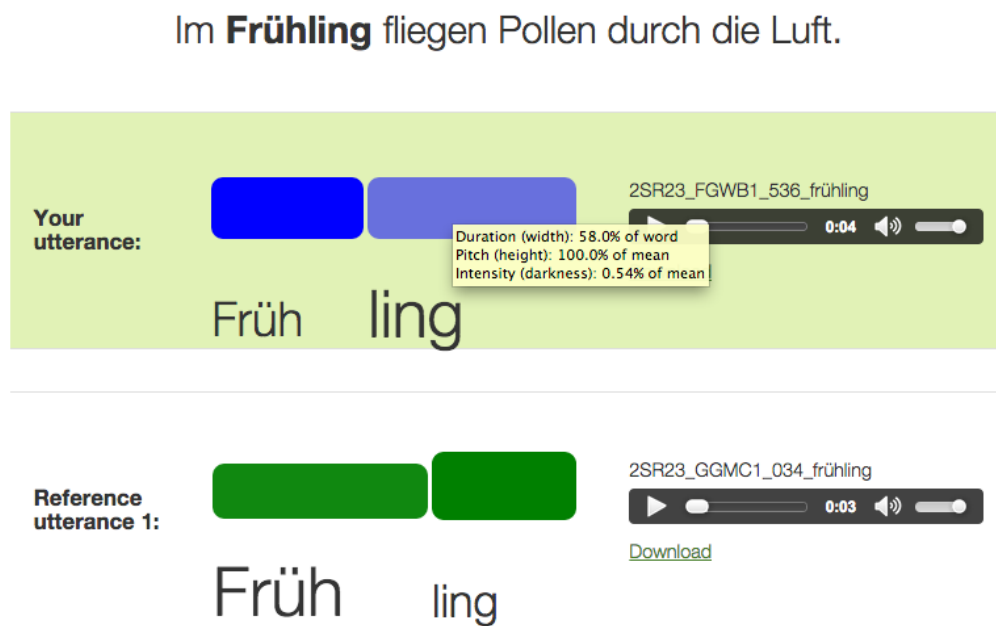**Figure 5.2:** Screenshot of feedback via graphical abstractions of prosody

Figure 5.2 illustrates the display of such graphical abstractions in the prototype CAPT tool. Each syllable in an utterance is represented by a rectangle, the length of which corresponds to the duration of that syllable (as a percentage of the total word duration), the height of which represents the mean F0 in that syllable (normalized by dividing the absolute mean F0 for the syllable by the overall mean in the word), and the opacity of which corresponds to the mean intensity of that syllable (again normalized by dividing by the mean in the entire word). If the learner hovers their mouse over one of the rectangles, they are presented with the exact values for each of these features in a tooltip overlay, which can be seen as the small yellow box in the central area of fig. 5.2.

In the prototype CAPT tool, the mappings between graphical properties (width, height, and opacity of the rectangle) and prosodic features (duration, F0, and intensity, respectively) are hard-wired. However, researchers using the system to experiment with different feedback methods should ideally be able to change the mapping or omit one of the features if necessary, so a more flexible feature-mapping mechanism would be a worthwhile improvement to the system (see Section 6.2).

Stylized text

A related approach to the abstract graphical representations just described involves stylizing, or reshaping, the text of the word(s) pronounced to match the prosodic features of the learner's utterance. This is essentially the approach used by the work Sitaram et al. (2011), who used text stylization to help visualize prosody in the Project LISTEN reading tutor (see Section 2.2.4. Such text stylization is also often used to convey canonical prosody in pronunciation instruction materials (Behme-Gissel, 2005; Hirschfeld et al., 2007), e.g. by

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

Your utterance:	flie **gen**	2SR23_FGBA2_546_fliegen ▶ ⟶ 0:04 🔊 ⟶ Download
Reference utterance 1:	**flie** gen	2SR23_GGBA2_007_fliegen ▶ ⟶ 0:04 🔊 ⟶ Download

using larger text for the stressed syllable than for the unstressed syllable(s) in a given word. Familiarity with this type of presentation might make feedback via stylized text easier for learners to comprehend, so text stylization was another form of implicit visual feedback implemented in the system, as illustrated in fig. 5.3.

