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

March 29, 2015

Anjana Sofia Vakil
anjanav@coli.uni-saarland.de
Automatic diagnosis and feedback for lexical stress errors in non-native speech
March 29, 2015
Supervisors: Prof. Dr. Bernd Möbius and Dr. Jürgen Trouvain

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

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

Declaration 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, March 29, 2015

Abstract

The prosodic realization of lexical stress, the phenomenon by which certain syllable(s) in a word are accentuated more than others, is an important feature of the German phonological system, but one that can pose a considerable challenge to students learning German as a foreign language (L2). This challenge is particularly daunting for native (L1) speakers of French, as lexical stress is realized quite differently (or perhaps not at all) in French word prosody. As pronunciation training typically demands substantial individual attention from an instructor, which is not always feasible in classroom language-learning settings, Computer Assisted Pronunciation Training (CAPT) has emerged over recent decades as a promising way of using technology to deliver individualized pronunciation training in the absence of a human teacher. This thesis investigates how CAPT can be used to help L1 French speakers learning German as L2 improve their pronunciation with regard to lexical stress prosody.

In an effort to illuminate the nature of L1 French learners' lexical stress errors in German, the thesis describes the manual annotation of lexical stress errors in a learner speech corpus, and presents an analysis of the frequency and types of errors observed. A variety of methods for automatically diagnosing such errors in learner word utterances are then explored, including novel methods for diagnosis by means of classification using supervised machine learning, as well as by means of comparison with one or more reference utterances, i.e. the same word pronounced by a native German speaker. The ways in which diagnoses produced by these methods can be used to deliver diverse types of feedback on these errors are then explored, including types which have not previously been utilized in German CAPT.

In its most important contribution, this thesis describes a prototype CAPT tool, **[TODO de-stress]**, which integrates these various diagnostic and feedback methods via an easy-to-use web interface. This tool is designed not only to provide students with feedback on their lexical stress errors, but also to facilitate research on the efficacy of the various types of diagnosis and feedback explored in this thesis, and to enable L2 German teachers to create exercises best suited to their students' needs; it thus constitutes an important step towards the development of a comprehensive, intelligent CAPT system for French learners of German.

Zusammenfassung

Die prosodische Realisierung von lexikalischer Betonung, das Phänomen, bei dem eine oder mehrere bestimmte Silben eines Wortes akzentuierter produziert werden als andere, ist ein wichtiges Merkmal des phonologischen Systems des Deutschen, was allerdings eine bedeutende Herausforderung für Lerner des Deutschen als Fremdsprache (L2) darstellen kann. Diese Herausforderung ist besonders groß für Sprecher mit Französisch als Muttersprache (L1), da lexikalische Betonung anders (vielleicht überhaupt nicht) realisiert wird. Da Aussprachetraining üblicherweise viel individuelle Aufmerksamkeit für jeden Lerner erfordert, die im Fremdsprachenunterricht nicht immer zureichend gewährleistet werden kann, entwickelte sich über die letzten Jahrzehnte mit dem computergestütztem Aussprachetraining (Computer Assisted Pronunciation Training (CAPT)) eine neue Möglichkeit heraus, seine Aussprache auch ohne einen Lehrer/eine Lehrerin individuell zu trainieren. Diese Arbeit untersucht, wie CAPT-Systeme dabei behilflich sein können, die Realisierung lexikalischer Betonung von französischen Lernern des Deutschen zu verbessern.

In dem Bemühen der Natur der Fehler von lexikalischer Betonung durch französische Muttersprachler auf den Grund zu gehen, beschreibt die Arbeit die manuelle Annotation der lexikalischen Betonungsfehler in einem Lernerkorpus und präsentiert eine Analyse der verschiedenen Fehlertypen und deren Häufigkeiten. Unterschiedliche Methoden der automatischen Diagnose dieser Fehler in Lerneräußerungen werden untersucht, einschließlich neuer Methoden, wie beispielsweise der Klassifizierung durch überwachte maschinelle Lernverfahren oder wie Vergleiche mit einer oder mehreren Vergleichsäußerungen von einem deutschen Muttersprachler. Anschließend wird untersucht, inwiefern die Diagnosen dieser Methoden genutzt werden können, um verschiedene Arten von Feedback zu präsentieren. Dies bezieht ebenfalls Feedbackmethoden mit ein, die vorher noch in keinem deutschen CAPT-Stystem verwendet wurden.

Der wichtigste Beitrag dieser Arbeit ist die Beschreibung eines prototypischen CAPT Tools (**[TODO de-stress]**), das die verschiedenen diagnostischen Verfahren und Feedbackmethoden über ein leicht zu nutzendes Interface miteinander vereint. Dieses System ist so konzipiert, dass es nicht nur Feedback zur Realisierung von lexikalischer Betonung gibt, sondern ebenfalls eine Plattform zur Effizienzanalyse der verschiedenen Diagnose- und Feedbacktypen darstellt. Es ermöglicht außerdem Lehrern von Deutsch als Fremdsprache, speziell auf die Bedürfnisse der Schüler abgestimmte Aufgaben zu erstellen. Aus diesem Grund trägt es zur weiteren Entwicklung von umfassenden, intelligenten CAPT-Systemen für französische Lerner des Deutschen bei.

Übersetzt aus dem Englischen von Jeanin Jügler und Frank Zimmerer

Résumé

La réalisation prosodique de l'accent lexical, c.-à-d. l'accentuation plus ou moins forte d'une ou de plusieurs syllabes d'un mot, est un élément important de la phonologie allemande qui peut s'avérer très problématique pour les personnes qui apprennent l'allemand en tant que langue étrangère (L2). Les personnes de langue maternelle (L1) française sont particulièrement mal armées pour l'appréhender puisque la réalisation de l'accent est différente (voire inexistante) en prosodie lexicale française. Les exercices de prononciation demandent souvent une attention individuelle considérable de la part de l'enseignant qui n'est pas toujours envisageable dans le contexte des cours de langue en classe. L'Enseignement de la prononciation assisté par ordinateur (EPAO) apparaît depuis quelques décennies comme une utilisation prometteuse de la technologie pour l'enseignement individualisé de la prononciation en l'absence d'un enseignant de chair et d'os. Ce mémoire s'intéresse aux possibilités qu'offre l'EPAO pour aider les francophones qui apprennent l'allemand comme langue étrangère à mieux maîtriser l'accent lexical.

Afin de déterminer la nature des erreurs d'accent lexical des francophones qui apprennent l'allemand, ce mémoire décrit comment ces erreurs ont été recensées manuellement dans un corpus vocal et analyse la fréquence et les types d'erreurs observées. Il explore ensuite différentes méthodes pour diagnostiquer automatiquement ces erreurs, dont certaines méthodes inédites qui opèrent une classification par le biais de l'apprentissage sur ordinateur supervisé ou qui comparent les prononciations des apprenants avec un ou plusieurs énoncés de référence, prononcé(s) par une personne de langue maternelle allemande. Ce mémoire s'intéresse ensuite aux utilisations possibles de ces diagnostiques pour informer les étudiants de leurs erreurs, parfois d'une façon encore inédite en EPAO appliqué à l'allemand.

L'élément principal de ce mémoire reste cependant sa description d'un prototype d'outil d'EPAO, **[TODO de-stress]**, qui combine ces différents diagnostiques et méthodes de retour d'information, présentés à l'aide d'une interface web facile d'utilisation. Son intérêt ne réside pas seulement dans la communication aux apprenants de leurs erreurs d'accentuation lexicale, mais aussi dans le fait qu'il facilite la recherche sur l'efficacité des différents types de diagnostiques et de retours examinés dans ce mémoire. Il permet également aux professeurs d'allemand comme langue étrangère de créer des exercices adaptés aux besoins de leurs élèves. Ainsi, cet outil constitue une étape importante dans le développement d'un système EPAO complet et intelligent pour les francophones qui apprennent l'allemand.

Traduit de l'anglais par Marie Springinsfeld

Acknowledgments

First of all, I am extremely grateful to my thesis supervisors at Saarland University, **Prof. Dr. Bernd Möbius** and **Dr. Jürgen Trouvain**; this thesis could not have been written without their support, encouragement, and feedback.

The work reported in this thesis was partially supported by the Franco-German project *Individualized Feedback for Computer-Assisted Spoken Language learning* (IFCASL), funded jointly by the Deutsche Forschungsgemeinschaft (DFG) and the Agence Nationale de la Recherche (ANR). I thank the the entire **Project IFCASL team** at both Saarland University (Saarbrücken, Germany) and LORIA (Nancy, France) for giving me the opportunity to work with them on this project over the last two years, and for supporting my work at both institutions.

Among the researchers at LORIA, I would especially like to thank: **Dr. Yves Laprie** for providing invaluable information on the automatic processing of prosody and for generously sharing his office with me; **Dr. Julie Busset** for allowing me to work closely with her on refactoring the *JSnoori* software to facilitate integration with programs such as *de-stress*; **Dr. Anne Bonneau** for advising me on feedback on prosody in nonnative speech and for welcoming me in Nancy; and **Dr. Dominique Fohr** and **Dr. Odile Mella** for elucidating the automatic segmentation of nonnative speech.

I am also extremely grateful to IFCASL team members **Jeanin Jügler** and **Dr. Frank Zimmerer** at Saarland University for all their help, and especially for assisting with the lexical stress error annotation project and providing the German translation of the thesis abstract.

I would additionally like to thank the **annotators** (not named to preserve anonymity) who donated their time to the error annotation project; without the labeled data produced through their efforts, much of the work reported in this thesis would not have been possible.

My heartfelt thanks also go to **Marie Springinsfeld** for generously providing the French translation of the thesis abstract, and to **Max Rabkin** and **Nathaneil Ward** for kindly helping to proofread this document.

Finally, I would like to express my gratitude to my mother, **Eileen Julian**, and to my wonderful friends in Saarbrücken and around the world. I could not have completed this thesis or the studies leading up to it without their loving support.

Contents

List of Figures

List of Tables

1 Introduction

For students with French as their first language (L1) who are learning German as a second language (L2), the sound system of the L2 can pose a variety of difficulties, one of the most important and interesting of which is the way in which certain syllables in German words are accentuated more than others, a phenomenon referred to as lexical stress. Learning to navigate German lexical stress is especially challenging for L1 French speakers, because this phenomenon is realized very differently (perhaps not at all) 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 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 made by these learners can be reliably detected automatically, and on the other, an exploration of the types of multimodal feedback that can be automatically delivered based on this error detection. The outcome of these investigations and primary contribution of this thesis project is a prototype CAPT tool called **de-stress**: the German (**de**) **S**ystem for **T**raining and **R**esearch on **E**rrors in **S**econd-language **S**peech. Implemented as a web application in the Java-based language Groovy[1] using the Grails web framework,[2] **[TODO de-stress]** is open-source and publicly available online.[3] Figure 1.1 provides a conceptual overview of the tool, which provides a variety of options for diagnosis of and feedback on lexical stress errors, and could potentially be a useful component of a fully-fledged intelligent CAPT system (a type of Intelligent Tutoring System, or ITS).

Via a simple web interface, the system presents a student with a German sentence, 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.

[1] `http://groovy-lang.org`
[2] `https://grails.org`
[3] `https://github.com/vakila/de-stress`

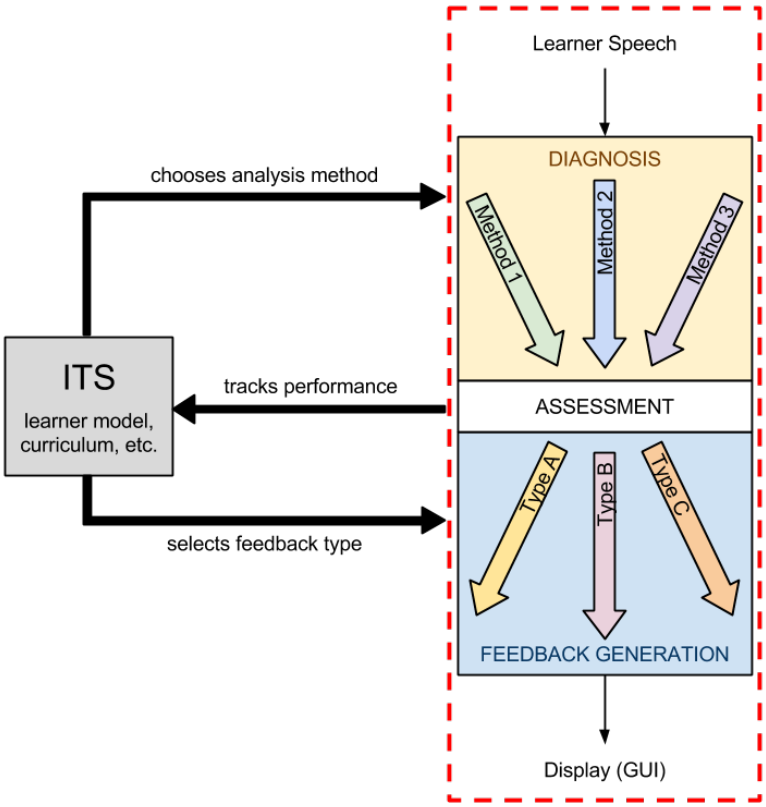


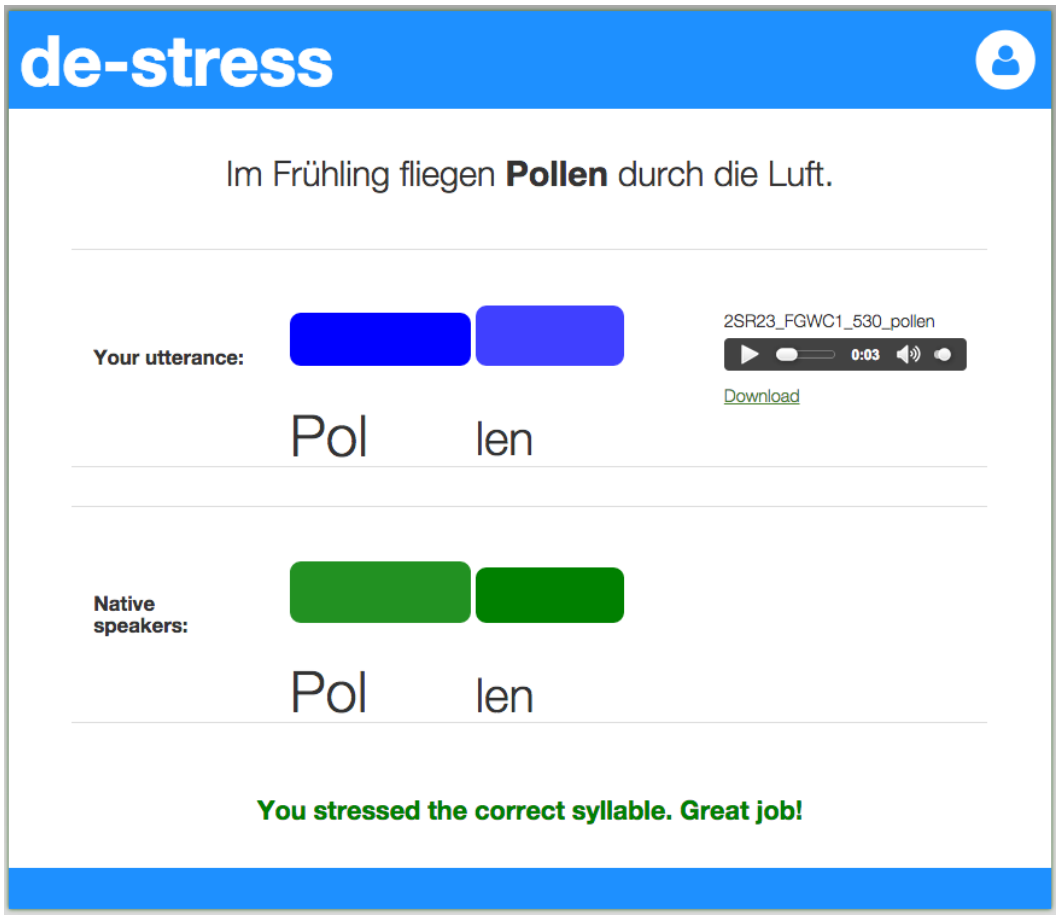
Figure 1.1: Conceptual diagram of the prototype lexical stress CAPT tool (demarcated by dashed line) and its possible function in the context of a more comprehensive Intelligent Tutoring System (ITS).

The student's utterance is subsequently analyzed for lexical stress errors using a variety of diagnostic approaches (see Chapter 4), and finally the student is presented with one or more types of feedback on their realization of lexical stress in the analyzed utterance (see Chapter 5). Figure 1.2 presents a screenshot of the interface presenting such feedback.

In addition to this student-facing interface, an administrative interface allows a language teacher or a researcher of L2 language acquisition to 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, **[TODO de-stress]** 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. The administrative interface for exercise creation can be seen in fig. 1.3.

Both instructional and research applications have thus motivated the development of **[TODO de-stress]**. 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

Figure 1.2: The student-facing interface of [TODO de-stress]



to which learners in which situations, this tool could become a useful component of a fully-fledged intelligent CAPT system, in which models of relevant aspects of the learning context (e.g. the student's skill level, progress, or personal preferences; the current learning objective or position in a sequence of exercises; etc.) are used to automatically decide which modules of the tool to activate, as fig. 1.1 illustrates.

1.2 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)[4] The project brings together researchers at Saarland University (Saarbrücken, Germany) and LORIA (Nancy, France), and is supported by the Deutsche Forschungsgemeinschaft (DFG) and the Agence Nationale de la Recherche (ANR).

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

[4] `http://www.ifcasl.org`

Figure 1.3: The interface of [TODO de-stress] for teachers/researchers

project objectives has been the compilation of a bidirectional learner speech corpus for the French-German language pair. Comprising phonetically diverse utterances in French and German spoken by both native speakers and nonnative speakers with the other language as L1 (Fauth et al., 2014; Trouvain et al., 2013), this is the first known corpus of L2 speech in both directions of this language pair.

As described by Fauth et al. (2014) and Trouvain et al. (2013), the IFCASL corpus contains 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 approximately half of the L2 stimuli, subjects were allowed to listen to a recording of the text as uttered by a native speaker before recording their utterance. An even gender distribution was maintained among the speakers recorded, and both children (adolescents under 18 years of age) and adults participated in the recordings, the majority of participants being adults. While children were uniformly of beginner proficiency in their L2, adults of both beginner and advanced levels in the L2 were recorded. Additional details about the corpus and its participants are given in Section 3.1.

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. (While the IFCASL project as a whole is also concerned with L1 German speakers learning French as L2, and the collected recordings are evenly distributed between the two languages, this thesis focuses exclusively on French L1 speakers learning German as L2.) As the IFCASL corpus had not been annotated for lexical stress errors, one major contribution of this thesis project

is the collection of error annotations for a subset of the German-language utterances by L1 French speakers; the details of this annotation effort are described in Chapter 3.

Furthermore, **[TODO de-stress]** has been designed with a view to contributing to the overall set of software developed in the context of the IFCASL project, such that it is as compatible as possible with the other tools developed and used by the IFCASL team, especially the **[TODO JSnoori]** speech processing software developed at LORIA (see Section 2.2.2).

1.3 Thesis overview

Chapter 2: Background and related work introduces Computer-Assisted Pronunciation Training (CAPT) and describes 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: Lexical stress errors by French learners of German describes the annotation and analysis of lexical stress errors in nonnative German speech produced by native French speakers, which constitutes an important contribution of this thesis. The annotation process and the resultant findings with respect to the distribution of lexical stress errors in L2 German speech are presented in detail in this chapter.

Chapter 4: Diagnosis of lexical stress errors details how **[TODO de-stress]** 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. The novel diagnostic approaches described in this chapter comprise another major contribution of this work.

Chapter 5: Feedback on lexical stress errors describes the final contribution of this project: the feedback module of **[TODO de-stress]**, which is capable of delivering a variety of feedback types simultaneously or in isolation. This chapter explains the types of feedback available, and how these are generated from the diagnosis of the learner's speech described in Chapter 4.

Chapter 6: Conclusion and outlook summarizes the ways in which this thesis project has contributed to the advancement of CAPT for French learners of German, and outlines some possible directions for future extensions of this work.

2 Background and related work

As stated in the previous chapter, the work reported in this thesis aims to make progress towards the development of a Computer Assisted Pronunciation Training (CAPT) system for learners of German as a foreign language (L2), with a particular focus on learners whose native language (L1) is French. This work therefore draws from and builds upon research in diverse but related fields, including L2 education, language technology, and phonetics/phonology.

This chapter sets the stage for the original work described in the remainder of the thesis by summarizing some relevant previous research on pronunciation in foreign language teaching and CAPT, describing the phenomenon of lexical stress as it relates to the comparative word prosody of German and French, and motivating this work's focus on lexical stress errors as a type of error well-suited to correction via CAPT.

2.1 Pronunciation in foreign language education

In the foreign language classroom, and the German language classroom in particular, less focus has traditionally been placed on pronunciation than other aspects of language education, such as grammar and vocabulary (Hirschfeld and Trouvain, 2007). However, even when pronunciation is taught in the classroom, a number of factors may limit the effectiveness of that training. 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 (Derwing and Munro, 2005; Hirschfeld and Trouvain, 2007). 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 (Neri et al., 2002). 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 (ibid.).

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 (Derwing and Munro, 2005; Field, 2005; Munro and Derwing, 1999; Neri et al., 2002; Witt, 2012). As Field (2005) and others point out, the exact definition of the term *intelligibility* is a topic of debate in the literature, as is the question of whether and how it differs from the notion of comprehensibility; here, let us follow Munro and Derwing (1999, p. 289) in

understanding intelligibility broadly as "the extent to which a speaker's message is actually understood by a listener."

Research on the impact of various types of pronunciation errors on intelligibility tends to indicate that errors on the prosodic (suprasegmental) level may hinder intelligibility more than segmental errors. In a study of nonnative English speakers from various native language (L1) backgrounds, Anderson-Hsieh et al. (1992) found that overall pronunciation ratings, which reflect intelligibility, correlated more strongly with measures of prosodic accuracy than with accuracy with respect to segmental sounds or syllable structure. Later research by Hahn (2004) points to the particular importance of primary sentence stress, e.g. the accentuation of new or contrasting words in a given sentence, to the intelligibility of L2 English speakers, especially with respect to the way in which these speakers (fail to) use intonation to convey stress. Though there exists relatively little research on the impact of various pronunciation errors on intelligibility in German specifically, some studies suggest that prosodic errors, and lexical stress errors especially, may hinder intelligibility of L2 speech more than other types of errors (Hirschfeld and Trouvain, 2007; Hirschfeld, 1994).

