

PHONETIC ACCOMMODATION OF HUMAN INTERLOCUTORS
IN THE CONTEXT OF HUMAN-COMPUTER INTERACTION

DISSERTATION

zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes



vorgelegt von
IONA GESSINGER
aus Stuttgart, Deutschland

Saarbrücken, im Januar 2022

Iona Gessinger: *Phonetic Accommodation of Human Interlocutors in the Context of Human-Computer Interaction* © Januar 2022

DEKAN DER FAKULTÄT P: Univ.-Prof. Dr. Augustin Speyer

ERSTGUTACHTER: Univ.-Prof. Dr. Bernd Möbius

ZWEITGUTACHTER: Univ.-Prof. Dr. Volker Dellwo

Tag der letzten Prüfungsleistung: 9. Dezember 2021

In liebevoller Erinnerung an Ingrid Amann.
1932 – 2020

ACKNOWLEDGEMENTS

I would like to thank

my supervisor, Bernd Möbius
for guiding my way and supporting me throughout this process,

Eran Raveh, Sébastien Le Maguer, Bistra Andreeva, and Ingmar Steiner
without whom some of this work would not have been possible,

Nauman Fakhra and Christine Mangold
who were instrumental in bringing Mirabella to life,

Johannah O'Mahony, Jens Neuerburg, and Katie Ann Dunfield
for much help in the lab and in working through the data,

Antje Schweitzer, Natalie Lewandowski, Nicolas Becker, and Katrin Menzel
for the expertise they shared with me,

Erika Brandt, Jürgen Trouvain, Cristina Deeg, Volker Dellwo, Katja Häuser,
Michael Hedderich, Yuri Bizzoni, Jan Michalsky, and Zaher Alchihabi
for valuable support at various levels,

Nicole Amann-Gessinger and Hartmut Gessinger
for everything.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project “Phonetic Convergence in Human-Computer Interaction” (project number: [278805297](#); IDs: MO 597/6-1,2 and STE 2363/1-1).

As an associate member of SFB1102 “Information Density and Linguistic Encoding” and its integrated graduate program, I enjoyed many opportunities for scientific training and exchange (project number: [232722074](#)).

ABSTRACT

Phonetic accommodation refers to the phenomenon that interlocutors adapt their way of speaking to each other within an interaction. This can have a positive influence on the communication quality. As we increasingly use spoken language to interact with computers these days, the phenomenon of phonetic accommodation is also investigated in the context of human-computer interaction: on the one hand, to find out whether speakers adapt to a computer agent in a similar way as they do to a human interlocutor, on the other hand, to implement accommodation behavior in spoken dialog systems and explore how this affects their users. To date, the focus has been mainly on the global acoustic-prosodic level.

The present work demonstrates that speakers interacting with a computer agent also identify locally anchored phonetic phenomena such as segmental allophonic variation and local prosodic features as accommodation targets and converge on them. To this end, we conducted two experiments. First, we applied the shadowing method, where the participants repeated short sentences from natural and synthetic model speakers. In the second experiment, we used the Wizard-of-Oz method, in which an intelligent spoken dialog system is simulated, to enable a dynamic exchange between the participants and a computer agent — the virtual language learning tutor *Mirabella*. The target language of our experiments was German.

Phonetic convergence occurred in both experiments when natural voices were used as well as when synthetic voices were used as stimuli. Moreover, both native and non-native speakers of the target language converged to *Mirabella*. Thus, accommodation could be relevant, for example, in the context of computer-assisted language learning.

Individual variation in accommodation behavior can be attributed in part to speaker-specific characteristics, one of which is assumed to be the personality structure. We included the *Big Five* personality traits as well as the concept of *mental boundaries* in the analysis of our data. Different personality traits influenced accommodation to different types of phonetic features. Mental boundaries have not been studied before in the context of phonetic accommodation. We created a validated German adaptation of a questionnaire that assesses the strength of mental boundaries. The latter can be used in future studies involving mental boundaries in native speakers of German.