When using geometric shapes to visualize prosody, different prosodic features can be conveyed simultaneously by mapping each to a different geometric property, as described in Section 5.1.1 However, when dealing with text, visualizing multiple prosodic features at the same time is more difficult. First of all, noticing a clear difference between two syllables in terms of textual features such as height, font weight, or letter spacing is not as easy as comparing the height and width of two rectangles, given the inherent geometric variability of the different letters of the alphabet. Secondly, if text is stretched or skewed too dramatically, it becomes more difficult to read, which may be distracting for learners using the system.

Therefore, in the CAPT tool developed here, the text of a given syllable is stylized with a simple mapping between font size (as a multiple of the default size) and duration (as a fraction of the word duration). Duration was chosen as the prosodic feature to visualize based on its relative importance for the perception of lexical stress in German (see Chapter 4). Font size was chosen as the textual feature to manipulate because it can be changed without distorting the text, i.e. without risking decreased legibility. Of course, other mappings between prosodic and textual features could be imagined, and once again the addition of other options than those currently implemented in the system could be worthwhile (see Section 6.2), though this might be less useful for text stylization than for graphical visualizations, for the reasons mentioned in the previous paragraph.

5.1.2 Auditory

[TODO Is an intro necessary here? What should/can be said?]

Student & reference audio

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. This type of implicit auditory feedback is perhaps the most simple feedback type to deliver, so the CAPT tool naturally offers learners the ability to listen to their own utterance as well as the reference utterance(s), and to download a wave file of any utterance for later reference if they so choose. Though learners may not always be able to detect errors in their pronunciation or possibilities for improvement from such implicit feedback alone, in conjunction with visual feedback of the types mentioned above this auditory feedback may help them improve their sensitivity to the stress patterns audible in the utterance(s). Therefore, as seen in figs. 5.2 and 5.3, these audio recordings are always accessible alongside the visual (and other types of) feedback presented in the CAPT tool.

Resynthesized audio

As described in Section 2.2, previous work on delivering lexical stress feedback has revealed that learners sometimes benefit from prosodically modified implicit auditory feedback, either in the form of a learner utterance modified to reflect the "correct" prosody of a native reference utterance (Bonneau and Colotte, 2011). Thanks to the speech resynthesis capabilities of Jsnoori, this is another feedback option offered by the prototype CAPT tool.

To modify the learner's utterance to match a reference utterance, Jsnoori uses the technique of Time Domain Pitch Synchronous Overlap and Add, or TD-PSOLA (**Moulines1990**). This technique uses the general strategy of creating a new, modified version of the original signal by using a windowing function to break the signal into a series of overlapping frames, where the spacing between those frames is pitch-synchronous, i.e. corresponds to the F0 period of (that part of) the signal. The duration of a region of the signal (e.g. a vowel) can then be decreased or increased by removing or duplicating frames in that region while keeping the spacing between frames consistent, and the perceived pitch of a region can be lowered or raised by increasing or decreasing the spacing between frames. The implementation of this signal-transformation technique in Jsnoori includes an improved method for detecting pitch marks in the original signal (**Laprie1998**; **Colotte2002**), as the accuracy of pitch marking is vitally important to the quality of resynthesis obtained via TD-PSOLA.

To apply this technique in the context of prosody-oriented CAPT (see **????**), the target prosody for that utterance is first established via analysis and comparison of F0 and duration in the learner and reference utterances, resulting in the computation of target F0 contours and relative phone durations for (the relevant sections of) the learner's utterance. The learner's signal is then transformed to match the targets for F0 and duration by means of the removal/addition and re-spacing of frames, as described in the previous paragraph. The resulting signal maintains the individuality of the learner's voice, yet replaces their original "incorrect" prosody with the "correct" prosody of the reference speaker.

Like the Jsnoori software itself, the prototype CAPT tool offers the learner the opportunity to listen to this resynthesized utterance alongside their original utterance and that of the the native speaker. The hope is that by offering this this implicit auditory feedback as one of the many feedback types teachers/researchers can choose to present to the learner, the CAPT tool will facilitate further research into the use of this type of speech resynthesis for L2 language teaching.
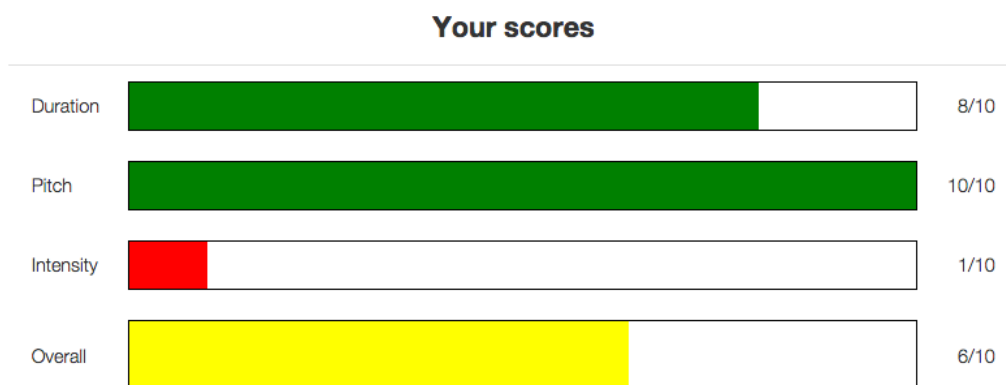
5.2 Explicit feedback