To reduce these and other types of errors in learner speech, listening exercises designed to help learners perceive phonological phenomena in the target language have been found to be valuable, with some work suggesting that perception training can have a positive impact on the intelligibility of L2 speech, even in the absence of exercises or feedback concerned with production (Derwing and Munro, 2005; Hirschfeld and Trouvain, 2007). However, other research suggests that production-oriented training with corrective feedback on pronunciation errors produces bigger performance gains. As Neri et al. (2002) point out, corrective feedback from human instructors has been shown to help adult L2 learners notice errors in their pronunciation more than implicit feedback in the form of exposure to speech from L1 speakers of the target language, enabling them to start working towards correcting these errors and thus improving their pronunciation. This supports the claims of Mehlhorn (2005) regarding the importance of individualized pronunciation coaching as a way to help L2 German learners become more cognizant of their pronunciation deviations as they take autonomous control of their own pronunciation learning. In a study of L2 German speakers from a variety of L1 backgrounds, Dlaska and Krekeler (2013) compared the effects of implicit and explicit feedback on the improvement in comprehensibility of the L2 learners' speech, and found that individualized, explicit, corrective feedback from a human language instructor led to more substantial improvements in comprehensibility than the implicit feedback of listening to the learner's own recorded utterance or that of a model native speaker; their findings seem to point to the difficulties learners may have in identifying their own errors, and confirm the importance of offering explicit feedback in L2 pronunciation instruction.

It is thus generally evident that individualized corrective feedback is quite important for L2 pronunciation learning, including L2 German learning specifically; however, there remains much to be learned about exactly when and how feedback can be most effective. This is the motivation behind the diversity of feedback types implemented in the **[TODO de-stress]** system (see Chapter 5); as described in the previous chapter, this system is intended to facilitate finer-grained research on the effects of such feedback on the acquisition of L2 German word prosody by L1 French speakers.

2.2 Computer-Assisted Pronunciation Training

In recent decades, the educational value of speech technologies has been well demonstrated (Eskenazi, 2009), with Computer-Assisted Pronunciation Training[1] (CAPT) emerging as one important educational application for foreign-language education (FLE) (Delmonte, 2011; Neri et al., 2002; Witt, 2012). 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 (Neri et al., 2002). 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 in Section 2.1. However, as Derwing and Munro (2005) point out, for CAPT to be effective, it must be informed by research on not only speech technology, but both speech production and perception on the one hand, and language acquisition and pedagogy on the other. With this in mind, an overarching aim of this thesis is to take initial steps toward the development of German CAPT technology which takes into account information from these diverse fields.

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 may 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 (Sections 2.2.2–2.2.5); comprehensive overviews and comparisons of these and many other systems are given by Delmonte (2011), Eskenazi (2009), Neri et al. (2002), and Witt (2012).

Additionally, both the diagnosis and feedback modules of the **[TODO de-stress]** CAPT tool developed in this thesis project 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 1.2). One relevant body of work at LORIA has 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; this work is summarized in Section 2.2.1. Furthermore, the group has developed the Snorri suite of speech analysis software, described in Section 2.2.2, the most current version of which, JSnoori, has provided the foundation for many elements of **[TODO de-stress]**.

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

2.2.1 Automatic processing of learner speech

An important prerequisite of any useful CAPT system is the capability to process learner speech automatically. This may be attempted via the application of general techniques for automatic speech recognition or automatic alignment of read speech, but given the drastic ways in which L2 speech typically differs from that of L1 speakers, additional measures must usually be taken to ensure that learners' utterances can be processed reliably and robustly. Over the past decade or more, researchers at LORIA have been advancing the state of the art for nonnative speech processing, particularly in the context of CAPT applications (e.g. Bonneau et al., 2012; Bouselmi et al., 2005, 2012; Jouvet et al., 2011; Mesbahi et al., 2011; Orosanu et al., 2012); this section provides an overview of that work.

One approach to improving recognition accuracy on L2 speech explored by these researchers involves modifying or adapting the L1 acoustic models used for recognition to better cope with L2 speech data. Bouselmi et al. (2005) found that merging acoustic models trained on native speech in the target language (English, in their evaluation) with models trained on speech in the learner's L1 (e.g. French) resulted in better recognition accuracy on L2 speech, and in later work (2012) achieved additional improvements in accuracy by modeling nonnative pronunciation errors (e.g. the insertion, deletion, or substitution of a phoneme) via the automatic extraction of non-standard pronunciations from a corpus of L2 speech. Whereas Bouselmi et al. (2012) explored the use of L2 pronunciation variants at the level of acoustic modeling, other research at LORIA has shown that augmenting the pronunciation lexicon with nonnative pronunciation variants can also be an effective way of improving the accuracy of L2 speech recognition (Bonneau et al., 2012; Jouvet et al., 2011; Mesbahi et al., 2011). Mesbahi et al. (2011) evaluated the segmentation boundaries (i.e. starting and ending points of each phone in the utterance) automatically produced using a lexicon with nonnative pronunciation variants, and found that these were for the most part quite close to (within 15 milliseconds of) the boundaries placed manually by trained scientists, and concluded from this that such automatically-produced segmentations are reliable enough for use in CAPT.

Though these adaptation techniques thus reduce the number of errors in the automatic segmentation of L2 speech, some errors still remain as a result of more profound deviations between the expected utterance and that actually produced by the learner. For example, learners may insert or delete not only individual phones, but entire syllables; they may hesitate or repeat words or parts thereof; and they may pronounce words or phrases entirely different from the expected text. Bonneau et al. (2012) and Orosanu et al. (2012) explored methods for detecting errors like these by comparing two segmentations: one produced by forced alignment (see Section 4.1), in which the recognition task is constrained by the text of the expected utterance, and the other produced by unconstrained phone-level recognition (phone decoding). By detecting deviations between these two segmentations in a pre-processing step, a CAPT system would be able to catch aberrant utterances before automatic segmentation is attempted, so that the learner can be asked to record their utterance again.

2.2.2 The Snorri suite and JSnoori

Another contribution of the LORIA group vital to this thesis project is the Snorri suite of software, which includes the now-outdated original Snorri developed in 1987 (Fohr and Laprie, 1989) and a Unix adaptation developed shortly thereafter, a later Windows port of Snorri appropriately called WinSnorri (Laprie, 1999), and most recently a partial Java port of WinSnorri, JSnoori[3] (Project-Team PAROLE, 2013). JSnoori is the incarnation of the Snorri suite which is still under active development and is most relevant to this work.

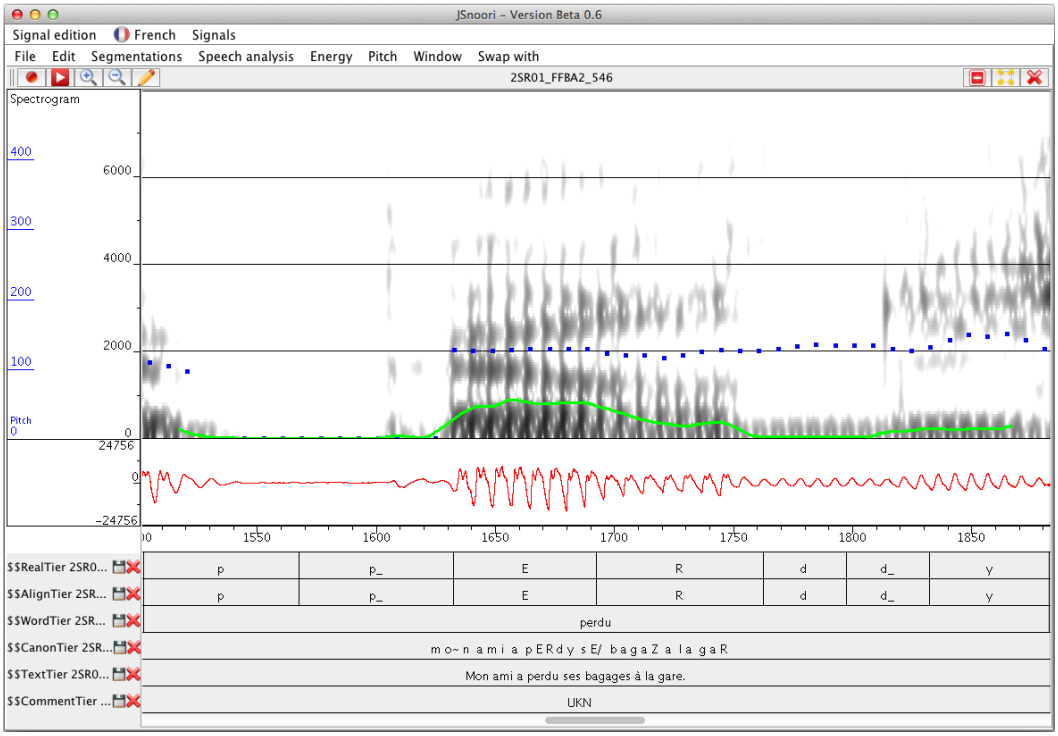
The screenshot in fig. 2.1 depicts JSnoori's interface for speech analysis. In this example, a recording of a native speaker utterance of the French sentence "Mon ami a perdu ses bagages à la gare" (*My friend lost his luggage at the train station*) is under analysis, with only the portion of the signal corresponding to the word "perdu" (*lost*) visible. At the bottom of the screen, in grey, the phone-, word- and sentence-level segmentations of the utterance are displayed. The oscillogram (waveform) of the selected portion of the speech signal is visible as a red contour above the segmentations. In the upper portion of the screen, a spectrogram of the selected signal segment is presented, with overlaid, colored contours corresponding to the fundamental frequency (blue dots) and intensity (green line) of the signal. Controls at the top of the window allow the user to listen to the selected recording and utilize the various speech analysis capabilities of the software.

Like its predecessors Snorri and WinSnorri, JSnoori offers users a wide array of signal processing and visualization features, as described by Fohr and Laprie (1989) and Laprie (1999). In addition to basic signal recording, editing, and playback, the software can also be used to segment and annotate speech recordings manually or automatically (see Section 2.2.1). Users can choose among various types of spectral visualizations, including wide- or narrow-band spectrograms and cepstrally smoothed spectrograms. Sophisticated pitch detection algorithms (see Section 4.2.2) enable the analysis and visualization of fundamental frequency (F0) in the signal. JSnoori also allows users to prosodically modify a given speech signal by altering the duration or pitch contour of that signal or a portion thereof; see Section 5.1.2 for a more detailed explanation of the techniques used to accomplish this modification.

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 and Colotte, 2011; Bonneau et al., 2004; Henry et al., 2007), who have used this software to assess lexical stress in L1 French speakers' pronunciation of English words, and deliver feedback on learners' errors. In its role as a CAPT tool, JSnoori (like its predecessor, WinSnoori) takes as input a learner utterance, a native reference utterance, and segmentations of each, performs an acoustic comparison of the two utterances, and delivers feedback in the form of e.g. annotated displays of the speech signal and spectrogram of each utterance. 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 (see Section 5.1.2 for details).

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

Figure 2.1: Screenshot of the speech-analysis interface of JSnoori.



A pilot experiment conducted by Bonneau and Colotte (2011) used WinSnoori to investigate the effect of these types of visual and auditory feedback on the prosodic realization of English lexical stress by L1 French learners, and seemed to indicate that this feedback has a beneficial effect for learners with a low proficiency level in the L2 (i.e. beginners), whereas more advanced learners seem to improve their pronunciation just as well when simply allowed to listen to a reference utterance recorded by a native speaker as a form of implicit feedback. However, as all of the feedback types available were presented simultaneously to learners in the experimental condition, further research is needed to determine whether there are differences in efficacy among the individual feedback types; enabling such research is the motivation behind the modular feedback component of **[TODO de-stress]** (see Chapter 5).

As described in Chapters 4 and 5, the JSnoori software is vital to this thesis project; **[TODO de-stress]** 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; Probst et al. (2002) found that selecting a "golden speaker" whose voice closely matched the learner's improved learning gains, a finding which informed the implementation of various reference speaker selection methods in the diagnosis module of **[TODO de-stress]** (see Section 4.3.3). Fluency also implements an error-catching step to reject utterances which do not match the expected text (Eskenazi et al., 2000), in the same vein as that of Mesbahi et al. (2011) and Orosanu et al. (2012). Eskenazi et al. (2007) report that Fluency's commercial spin-off, NativeAccent™, has been shown to help real-world users significantly improve their pronunciation skills.

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. As this section explains, ideas and techniques from the Reading Tutor have influenced both the diagnosis and feedback modules of the proposed CAPT tool.

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, a result which motivated **[TODO de-stress]**'s implementation of a classification-based alternative to the more typical comparison-based method of error diagnosis (see Section 4.4).

The Reading Tutor's 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). Inspired by this work, a similar method of feedback via text stylization has been implemented in **[TODO de-stress]** (see Section 5.1.1).

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. In contrast, relatively little work has been done on prosody-oriented CAPT in German. However, work by Jilka and Möhler (1998) on L2 German speech produced by L1 English speakers also suggests that manipulating the intonation contours of learner utterances may be an effective way to provide corrective

feedback on intonational errors contributing to a perceived foreign accent in German, and Bissiri et al. (2009; 2006) 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. Based on these findings, feedback via resynthesis may be a very useful element of a German CAPT system, and such feedback is therefore one of the types implemented in **[TODO de-stress]** (see Section 5.1.2).

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. Unfortunately, as Neri et al. (2002) point out, the necessity of having an instructor help the learner interpret the visual and auditory feedback provided by this software negates one primary advantage of CAPT: the ability to offer learners feedback in situations where individualized attention from a human instructor is limited, which is often the case in foreign language classrooms, as explained in Section 2.1.

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 seem to be "deaf" to lexical stress, insofar as they have significant and lasting difficulty perceiving lexical stress in Spanish, another language in which stress serves a contrastive function. 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. Indeed, with regard to German specifically, Hirschfeld and Trouvain (2007) report that lexical stress errors are commonly observed among L2 speakers with French as L1.

In short, based on prior work on French learners of variable-stress languages, it can reasonably be expected that L1 French learners of German as L2 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 influence 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.

The selection of an error type to address is guided here by a set of three criteria that such an error must meet; similar criteria are proposed by Cucchiarini et al. (2009) and Neri et al. (2002). First, 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. Second, 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. 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. 2.2, the best error to target with the CAPT system will fulfill all three criteria, rather than only one or two. 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 /aː/ as [a]) may have a greater impact on intelligibility in some cases, but may be more difficult to reliably identify automatically.

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 presents the reasoning behind 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 the perceived intelligibility of L2 speakers than segmental errors (Derwing and Munro, 2005; Hahn, 2004; Witt, 2012), and several studies have found lexical stress errors in particular to have an impact on L2 intelligibility in free-stress languages like

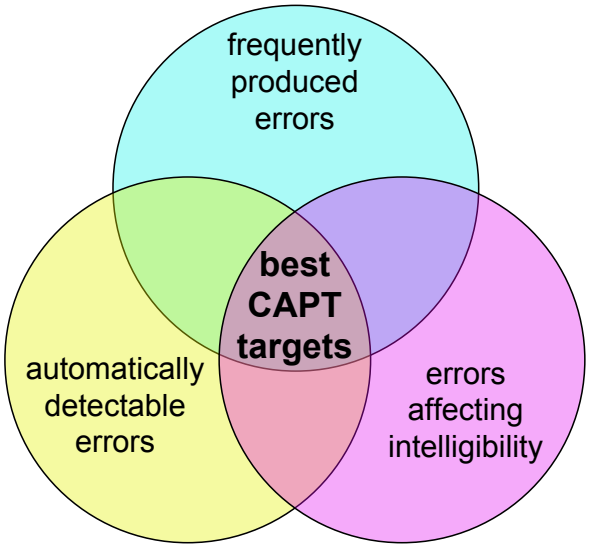


Figure 2.2: Criteria for selecting errors to target in a CAPT system.

English, Dutch, and German (Cutler, 2005; Field, 2005; Hirschfeld, 1994). 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 (Hahn, 2004; 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. Bonneau and Colotte, 2011; Wik et al., 2009).

Furthermore, in recent years some researchers have reported success in applying machine learning methods to the classification of lexical stress patterns in English words. Kim and Beutnagel (2011) experimented with various classifiers to identify stress patterns in 3- and 4-syllable English words, and reported accuracy in the 80-90% range for high-quality recordings of L1 English speech; in pilot experiments with low-quality recordings, however, the authors obtained lower accuracy: 70-80% on L1 speech and 50-60% on utterances by L2 speakers. Shahin et al. (2012) trained Neural Networks to classify stress patterns in bisyllabic words uttered by L1 English children, with the intended application of treating childhood dysprosody, and reported classification accuracy over 90% for some stress patterns.

As lexical stress errors thus fulfill the aforementioned criteria for targeting with CAPT, such errors are the focus of the proposed CAPT system. 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

As this chapter has shown, CAPT is an important application of speech technology for language learning, as it can help overcome typical difficulties with regard to the teaching of L2 pronunciation, and especially prosody, in the German language classroom (Section 2.1). The previous research on CAPT and related technologies summarized in Section 2.2 has demonstrated numerous ways in which prosodic errors in L2 speech can be automatically detected, setting the stage for the exploration of diagnostic methods presented in Chapter 4. Similarly, the diverse feedback methods explored in the work described in this chapter have provided the inspiration and technological foundations underlying the modular feedback functionality of **[TODO de-stress]**, which is the subject of Chapter 5.

Furthermore, this chapter has introduced the phenomenon of lexical stress in the context of L2 German acquisition by L1 French speakers (Section 2.3), and has motivated the selection of this phenomenon as the focus of this thesis project (Section 2.4). This discussion of lexical stress provides the context for the original work presented in the following chapter, which investigates the prosodic realization of lexical stress by French learners of German in more detail.

3 Lexical stress errors by French learners of German

As the previous chapter has shown, lexical stress is a challenging phenomenon for native (L1) French speakers to realize correctly in their nonnative (L2) German prosody, such that pronunciation errors with respect to lexical stress are expected to be produced frequently by this group of German learners. However, empirical studies of this type of error as produced by French learners of German are quite few in number (see Sections 2.3.3 and 2.4.2), so one objective of this thesis project was to research to what extent the expected types of lexical stress errors by French speakers of German are actually produced. As the IFCASL corpus (see Section 1.2) provides valuable data on L2 German speech produced by L1 French speakers, it is a perfect starting point for such investigations; however, given that the existing corpus annotation does not include information on lexical stress errors, a subset of this corpus had to be manually annotated for such errors; the annotation and subsequent analysis of lexical errors in this data, presented in this chapter, constitute a major contribution of this thesis project.

The first sections of this chapter describe the selection of material to be annotated (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).

Given the error judgments thus collected, different annotators' judgments of the same utterances were compared to determine the reliability of the annotation, i.e. the agreement between annotators in terms of the labels they assigned to each utterance. 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 nonnative German speakers identify errors? (Section 3.4.2)
- Are there differences in how annotators with different levels of expertise (annotation experience or 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 single "gold-standard" label for each utterance had to be selected; Section 3.5 describes how this was accomplished in cases of disagreement.

Table 3.1: Number of speakers in the portion of the IFCASL-FG corpus annotated for lexical stress, in terms of speakers' age, gender, and proficiency level (Fauth et al., 2014)

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

Finally, given the gold-standard labels for each utterance, the distribution of lexical stress errors in the annotated data was analyzed; the following questions guided this analysis, which is detailed in Section 3.6.

- Are lexical stress errors observed frequently in the IFCASL data? (Section 3.6.1)
- Are lexical stress errors observed more frequently with certain word types than with others? (Section 3.6.2)
- 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 3.6.3 and 3.6.4)

In addition to enabling this error distribution analysis, the annotated data compiled as described in this chapter also enables the use of supervised machine learning methods for the automatic detection of lexical stress errors, which is the subject of Section 4.4).

3.1 Data

The IFCASL corpus (Fauth et al., 2014; Trouvain et al., 2013), introduced in Section 1.2, contains recordings of native and nonnative speech in French and German, and is thus a invaluable resource for research on pronunciation errors in this language pair, with the subset of the corpus containing L2 German speech by L1 French speakers (henceforth IFCASL-FG) being most relevant to the work reported in this thesis. As described in Section 1.2, speakers of varying ages, genders, and German proficiency levels are represented in the IFCASL-FG corpus; the exact number and characteristics of these speakers are presented in table 3.1.

The subset of IFCASL-FG selected for the lexical stress error annotation undertaken in this work, henceforth referred to simply as "the dataset," 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).

Though previous work on lexical stress errors in L2 speech has often dealt with words uttered in isolation (e.g. Bonneau and Colotte, 2011), this work deals with word utterances

extracted from longer utterances of complete sentences; as Neri et al. (2002, p. 6) point out, for relevance to real communication it is important that CAPT systems address connected speech. The word's position in the carrier phrase/sentence was not taken into account; although it could be hypothesized that phrase position would have an effect on lexical stress realization by native French speakers, given the phenomenon of phrasal accent in French (see Section 2.3.2), Michaux and Caspers (2013) found no effect of phrase position on French speakers' realization of words in Dutch. Given the similarities in the lexical stress systems of Dutch and German with respect to French, no such phrase-level effect is therefore expected here, though investigation of the effect of phrase or sentence position on stress realization could be an interesting direction for future work.

In the IFCASL corpus recordings, sentences containing these words were read aloud by both L1 and L2 (L1 French) German speakers, as mentioned in Section 1.2. Here, only the L2 utterances were manually annotated; it is assumed that the L1 German speakers always realize lexical stress correctly, so the utterances of the selected word types in the native German subset of the IFCASL corpus (IFCASL-GG) were not manually annotated, but rather automatically labeled as correct realizations (see Section 4.4.1).

To compile the dataset, 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 nonnative speech, is a challenging problem under active research (see Section 2.2.1), 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 dataset described here.

3.2 Annotators

A total of 15 annotators participated in the annotation of this dataset, 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) are native German speakers, and three are nonnative (L2) speakers: two are native speakers of American English, and one is a native Hebrew

Table 3.2: The twelve bisyllabic initial-stress words types selected from the IFCASL corpus for lexical stress error annotation. Canonical pronunciations for each word type are given in IPA notation. The rightmost column lists the number of tokens (utterances) of each word type included in the annotated dataset.

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

speaker. The L2 German speakers all have some knowledge of German as L2, though the exact German proficiency levels of these annotators are unknown.

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 little, if any, 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.

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 each annotator, along with the number of tokens labeled for each type. Some judgments by annotators 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.

Word types were assigned to ensure that each 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.

Table 3.3: Annotators participating in the lexical stress error annotation. The anonymous identifier (ID), native language (L1) and expertise level of each annotator are presented, along with the word types each was asked to annotate and the number of usable token annotations by that participant for each word.

ID	L1	Expertise	Word types annotated (nb. of usable tokens)
A	German	expert	Flagge (55), Ringen (55), Tschechen (56)
H	German	expert	fliegen (56), Frühling (56), Pollen (55)
B	German	intermediate	halten (56), Mörder (56), Tatort (56)
D	German	intermediate	Flagge (49), Pollen (53), Ringen (49)
F	German	intermediate	fliegen (56), Frühling (56), Tatort (56)
I	German	intermediate	halten (56), Ringen (55), Tschechen (56)
J	German	intermediate	Frühling (56), Mörder (56), Tatort (56)
M	German	intermediate	Frühling (56), Ringen (54), tragen (55)
O	German	intermediate	manche (56), Mörder (56), tragen (55)
C	German	novice	E-mail (56), halten (56), Pollen (55)
L	German	novice	E-mail (56), Flagge (54), Tatort (56)
N	German	novice	fliegen (56), manche (56), Tschechen (56)
G	Hebrew	intermediate	Flagge (20), fliegen (0), Pollen (0)
E	English (US)	intermediate	halten (56), Mörder (56), Tschechen (56)
K	English (US)	intermediate	E-mail (56), Flagge (55), fliegen (56), manche (56), Pollen (55), tragen (55)

Table 3.4: Number of annotators assigned to each word type, in terms of the native language and expertise groups described in this section. The rightmost column gives the total number of annotators assigned to each word type.