KURZZUSAMMENFASSUNG

Bei phonetischer Akkommodation handelt es sich um das Phänomen, dass Gesprächspartner¹ ihre Sprechweise innerhalb einer Interaktion aneinander anpassen. Dies kann die Qualität der Kommunikation positiv beeinflussen. Da wir heutzutage immer öfter mittels gesprochener Sprache mit Computern interagieren, wird das Phänomen der phonetischen Akkommodation auch im Kontext der Mensch-Computer-Interaktion untersucht: zum einen, um herauszufinden, ob sich Sprecher an einen Computeragenten in ähnlicher Weise anpassen wie an einen menschlichen Gesprächspartner, zum anderen, um das Akkommodationsverhalten in Sprachdialogsysteme zu implementieren und zu erforschen, wie dieses auf ihre Benutzer wirkt. Bislang lag der Fokus dabei hauptsächlich auf der globalen akustisch-prosodischen Ebene.

Die vorliegende Arbeit zeigt, dass Sprecher in Interaktion mit einem Computeragenten auch lokal verankerte phonetische Phänomene wie segmentale allophone Variation und lokale prosodische Merkmale als Akkommodationsziele identifizieren und in Bezug auf diese konvergieren. Dabei wendeten wir in einem ersten Experiment die Shadowing-Methode an, bei der die Teilnehmer kurze Sätze von natürlichen und synthetischen Modellsprechern wiederholten. In einem zweiten Experiment ermöglichten wir mit der Wizard-of-Oz-Methode, bei der ein intelligentes Sprachdialogsystem simuliert wird, einen dynamischen Austausch zwischen den Teilnehmern und einem Computeragenten — der virtuellen Sprachlerntutorin *Mirabella*. Die Zielsprache unserer Experimente war Deutsch.

Phonetische Konvergenz trat in beiden Experimenten sowohl bei Verwendung natürlicher Stimmen als auch bei Verwendung synthetischer Stimmen als Stimuli auf. Zudem konvergierten sowohl Muttersprachler als auch Nicht-Muttersprachler der Zielsprache zu *Mirabella*. Somit könnte Akkommodation zum Beispiel im Kontext des computergestützten Sprachenlernens zum Tragen kommen.

Individuelle Variation im Akkommodationsverhalten kann unter anderem auf sprecherspezifische Eigenschaften zurückgeführt werden. Es wird vermutet, dass zu diesen auch die Persönlichkeitsstruktur gehört. Wir bezogen die *Big Five* Persönlichkeitsmerkmale sowie das Konzept der *mental*en Grenzen in die Analyse unserer Daten ein. Verschiedene Persönlichkeitsmerkmale beeinflussten die Akkommodation zu unterschiedlichen Typen von phonetischen Merkmalen. Die mentalen Grenzen sind im Zusammenhang mit phonetischer Akkommodation zuvor noch nicht untersucht worden. Wir erstellten eine validierte deutsche Adaptierung eines Fragebogens, der die Stärke der mentalen Grenzen erhebt. Diese kann in zukünftigen Untersuchungen mentaler Grenzen bei Muttersprachlern des Deutschen verwendet werden.

¹ Aus Gründen der Lesbarkeit wird bei Personenbezeichnungen die männliche Form gewählt, es ist jedoch immer die weibliche Form mitgemeint.

AUSFÜHRLICHE ZUSAMMENFASSUNG

Das Konzept der phonetischen Akkommodation bezieht sich auf artikulatorische und stimmliche Anpassungen, die ein Sprecher² als Reaktion auf den Sprachinput eines anderen Sprechers vornimmt. Infolgedessen kann die Sprache der Gesprächspartner einander ähnlicher oder weniger ähnlich werden. Das erste Verhalten wird als *Konvergenz* und das zweite als *Divergenz* bezeichnet. Dieses Phänomen wurde für die gesprochene Interaktion zwischen Menschen dokumentiert (z.B. Pardo, 2006; Levitan und Hirschberg, 2011; Lewandowski, 2012) und trägt zu deren Erfolg und Qualität bei (z.B. Lee u. a., 2010a; Manson u. a., 2013; Borrie u. a., 2015).