5.2.1 Skill bars

One way in which the diagnosis of the learner's utterance in terms of duration, F0, and energy scores (calculated as described in Section 4.3.1) is made explicit to the learner is by means of graphical skill bars of the type often used in tutoring systems (e.g. Long and Aleven, 2011; Long and Aleven, 2013). As illustrated in figs. 5.4 and 5.5, these bars provide explicit visual feedback for each "skill" (feature type) by displaying the score for that feature (as an integer out of 10, visible on the right hand side of the bar), as well as by graphically representing this score with both the length of the filled region of the bar (as a fraction of the total bar width corresponding to the score) and the color of that region (green for scores above 0.7, red for scores below 0.25, and yellow for intermediate scores). The bottom-most bar represents the overall score, computed as a weighted average of the three individual scores, using the weights assigned to each score by the researcher/teacher when configuring the diagnosis method **[TODO ??]**.

Figure 5.4 illustrates a case where each of the three scores is given equal weighting, while fig. 5.5 shows the same scores in the case where duration is prioritized over F0, which is in turn given a higher weight than intensity. **[TODO learner should be informed of the weighting - take new screenshot]** Although the teacher or researcher setting up the exercise may (justifiably) choose to prioritize a feature, such as duration, in this way, one drawback of the skill bar visualization is that the equal size of the bars for each feature score seems to convey that all are equally important. Modifying the size of each bar based on the weight accorded to its feature may be a simple way to improve the effectiveness of this feedback type.

A potential problem with this feedback method relates to the fact, discussed in Section 4.3.1, that the scores output by Jsnoori's diagnostic tools are actually discrete, ordinal values, despite the use of numbers between 0 and 1 to represent these values. Therefore, the visual representation of these scores as if they belonged to a true interval or continuous variable is not necessarily justified, and this feedback may be somewhat confusing to the learner. It would be worthwhile to investigate this hypothesis through the type of in-vivo studies the CAPT tool has been designed to facilitate.

Figure 5.4: Screenshot of explicit feedback via skill bars, where all three skills (prosodic feature types) contribute equally to the overall score

Your scores

Duration		8/10
Pitch		10/10
Intensity		1/10
Overall		6/10

Figure 5.5: Screenshot of explicit feedback via skill bars, where duration contributes more than F0, which contributes more than intensity (60%, 30% and 10% of overall score, respectively) **[TODO learner should be informed of the weighting - take new screenshot]**

Your scores

Duration		8/10
Pitch		10/10
Intensity		1/10
Overall		8/10
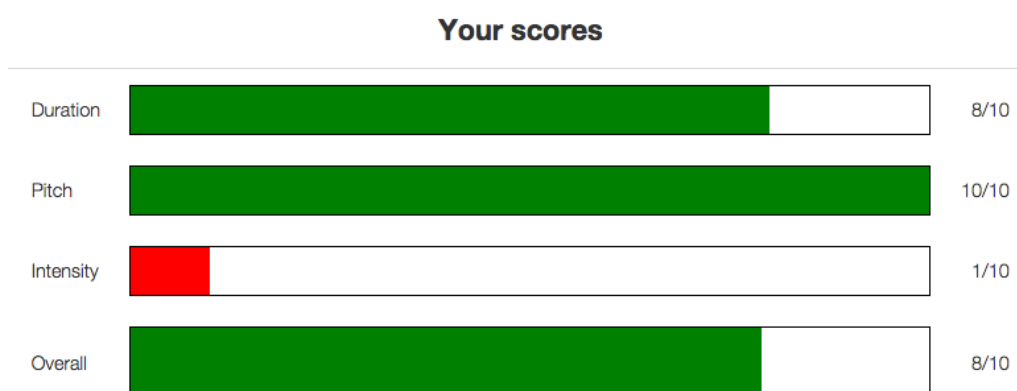
Table 5.1: Messages used to deliver explicit verbal feedback of feature-specific scores