	Native	Nonnative	Expert	Intermediate	Novice	Total
E-mail	2	1	0	1	2	3
Flagge	3	2	1	3	1	5
fliegen	3	1	1	2	1	4
Frühling	4	0	1	3	0	4
halten	3	1	0	3	1	4
manche	2	1	0	2	1	3
Mörder	3	1	0	4	0	4
Pollen	3	1	1	2	1	4
Ringen	4	0	1	3	0	4
Tatort	4	0	0	3	1	4
tragen	2	1	0	3	0	3
Tschechen	3	1	1	2	1	4

3.3 Annotation method

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

- [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 K who participated in two sessions (separated by several days) and annotated a total of six word types.

The lexical stress error annotations collected in this manner for each token (utterance) in the dataset enable two important contributions of this thesis, described in the following sections. First, the multiple annotations for each token from different annotators permit an analysis of inter-annotator agreement with regard to the identification of lexical stress errors; Section 3.4 presents this analysis, along with statistics on the relative frequencies with which the five labels were selected by annotators from the different L1 and expertise groups described in Section 3.2. Secondly, and perhaps more importantly, the errors identified by these annotators in the data extracted from the IFCASL corpus enable an analysis of the frequency of lexical stress errors in the speech of L1 French learners of German as L2; this is the subject of Section 3.6.

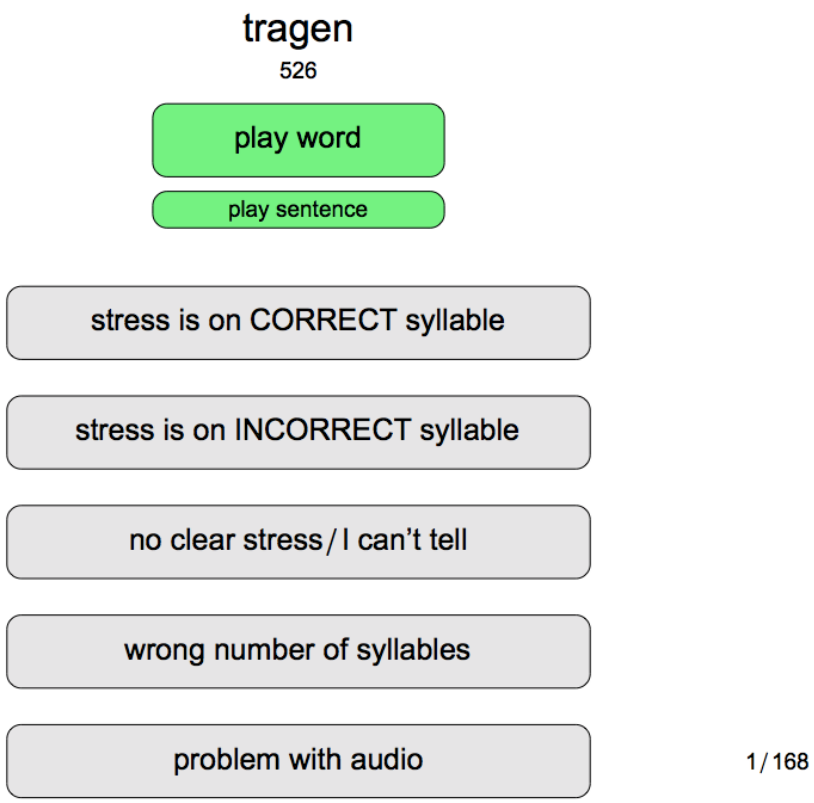


Figure 3.1: A screenshot of the graphical annotation tool scripted in Praat. Green buttons allow the annotator to listen to the word and sentence utterances. Gray buttons allow the annotator to record their judgment of stress accuracy; from top to bottom, the buttons correspond to the labels [correct],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

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. For 268 of these utterances, i.e. approximately 40% of the dataset, there was perfect agreement among annotators as to which of the five possible labels was most appropriate. This means that for the majority of the dataset (400 utterances, or approximately 60%), at least one annotator's judgment diverged from that of the other(s) who labeled the same utterance.

To make sense of these differences, agreement in label assignments was calculated for each pair of annotators who overlapped, i.e. labeled any of the same tokens. Two metrics were used to quantify agreement between a pair of annotators: the simple percentage of observed agreement, and Cohen's Kappa (κ) statistic; the following paragraphs describe how these are computed. In the analysis presented in Sections 3.4.1–3.4.3, both metrics are presented together in the hopes of providing a more comprehensive picture of inter-annotator agreement than either can convey alone.

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. 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 statistic (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 measure of inter-annotator agreement. A κ

Table 3.5: Overall pairwise agreement between annotators

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

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.

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, the mean and median percentage agreement are near 55%, indicating that, roughly speaking, annotators agree just slightly more often than they disagree. Turning to the κ values, given 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 highly reliable annotations, and signifies "fair" agreement between annotators according to the thresholds proposed by Landis and Koch (1977). Examination of the minimum and maximum values reveals that while some pairs of annotators seem to exhibit "substantial" agreement, indicating reasonably reliable judgments, other pairs have "poor" 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.

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. Figure 3.2 illustrates the pairwise agreements involving each of the 15 annotators. **[TODO table of percent/kappa min/avg/max by annotator, since graphs are difficult to read precisely?]** Evidently, there is noticeable variation from annotator to annotator, with some annotators (e.g. C, with a mean percentage agreement of **[TODO X]** and a mean κ of **[TODO Y]**) appearing generally more reliable than others (e.g. F, with mean %=**[TODO X]** and mean κ=**[TODO Y]**). **[TODO stat. significance?]** However, these figures should be interpreted with caution because they do not account for differences in the number of overlapping annotators/tokens available for each annotator. Nonetheless, the observed variability among

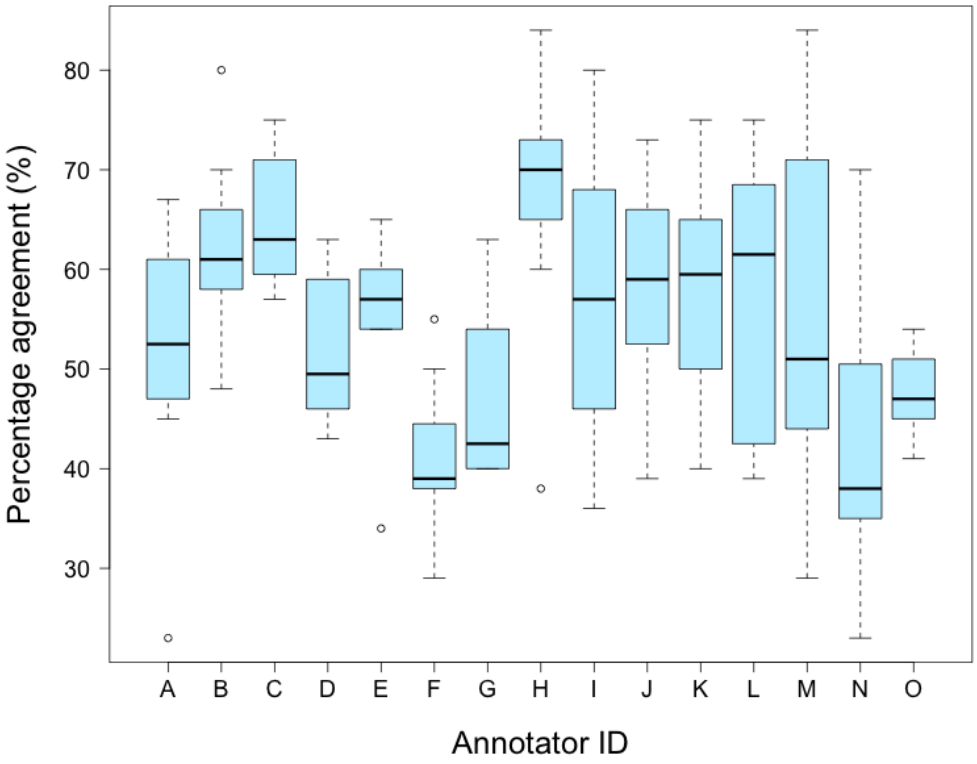
Table 3.6: [TODO caption]

ID	Pairs	Pairwise % Agreement			Pairwise κ		
		Min.	Avg.	Max.	Min.	Avg.	Max.
A	8	23.2%	51.4%	67.3%	-0.01	0.22	0.34
B	7	48.2%	62.4%	80.4%	0.17	0.33	0.61
C	7	57.1%	65.4%	75.0%	0.28	0.39	0.54
D	7	42.9%	52.0%	63.2%	0.22	0.29	0.47
E	6	33.9%	54.7%	65.2%	0.12	0.22	0.36
F	7	28.6%	41.2%	55.4%	0.04	0.13	0.25
G	4	40.0%	47.0%	63.2%	0.10	0.23	0.47
H	7	38.4%	66.9%	83.9%	0.13	0.29	0.47
I	7	35.7%	57.4%	80.4%	0.13	0.33	0.61
J	7	39.3%	58.3%	73.2%	0.11	0.20	0.32
K	10	40.0%	57.4%	75.0%	0.10	0.31	0.51
L	8	39.3%	57.4%	75.0%	0.08	0.30	0.54
M	8	28.6%	55.6%	83.9%	0.04	0.26	0.51
N	7	23.2%	43.0%	69.6%	-0.01	0.13	0.22
O	6	40.5%	47.4%	53.6%	0.13	0.20	0.28

annotators cannot be ignored, and Sections 3.4.2 and 3.4.3 attempt to investigate whether differences in annotators' L1 and level of expertise could potentially explain some of this variability.

It is also of interest to analyze the overall inter-annotator agreement for each word type in the dataset (see table 3.2 for the list of annotated word types). **[TODO table of min/med/max/mean by word?]** As fig. 3.3 illustrates, there are noticeable differences in agreement among word types. Annotators exhibit relatively high agreement ("moderate" according to Landis and Koch, 1977) on certain words, such as *E-mail*, *halten*, and *Pollen*), while for other words (e.g. *manche* and *Tschechen*) agreement values are closer to chance ("slight" agreement in Landis and Koch's schema). No clear explanation for these variations readily presents itself, but one reasonable hypothesis might be that annotators may agree more often for words which are particularly easy or difficult for learners, i.e. words for which learners tend to make few/many errors, as a result of a high imbalance between correct and erroneous utterances. However, as the error distribution analysis in Section 3.6.2 will show, this does not seem borne out by the data. The highest-agreement word types do not seem to have particularly imbalanced proportions of [correct] utterances; for *E-mail*, *halten*, and *Pollen* approximately 60-65% of utterances were [correct], which is close to the overall average of 63.77%. Furthermore, some of the word types with unusually high or low proportions of [correct] utterances (e.g. *Frühling*, with over 85%, or *Tatort*, with under 40%), seem to have agreement relatively close to the overall average. Therefore, further work will be needed to shed light on whether the agreement differences observed among different word types might simply be a consequence of individual variation among the annotators assigned to each type, or whether other phenomena (e.g. frequency effects, segmental pronunciation errors, or interference from French) might be at play.

(a) Pairwise percentage agreement by annotator



(b) Pairwise κ by annotator

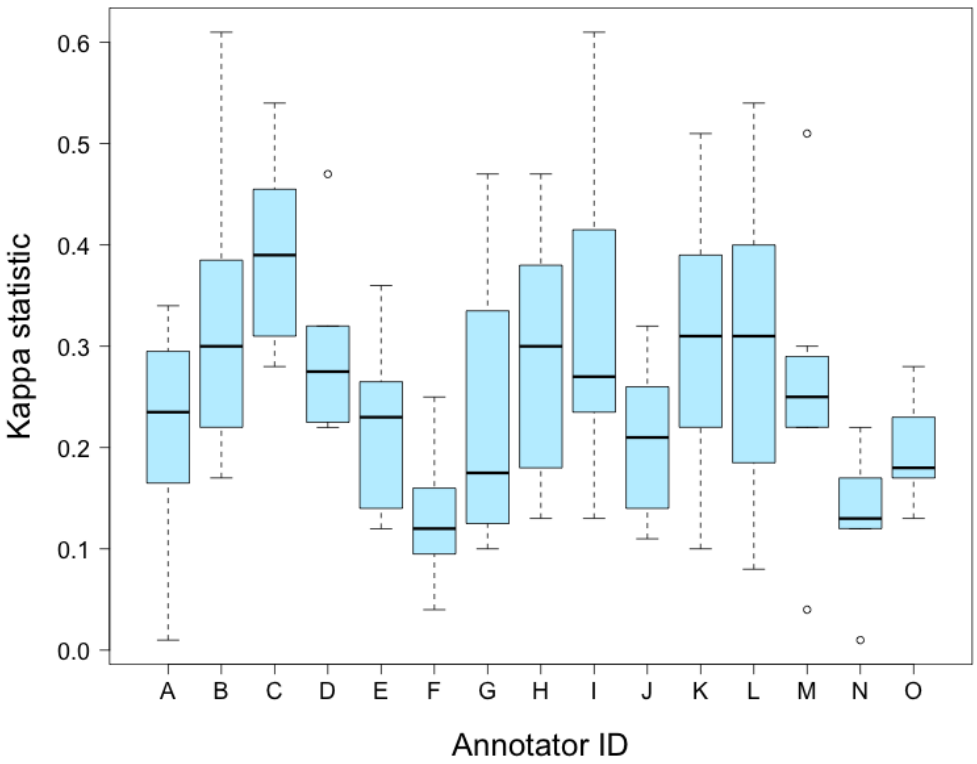
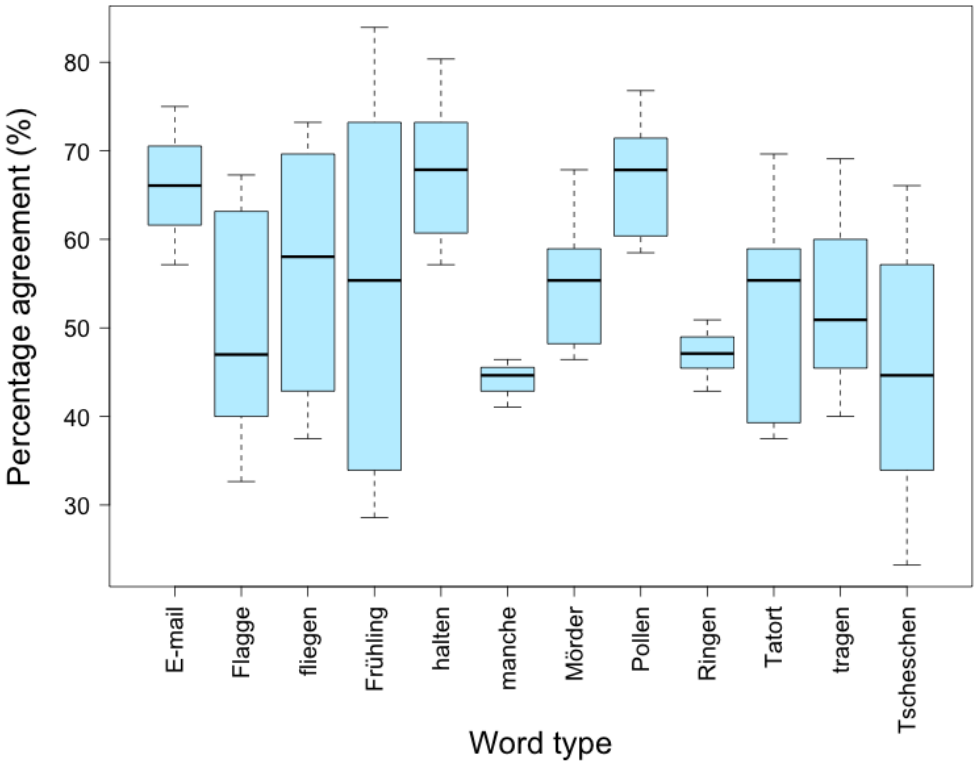


Figure 3.2: **[TODO Redo: smaller height, bigger labels, diff box styles to show L1/expertise of each annotator?]** Each annotator's pairwise agreement with all other annotators with whom they overlapped. In these box-and-whisker plots, black horizontal lines represent median values, and boxes extend from the first quartile to the third quartile, thus representing the Inter Quartile Range (IQR). Whiskers (dashed lines) extend to the minimum and maximum values within 1.58 IQR of the first and third quartiles, respectively, and roughly represent a 95% confidence interval around the median. Circles depict outlying values that fall outside of the IQR by a distance of more than approximately

(a) Pairwise percentage agreement by word type



(b) Pairwise κ by word type

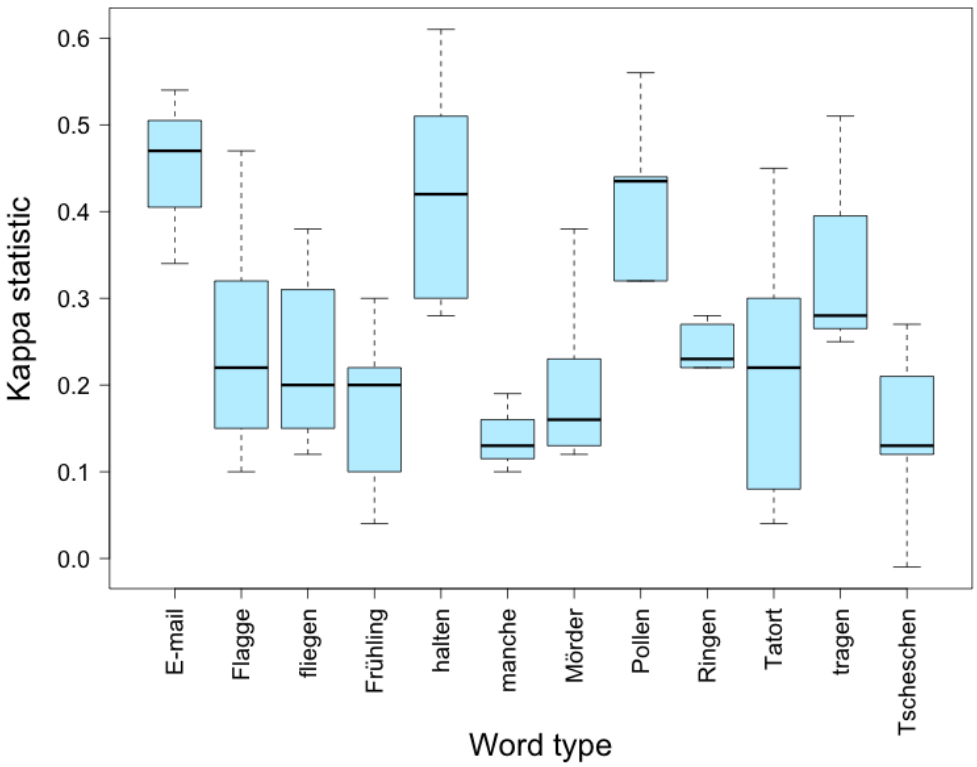


Figure 3.3: Pairwise agreement between annotators for each word type. See fig. 3.2 on p. 29 for an explanation of the box-and-whisker plots displayed here.

To put the agreement values reported in this section in context, it is worthwhile to compare them with those obtained in recent work by Michaux and Caspers (2013) on L2 Dutch speech by L1 French speakers. In this study, three expert annotators were asked to indicate which syllable they perceived as stressed in Dutch word utterances by the L1 French speakers. Two of the annotators were L1 Dutch speakers, the other an L1 French speaker who was nevertheless "highly proficient" in Dutch (p. 90). The authors report an average κ of 0.71 between annotators, representing "substantial" agreement in the schema of Landis and Koch (1977), and a much higher value than those observed in the present work. One plausible explanation for this difference may be the fact that while the three annotators participating in the Michaux and Caspers (2013) study were all "phonetically trained" (p. 90), i.e. experts (see Section 3.2), the annotation project described here collected judgments from a larger number of annotators, and only two of the 15 participating annotators had extensive phonetic training. The relationship between annotator expertise and inter-annotator agreement is explored further in Section 3.4.3.

The following sections present detailed analyses of the agreement between annotators with different native languages (Section 3.4.2) and different levels of expertise (Section 3.4.3). However, on the whole, it already seems evident that inter-annotator agreement in this lexical stress error annotation task is relatively low. This may simply signal low reliability among the particular annotators participating in this study, but it may also be a preliminary indication of the considerable difficulty of the task of diagnosing errors in L1 French speakers' realizations of lexical stress in German; this notion will be revisited in Section 4.4.

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 nonnative German speakers identify errors?

To answer this question, it is useful to look at the inter-annotator agreement between native and nonnative 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 nonnative speaker, as well as agreement between pairs in which both annotators were native speakers. Due to the small size of the nonnative group (3 annotators) and the aforementioned technical problems with annotator G's data (see Section 3.2), there was very little overlap between nonnative annotators (only one pairwise comparison), preventing meaningful analysis of agreement within the nonnative group. The precise mean, maximum, median, and minimum pairwise values for the two agreement metrics are listed in table 3.7, for both native-nonnative pairs and native-native pairs.

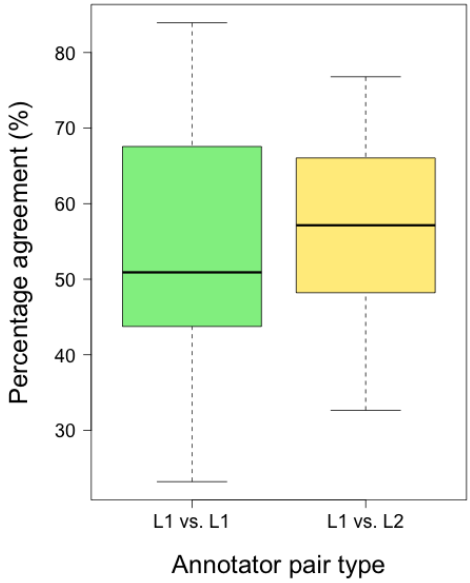
Looking at these statistics, we see little difference between the two types of pairs; in particular, the mean percentage agreement and κ values for native-nonnative and native-native pairs are quite close. **[TODO Stat significance?]** If anything, it would appear that agreement

Table 3.7: Inter-annotator agreement between native and nonnative 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

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 agreement statistics do not tell us much about difference between how the two groups of annotators judge lexical stress accuracy.