Neben einem Erklärungsansatz, der von einer automatischen Perzeptions-Produktions-Integration als Quelle der Konvergenz ausgeht (Pickering und Garrod, 2004; Pickering und Garrod, 2013; Pickering und Garrod, 2021), wird angenommen, dass die Akkommodation zur Regulierung der sozialen Distanz zwischen den Sprechern dient, wobei Konvergenz die Distanz verringert und Divergenz sie erhöht (Giles, 1973; Giles u. a., 1991; Shepard u. a., 2001; Gallois u. a., 2005; Giles, 2016). Zu den sozialen Faktoren, die nachweislich die Stärke und Richtung der phonetischen Akkommodation beeinflussen, gehören zum Beispiel die wahrgenommene Attraktivität und Sympathie eines Gesprächspartners (z.B. Babel u. a., 2014; Schweitzer und Lewandowski, 2014; Michalsky und Schoormann, 2017) und die hierarchische Beziehung zwischen den Gesprächspartnern (z.B. Gregory und Webster, 1996).

Als konkretes Beispiel für ein Szenario, das vom dynamischen Prozess der phonetischen Akkommodation profitieren kann, betrachten wir den Fall des Sprachenlernens in einer Klassenzimmerumgebung: Der Lehrer stellt eine fehlerhafte Aussprache des Lerners fest und weicht von dieser ab — möglicherweise bewusst. Für ein kontinuierliches Merkmal, z.B. die Vokalqualität, kann dies bedeuten, eine extremere Version zu produzieren. Für ein kategoriales Merkmal, z.B. die Wortendung ⟨-ig⟩ im Deutschen als allophone Varianten [iç] oder [ik], kann dies bedeuten, die bevorzugte Version zu betonen. Der Lerner nähert sich dann der Aussprache des Lehrers an — vor allem, wenn der Lehrer als sympathisch und hierarchisch überlegen wahrgenommen wird. Während zweifellos auch andere Faktoren zum Lernprozess beitragen, ist dies ein denkbare Szenario für Akkommodation in der Mensch-Mensch-Interaktion (HHI).³

Die zunehmende Häufigkeit von gesprochenen Interaktionen mit Computern in unserem Alltag wirft die Frage auf, ob eine solche dynamische phonetische Anpassung auch für die Mensch-Computer-Interaktion (HCI) relevant ist. Ein Aspekt, der eine solche Relevanz implizieren würde, ist der vermutete Beitrag der Akkommodation zum kommunikativen Erfolg und zur Dialogqualität. Denn es ist eines der übergreifenden Ziele in der HCI-Forschung, die Kommunikation mit einem Computer für den menschlichen Benutzer so angenehm wie möglich zu gestalten.

Parallel zum obigen Beispiel stellt das computergestützte Sprachenlernen (CALL) eine Situation dar, in der phonetische Akkommodation, insbesondere Konvergenz, für den Benutzer eines Sprachdialogsystem (SDS) wünschenswert wäre. CALL bietet eine niedrighwellige Möglichkeit, mit dem Erwerb einer Fremdsprache zu beginnen. Auf dem CALL-Markt findet man die ganze Bandbreite von einfachen, kostenlosen Anwendungen bis hin zu ausgefeilten Lernprogrammen. Einerseits wird oft argumentiert, dass eine CALL-Anwendung einen menschlichen Lehrer nicht vollständig ersetzen kann, insbesondere wenn es um phonetische Aspekte

² Aus Gründen der Lesbarkeit wird bei Personenbezeichnungen die männliche Form gewählt, es ist jedoch immer die weibliche Form mitgemeint.

³ Zur besseren Orientierung im Hauptteil der Dissertation führen wir in dieser Zusammenfassung bereits die englischen Abkürzungen ein.

der Kommunikation geht (z.B. Lee u. a., 2014). Andererseits ermöglichen Fortschritte in allen Bereichen der HCI-Forschung, insbesondere bei der automatischen Spracherkennung und der Text-to-Speech-Synthese, eine zunehmend erfolgreiche gesprochene Kommunikation zwischen menschlichen Lernern und virtuellen Lehrern. Unter der Voraussetzung, dass die Sprachausgabe der CALL-Anwendung von muttersprachlicher Qualität ist, würde eine Annäherung an diese zu einer Verbesserung der Produktion der gelernten Sprache auf Seiten des Benutzers führen. Insbesondere die Aussprache von Sprachsegmenten und die Realisierung prosodischer Phänomene, wie zum Beispiel die Frageintonation, bieten sich in diesem Zusammenhang als Ziele für Akkommodation an, da es für diese Merkmale klar definierte Standardrealisierungen gibt.