Feature type	Score	Message
Duration (timing)	[TODO]	[TODO]
F0 (pitch)	[TODO]	[TODO]
Intensity (energy)	[TODO]	[TODO]

5.2.2 Verbal feedback

Another way to explicitly deliver feedback on the learner's diagnosis is by verbalizing this diagnosis with one or more appropriate messages. If individual scores have been computed for the three prosodic feature types (duration, F0, and intensity) as described in Section 4.3.1, these are verbalized using the corresponding message from the set listed in table 5.1. If the learner's utterance has been diagnosed via classification and assigned one of the possible stress-accuracy labels (see Section 4.4), that classification is verbalized with one of the following messages:

- [correct]: "You stressed the correct syllable. Great job!"

- [none]: "It sounds like you pronounced both syllables with equal stress. Next time, try to use duration, pitch, and loudness to make the first syllable sound more important than the second syllable."

- [incorrect]: "It sounds like you stressed the incorrect syllable. Remember that the stress in <WORD> should be on the first syllable."

5.3 Self-assessment

Research on the efficacy of computer-based and intelligent tutoring systems has often pointed to the fact that encouraging and assisting learners to develop their metacognition, i.e. their understanding of their own learning habits, goals, challenges, etc., can lead to greater engagement and thus to increased motivation to continue using the tutoring system, and may also help learners reach their educational goals faster by optimizing their own learning strategies (see e.g. ????). One important metacognitive process is that of self-assessment, i.e. a learner's own evaluation of their work (in this context, their pronunciation) without reference to feedback from a human or machine tutor. Self-assessment is a valuable way for learners to give themselves feedback on their own performance, and asking learners to self-assess before presenting any feedback from the system may help them to internalize the system's feedback as well as to improve their own self-assessment skills over time.

In the prototype CAPT tool, learners can optionally be asked to assess their own pronunciation by filling out a short questionnaire before any feedback is delivered. This questionnaire, seen in fig. 5.6, asks learners to listen to their utterance and that of the reference speaker(s) and assess whether they have placed stress on the correct syllable, whether stress is clearly realized in their utterance, and how they can improve their stress production going forward. The learner's responses to these questions are not evaluated in any way, but instead are

Figure 5.6: Self-assessment questionnaire presented to learner before feedback delivery

Self-assessment

Listen to your utterance and the reference utterance(s).
Then answer these questions:

Which syllable did you stress?

○ The first syllable (correct)

○ The second syllable (incorrect)

○ Neither syllable (incorrect)

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

○ Yes

○ No

○ I don't know

What could you work on for next time?

[TODO logged in the system] so that they can later reflect on their self-assessed progress and refer to their self-directed advice.

The hope is that the inclusion of the self-assessment option alongside the other feedback options will facilitate research on whether requiring learners to complete this type of self-assessment activity has any influence on their self-regulated learning habits and, perhaps by extension, the improvement of their L2 prosody.

5.4 Controlling feedback in the system

As mentioned in Section 4.5 and illustrated in **??**, a researcher (or teacher) creating a new exercise in the CAPT tool is asked to select a FeedbackMethod for that exercise, which captures the researcher's choices about which feedback types to present to a student after their utterance has been evaluated by the corresponding DiagnosisMethod. Figure 5.7 presents a screenshot showing the various options available to the researcher, as have been described in this chapter.

Noticeable in fig. 5.7 is the FeedbackMethod parameter "Requires Scorer Type"; this parameter captures the fact that not all feedback types are compatible with all of the diagnosis methods offered by the system (see **??**). For example, if the chosen classification method diagnoses learner errors by classification and outputs a single class label instead of individual scores for each of the three feature types (duration, F0, and intensity), it does not make sense to present feedback in the form of skill bars (**??**). Therefore, each FeedbackMethod is aware of which type of scores it requires, so that the researcher never accidentally chooses

Figure 5.7: Screenshot of the researcher-facing interface to create a FeedbackMethod [TODO Update screenshot]

Create FeedbackMethod

Name *

Description *

Requires Scorer Type

Show Skill Bars ☐

Play Feedback Signal ☐

Display Shapes ☐

Style Text ☐

Self Assessment ☐

Create

Table 5.2: Compatibility of the various diagnosis and feedback types available in the proto-type CAPT tool. Compatible combinations are indicated with "+", incompatible combinations with "−".