(a) Pairwise % agreement by L1 group



(b) Pairwise κ by L1 group

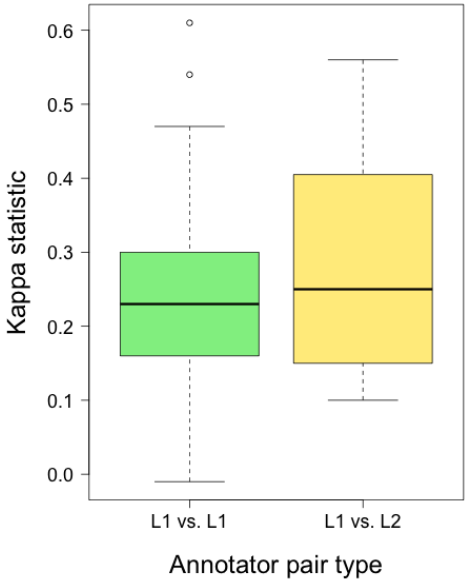
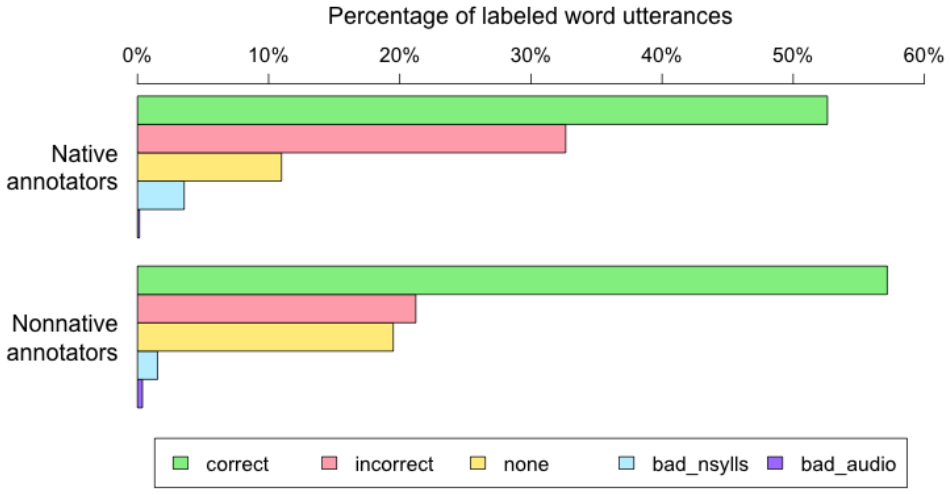


Figure 3.4: [TODO larger labels?] Pairwise agreement between annotators based on L1 group (L1 = native German speaker, L2 = nonnative speaker). See fig. 3.2 on p. 29 for an explanation of the box-and-whisker plots displayed here.

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, the native and nonnative speakers judged utterances as having correct lexical stress with approximately the same frequency: 52.7% of native annotators' judgments were [correct], vs. 57.3% for nonnative annotators. However, nonnative speakers seemed 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.

Figure 3.5: Distribution of labels assigned by native and nonnative annotators, as a percentage of the total number of utterances labeled by that annotator group **[TODO add exact values?]**



Though the differences between native and nonnative annotators are interesting from the perspective of L2 perception of lexical stress, 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 nonnative speech is of more relevance to this work than the way it is perceived by nonnative speakers. Therefore, the remainder of this chapter is concerned exclusively with the judgments of native annotators, which is to say that judgments by nonnative annotators are not included in the analyses that follow.

3.4.3 Expert vs. intermediate vs. novice annotators

This section brings us to the last of the questions raised at the beginning of the chapter concerning inter-annotator agreement in the stress-annotated data, namely:

Are there differences in how annotators with different levels of expertise 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 (nonnative) speech might have a positive impact on an annotator's ability to reliably judge the accuracy of lexical stress realizations by nonnative 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

Table 3.8: Pairwise agreement between annotators based on their level of expertise: expert (Exp), intermediate (Int), or novice (Nov).

	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

of nonnative lexical stress errors by non-experts not be ignored in favor of perception of such errors by experts.

Just as the previous section analyzed native vs. nonnative 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:

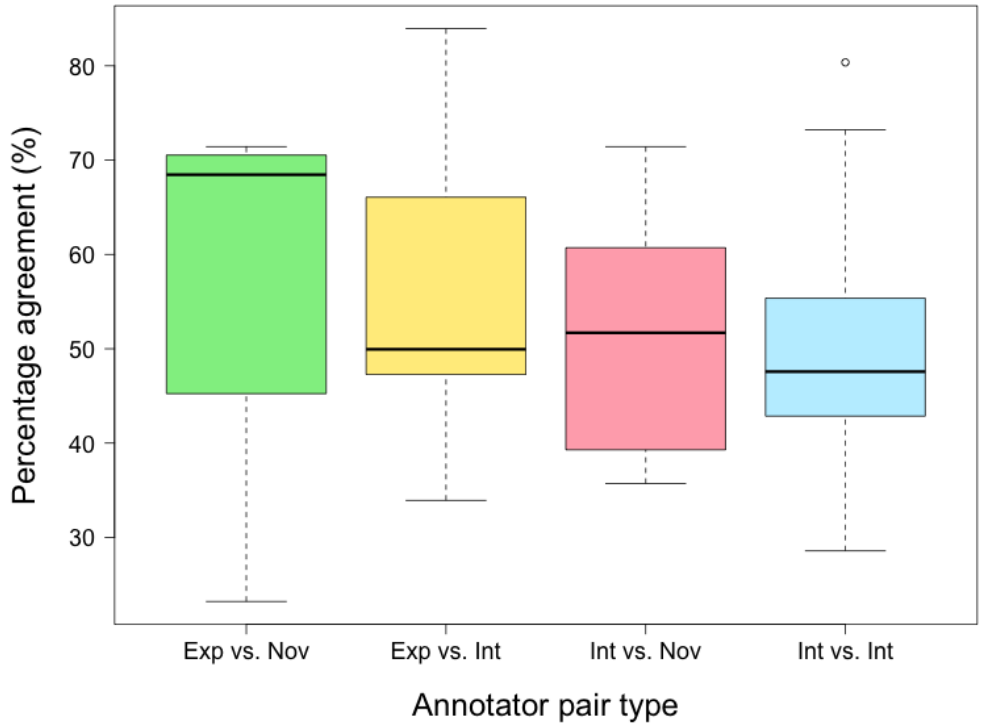
- Expert annotator vs. novice annotator (Exp vs. Nov)
- Expert annotator vs. intermediate annotator (Exp vs. Int)
- Novice annotator vs. intermediate annotator (Nov vs. Int)

Additionally, pairwise agreement was tallied for pairings between two intermediate annotators (Int vs. Int), 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, 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 3.8 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 **[TODO Stat significance?]**; 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.

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

(a) Pairwise % agreement by expertise group



(b) Pairwise κ by expertise group

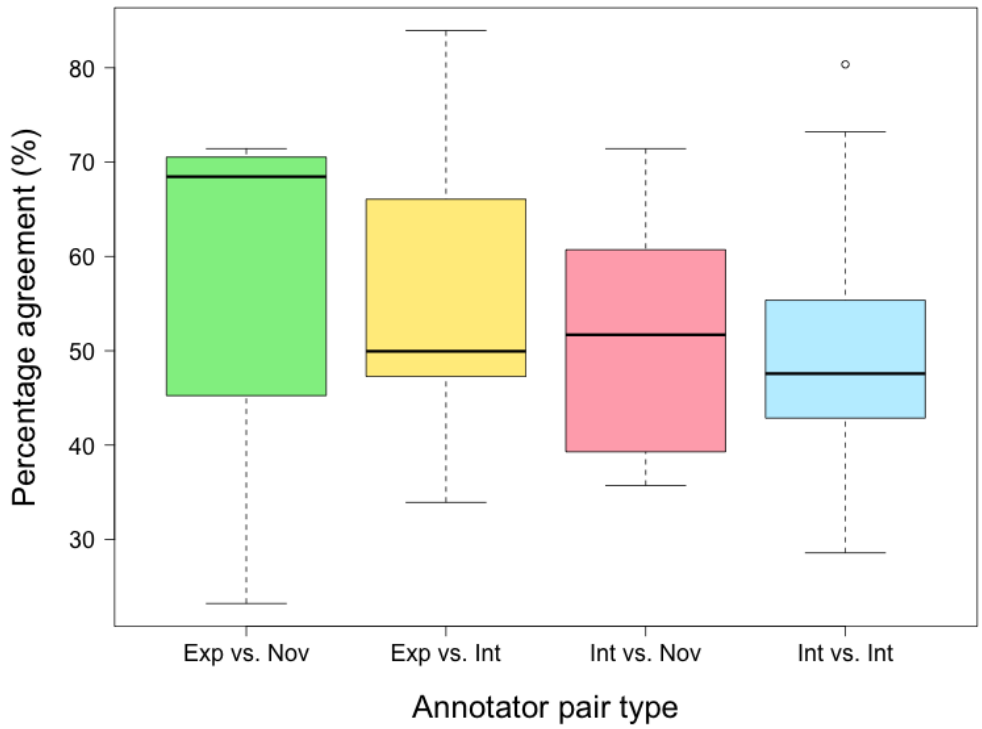
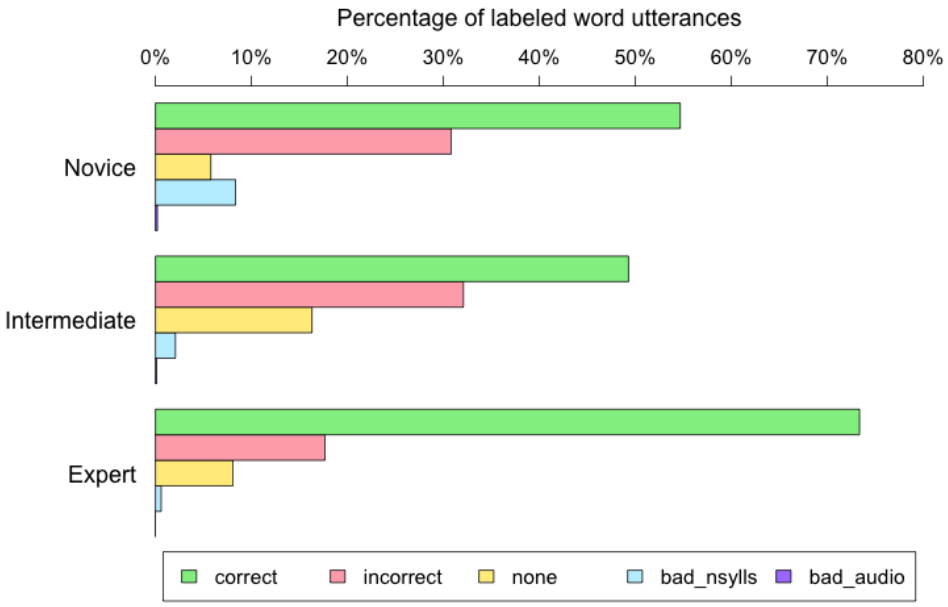


Figure 3.6: Pairwise agreement between annotators based on their level of expertise: expert (Exp), intermediate (Int), or novice (Nov). See fig. 3.2 on p. 29 for an explanation of the box-and-whisker plots displayed here.

Figure 3.7: Distribution of labels assigned by annotators of different expertise levels, as a percentage of the total number of utterances labeled by that annotator group



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. One could speculate that experts' familiarity with nonnative speech and knowledge of possible inter-speaker variations in lexical stress realization may be the cause for this willingness to accept a high proportion of utterances as correct.

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, it seems plausible that if experts' extensive experience with nonnative speech could be an explanation for their aforementioned "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.

3.5 Choosing gold-standard labels

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 obser-

vations about the difficulty of reliably assessing lexical stress accuracy, as well as insight into the differences in how judgments by annotators with different native languages and levels of expertise compare. However, if the annotations are to be used to analyze the distribution of errors in learner speech, or to train an automated error-diagnosis system (see Section 4.4), each token in the dataset must ultimately be assigned a single "gold-standard" label from the set of possible labels (see Section 3.3), henceforth referred to as $S = \{$[correct],[incorrect],[none],[bad_nyslls],[bad_audio]$\}$. In error analysis and classifier training, the gold-standard label will represent the ground truth as to whether a given utterance contains a lexical stress error.

In some cases, assigning a gold-standard label was trivial, while in others a decision had to be made between competing candidate labels. This section describes the procedure by which a single label was chosen for each word token (utterance) in the data set described in Section 3.1. In the remainder of this section, the gold-standard label chosen for a given word token t will be referred to as $s_{\text{gold}}(t)$.

To prepare for gold-standard labeling, all available annotations for t were tallied by their label type $s \in S$, resulting in a set $S_t \subseteq S$ of labels assigned to t by any of the native annotators who labeled this token (nonnative 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, making it possible to determine which label(s) received the highest number of votes and may thus be considered the best candidate(s) for $s_{\text{gold}}(t)$. The set of highest-voted candidate labels will be referred to as $S_{t\text{-max}} \subseteq S_t$.

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 dataset; this procedure is outlined in table 3.9. At each step i in the procedure, any tokens whose set S_t of observed labels fits condition C_i are assigned the gold-standard label described in column $s_{\text{gold}}(t)$; the number of tokens matching C_i is given as $N(C_i)$, and $N(C_{1\ldots i})$ represents the total number of tokens which have been assigned a gold-standard label at the end of step i in the labeling procedure (i.e. the number of tokens matching C_i or any previous condition).

As mentioned in Section 3.4.1, 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, making it easy to assign this label (s_a) as the gold standard for the utterance. Condition 1 (C_1) in table 3.9 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 (s_m) as $s_{\text{gold}}(t)$ is logical; these are captured by C_2.

Therefore, for a total of 533 tokens (approximately 80% of the word utterances in the dataset), the choice of $s_{\text{gold}}(t)$ was essentially uncontroversial. For the remaining utterances, however, choosing gold-standard labels was a less straightforward task, and the decisions made in steps 3-7 are somewhat more controversial.

Table 3.9: Procedure for choosing a gold-standard label $s_{\text{gold}}(t)$ for a given token t. At step i, tokens matching the condition C_i are assigned the label in column $s_{\text{gold}}(t)$. The rightmost columns $N(C_i)$ and $N(C_{1\ldots i})$ 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)	$s_{\text{gold}}(t)$	$N(C_i)$	$N(C_{1\ldots i})$
1.	$\lvert S_t\rvert = 1$, i.e. $S_t = \{s_a\}$	s_a	268	268
2.	$\lvert S_{t\text{-max}}\rvert = 1$, i.e. $S_{t\text{-max}} = \{s_m\}$	s_m	265	533
3.	$s_{\text{exp}} \in S_{t\text{-max}}$	s_{exp}	51	584
4.	$\lvert S_{t\text{-max}}\rvert = 2$, [bad_nsylls] $\in S_{t\text{-max}}$, i.e. $S_{t\text{-max}} = \{$[bad_nsylls]$, s_o\}$	s_o	17	601
5.	$\lvert S_{t\text{-max}}\rvert = 3$	[none]	6	607
6.	$S_{t\text{-max}} = \{$[none]$, s_{\text{cert}}\}$, $s_{\text{cert}} \in \{$[correct],[incorrect]$\}$	s_{cert}	21	628
7.	$S_{t\text{-max}} = \{$[correct], [incorrect]$\}$	[correct]	40	668

In a third step, if either of the two expert annotators had labeled one of the remaining tokens, the expert's judgment (s_{exp}) was taken as $s_{\text{gold}}(t)$; 51 tokens met this condition (C_3).

Next, in step 4, if there were exactly two labels in $S_{t\text{-max}}$ and one of them was [bad_nsylls], the other label (s_o) 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 $S_{t\text{-max}}$ 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 $S_{t\text{-max}}$ 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 (s_{cert}), based on the assumption that if at least one annotator was able to

categorically classify the given utterance as correctly or incorrectly realizing lexical stress, other native-speaking listeners might be inclined to make the same judgment.

The remaining 40 utterances were captured by the seventh and final condition, C_7, in which $S_{t\text{-max}}$ contained exactly two labels: [correct] and [incorrect]. In these cases, the learner's utterance was assessed generously and the [correct] label was chosen as $s_{\text{gold}}(t)$, to capture the fact that as mentioned in Section 3.4.1, assessing the accuracy of a lexical stress realization seems to be a somewhat difficult task, and if at least one of the native speakers who heard the given utterance were willing to accept its stress realization as correct, the learner should not be "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 Section 4.4.

3.6 Results

Given the gold-standard 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:

- Are lexical stress errors observed frequently in the IFCASL data? (Section 3.6.1)
- Are lexical stress errors observed more frequently with certain word types than with others? (Section 3.6.2)
- 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 3.6.3 and 3.6.4)

In the hope of providing tentative answers to these questions, this section describes and analyzes the distribution of errors in the dataset 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 3.5.

3.6.1 Overall frequency of lexical stress errors

The overall distribution of the lexical stress accuracy judgments observed in the annotated dataset is detailed in table 3.10 and illustrated in Section 3.6.1. 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

Table 3.10: Overall frequency of lexical stress errors in the annotated data

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%

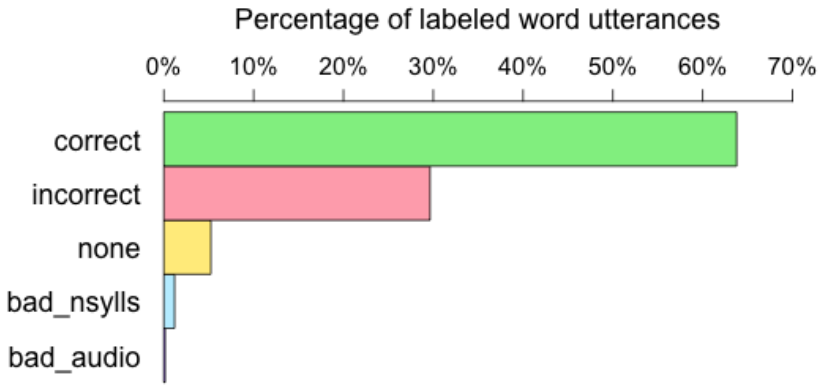


Figure 3.8: Overall distribution of lexical stress errors in the annotated data

either syllable (corresponding to the [none] label, 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.

3.6.2 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 table 3.11 and illustrated in fig. 3.9.

As should be expected, 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*, a much higher

Figure 3.9: Distribution of errors by word type, as a percentage of the total number of labeled tokens (utterances) of that word type (see table 4.9 for precise values)

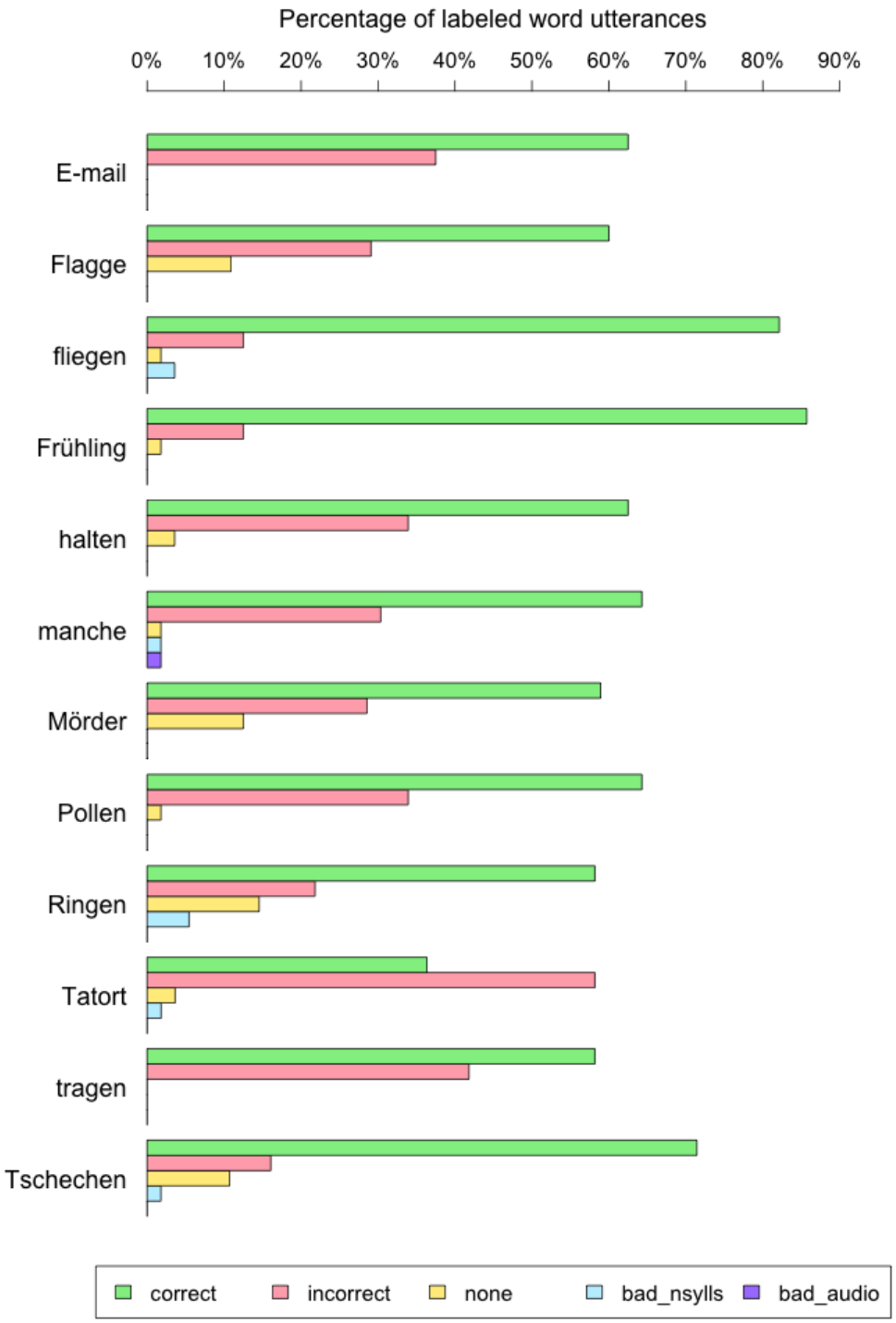


Table 3.11: Errors by word type: Number of tokens (utterances) assigned to each of the five categories, listed by word type. Equivalent percentages of each word type's total number of tokens are given in parentheses.

Word	correct	incorrect	none	bad_nsylls	bad_audio
E-mail	35 (62.5%)	21 (37.5%)	0	0	0
Flagge	33 (60.0%)	16 (29.1%)	6 (10.9%)	0	0
fliegen	46 (82.1%)	7 (12.5%)	1 (1.8%)	2 (3.6%)	0
Frühling	48 (85.7%)	7 (12.5%)	1 (1.8%)	0	0
halten	35 (62.5%)	19 (33.9%)	2 (3.6%)	0	0
manche	36 (64.3%)	17 (30.4%)	1 (1.8%)	1 (1.8%)	1 (1.79%)
Mörder	33 (58.9%)	16 (28.6%)	7 (12.5%)	0	0
Pollen	36 (64.3%)	19 (33.9%)	1 (1.8%)	0	0
Ringen	32 (58.2%)	12 (21.8%)	8 (14.6%)	3 (5.5%)	0
Tatort	20 (36.4%)	32 (58.2%)	2 (3.6%)	1 (1.8%)	0
tragen	32 (58.2%)	23 (41.8%)	0	0	0
Tschechen	40 (71.4%)	9 (16.1%)	6 (10.7%)	1 (1.8%)	0

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. With "x" and "-" indicating stressed and unstressed syllables, respectively, the sentence's rhythm can be represented as:

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

As a result of this regularity, correctly realizing the prosody of each word in this sentence may present less of a challenge to L1 French speakers than a less regular sentence, and may thus explain their uncharacteristically flawless productions of the words therein. The fact that no fewer than the expected proportion of errors were observed in utterance of *Pollen* would seem to contradict this speculative explanation; however, unlike the other words, *Pollen* is doubly challenging for L1 French speakers, insofar as its first vowel is a short ɔ, as opposed to the long o: in the word *Polen* (meaning *Poland* or *Poles* in English), with which *Pollen* forms a minimal pair.[1] Differentiating between long and short vowels when speaking German may be another pronunciation challenge for French speakers **[TODO (Zimmerer and Trouvain, 2015)]**, as vowel length does not serve a contrastive function in French (Peperkamp and

[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 dataset to be annotated for lexical stress errors.