Die vorliegende Arbeit untersucht die phonetische Akkommodation von menschlichen Gesprächspartnern in Interaktion mit einem Computeragenten. Frühere Studien in diesem Kontext konzentrierten sich auf globale akustisch-prosodische Merkmale wie Grundfrequenz, Intensität und Sprechgeschwindigkeit (z.B. Bell u. a., 2003; Oviatt u. a., 2004; Staum Casasanto u. a., 2010; Gijssels u. a., 2016). Wir erweitern die bestehende Literatur durch die Untersuchung von eher lokal verankerten phonetischen Merkmalen. Der Kern der vorliegenden Arbeit ist im CALL-Bereich angesiedelt und besteht aus einer Interaktion mit einer virtuellen Tutorin zum Erlernen der deutschen Sprache, die *Mirabella* heißt.

Zunächst stellt sich jedoch die Frage, was HCI — genauer gesagt, eine Interaktion mit einem SDS — aus der Perspektive des menschlichen Benutzers ausmacht. Wir nähern uns der Interaktion mit einem Computer in zwei Schritten an:

Die erste Komponente von gesprochener HCI ist ihre Schnittstelle, nämlich die gesprochene Sprache selbst. Bei einem SDSs findet normalerweise synthetische Sprache Verwendung. Daher untersuchten wir, ob Versuchspersonen phonetische Merkmale von synthetischen Stimmen in ähnlicher Weise übernehmen wie von natürlichen (d.h. menschlichen) Stimmen. Zu diesem Zweck führten wir ein *Shadowing-Experiment* mit natürlicher und synthetischer Sprache durch.

Die zweite, vielleicht entscheidendere HCI-Komponente ist die Überzeugung des Benutzers, tatsächlich mit einem Computer zu interagieren (Branigan u. a., 2010). Um eine überzeugende Interaktion mit einem Computeragenten zu ermöglichen und gleichzeitig die untersuchten phonetischen Merkmale kontrollieren zu können, führten wir ein *Wizard-of-Oz (WOz)-Experiment* durch, in dem das SDS *Mirabella* simuliert wurde. Auch hier verwendeten wir natürliche und synthetische Sprache.

Ein Aspekt der Akkommodationsforschung, der zunehmend an Relevanz gewinnt, ist die Untersuchung der individuellen Unterschiede zwischen Sprechern, um die häufig beobachtete Variation im akkommodierenden Verhalten zu erklären (z.B. Pardo u. a., 2018; Weise u. a., 2019).

Zu den untersuchten potenziellen Einflussfaktoren gehören das phonetische Talent in der Sprachproduktion, -perzeption und -imitation sowie kognitive Maße wie Arbeitsgedächtnis, Aufmerksamkeitsspanne und mentale Flexibilität (z.B. Lewandowski und Jilka, 2019; Yu u. a., 2013). Der Einfluss der Sprecherpersönlichkeit auf die Akkommodation wird ebenfalls untersucht, typischerweise innerhalb des *Big Five* Persönlichkeitsparadigmas (Costa und McCrae, 1992).

Im Rahmen des WOz-Experiments untersuchten wir ebenfalls die *Big Five* Persönlichkeitsfaktoren. Darüber hinaus betrachten wir ein Konzept, das im Kontext der phonetischen Akkommodation noch nicht untersucht wurde, nämlich das der *mentalen Grenzen* (Hartmann u. a., 1987; Hartmann, 1989; Hartmann, 1991). Die Durchlässigkeit der mentalen Grenzen, die anhand des Boundary Questionnaire (BQ) quantifiziert werden kann, veranschaulicht zum einen den Grad der Vernetzung innerhalb des menschlichen Geistes und zum anderen zwischen dem Geist und der ihn umgebenden Welt. Die mentalen Grenzen wurden als potenzieller Einflussfaktor auf die Aussprache von Fremdsprachen vorgeschlagen (z.B. Guiora u. a., 1972; Ehrman, 1999; Więckowska, 2011; Baran-Lucarz, 2012) — letztlich eine Manifestation der phonetischen Akkommodation

— und sind daher auch ein möglicher Indikator für die phonetische Akkommodation in anderen Kontexten.

Wir stellen eine deutsche Adaptierung der empirisch abgeleiteten Kurzversion des Boundary Questionnaire (BQ-Sh; Rawlings, 2001) vor, um die Berücksichtigung dieses Persönlichkeitsfaktors in der Akkommodationsforschung mit deutschen Muttersprachlern zu ermöglichen. Als ersten Anwendungsfall analysierten wir die WOz-Daten im Hinblick auf den Einfluss mentaler Grenzen auf die phonetische Akkommodation.