	Diagnosis method	
Feedback type	Comparison	Classification
Graphical abstraction	+	+
Stylized text	+	+
Student & reference audio	+	+
Resynthesized audio	+	−
Skill bars	+	−
Verbal feedback	+	+
Self-assessment	+	+

an incompatible pairing between DiagnosisMethod and FeedbackMethod when creating a new exercise. Table 5.2 summarizes the (in)compatibility between the various types of diagnosis and feedback offered by the CAPT tool.

5.5 Summary

[TODO]

Conclusion and outlook

<div style="text-align: right">6</div>

6.1 Thesis summary

[TODO here or in intro? or somewhere else?] This thesis project explores 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 different diagnostic approaches to study their effects, is one of the primary contributions of this work.

6.2 Future work

6.2.1 Processing non-native speech

Evaluation of segmentation accuracy

[TODO]

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.

Coping with segmentation errors

[TODO]

As mentioned in Section 4.1, 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.

Features of difficult-to-classify utterances

Would be a good idea to analyze the utterances that had low inter-annotator agreement (and those that were misclassified by the automatic system) to find features that differentiate "clear" utterances from "unclear" ones. (cf what Michaux and Caspers (2013) plan to do)

Pitch in Jsnoori

Soon to implement YIN (Cheveigné and Kawahara, 2002) temporal method, and then combine results with those of Martin's or another spectral method such as SWIPE (Camacho, 2007) (combining with NNs or something else)

6.2.2 Diagnosis

Multiple references

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. Alternatively, the learner's utterance could perhaps be compared directly with some unified representation of all the reference utterances; for example, if we represent each reference utterance as a point in n-dimensional space, with each dimension representing a relevant feature, the references will form a cluster which can serve as a representation of the variation permissible in native speech. By plotting the learner's utterance in the same space, it could be possible to distinguish how well (or poorly) this utterance fits into that cluster, and thereby produce a diagnosis.

Auto reference selection

.... 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)

Relevant features may include.... spectral and duration-based features, and/or other features informed by research on speaker identification (e.g. Shriberg et al., 2005). **[TODO examples of speaker ID features]**

No-reference diagnosis

Other 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 **[TODO fancier]** classification-based machine learning approaches ...online, semi-supervised learning which could allow the system to benefit from each new utterance recorded as a learner continues to use the system.

6.2.3 Feedback

Mapping prosodic features to graphical properties

Shapes: 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.

Text: 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.

Animation to convey pronunciation targets

...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.

Alternative prosodic modifications of learner speech

- Based on generalized contours instead of a concrete example - possible?
- Resynthesis of native reference utterance to emphasize/exaggerate stressed syllables, cf. (Bissiri et al., 2006; Bissiri and Pfitzinger, 2009)

References

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. 67).

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, 79).

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, 79).

Boersma, Paul and David Weenink (2014). *Praat: doing phonetics by computer* (cit. on pp. 15, 18).

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, 12, 44, 45, 49, 69).

Bonneau, Anne, Matthieu Camus, Yves Laprie, and Vincent Colotte (2004). "A computer-assisted learning of English prosody for French students". In: *Proceedings of InSTIL/ICALL2004 – NLP and Speech Technologies in Advanced Language Learning Systems*. Venice (cit. on pp. 7, 49).

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 pp. 15, 41, 78).

Bouselmi, Ghazi, Dominique Fohr, Irina Illina, and Jean Paul Haton (2005). "Fully automated non-native speech recognition using confusion-based acoustic model integration". In: *Eurospeech—9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)* (cit. on p. 41).

Bouselmi, Ghazi, Dominique Fohr, and Irina Illina (2012). "Multilingual recognition of non-native speech using acoustic model transformation and pronunciation modeling". In: *International Journal of Speech Technology* 15.2, pp. 203–213 (cit. on p. 41).

Camacho, A (2007). "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music". In: (cit. on p. 78).

Cheveigné, A De and H Kawahara (2002). "YIN, a fundamental frequency estimator for speech and music". In: *... Journal of the Acoustical Society of ...* (Cit. on p. 78).

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. 10).

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. 9, 11, 12, 42, 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 pp. 5, 11).