Dupoux, 2002). 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. It also seems plausible that frequency effects could be at play (*Pollen* being ostensibly less frequent than *Frühling* or *fliegen*), or that there may be an influence from the word's position in the phrase/sentence (or interpreted position, in the case of L2 speakers inferring incorrect phrase boundaries).

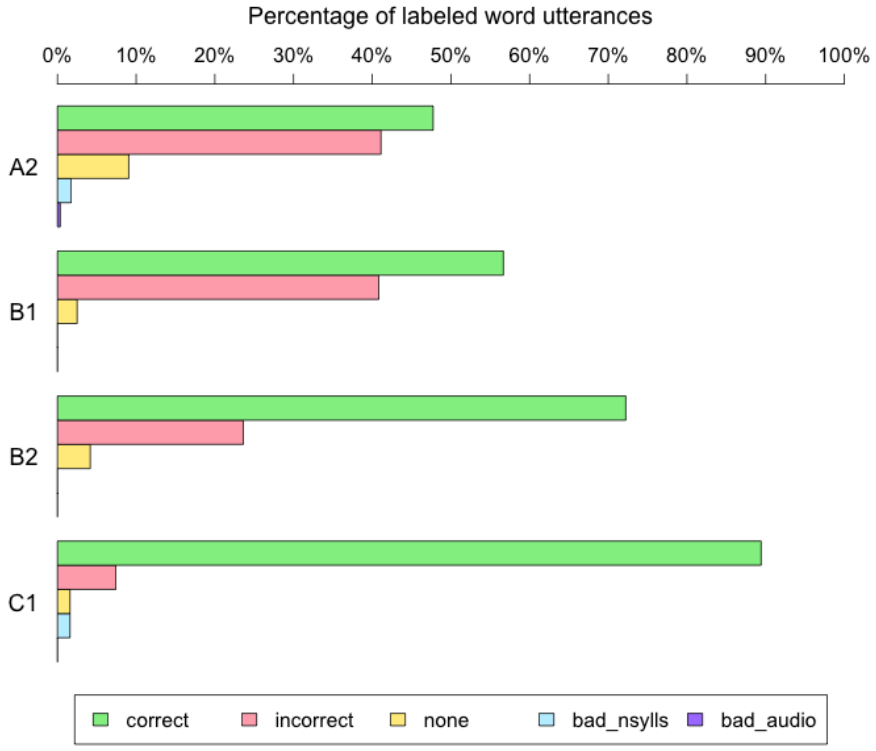
As for the uncharacteristically large proportion of errors in *Tatort* (*crime scene*), it may be the case that learners have difficulty identifying the boundary between syllables/morphemes in this word, which is a compound of the words *Tat* (*act*) and *Ort* (*place*). As *Tatort* is quite possibly an unfamilar word to the learners, especially those with lower German proficiency, perhaps 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 be more prominent. However, once again it should be noted that this purely speculative explanation is not (and cannot be) verified by the data collected here.

3.6.3 Errors by L2 proficiency level

As Section 3.1 stated, the L1 French speakers whose recordings comprise the dataset span four levels of L2 German proficiency: A2 (elementary), B1 (intermediate), B2 (upper intermediate), and C1 (advanced). The rightmost column of table 3.12 gives the number and proportion of utterances from speakers of each level in the dataset, 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 fig. 3.10a. Because the total number of utterances by speakers of each of the two intermediate (B) levels in the corpus is lower than the number by speakers of the lowest (A2) and highest (C1) levels, the judgments have also been grouped into two broader categories for easier comparison: beginners (A2 and B1) and advanced speakers (B2 and C1). The breakdown of stress errors by these groups is given in the lower portion of table 3.12 and illustrated in fig. 3.10b.

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

(a) Each level



(b) Level groups (Beginner is A2 + B1, Advanced is B2 + C1)

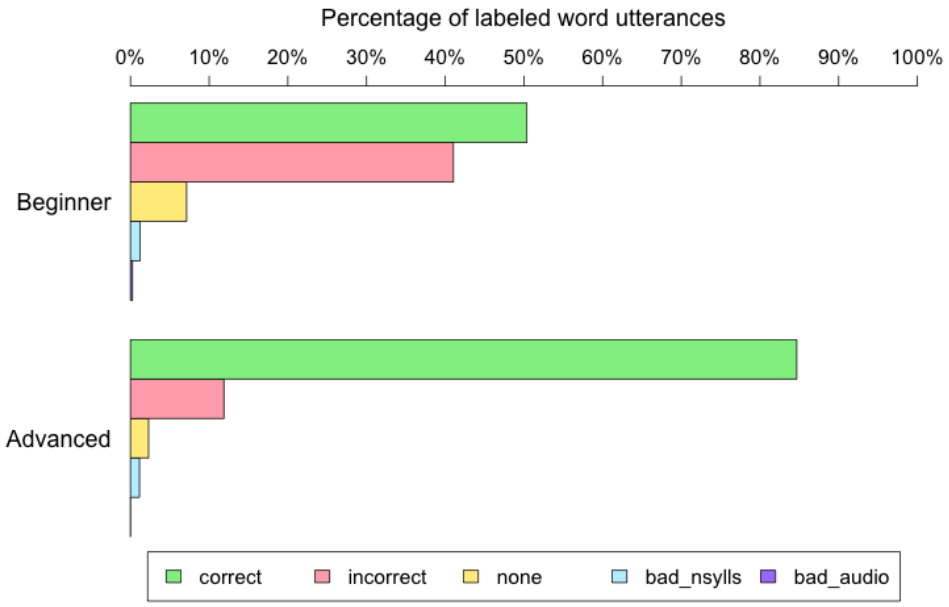


Figure 3.10: Distribution of errors by speaker's L2 German proficiency level, as a percentage of the total number of labeled tokens (utterances) from speakers of that level/group (see table 3.12 for precise values)

Table 3.12: Errors by L2 proficiency level: Number of tokens (utterances) assigned to each of the five categories, listed by the L2 German proficiency level of the speaker. Equivalent percentages of the total number of tokens for each level/level group are given in parentheses.

(a) Individual levels

	A2	B1	B2	C1
correct	137 (47.7%)	68 (56.7%)	52 (72.2%)	169 (89.4%)
incorrect	118 (41.1%)	49 (40.8%)	17 (23.6%)	14 (7.4%)
none	26 (9.1%)	3 (2.5%)	3 (4.2%)	3 (1.6%)
bad_nyslls	5 (1.7%)	0	0	3 (1.6%)
bad_audio	1 (0.4%)	0	0	0
Total (% of dataset)	287 (43.0%)	120 (18.0%)	72 (10.8%)	189 (28.3%)

(b) Level groups

	Beginner (A2+B1)	Advanced (B2+C1)
correct	205 (50.37%)	221 (84.67%)
incorrect	167 (41.03%)	31 (11.88%)
none	29 (7.13%)	6 (2.30%)
bad_nyslls	5 (1.23%)	3 (1.15%)
bad_audio	1 (0.25%)	0
Total (% of dataset)	407 (60.93%)	261 (39.07%)

will only be useful if that error is produced with considerable frequency by the learners using the system; therefore, it would seem from the frequency of lexical stress errors in their speech that learners of lower proficiency levels may benefit more from a CAPT system targeting such errors than learners of higher proficiency. This conforms with findings of Michaux (2012), which also seemed to indicate that, as might be expected, beginners make more lexical stress errors than advanced learners.

3.6.4 Errors by speaker age and gender

Given that the IFCASL corpus (Fauth et al., 2014; Trouvain et al., 2013), and by extension the dataset annotated for lexical stress errors, contains recordings from speakers of both genders and from adult speakers (18-30 years old) as well as children (adolescents ages 15-16) (see table 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 3.13 and illustrated in fig. 3.11.

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 groups (43 adults vs. 13 children): of the 668 tokens annotated in total, 513 (over three-fourths) were from adult speakers while only 155 (less than one-fourth) were utterances by children. Furthermore, it must be highlighted that there is a strong interaction between age and proficiency level: as seen in table 3.1, all of the child speakers represented in the data are

Table 3.13: Errors by speaker age and gender: Number of tokens (utterances) assigned to each of the five categories, listed by the age/gender of the speaker. Equivalent percentages of the total number of tokens for each age/gender group are given in parentheses.

(a) Individual age/gender categories. Boys and Girls are males and females under the age of 18, respectively, and Men and Women are males and females over 18, respectively.

	Boys	Girls	Men	Women
correct	48 (36.64%)	6 (25.00%)	184 (73.02%)	188 (72.03%)
incorrect	60 (45.80%)	17 (70.83%)	61 (24.21%)	60 (22.99%)
none	17 (12.98%)	1 (4.17%)	7 (2.78%)	10 (3.83%)
bad_nsylls	5 (3.82%)	0	0	3 (1.15%)
bad_audio	1 (0.76%)	0	0	0
Total (%) of corpus)	131 (19.61%)	24 (3.59%)	252 (37.72%)	261 (39.07%)

(b) By age, regardless of gender. Children are adolescents under age 18, adults are over 18. Adult beginners are adults with a German proficiency level of A2 or B1.

	Children	Adults	Adult beginners
correct	54 (34.84%)	372 (72.51%)	151 (59.92%)
incorrect	77 (49.68%)	121 (23.59%)	90 (35.71%)
none	18 (11.61%)	17 (3.31%)	11 (4.37%)
bad_nsylls	5 (3.23%)	3 (0.58%)	0
bad_audio	1 (0.65%)	0	0
Total (%) of corpus)	155 (23.20%)	513 (76.80%)	252 (37.72%)

(c) By gender, regardless of age.

	Males	Females
correct	232 (60.57%)	194 (68.07%)
incorrect	121 (31.59%)	77 (27.02%)
none	24 (6.27%)	11 (3.86%)
bad_nsylls	5 (1.31%)	3 (1.05%)
bad_audio	1 (0.26%)	0
Total (% of corpus)	383 (57.34%)	285 (42.66%)

(a) Individual age/gender groups

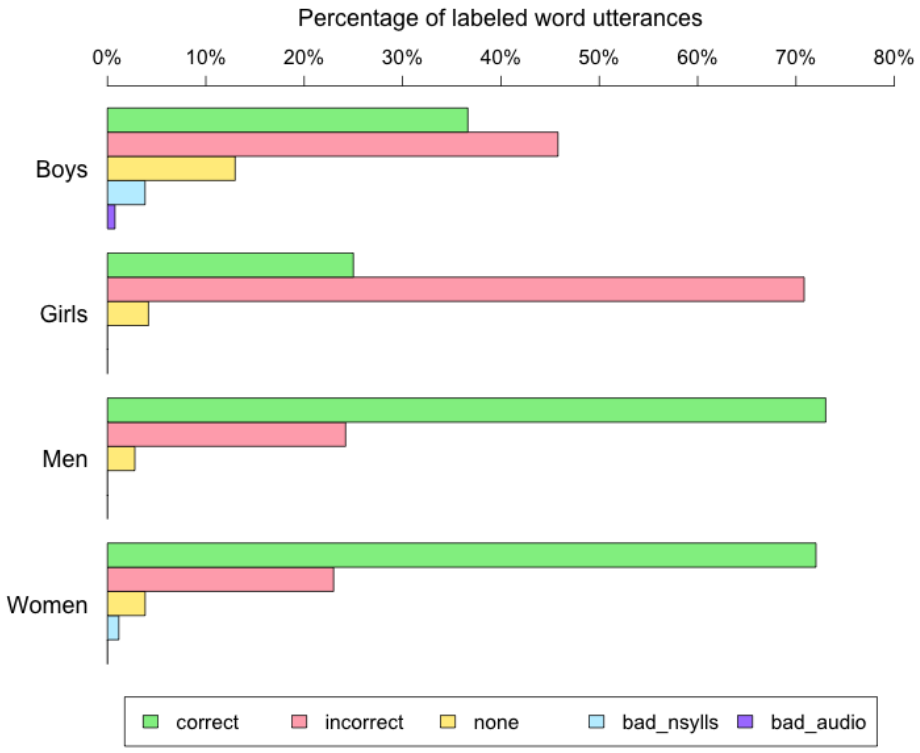
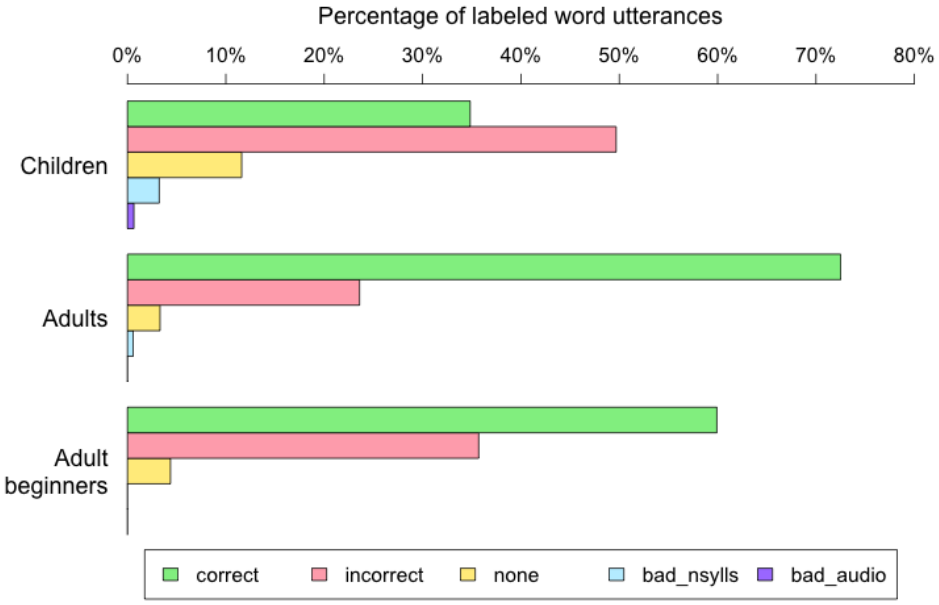


Figure 3.11: Distribution of errors by speaker's age and gender, as a percentage of the total number of labeled tokens (utterances) from speakers of that age/gender group (see table 3.13 for precise values)

(b) By age, regardless of gender



(c) By gender, regardless of age

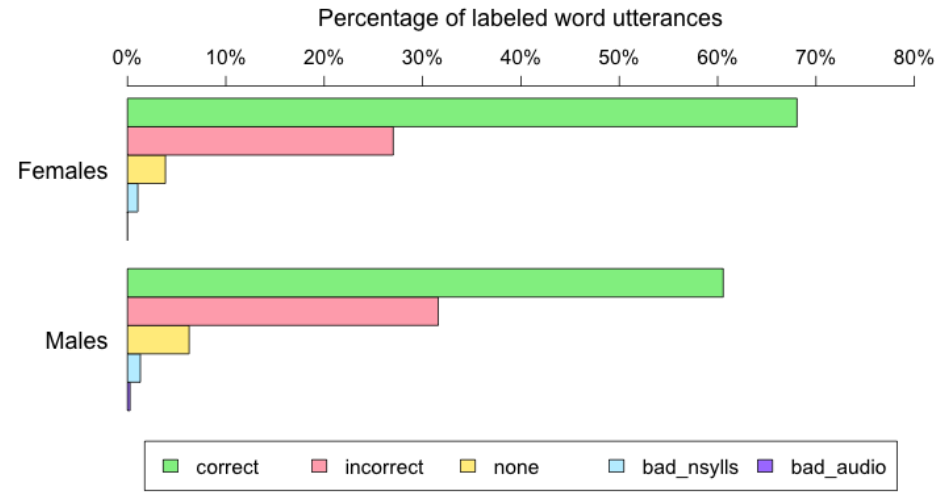


Figure 3.11: (continued) Distribution of errors by speaker's age and gender, as a percentage of the total number of labeled tokens (utterances) from speakers of that age/gender group (see table 3.13 for precise values)

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; instead of comparing the children's error distribution to that of all adults, it may be helpful to restrict the comparison to adults of the lower proficiency levels. Therefore, in addition to the results for all adults, table 3.13b and fig. 3.11b display results for the group of adult beginners (level A2 or 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 the results reported in the previous section seem to indicate that beginners may benefit more from a lexical stress CAPT system than advanced learners, it may also be the case that children in particular stand to gain more from such a system than adult beginners.

Coming now to the question of whether there is any difference in error distribution between speakers of different genders, a brief glance at table 3.13c and fig. 3.11c 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%). Therefore, it would seem that the error distribution observed in the annotated dataset provides no indication of a meaningful difference in the way speakers of different genders realize lexical stress in their L2 German.

3.7 Summary

In an effort to shed light on the nature of lexical stress errors in the speech of L1 French learners of German as L2, this chapter has described original efforts to annotate such errors in a small corpus of learner speech, study the frequency and type of errors made by learners, and analyze differences in how these errors are perceived by L1 and L2 German speakers with varying levels of phonetics/phonology expertise.

As described in Sections 3.1–3.3, lexical stress realizations in utterances of bisyllabic, initial-stress words (Section 3.1) were evaluated by multiple annotators from different L1 and phonetic training backgrounds (Section 3.2). Annotators were asked to use a graphical annotation tool to label each recorded word utterance as correctly or incorrectly realizing lexical stress (i.e. the speaker clearly stressed the correct or incorrect syllable), failing to

clearly realize stress (i.e. the speaker did not seem to stress either syllable), or having technical or other problems which prevented the assessment of lexical stress (Section 3.3).

Analysis of the labels assigned by different annotators to the same utterances (Section 3.4) revealed that inter-annotator agreement was relatively low, with only "fair" agreement (Landis and Koch, 1977), on average, between each pair of annotators who labeled the same utterances. Considerable variation was observed among individual annotators (Section 3.4.1), which did not seem to be explained by differences among annotators in terms of their L1 (Section 3.4.2) or level of experience with phonetics/phonology or speech data annotation (Section 3.4.3). **[TODO Stat significance?]** However, some other differences were observed in how different groups of annotators labeled lexical stress errors; specifically, L2 German speakers annotated a higher proportion of utterances as having unclear stress compared to L1 speakers (Section 3.4.2), and expert annotators judged substantially higher proportion of utterances as correctly realizing stress compared to intermediate or novice annotators (Section 3.4.3). As described in Section 3.4.1, there also seemed to be variability in inter-annotator agreement with respect to the different word types represented in the dataset, though further work is needed to discern the factors responsible for this observation (see Section 6.2).

The multiple, often conflicting, error annotations from different annotators were consolidated into a single gold-standard annotation for each utterance in the dataset (see Section 3.5), which served as the basis for an analysis of the frequency and type of errors produced by learners, as presented in Section 3.6. This analysis seemed to confirm the expectation set out in Sections 2.3.3 and 2.4.2 that lexical stress errors are quite frequent in the speech of French speakers of German, with correct lexical stress realizations observed in only approximately two-thirds of utterances, on average (Section 3.6.1). A finer-grained analysis revealed that the frequency of such errors appeared considerably lower in the speech of advanced learners than that of beginners (Section 3.6.3), and that children seemed to make more errors than adult beginners (Section 3.6.4); no substantial difference was observed between speakers of different genders. **[TODO stat significance?]** As in the case of inter-annotator agreement, considerable variation was observed in the frequency of errors in utterances of different word types (Section 3.6.2), though once again the factors underlying this variability are not immediately evident and should be investigated in future work (see Section 6.2).

The error annotation and analysis described in this chapter thus contribute considerably to our understanding of the difficulties L1 French speakers may have realizing lexical stress in their L2 German speech, and the fact that learners, especially beginners and children, seem to struggle with lexical stress production justifies the selection of lexical stress errors as the focus of this thesis project. Additionally, the analysis of inter-annotator agreement presented in this chapter, specifically the finding that the observed agreement was generally rather low, constitutes an important discovery with respect to the task of identifying such errors in learner speech: though further research is needed to determine why and under which conditions this is the case (see Section 6.2), it would seem that diagnosing lexical stress errors may be a challenging task for at least some L1 and L2 German speakers. If true, this has important implications for the development and evaluation of automatic error diagnosis systems, the subject of the following chapter, and these implications will be discussed further in Section 4.4.

4 Diagnosis of lexical stress errors

In order to provide learners with useful feedback on their lexical stress errors, a Computer-Assisted Pronunciation Training (CAPT) system 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 nonnative (L2) utterance;

(b) An analysis of how lexical stress is realized in the given utterance;

(c) A representation of how native (L1) 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 (Section 4.5).

4.1 Automatic segmentation of nonnative speech

Segmentation and 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 (p. 55) for examples of multi-level segmentations of word utterances from the IFCASL corpus (see Section 3.1).

A reasonably accurate segmentation of an L2 learner's utterance is indispensable for an analysis of the accuracy of their pronunciation, 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 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 (Fauth et al., 2014; Fohr and Mella, 2012; Fohr et al., 1996; Mesbahi et al., 2011), a constrained type of speech recognition in which the task is not to determine what was said (i.e. the best phone or word sequence corresponding to a recorded utterance), but rather to find the best alignment between the phone/word sequence, known in advance, and the corresponding segments of the given speech signal. This technique 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 this context. 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; one example is the lexicon of the MARY text-to-speech synthesis system (Schröder and Trouvain, 2003). To account for differences in how different speakers (especially nonnative 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 nonnative) that might be encountered. As stated in Section 2.2.1 researchers at LORIA have found the inclusion of nonnative pronunciation variants to lead to improvements in the accuracy of automatic segmentation of nonnative speech (Bonneau et al., 2012; Jouvet et al., 2011; Mesbahi et al., 2011; Orosanu et al., 2012).

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 (Kohler, 1996). However, research by Bouselmi et al. (2005, 2012) has shown that even more accurate segmentation of learners' utterances can be obtained by using acoustic models adapted to L2 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.

Given these resources, the JSnoori software, which **[TODO de-stress]** uses for speech processing, is capable of automatically segmenting a learner's utterance almost instantly. 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, In its current implementation, **[TODO de-stress]** 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, **[TODO de-stress]** can easily be extended to perform on-the-fly segmentation of learner utterances.

Although the IFCASL corpus also contains manually-corrected versions of the majority of the forced-alignment segmentations, **[TODO de-stress]** 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-fledged CAPT system, which would need to perform segmentation on the fly without recourse to manual verification. Indeed, forced alignment is not a perfect method; because of the constraints put on the recognition system, it will always find a match between the given text and audio, even if they do not correspond (e.g. if a completely different sentence/word was uttered). Therefore, inaccuracies in the phone boundaries determined using this technique must be expected, especially when the alignment is performed on nonnative utterances using an acoustic model trained on native speech. As discussed in Section 2.2.1, however, Mesbahi et al. (2011) compared manual segmentations of L2 speech with automatic segmentations produced via these techniques, and concluded that the automatically-produced segment boundaries were accurate enough to be useful in CAPT for prosodic feedback. Nonetheless, a fully-fledged CAPT system extending **[TODO de-stress]** would ultimately have to cope with any problems that may result from using imperfect automatic segmentations as a starting point for analysis; this could be accomplished using techniques for detecting incorrect utterances such as those developed by Bonneau et al. (2012) and Orosanu et al. (2012), as discussed in Section 2.2.1.

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 syllable-level analysis is important for the diagnosis of lexical stress errors, syllable 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

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 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.

Table 4.1: Features used for duration analysis, and their values for the sample utterances of "Flagge" in fig. 4.1. S0 refers to the word's first syllable, S1 to the second syllable; similarly, V0 and V1 refer to the nucleus (vowel) of the first and second syllable, respectively.

Feature name	Description	Value	
		(a) G	(b) F
REL-SYLL-DUR	Duration of S1/duration of S0	0.50	1.00
REL-V-DUR	Duration of V1/duration of V0	0.56	1.71

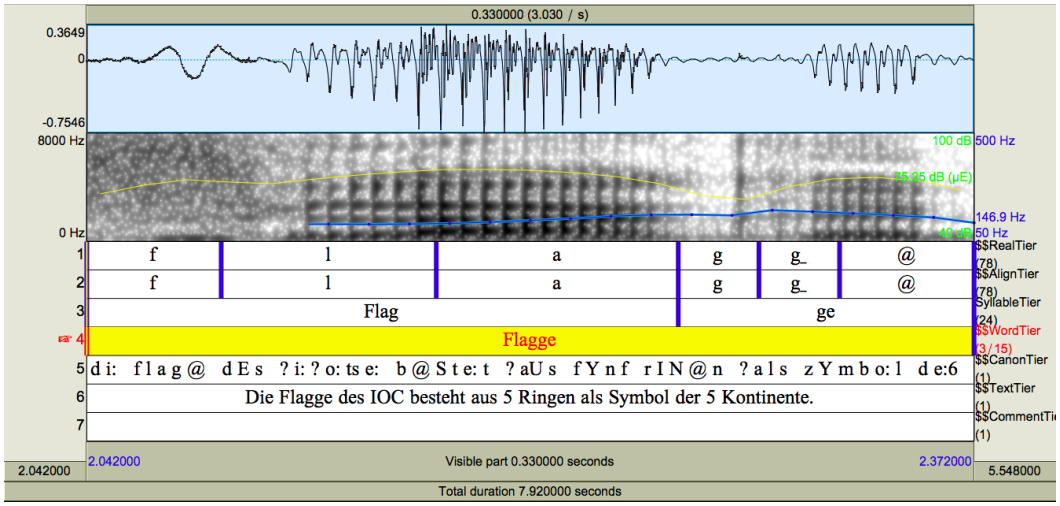
Throughout this section, the features discussed are illustrated with their values for a word from two sample utterances of a German word selected from the IFCASL corpus: one by a native German speaker (assumed to represent a correct realization of lexical stress), the other an utterance from an L1 French speaker which was assigned a gold-standard label of [incorrect] (see Chapter 3). The oscillogram, waveform, and segmentations for these samples are shown in fig. 4.1.

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

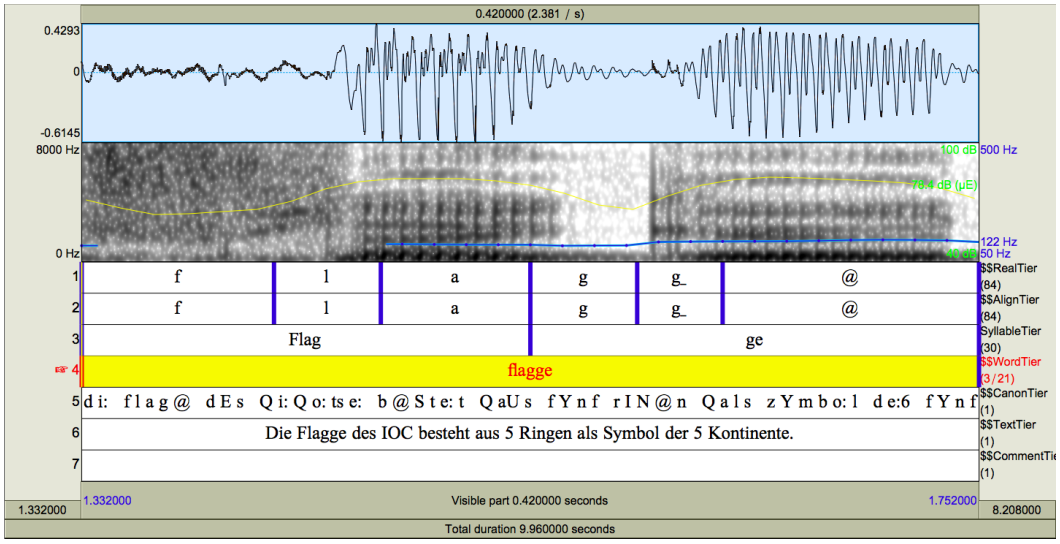
4.2.1 Duration

Analysis of duration (timing) is extremely important for detecting stress patterns; indeed, 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), duration analysis in this work takes into account the duration of each syllable in the word to be analyzed, and of the vowels at the nucleus of each syllable. 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 4.1, along with the values computed for each feature from the sample utterances shown in fig. 4.1.



(a) L1 German speaker (G)



(b) L1 French speaker (F)

Figure 4.1: Two sample utterances of the word "Flagge" from the IFCASL corpus, used to illustrate the features discussed in this section. In these Praat screenshots, the oscillogram and spectrogram of the signal are visible in the top portion of the image, while the bottom portion displays the corresponding segmentations. Segmentations at the phone, syllable, and word level are displayed, along with the canonical phone sequence and orthographic form of the uttered sentence.

4.2.2 Fundamental frequency

As described in Section 2.3, the fundamental frequency (F0) of an utterance, which corresponds at the perceptual level to its pitch, also provides a strong indication of how lexical stress is realized in that utterance, and F0 features are another crucial component of the system's prosodic analysis.

The pitch (F0)[1] contour of a given utterance is estimated 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 offset by 8 milliseconds). JSnoori also implements several improvements to pitch detection beyond Martin's algorithm (Di Martino and Laprie, 1999), including 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.

Thanks to these techniques, JSnoori is capable of efficient, generally accurate detection of F0 within the range of 65-800 Hertz (Hz), with each pitch point in the contour being subsequently converted from Hz 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 table 4.2. 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.

Much of the work on assessing nonnative 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).

[1] As stated previously, pitch and F0 are not identical, the latter being an acoustic (objective) property of the speech signal and the former the perceptual (subjective) correlate thereof. However, in JSnoori and much of the literature, the term "pitch" is often used to refer to an aspect of the speech signal which can be measured objectively. For consistency with JSnoori's terminology, in this section the terms will be used somewhat interchageably.

Table 4.2: Features used for fundamental frequency (F0) analysis, and their values for the sample utterances of "Flagge" in fig. 4.1. S0 refers to the word's first syllable, S1 to the second syllable; similarly, V0 and V1 refer to the nucleus (vowel) of the first and second syllable, respectively.

Feature name	Description	Value	
		(a) G	(b) F
REL-SYLL-F0-MEAN	Mean F0 in S1/mean F0 in S0	1.11	1.09
REL-V-F0-MEAN	Mean F0 in V1/mean F0 in V0	0.95	1.10
REL-SYLL-F0-MAX	Maximum F0 in S1/max. F0 in S0	1.10	1.12
REL-V-F0-MAX	Max. F0 in V1/max. F0 in V0	0.96	1.12
REL-SYLL-F0-MIN	Minimum F0 (> 0) in S1/min. F0 in S0	1.04	1.04
REL-V-F0-MIN	Min. F0 in V1/min. F0 in V0	0.97	1.18
REL-SYLL-F0-RANGE	F0 Range (max. F0$-$min. F0) in S1/ F0 range in S0	1.27	1.43
REL-V-F0-RANGE	F0 Range in V1/F0 range in V0	0.91	0.82
F0-MAX-INDEX	$\begin{cases} 0 & \text{if max. F0 in S0} > \text{max. F0 in S1} \\ 1 & \text{otherwise} \end{cases}$	1	1
F0-MIN-INDEX	$\begin{cases} 0 & \text{if min. F0 in S0} < \text{min. F0 in S1} \\ 1 & \text{otherwise} \end{cases}$	0	0
F0-MAXRANGE-INDEX	$\begin{cases} 0 & \text{if F0 range in S0} > \text{F0 range in S1} \\ 1 & \text{otherwise} \end{cases}$	1	1

Table 4.3: Features used for intensity analysis, and their values for the sample utterances of "Flagge" in fig. 4.1. S0 refers to the word's first syllable, S1 to the second syllable; similarly, V0 and V1 refer to the nucleus (vowel) of the first and second syllable, respectively.

Feature name	Description	Value (a) G	Value (b) F
REL-SYLL-ENERGY-MEAN	Mean energy in S1/S0 mean en.	0.39	0.38
REL-SYLL-ENERGY-MAX	Maximum en. in S1/S0 max. en.	0.52	0.45
REL-VOWEL-ENERGY-MEAN	Mean en. in V1/mean en. in V0	0.43	0.43
REL-VOWEL-ENERGY-MAX	Max. en. in S1/max. en. in S0	0.53	0.61
ENERGY-MAX-INDEX	$\begin{cases} 0 & \text{if S0 max. en.} > \text{S1 max. en.} \\ 1 & \text{otherwise} \end{cases}$	0	0

4.2.3 Intensity

Research on lexical stress prosody has generally indicated that intensity is the least important of the three features, i.e. corresponds least closely to lexical stress patterns (Cutler, 2005). Indeed, existing lexical stress assessment tools may not take intensity into account, as was the case with the prosodic diagnosis functionality of JSnoori at the start of this thesis project. However, intensity can nonetheless have an impact on the perception of lexical stress, especially in combination with pitch or duration, or both (ibid.); Therefore, in addition to duration and fundamental frequency, JSnoori's learner speech analysis module was modified to take intensity (energy)[2] into account.

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 Hz in spectra extracted from the signal by Fast Fourier Transform, using Hamming windows 20 milliseconds long offset by 4 milliseconds. Energies below a "silence threshold" of 60 decibels are not counted 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 using JSnoori, **[TODO de-stress]** analyzes a given learner utterance and diagnoses their lexical stress error(s) (or lack thereof) by comparing this L2 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.

[2] For compatibility with the terminology used in JSnoori, these terms will be used interchangeably in this section.

4.3 Diagnosis by direct comparison

One 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 an approach commonly taken in CAPT systems and research (see e.g. Bonneau and Colotte, 2011; Delmonte, 2011; Eskenazi, 2009). 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; an error can then be diagnosed when the L2 learner's utterance and that of the native speaker differ substantially with respect to the relevant features. This section describes the various techniques for diagnosis by comparison implemented in **[TODO de-stress]**.

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 Section 2.2.2, this is the approach used to evaluate learner speech in JSnoori and its predecessor WinSnoori (Bonneau and Colotte, 2011; Bonneau et al., 2004; Henry et al., 2007). In **[TODO de-stress]**, 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. 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).

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). 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, 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, assuming that the syllable with the highest F0 peak is the syllable that has been stressed; as noted in Section 4.2.2, this assumption may be less controversial for English, the target language in mind during development of JSnoori, than for German, in which the stressed syllable may be realized with a higher or lower F0 than the unstressed one. 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 with the analogous difference in the reference utterance; if the difference between these differences is greater than a threshold, the learner is considered not to have expressed stress strongly enough, receiving a score of 0.8. Otherwise, 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, such that the syllable with the higher energy peak is assumed to contain the stress. 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 **[TODO de-stress]** 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, the hope being that the variability between these reference utterances will capture more general traits of native pronunciation.

In **[TODO de-stress]**, 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. In the current implementation, this combination consists of simply 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. More sophisticated ways of combining several single-reference scores into one multiple-reference score are certainly conceivable, and this could be an interesting direction for future work.

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, **[TODO de-stress]** therefore enables the choice of a reference from a set of available speakers. When designing an exercise, the researcher can either manually select a reference utterance for all students who will complete that exercise, or enable each student to manually select their own reference.

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. (ibid.) 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; **[TODO de-stress]** follows Probst et al. (ibid.) in using F0 mean and range for that representation, where each 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

The comparison-based diagnostic approach explored in the previous section is useful, as demonstrated by its frequent appearance in CAPT research and systems, yet it is not without limitations. First of all, although comparisons to multiple reference speakers and careful selection of the best reference speaker for a learner might help mitigate some of the interference of speaker- or utterance-dependent features of the reference with the assessment of the learner's utterance, this assessment is still closely tied to a small number of L1 utterances, which may still not provide a general enough representation of the correct realization of lexical stress in the target language. As mentioned in Section 2.2.4, 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; similar results could reasonably be expected in the context of CAPT. A second limitation of the comparison-based approach is that and therefore it limits the exercises available in a CAPT system to those using only words/sentences for which recordings by L1 speakers are available.

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, may be a way to overcome these shortcomings of the comparison approach: 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. One logical approach to such generalized modeling involves using machine learning algorithms to classify a given utterance as correct or incorrect with respect to lexical stress, based on prosodic features of that utterance.

This classification-based diagnostic approach to identifying lexical stress errors is the one which has ostensibly been least explored in CAPT research. As mentioned in Section 2.4.3, some researchers have successfully used machine learning to classify (native) English utterances based on their stress patterns (Kim and Beutnagel, 2011; Shahin et al., 2012), but with the exception of one pilot experiment by Kim and Beutnagel, it seems that the use of this technique to classify L2 speech, and particularly L2 German speech, has yet to be researched.

Given the relative novelty of this type of diagnosis for prosodic errors in CAPT, an investigation of the feasibility of lexical stress error diagnosis by classification was one objective of this thesis project, and constitutes one of its main contributions. To this end, a series of classification experiments were 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,
- which of the prosodic features discussed in Section 4.2 are most useful for classification, and
- whether a classification-based approach can lead to reasonably accurate diagnosis for words or speakers not seen in the training data.

The sections that follow describe these experiments and their findings.

4.4.1 Data and method

[TODO use dataset, IFCASL-FG and IFCASL-GG if appropriate]

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-FG 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 L2 utterances and their gold-standard labels from the annotated dataset (see Section 3.5), the corresponding L1 utterances of the selected word types from the the IFCASL-GG 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.

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) (Breiman et al., 1984); though WEKA implements a wide variety of other classifiers, some of which could conceivably offer better performance, CARTs were chosen for their simple and efficient training process and ease of interpretation by humans. In future work (see Section 6.2), it would be interesting to compare different classification algorithms to see if other classifiers are more effective for this type of data, along the lines of the experiments by Kim and Beutnagel (2011).

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. 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, averaging results from each test set to obtain the overall performance statistics. The features and divisions of data used in each experiment are described in the sections below.

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_1 measure, also known simply as F-measure or F-score: a metric which combines Precision and Recall by taking their harmonic mean, weighting both equally. It is computed with the formula:

$$F_1 = 2PR/(P+R)$$

- F_2 measure: a metric similar to F_1 measure, but in which Recall is accorded twice as much importance as Precision. F_2 is computed as:

$$F_2 = (1+2^2) \cdot PR/(2^2 \cdot P + R) = 5PR/(4P+R)$$

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 (Neri et al., 2002). 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. For this reason, the results in this section report both the commonly used F_1 measure, which weights precision and recall evenly, as well as F_2, which prioritizes recall over precision.

4.4.2 Feature performance

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 nonnative 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 nonnative, 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 nonnative 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

Table 4.5 lists the results of experiments with the prosodic features described in table 4.4a. As seen in the first three rows oftable 4.5, the results obtained using features representing each of the three acoustic correlates of lexical stress, duration, F0, and intensity (energy), conform with findings from other research (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 a 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

Table 4.4: Feature sets used in classification experiments

(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

Set name	Feature(s)
WORD	The word being uttered (e.g. *Tatort*)
LEVEL	Speaker's L2 German skill level (A2/B1/B2/C1)
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, quantified by percent accuracy (% acc.) and Kappa agreement (κ) with respect to the gold-standard labels, as well as precision (P), recall (R) and F_1 and F_2 measures for the [correct] class. The best values achieved for each metric are displayed in **bold**.

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

to the duration features, in that they tend to make classifiers more conservative in labeling utterances as [incorrect].

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 3.6, 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 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 3.6.3). If the intended application of the classifier were the assessment of a learner's proficiency level, including this feature would obviously be

Table 4.6: Results of experiments with speaker and word features, quantified by percent accuracy (% acc.) and Kappa agreement (κ) with respect to the gold-standard labels, as well as precision (P), recall (R) and F_1 and F_2 measures for the [correct] class. The best values achieved for each metric are displayed in **bold**.

(a) 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

(b) 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

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. 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, and as Neri et al. (2002, pp. 15-16) point out, for any automatic pronunciation scoring system to be useful, there must be a strong correlation between how the system evaluates a given utterance and how human raters would evaluate it. Here, classification performance was evaluated against the gold-standard annotation of the dataset; yet, as described in Section 3.5, the gold-standard labels represent a consolidation of multiple, often conflicting labels from different annotators (see Section 3.4). Considered in the light of the relatively low inter-annotator agreement observed when humans were asked to diagnose lexical stress errors, then, the classification accuracy and κ scores do not seem so unimpressive. 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%). It seems possible, given the low inter-annotator agreement observed in the annotation described in Chapter 3, that 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. If this is the case, 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 Performance on unseen speakers and words

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 nonnative speakers). However, as stated at the beginning of Section 4.4, 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.

Table 4.7: Results of experiments with unseen speakers, averaged over all 56 held-out speakers. Performance is quantified by percent accuracy (% acc.) and Kappa agreement (κ) with respect to the gold-standard labels, as well as precision (P), recall (R) and F_1 and F_2 measures for the [correct] class. The best values achieved for each metric are displayed in **bold**.

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

Unseen speakers

To create training and testing data for experiments with utterances from unseen speakers, the entire set of labeled nonnative utterances (see Section 4.4.1) was divided into 56 subsets, each containing all utterances from one of the 56 unique nonnative 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, accuracy drops from 71.87% to 70.22%, in κ from 0.34 to 0.24, and 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 nonnative 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

Table 4.8: Results of experiments with unseen words, averaged over all 12 held-out word types. Performance is quantified by percent accuracy (% acc.) and Kappa agreement (κ) with respect to the gold-standard labels, as well as precision (P), recall (R) and F_1 and F_2 measures for the [correct] class. The best values achieved for each metric are displayed in **bold**.

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**

in performance observed with the inclusion of speaker-related features (see Section 4.4.2), along with the apparent interplay 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.

In comparison with the best results obtained on seen words as reported in the previous section (69.77% accuracy and κ =0.29 for the DUR+F0 feature set, and 71.87%/κ =0.34 for 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

Table 4.9: Best classification results (using feature set DUR+F0) on unseen words, by word type. Performance is quantified by percent accuracy (% acc.) and Kappa agreement (κ) with respect to the gold-standard labels, as well as precision (P), recall (R) and F_1 and F_2 measures for the [correct] class. **[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	−0.1	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

(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 **[TODO de-stress]** 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 [TODO de-stress]

In the **[TODO de-stress]** CAPT tool, 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. 4.2. When creating a new exercise, in which a learner is asked to read one of the sentences from the IFCASL corpus (see Section 1.2) 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 object for that exercise (see fig. 1.3 on p. 4). The DiagnosisMethod 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. 4.3. As illustrated in fig. 1.3, this DiagnosisMethod is combined with a (compatible) FeedbackMethod when the exercise is created; see Section 5.4 for a description of FeedbackMethod options and their compatibility with the various possible DiagnosisMethod options.

Figure 4.2: Overview of diagnosis options in [TODO de-stress] [TODO Redo as Tikz flowchart]

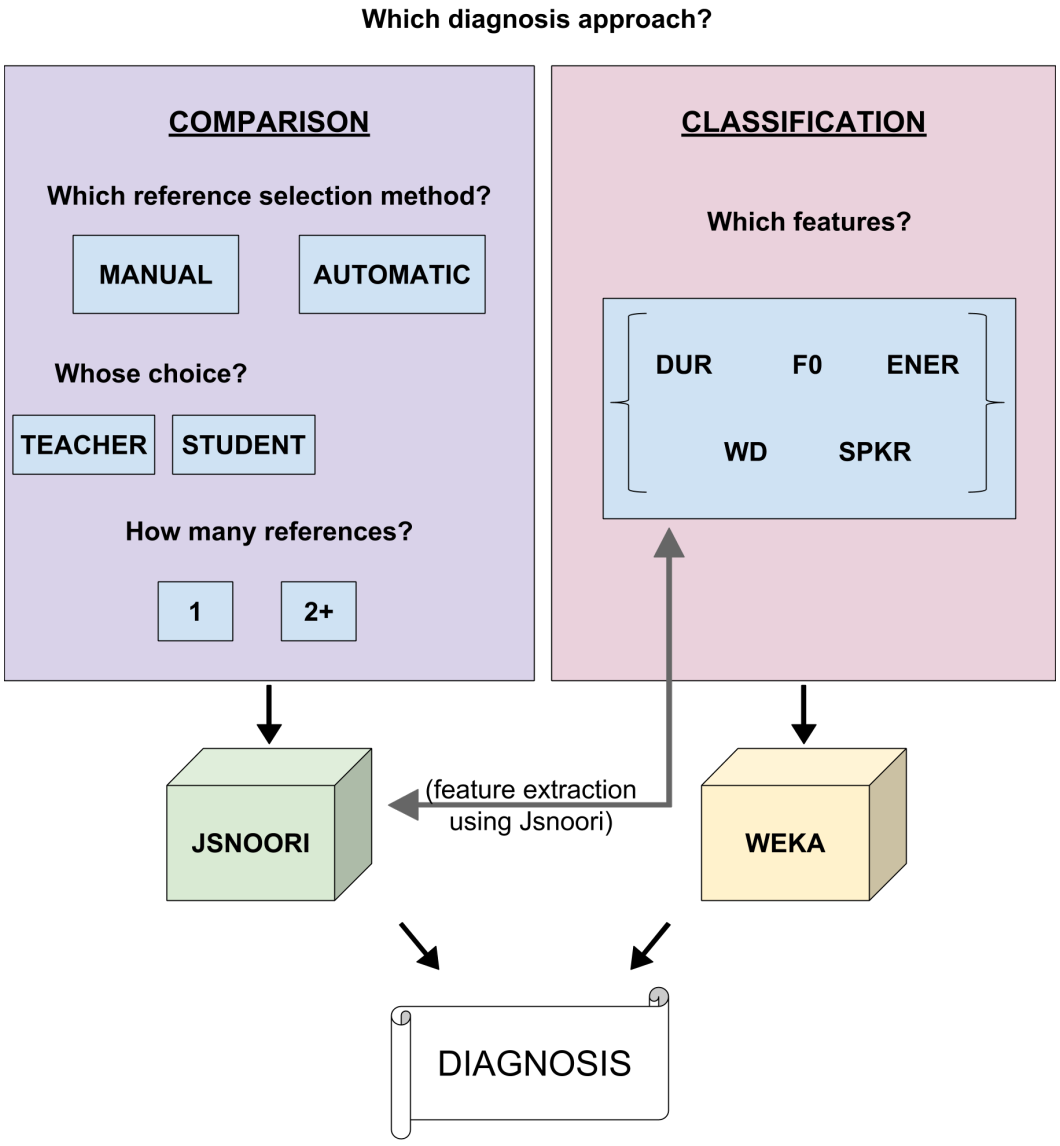
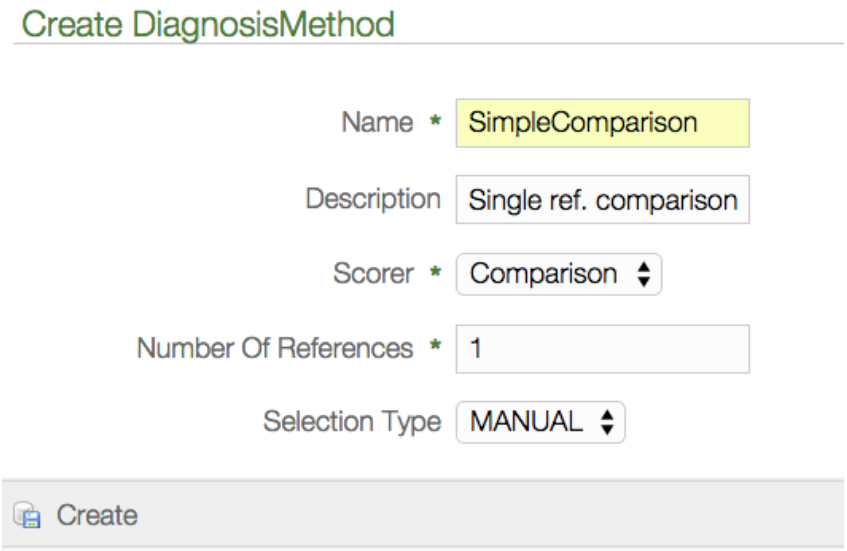


Figure 4.3: Screenshot of the researcher-facing interface to create a DiagnosisMethod



4.6 Summary

This chapter has explored an array of methods by which lexical stress errors in the speech of French learners of German can be automatically diagnosed, as implemented in the modular diagnostic component of the **[TODO de-stress]** CAPT tool.