Di Martino, J and Y Laprie (1999). "An efficient F0 determination algorithm based on the implicit calculation of the autocorrelation of the temporal excitation signal". In: *6th European Conference on Speech Communication & Technology (EUROSPEECH'99)*, p. 4 (cit. on p. 45).

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 pp. 9, 44).

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. 8, 52).

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. 10, 12).

Eskenazi, Maxine (2009). "An overview of spoken language technology for education". In: *Speech Communication* 51.10, pp. 832–844 (cit. on pp. 6, 77).

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, 40, 41).

Fohr, D and Y Laprie (1989). "Snorri: an interactive tool for speech analysis." In: *EU-ROSPEECH* (cit. on pp. 6, 7).

Fohr, Dominique and Odile Mella (2012). "CoALT: A Software for Comparing Automatic Labelling Tools." In: *LREC*, pp. 325–332 (cit. on pp. 40, 77).

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. 40).

Hahn, L. D. (2004). "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals". In: *TESOL quarterly* 38.2, pp. 201–223 (cit. on pp. 5, 11).

Hall, M, E Frank, and G Holmes (2009). "The WEKA data mining software: an update". In: *ACM SIGKDD . . .* (Cit. on p. 53).

Henry, Guillaume, Anne Bonneau, and Vincent Colotte (2007). "Tools devoted to the acquisition of the prosody of a foreign language". In: *International Congress of Phonetic Sciences (ICPhS 2007)*. August, pp. 1593–1596 (cit. on pp. 7, 49, 65).

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, 11).

Hirschfeld, Ursula and Kerstin Reinke (1998). *Phonetik Simsalabim: Ein Übungskurs für Deutschlernender (Begleitbuch)*. Langenscheidt (cit. on p. 66).

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. 67).

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).

Jouvet, Denis, Larbi Mesbahi, Anne Bonneau, Dominique Fohr, Irina Illina, and Yves Laprie (2011). *Impact of Pronunciation Variant Frequency on Automatic Non-Native Speech Segmentation*. en (cit. on p. 41).

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. 12, 52).

Landis, J R and G G Koch (1977). "The measurement of observer agreement for categorical data." In: *Biometrics* 33.1, pp. 159–174 (cit. on pp. 20, 61).

Laprie, Y (1999). "Snorri, a software for speech sciences". In: *MATISSE-ESCA/SOCRATES Workshop on Method . . .* (Cit. on p. 7).

Long, Y and V Aleven (2011). "Students' Understanding of Their Student Model". In: *Artificial Intelligence in Education* (cit. on p. 70).

– (2013). "Supporting students' self-regulated learning with an open learner model in a linear equation tutor". In: *Artificial intelligence in education* (cit. on p. 70).

Martin, P (1982). "Comparison of pitch detection by cepstrum and spectral comb analysis". In: *Acoustics, Speech, and Signal Processing, IEEE . . .* (Cit. on p. 45).

Martin, Philippe (2004). "WinPitch LTL II, a multimodal pronunciation software". In: *In-STIL/ICALL Symposium 2004* (cit. on pp. 8, 65).

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, 40–42, 78).

Michaux, MC (2012). "Exploring the production and perception of word stress by French-speaking learners of Dutch". In: *Workshop on Crosslinguistic Influence in Non-Native Language Acquisition* (cit. on pp. 10, 11, 34).

Michaux, MC and J Caspers (2013). "The production of Dutch word stress by Francophone learners". In: *Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013)*, pp. 89–94 (cit. on pp. 10, 14, 20, 32, 78).

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. 8).

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, 66).

Ney, H (1981). "A dynamic programming technique for nonlinear smoothing". In: *Acoustics, Speech, and Signal Processing, IEEE . . .* (Cit. on p. 45).

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, 15, 41, 78).

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, 51).

Project-Team PAROLE (2013). *Activity Report 2013*. Tech. rep. Nancy: LORIA (cit. on p. 7).

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. 79).

Secrest, B and GR Doddington (1983). "An integrated pitch tracking algorithm for speech systems". In: *Acoustics, Speech, and Signal . . .* (Cit. on p. 45).

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. 12, 52).

Shriberg, E., L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke (2005). "Modeling prosodic feature sequences for speaker recognition". In: *Speech Communication* 46.3-4, pp. 455–472 (cit. on p. 79).

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. 8, 67, 79).

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. 8).

Wik, P, R Hincks, and JB Hirschberg (2009). "Responses to Ville: A virtual language teacher for Swedish". In: (cit. on p. 12).

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, 12).