The first requirement for error diagnosis, accurate automatically-produced segmentations of the words, syllables, and phones of a learner's utterance (or that of a native speaker), can be obtained through forced alignment with the text of the utterance, as described in Section 4.1. Forced alignment for German utterances being currently under development in the JSnoori software used by **[TODO de-stress]** for speech processing, the current version of **[TODO de-stress]** mocks this alignment step by using automatically-produced segmentations for utterances from the IFCASL corpus (Fauth et al., 2014; Trouvain et al., 2013; see also Section 1.2).

Analysis of the lexical stress realization of a given utterance, based on a set of prosodic features, is a second requirement for diagnosis, and Section 4.2 has presented the set of features which can be used for such analysis in **[TODO de-stress]**. These features capture the three acoustic properties most strongly correlated with the prosodic realization of lexical stress, namely duration, F0, and intensity. In **[TODO de-stress]**, such features are extracted from a given utterance using the forced-alignment segmentations of that utterance and the speech processing capabilities of JSnoori.

Using these features to represent a given L2 utterance as well as corresponding utterance(s) by L1 speakers, it is possible to assess the learner's utterance via one of two primary strategies: comparison or classification. Section 4.3 has explored the possibilities for comparison-based diagnosis, in which features of the relevant segments of a learner's utterance are compared to the analogous features in a L1 utterance, and an error is diagnosed when the utterances differ considerably with respect to the relevant features. As described in Section 4.3.1, JSnoori

uses this comparison-based approach to score each L2 utterance with respect to duration, F0, and prosody, and **[TODO de-stress]** can use these scores as one form of diagnosis and a starting point for the delivery of certain types of feedback, as will be discussed in Chapter 5. In addition to diagnosing an L2 utterance in comparison with a single L1 utterance, **[TODO de-stress]** can combine scores with respect to multiple L1 utterances into a single score (see Section 4.3.2), thus helping to reduce some of the risk of the diagnosis "over-fitting" to speaker- or utterance-dependent features of a single reference. Another option available in **[TODO de-stress]** relates to the method of choosing the reference utterance(s) for a given learner utterance (see Section 4.3.3); though many existing CAPT systems, including JSnoori, require manual selection of reference utterances, **[TODO de-stress]** also offers an automatic selection option in which the reference is selected by choosing the L1 speaker(s) whose voice most closely resembles that of the learner in terms of F0 mean and range.

The second diagnosis strategy explored in this chapter (Section 4.4), classification of lexical stress errors using machine learning algorithms, is a more novel approach to lexical stress error identification in CAPT, in which a learner's utterance is compared to the more abstract model of L1 speech represented by a classifier trained on a large number of L1 utterances. The experiments described in Sections 4.4.2 and 4.4.3 constitute original contributions to the understanding of how, and how effectively, classification-based diagnosis can be used to identify (in)correct realizations of lexical stress. As described in Section 4.4.2, the features seemingly most useful for classification relate to the duration and F0 of the utterance(s), unsurprising considering that these have been shown to be most closely linked to lexical stress in German (Cutler, 2005; Dogil and Williams, 1999). Features capturing the word type of the utterance as well as the age, gender and proficiency level of the speaker were also found to be quite valuable for error classification; combining these features with all three prosodic feature types resulted in the highest overall accuracy observed on this dataset (70.65% accuracy, κ =0.31). As the observed agreement between the classifier's labels and the gold standard slightly exceeded the overall inter-annotator agreement observed when humans were asked to perform this error diagnosis task (see Section 3.4), these results seem encouraging. Unsurprisingly, slightly lower accuracy was observed when classifying utterances of word types or speakers not represented in the training data; however, the fact that accuracy on unseen words still remained comparable to the human inter-annotator agreement statistics seems to confirm the expectation that classification-based diagnosis may be a useful way to create CAPT systems which are are not limited to words/sentences for which recorded L1 utterances are available.

While comparison-based diagnosis is not a new approach to identifying lexical stress errors in CAPT, the availability of multiple-reference comparison and automatic selection of reference speakers in the system, as well as the ability for researchers or instructors to configure the comparison method, make **[TODO de-stress]** an important addition to the current CAPT landscape. The inclusion of a classification-based alternative to diagnosis is also a novelty for such a CAPT system, and this chapter's investigation of how L2 lexical stress errors can be diagnosed by classification is one of the major contributions of this thesis. By enabling researchers and instructors to choose among the various diagnosis options described in this chapter, **[TODO de-stress]** will facilitate much needed future work exploring which diagnostic methods are most useful in which learning contexts, and which types of feedback (described in the following chapter) can best convey these diagnoses to the learner.

5 Feedback on lexical stress errors

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 **[TODO de-stress]**. 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 **[TODO de-stress]** 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 1.1 and illustrated in fig. 1.1.

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, guided by the notion that to maximize its utility in future feedback research, **[TODO de-stress]** 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

Implicit feedback describes a category of feedback types which present the uttered (incorrect) L2 pronunciation and/or target (correct) L1 pronunciation of a given utterance to the L2 learner, offering them the opportunity to notice features of the correct pronunciation and/or the difference between their speech and the target without explicitly drawing their attention to the error(s) they have made or to the means by which they can correct those errors (cf. explicit feedback, Section 5.2). Implicit feedback on prosodic errors is often delivered in the L2 language classroom in the form of repetitions or recasts of learner utterances (Dlaska and Krekeler, 2013), and CAPT systems often present this type of feedback by allowing the learner to listen to recordings of their (L2) utterances or those of reference (L1) speakers, as well as by visualizing the relevant prosodic features of these utterances (see Section 2.2).

Some past research on the use of implicit feedback in L2 pronunciation training suggests that this type of feedback may be sufficient to help (some types of) learners notice and correct their errors (e.g. Neri et al., 2002, p. 8; Bonneau and Colotte, 2011; see also Chapter 2), As implicit feedback is generally simpler to deliver automatically than explicit corrective feedback, it is an attractive option for CAPT, and has therefore been included among the feedback options offered by **[TODO de-stress]**. The remainder of this section describes the various types of visual and implicit feedback available, and how they are generated.

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. Henry et al., 2007; Martin, 2004), 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, this is the case in JSnoori, as seen in fig. 2.1 in Section 2.2.2. 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 **[TODO de-stress]** 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 5.1 illustrates the display of such graphical abstractions in **[TODO de-stress]**. 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 a small yellow box in fig. 5.2a.

In the case of diagnosis via comparison (see Section 4.3), the features used to create the graphical representation of the target (correct) pronunciation are computed directly from the reference utterance(s). In the case of diagnosis via classification (see Section 4.4), however, no individual reference utterances are used, so a more abstract representation of how L1 speakers realize the word in question must be created. This is achieved by extracting the relevant duration, F0, and intensity features in every native utterance of the target word available in the dataset, and computing the average values for each syllable in the word. These averages are then used as the target values presented in the "reference" visualization, as seen in fig. 5.2b.

The mappings between graphical properties (width, height, and opacity of the rectangle) and prosodic features (duration, F0, and intensity, respectively) are currently hard-wired in **[TODO de-stress]**. 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 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 text-stylization feedback offered by **[TODO de-stress]**, the text of a given syllable is reshaped with a simple mapping between font size (as a multiple of the default size) and duration (as a fraction of the word duration); duration features for the learner and reference visualizations are computed in exactly the same way as for the graphical

Figure 5.1: Screenshots of feedback via graphical abstractions of prosody

(a) Graphical feedback generated from a comparison-based diagnosis using two reference utterances.

(b) Graphical feedback generated from a classification-based diagnosis. Also visible in this screenshot is the combination of graphical abstractions with the stylized text feedback described in Section 5.1.1

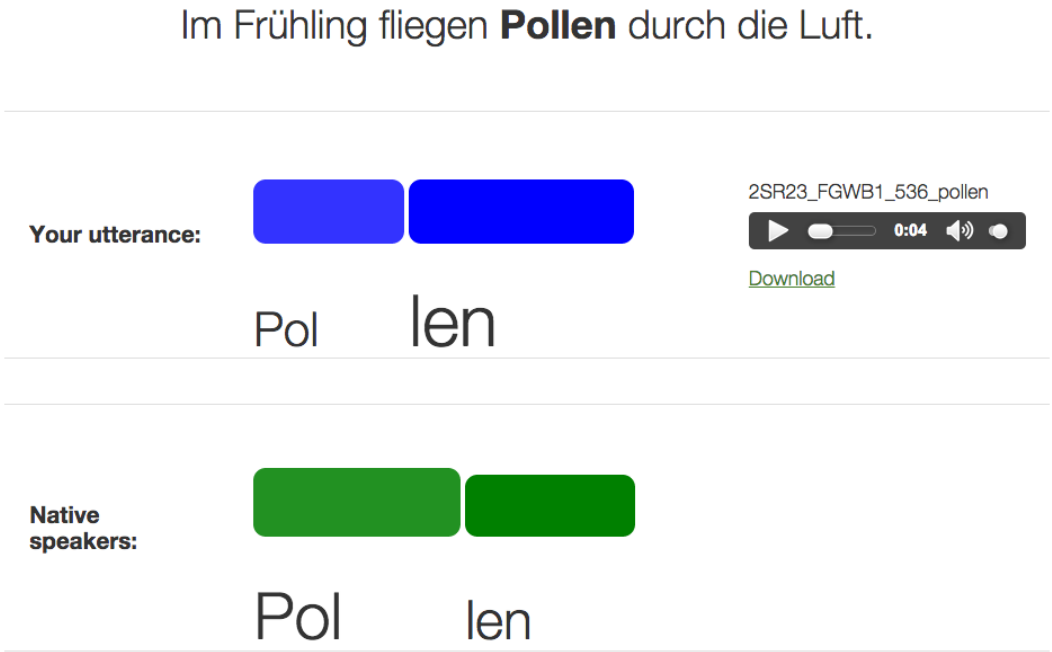


Figure 5.3: Screenshot of feedback on syllable duration via text stylization

abstraction feedback described in the previous section. 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

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 **[TODO de-stress]** 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. 5.1 and 5.3, these audio recordings are always accessible alongside the visual (and other types of) feedback presented.

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 available in **[TODO de-stress]**.

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 (Moulines and Charpentier, 1990). 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 (Colotte and Laprie, 2002; Laprie and Colotte, 1998), 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 Bonneau and Colotte, 2011; Henry et al., 2007), 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 JSnoori itself, **[TODO de-stress]** 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 implicit auditory feedback as one of the many feedback types teachers/researchers can choose to present to the learner, **[TODO de-stress]** will facilitate further research into the use of this type of speech resynthesis for L2 language teaching.

5.2 Explicit feedback

In contrast to implicit feedback, explicit feedback involves not only presenting the learner with the erroneous (L2) or target (L1) pronunciation, but directly calling the learner's attention to the errors in their production and possibly giving them clear instructions on how to correct those errors. In some situations, explicit corrective feedback has been empirically shown to help learners improve their pronunciation more than implicit feedback (e.g. Dlaska

and Krekeler, 2013). Furthermore, this type of feedback may be more motivating for learners, encouraging them to continue working on their pronunciation, and thus indirectly contributing even further to learning gains (ibid.). Therefore, any CAPT system should ideally be able to provide such feedback, and **[TODO de-stress]** is no exception; this section describes the explicit feedback options offered by the system, and how these are generated based on the error analysis provided by the tool's diagnostic module (see Chapter 4).

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, 2013). As illustrated in figs. 5.5a and 5.5b, 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 (see Section 4.5).

Figure 5.5a illustrates a case where each of the three scores is given equal weighting, while fig. 5.5b shows the same scores in the case where duration is prioritized over F0, which is in turn given a higher weight than intensity. 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.

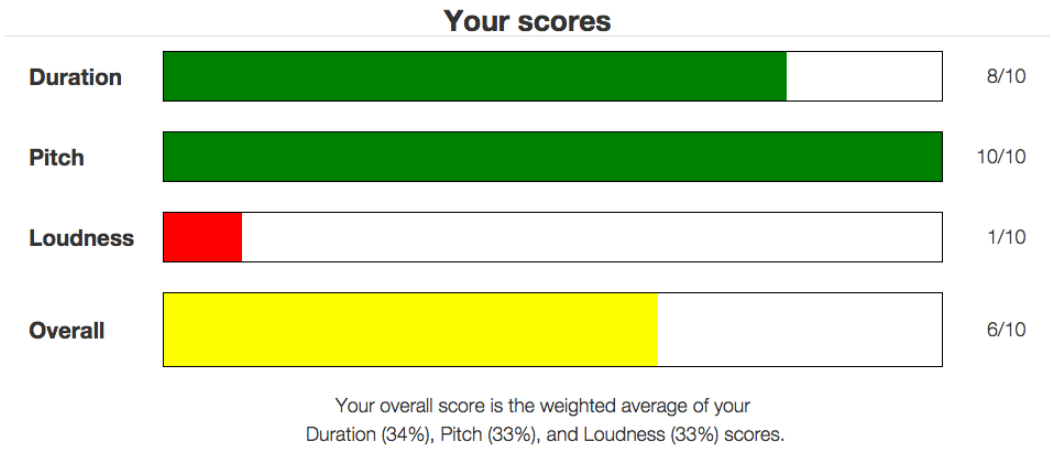
A potential problem with this feedback method relates to the fact, discussed in Section 4.3.1, that the scores output by JSnoori's diagnostic tools are actually discrete, ordinal values, despite the use of numbers between 0 and 1 to represent these values. Therefore, the visual representation of these 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 **[TODO de-stress]** has been designed to facilitate.

5.2.2 Verbal feedback

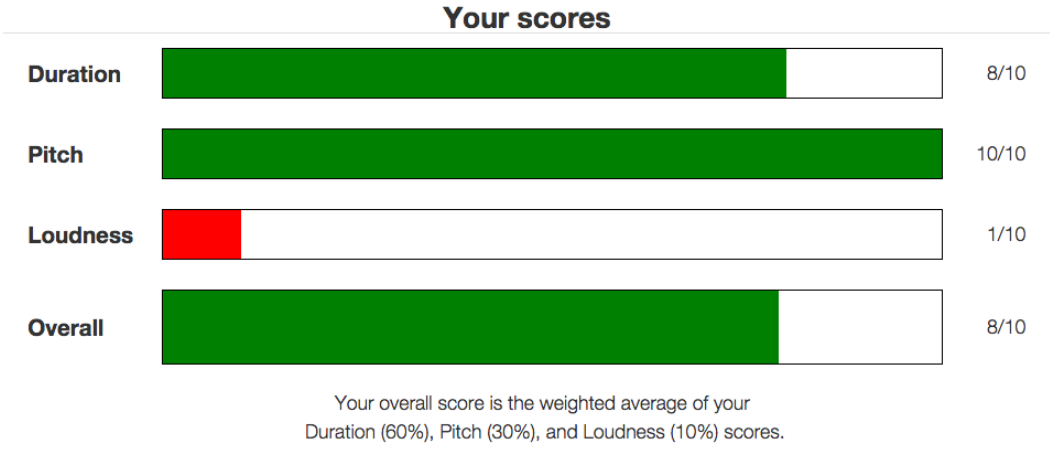
Another way to explicitly deliver feedback on the learner's diagnosis is by verbalizing this diagnosis with one or more appropriate messages. If individual scores have been computed for the three prosodic feature types (duration, F0, and intensity) as described in Section 4.3.1, these are verbalized 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

Figure 5.4: Screenshots of explicit feedback via skill bars

(a) Where all three skills (prosodic feature types) contribute equally to the overall score



(b) Where duration contributes more than F0, which contributes more than intensity (60%, 30% and 10% of overall score, respectively)



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 **[TODO 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. Long and Aleven, 2011, 2013). 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 its benefits in the CAPT context specifically have been pointed out by Neri et al. (2002) and Mehlhorn (2005). 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 a **[TODO de-stress]** exercise, 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 are 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.

Table 5.1: Messages used to deliver explicit verbal feedback of feature-specific scores

(a) Duration (timing)

Score	Message
0.0	"Sorry, I wasn't able to analyze duration in your utterance."
0.1	"I think you pronounced an incorrect number of syllables for this word."
0.3	"I think you pronounced an incorrect number of phones in at least one of the word's syllables."
0.5	"The wrong syllable has the longest vowel."
0.8	"The correct syllable's vowel is longest, good job! But it should be even longer compared to the unstressed syllable."
1.0	"No problems with duration, great job!"

(b) F0 (pitch)

Score	Message
0.0	"Sorry, I wasn't able to analyze pitch in your utterance."
0.1	"The wrong syllable has the highest pitch."
0.8	"The correct syllable has the highest pitch, good job! But it should be even higher compared to the unstressed syllable."
1.0	"Your pitch was pitch-perfect, great job!"

(c) Intensity (energy)

Score	Message
0.0	"Sorry, I wasn't able to analyze the loudness of your utterance."
0.1	"The wrong syllable is loudest."
0.8	"The correct syllable is loudest, good job! But it should be even louder compared to the unstressed syllable."
1.0	"No problems with loudness, great job!"

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

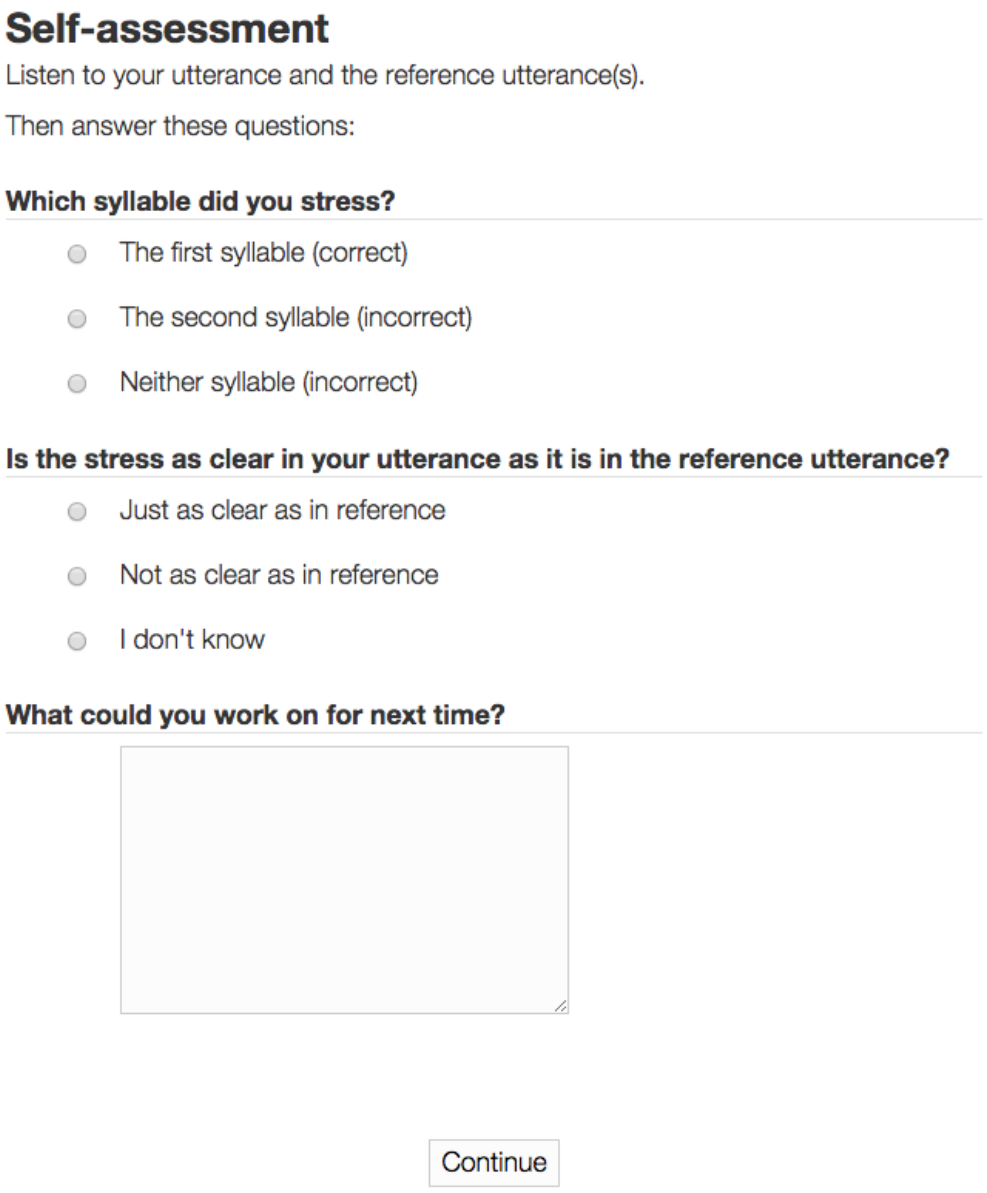
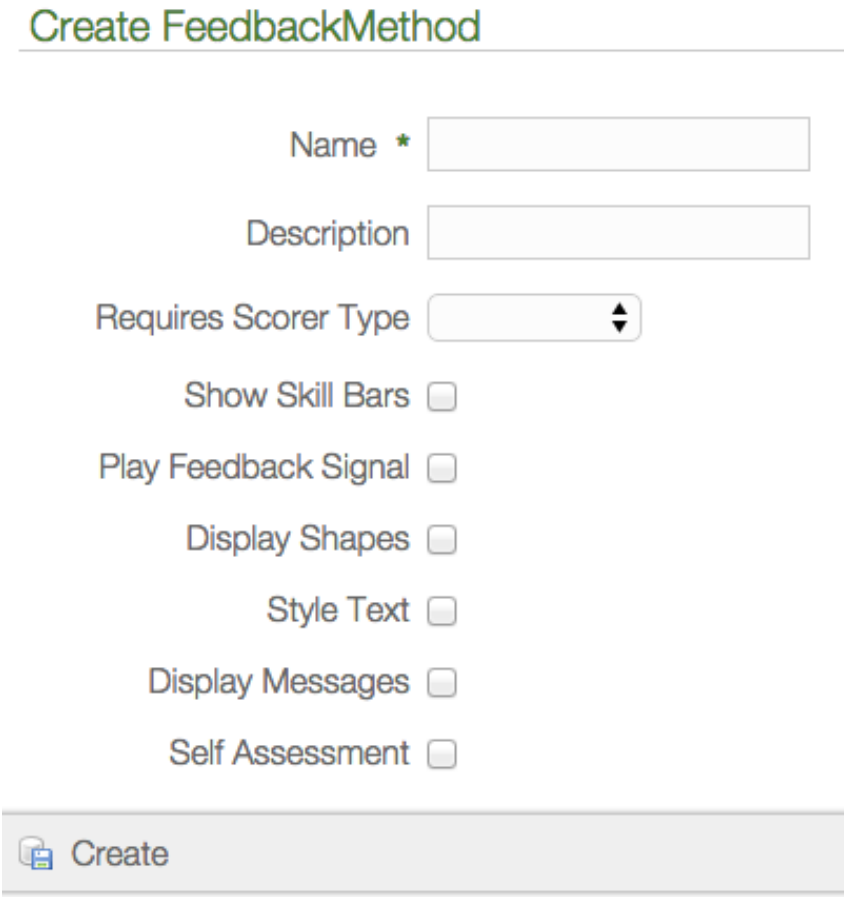


Figure 5.7: Screenshot of the researcher-facing interface to create a FeedbackMethod



5.4 Controlling feedback in [TODO de-stress]

As mentioned in Section 4.5 and illustrated in fig. 1.3, a researcher (or teacher) creating a new exercise in [TODO de-stress] 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 Chapter 4). For example, if the chosen classification method diagnoses learner errors by classification, and not by comparison with a reference utterance, it is not possible to present learners with implicit auditory feedback, either in the form of an original reference utterance or that of a resynthesized version of the learner's utterance which has been prosodically modified to match that reference. Therefore, each FeedbackMethod is aware of which type of DiagnosisMethod it requires, so that the researcher never accidentally chooses an incompatible pairing between DiagnosisMethod and

Table 5.2: Compatibility of the various diagnosis and feedback types available in **[TODO de-stress]**. Compatible combinations are indicated with "+", incompatible combinations with "−".

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

FeedbackMethod when creating a new exercise. Table 5.2 summarizes the (in)compatibility between the various types of diagnosis and feedback offered.

5.5 Summary

This chapter has presented the diverse array of feedback methods offered by **[TODO de-stress]**. The system can present learners with both explicit Section 5.2) as well as implicit (Section 5.1) feedback on their lexical stress errors. The latter may be delivered visually (Section 5.1.1), in the form of graphical or textual representations of prosody which may be easier to interpret than the more direct representations of speech signals more frequently seen in CAPT systems (see Section 2.2), or via the auditory channel (Section 5.1.2) in the form of original or prosodically modified utterances. Explicit feedback options include skill bars (Section 5.2.1) informing learners of the correctness of their pronunciation with regard to duration (timing), fundamental frequency (pitch), and intensity (loudness), as well as verbal feedback on this information via a series of error/success messages (Section 5.2.2). An additional feedback option offers learners the opportunity to self-assess their pronunciation (Section 5.3), in the hopes of encouraging them to develop a metacognitive understanding of their own learning process and take autonomous control of their learning.

Thanks to its modular implementation of the various feedback types, in combination with a simple configuration interface that allows researchers and/or instructors to easily pick and choose from the available options (Section 5.4), **[TODO de-stress]** is therefore a valuable platform for future empirical investigations into the impacts of these feedback types on the acquisition of L2 word prosody by L1 French learners of German. As established in Chapter 2 and further discussed in the following chapter, much work remains to be done to determine which feedback types are most effective in which learning contexts; by presenting a tool to facilitate such research, ostensibly the first of its kind for this type of error and aimed at this group of learners, this thesis has thus made an important contribution towards a more detailed understanding of the relative efficacy of feedback types in CAPT.

6 Conclusion and outlook

6.1 Thesis summary

This thesis has presented a series of investigations of how Computer Assisted Pronunciation Training (CAPT) can help native (L1) speakers of French learning German as a nonnative language (L2) improve their pronunciation with respect to the prosodic realization of lexical stress. Lexical stress, the phenomenon by which a given syllable is accorded a higher level of prominence than other syllables in a given word, serves a contrastive function in some languages (e.g. German) but not others (e.g. French). Lexical stress is very important in German, and may have a large impact on the intelligibility of L2 German speech (see Section 2.4.1); however, given that lexical stress is realized extremely differently (or not at all) in French (see Section 2.3.2), the correct prosodic realization of lexical stress in German is notoriously difficult for L1 French speakers (see Section 2.3.3).

Indeed, analysis of lexical stress errors in a small corpus of L2 German words uttered by L1 French speakers, described in Chapter 3, seemed to confirm that such errors are frequent in the speech of these learners. On the whole, errors were observed in approximately one-third of learner utterances (see Section 3.6.1), though such errors were produced much more frequently by learners of lower German proficiency than those with more advanced skills (see Section 3.6.3), with children making more errors than adult beginners (see Section 3.6.4). This evidence of a high frequency of production of lexical stress errors in the speech of such learners provides some justification for the selection of this type of error as the target of the prototype CAPT system developed in this work (see Section 2.4.2). However, the analysis of such errors must take into account the fact that when asked to annotate lexical stress errors in learner utterances, human annotators of varying backgrounds with respect to L1 and phonetics/phonology expertise exhibited generally low agreement (see Section 3.4); this may indicate that identifying such errors in learner speech is not a straightforward task, and may involve an element of subjectivity. If this is the case, it has important implications for treatment of such errors in CAPT, as reliable automatic detection is a requirement for a type of error to be addressed in a CAPT system (see Section 2.4.3), and the reliability of an automatic error diagnosis system should be evaluated in the context of the reliability of humans asked to perform the same task (see Section 4.4.2).

The evident difficulty of correct realization of lexical stress for L1 French speakers, combined with its importance to intelligibility in L2 German, seem to imply that this phenomenon is a particularly important one for such learners to work on when seeking to improve their pronunciation. In situations where language learners do not have access to a human German instructor, it is therefore advantageous to have CAPT tools which can automatically identify lexical stress errors in learners' L2 German utterances, and provide individualized corrective feedback on such errors to help learners overcome the hurdle this phenomenon presents.

With this underlying motivation, this thesis has explored the ways in which lexical stress errors can be automatically diagnosed in learner speech, and in which explicit and implicit feedback on these errors can be delivered to learners. It has also described the development of a prototype CAPT tool, **[TODO de-stress]**, which implements these diverse diagnosis and feedback methods in a modular architecture, in the hopes of facilitating future research on which of these are most beneficial to learners in which contexts.

Chapter 4 presented the approaches to diagnosis that have been explored in this thesis project and implemented in **[TODO de-stress]**. All diagnosis starts from a prosodic analysis of the learner's utterance in terms of duration, fundamental frequency (F0), and intensity (see Section 4.2), based on a segmentation of that utterance produced automatically by forced alignment (see Section 4.1). In a comparison-based diagnostic approach, the prosody of the learner's utterance is compared to that of a reference (L1) utterance of the same word, and scored with respect to each of the three dimensions (see Section 4.3.1). To capture the variability inherent in L1 pronunciations of the target word, multiple comparison scores can be combined (see Section 4.3.2), and the best reference speaker(s) for a given learner can be chosen automatically to maximize similarity between their voice and the learner's (see Section 4.3.3). An alternative diagnostic approach using classification by machine learning was also investigated (see Section 4.4); experiments with CART classifiers trained using different feature sets (see Section 4.4.2) revealed that duration features seem to be more useful for classification of L2 German stress errors than F0 features, which are in turn more useful than intensity features. The overall best results were obtained by combining prosodic features with features capturing the word type uttered and the German proficiency level of the learner. The agreement between the error diagnoses produced by the classifier and the gold-standard annotation was comparable to (if not slightly better than) that observed between human annotators in the manual annotation task, an encouraging finding with respect to the feasibility of reliable, automatic detection of such errors (see Section 4.4.2). Furthermore, only a slight drop in accuracy was observed when classifying utterances of word types which had been held out of the training data (see Section 4.4.3), highlighting the potential of a classification-based approach to enable the creation of CAPT systems which are not limited to only those words/sentences for which recorded L1 utterances are available. As little, if any, previous CAPT research has used classification to diagnose lexical stress errors in L2 German speech, particularly that of L1 French speakers, these investigations constitute a novel contribution to research on German CAPT.

As described in Chapter 5, another important contribution of this work is an exploration of the various types of feedback that can be presented to the learner, based on a comparison- or classification-based diagnosis of the learner's utterance using one of the methods described in Chapter 4. Inspired by both classroom teaching materials and existing prosody-aware CAPT systems, **[TODO de-stress]** offers a diverse array of feedback options. Learners can be given implicit feedback on their pronunciation in the form of graphical or textual visual representations of the prosody of the given utterance(s) (see Section 5.1.1), and via the auditory channel through the playback of original and prosodically-modified utterances (see Section 5.1.2 Explicit feedback can also be delivered, in the form of skill bars or verbal messages (see Section 5.2). A simple web interface gives instructors and researchers control over which feedback types are presented in a given exercise (see Section 5.4), thus making **[TODO de-stress]** a valuable new tool for research into the comparative efficacy of

various feedback types on pronunciation improvement, learner motivation, and other factors influencing the success of a CAPT system.

This thesis has thus presented original work taking steps towards the development of a comprehensive, intelligent CAPT system for French learners of German. Drawing from previous research on foreign-language pedagogy, phonetics/phonology, and speech technology, as well as original research on the frequency of lexical stress errors in learner speech and novel methods for diagnosing such errors, the **[TODO de-stress]** system developed in this thesis project is the first known CAPT tool dedicated to helping L1 French speakers improve their realization of lexical stress in German, and to facilitating research on the use of automatic, individualized feedback to help learners correct such errors. The following section suggests some possible directions for future work on improving **[TODO de-stress]** and using it to make further advancements in CAPT for German.

6.2 Future work

Although carrying out in vivo experiments with L2 German learners was outside the scope of this thesis project, the facilitation of such studies has been the primary motivation behind the development of the **[TODO de-stress]** CAPT tool described in this thesis. Therefore, the most logical and important extension of the work described here would be to conduct studies with L1 French learners of German using different combinations of diagnostic and feedback types, to determine if and when these types are effective in helping learners overcome difficulties with lexical stress prosody. Based on the findings of the error analysis described in Section 3.6, it may be prudent to focus studies on low-proficiency learners (especially children), as they seemingly stand to benefit most from such training. Once more is known about the conditions under which certain types of diagnosis/feedback are most effective, **[TODO de-stress]** could theoretically be transformed into a true Intelligent Tutoring System (ITS), in which models of the relevant aspects of the learning context could be used to automatically choose the best diagnostic/feedback method(s) for a given situation.

As an extension of the inter-annotator agreement analysis in Section 3.4, it would be interesting to analyze the utterances that had low inter-annotator agreement to find features that differentiate "clear" utterances (i.e. those labeled [correct] or [incorrect]) from "unclear" ones ([none]); Michaux and Caspers (2013) plan to undertake a similar analysis in their future work with French learners of Dutch. Additionally, further work is needed to shed light on the factors underlying the individual differences in reliability observed among annotators, as well as whether the differences in agreement on different word types might simply be a consequence of individual variation among the annotators assigned to each type, or whether other phenomena (e.g. frequency effects, segmental errors, or interference from French) might be at play.

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 a future version of **[TODO de-stress]**, a simple sentence- and/or word-level confidence measure for the segmentation of a given utterance could serve this function. While it is very difficult to compute such

a measure directly from the decoding scores of the forced aligner, it may be possible to evaluate the accuracy of the forced-alignment segmentations (cf. Mesbahi et al., 2011) to determine which types of boundaries the aligner typically places inaccurately (e.g. between a sonorant and a vowel), and then to calculate the proportion of error-prone boundaries in the given segmentation. Such a measure, while simplistic, could nevertheless provide some indication of when (not) to trust the automatic alignment, thus impacting decisions on how and whether to attempt error diagnosis/feedback. Another useful extension could be the inclusion of the type of error-filtering methods described by Bonneau et al. (2012) and Orosanu et al. (2012), which detect and reject utterances deviating from the expected text before alignment is attempted.

With respect to the diagnosis module of **[TODO de-stress]**, multiple directions for future improvements present themselves. First of all, for the selection of reference speaker(s) used in the comparison-based diagnosis approach, additional, more complex metrics of speaker similarity could be iconstructed using duration or intensity features, spectral features, and/or other features informed by research on speaker identification (e.g. Shriberg et al., 2005). Much work also remains to be done with respect to classification-based diagnosis, including evaluations using different machine learning algorithms and additional features which may be related to lexical stress in German (e.g. spectral features capturing aspects of vowel quality). To obtain a more fine-grained analysis of the learner's utterance in the classification approach, and thus enable the generation of additional types of feedback (e.g. skill bars) from classification-based diagnoses, different classifiers trained on only subsets of the available features could be combined in an ensemble configuration; for example, an ensemble of three classifiers trained exclusively on duration, F0, and intensity features respectively could produce scores analogous to JSnoori's three-dimensional pronunciation scores Section 4.3.1). It may also be of interest to explore techniques for online, semi-supervised learning, which could enable the classifier to improve its predictions by updating itself with new training instances resulting from the learner's continued use of the system over time.

Regarding future extensions of the feedback module, the graphical and text-based visualizations of prosody could be made more flexible with respect to the mapping(s) between prosodic features and graphical properties (e.g. size, opacity, color, etc.), with researchers and instructors being able to configure the mapping as they see fit, as in the system developed by Sitaram et al. (2011) (see Section 2.2.4). Another possible enhancement of these visualizations could be to add animation highlighting the difference between the learner's prosody and that of the reference(s) or target (e.g. by growing/shrinking a shape to convey a desired lengthening/shortening of that syllable), which might help the learner realize which corrections are necessary to make their speech more closely resemble the target. Regarding auditory feedback, it would be interesting to add a feedback option in which the L1 reference is modified to exaggerate the difference between stressed and unstressed syllables (cf. Bissiri and Pfitzinger, 2009; Bissiri et al., 2006). Finally, several additional types of feedback could be added to the variety currently implemented in **[TODO de-stress]**, such as metalinguistic feedback (e.g. reminders of stress placement rules in German), "gamification" (e.g. awarding points for correct realizations), or interactivity (e.g. allowing the learner to manipulate the prosody of an utterance via resynthesis).

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. 8).

Behme-Gissel, Helma (2005). *Deutsche Wortbetonung: ein Lehr- und Übungsbuch*. Iudicium (cit. on p. 79).

Bissiri, Maria Paola and Hartmut R. Pfitzinger (2009). "Italian speakers learn lexical stress of German morphologically complex words". In: *Speech Communication* (cit. on pp. 14, 94).

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. 14, 94).

Boersma, Paul and David Weenink (2014). *Praat: doing phonetics by computer* (cit. on pp. 21, 24).

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. 11, 12, 17, 18, 20, 54, 56, 59, 78, 82).

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. 11, 59).

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. 10, 52, 53, 94).

Bouselmi, Ghazi, Dominique Fohr, Irina Illina, and Jean Paul Haton (2005). "Fully automated non-native speech recognition using confusion-based acoustic model integration". In: *9th European Conference on Speech Communication and Technology (EUROSPEECH '05)* (cit. on pp. 10, 52).

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 pp. 10, 52).

Breiman, L, J Friedman, CJ Stone, and RA Olshen (1984). "Classification and regression trees". In: (cit. on p. 63).

Cohen, Jacob (1960). “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20.1, pp. 37–46 (cit. on p. 26).

Colotte, Vincent and Yves Laprie (2002). “Higher precision pitch marking for TD-PSOLA”. In: *Proceedings of the European Signal Processing . . .* (Cit. on p. 82).

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. 16).

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. 14, 15, 17, 54, 56, 58, 75).

Delmonte, Rodolfo (2011). “Exploring Speech Technologies for Language Learning”. In: *Speech and Language Technologies*. Ed. by Ivo Ipsic. InTech (cit. on pp. 9, 59).

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. 7–9, 16).

Di Martino, Joseph and Yves 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)*. Budapest, Hungary, p. 4 (cit. on p. 56).

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 pp. 8, 77, 82, 83).

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. 14, 54, 75).

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. 13, 62).

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. 15, 17).

Eskenazi, Maxine (2009). “An overview of spoken language technology for education”. In: *Speech Communication* 51.10, pp. 832–844 (cit. on pp. 9, 59, 93).

Eskenazi, Maxine and Scott Hansma (1998). “The Fluency pronunciation trainer”. In: *Proc. of Speech Technology in Language Learning*, pp. 77–80 (cit. on p. 12).

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 pp. 12, 13).

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. 13).

Fauth, Camille, Anne Bonneau, Frank Zimmerer, Jürgen Trouvain, Bistra Andreeva, Vincent Colotte, Dominique Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, Odile Mella, and Bernd Möbius (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. 4, 20, 45, 52, 53, 74).

Field, John (2005). “Intelligibility and the Listener: The Role of Lexical Stress”. In: *TESOL Quarterly* 39.3, p. 399 (cit. on pp. 7, 17).

Fohr, Dominique and Yves Laprie (1989). “Snorri: an interactive tool for speech analysis.” In: *First European Conference on Speech Communication and Technology (EUROSPEECH ’89)*. Paris, France, pp. 1613–1616 (cit. on p. 11).

Fohr, Dominique and Odile Mella (2012). “CoALT: A Software for Comparing Automatic Labelling Tools.” In: *LREC*, pp. 325–332 (cit. on p. 52).

Fohr, Dominique, Jean-François 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*, pp. 339–342 (cit. on p. 52).

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. 8, 16, 17).

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). “The WEKA data mining software: an update”. In: *ACM SIGKDD Explorations* 11.1, pp. 10–18 (cit. on p. 63).

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. 11, 59, 78, 82).

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 pp. 7, 8, 15).

Hirschfeld, Ursula (1994). *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender*. Vol. 57. Institut für Phonetik, JW Goethe-Universität (cit. on pp. 8, 15, 17).

Hirschfeld, Ursula and Kerstin Reinke (1998). *Phonetik Simsalabim: Ein Übungskurs für Deutschlernender (Begleitbuch)*. Langenscheidt (cit. on p. 78).

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. 79).

Jilka, Matthias and Gregor Möhler (1998). “Intonational foreign accent: speech technology and foreign language teaching”. In: *Proc. of the ESCA Workshop on Speech Technology in Language Learning*, pp. 115–118 (cit. on p. 13).

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”. In: *5th Language & Technology Conference - LTC’11*, pp. 145–148 (cit. on pp. 10, 52).

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. 18, 62, 63).

Kohler, Klaus J. (1996). “Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech”. In: *Fourth International Conference on Spoken Language Processing (ICSLP 96)* (cit. on p. 52).

Landis, J. Richard and Gary G. Koch (1977). “The measurement of observer agreement for categorical data.” In: *Biometrics* 33.1, pp. 159–174 (cit. on pp. 27, 28, 31, 50, 72).

Laprie, Yves (1999). “Snorri, a software for speech sciences”. In: *ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education (MATISSE)*. London, UK, pp. 89–92 (cit. on p. 11).

Laprie, Yves and Vincent Colotte (1998). “Automatic pitch marking for speech transformations via TD-PSOLA”. In: *European Signal Processing Conference (Eusipco)*, pp. 1–4 (cit. on p. 82).

Long, Yanjin and Vincent Aleven (2011). “Students’ Understanding of Their Student Model”. In: *Artificial Intelligence in Education (AIED)*. Springer-Verlag, pp. 179–186 (cit. on pp. 83, 85).

Long, Yanjin and Vincent Aleven (2013). “Supporting students’ self-regulated learning with an open learner model in a linear equation tutor”. In: *Artificial Intelligence in Education (AIED)*. Springer-Verlag, pp. 219–228 (cit. on pp. 83, 85).

Martin, P (1982). “Comparison of pitch detection by cepstrum and spectral comb analysis”. In: *Acoustics, Speech, and Signal Processing, IEEE* . . . (Cit. on p. 56).

Martin, Philippe (2004). “WinPitch LTL II, a multimodal pronunciation software”. In: *InSTIL/ICALL Symposium 2004* (cit. on pp. 14, 78).

Mehlhorn, Grit (2005). “Learner autonomy and pronunciation coaching”. In: *Proceedings of the Phonetics Teaching and Learning Conference, University College London* (cit. on pp. 8, 85).

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. 10, 13, 52, 53, 94).

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. 15, 17, 45).

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. 15, 21, 31, 93).

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. 13).

Moulines, Eric and Francis Charpentier (1990). “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”. In: *Speech Communication* 9.5-6, pp. 453–467 (cit. on p. 82).

Munro, Murray J. and Tracey M. Derwing (1999). “Foreign accent, comprehensibility, and intelligibility in the speech of second language learners”. In: *Language Learning* 49.Supplement s1, pp. 285–310 (cit. on p. 7).

Neri, A., C. Cucchiarini, H. Strik, and L. Boves (2002). “The pedagogy-technology interface in computer assisted pronunciation training”. In: *Computer Assisted Language Learning* (cit. on pp. 7–9, 14, 16, 21, 64, 69, 78, 85).

Ney, H (1981). “A dynamic programming technique for nonlinear smoothing”. In: *Acoustics, Speech, and Signal Processing, IEEE* . . . (Cit. on p. 56).

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. 10, 13, 52, 53, 94).

Peperkamp, Sharon and Emmanuel Dupoux (2002). "A typological study of stress 'deafness'". In: *Laboratory phonology* (cit. on p. 42).

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. 13, 61).

Project-Team PAROLE (2013). *Activity Report 2013*. Tech. rep. Nancy: LORIA (cit. on p. 11).

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. 52).

Secrest, B and GR Doddington (1983). "An integrated pitch tracking algorithm for speech systems". In: *Acoustics, Speech, and Signal . . .* (Cit. on p. 56).

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. 18, 62).

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. 94).

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. 13, 79, 94).

Trouvain, Jürgen, Yves Laprie, Bernd Möbius, Bistra Andreeva, Anne Bonneau, Vincent Colotte, Camille Fauth, Dominique Fohr, Denis Jouvet, Odile Mella, Jeanin Jügler, and Frank Zimmerer (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. 4, 20, 45, 53, 74).

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. 13).

Wik, P, R Hincks, and JB Hirschberg (2009). "Responses to Ville: A virtual language teacher for Swedish". In: (cit. on p. 18).

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, 7, 9, 16, 17).

Zimmerer, Frank and Jürgen Trouvain (2015). "Perception of French speakers' German vowels". In: *INTERSPEECH (submitted)* (cit. on p. 42).