Die Shadowing- und WOz-Experimente sowie die Adaptierung des BQ-Sh werden im Folgenden ausführlicher beschrieben.

SHADOWING-EXPERIMENT In diesem Experiment wiederholten Muttersprachler des Deutschen kurze deutsche Sätze, unmittelbar nachdem sie diese von einem weiblichen oder einem männlichen Modellsprecher gehört hatten, deren Stimmen entweder natürlich oder synthetisch waren. Die Annahme war, dass die Versuchspersonen ihre Sprache an die der Modellsprecher anpassen würden. Genauer gesagt, erwarteten wir Konvergenz in Richtung der Modellsprecher.

Zu den untersuchten phonetischen Phänomenen gehören allophone Kontraste und Schwa-Epenthese als Phänomene auf Segmentebene und die Realisierung von Tonakzenten als Phänomene der lokalen Prosodie. Darüberhinaus betrachteten wir die wortbasierte zeitliche Struktur und die Verteilung der spektralen Energie als globale Ähnlichkeitsmaße. Diese wurden zuvor im Zusammenhang mit Akkommodation nur als kombiniertes Merkmal untersucht (Lewandowski, 2012; Lewandowski und Jilka, 2019). Die allophonen Kontraste sind [ɛ:] vs. [e:] als Realisierung des Langvokals ⟨-ä-⟩, z.B. in *Mädchen*, und [ɪç] vs. [ɪk], das in der Wortendung ⟨-ig⟩ vorkommt, z.B. in *König*. Die Analyse einer so großen Anzahl von Merkmalen verschiedener phonetischer Ebenen ermöglicht eine umfassende Beurteilung des Akkommodationsverhaltens der Versuchspersonen.

Unseres Wissens nach, ist dies die erste Studie, welche die Akkommodation dieser phonetischen Phänomene beim Nachsprechen kurzer Äußerungen untersucht und sowohl natürliche als auch synthetische Stimuli als Akkommodationsziele verwendet.

Auf individueller Ebene stellten wir fest, dass die Versuchspersonen zu unterschiedlichen Teilmengen der untersuchten Merkmale konvergierten, während sie in anderen Fällen ihr Ausgangsverhalten beibehielten oder in seltenen Fällen sogar von den Modellstimmen divergierten. Dies bestätigt, dass Akkommodation in Bezug auf ein phonetisches Merkmal nicht unbedingt das Verhalten in Bezug auf ein anderes Merkmal vorhersagt. Dies war zuvor für akustisch-prosodische Merkmale in der HHI belegt worden (z.B. Cohen Priva und Sanker, 2018; Reichel u. a., 2018; Weise und Levitan, 2018).

Auf Ebene der Gruppe konvergierten die Versuchspersonen in der natürlichen Bedingung bezüglich aller untersuchten Merkmale zu den Modellsprechern, allerdings nur sehr schwach im Falle der Schwa-Epenthese. Die synthetischen Stimmen reduzierten zwar teilweise die Stärke der Effekte, die für die natürlichen Stimmen gefunden wurden, lösten aber auch akkommodierendes Verhalten aus. Das vorherrschende Muster für alle Stimmtypen war Konvergenz während der Interaktion, gefolgt von Divergenz nach der Interaktion. Wir konnten also zeigen, dass Sprecher eher lokal verankerte phonetische Merkmale sowohl von natürlichen als auch von synthetischen Stimmen übernehmen — auch wenn die Merkmale in kurze Äußerungen eingebettet sind.

Das Shadowing-Experiment bildet [Kapitel 3](#) dieser Dissertation. Das Kapitel basiert auf den folgenden veröffentlichten, von Experten begutachteten Konferenz- und Zeitschriftenartikeln: Gessinger u. a. (2016), Raveh u. a. (2017a), Gessinger u. a. (2017), Gessinger u. a. (2018) und Gessinger u. a. (2021b).

WIZARD-OF-OZ-EXPERIMENT Wir wendeten die WOz-Methode an, bei der die Versuchspersonen mit einem vermeintlich intelligenten SDS interagieren, während ein Experimentator die Ausgabe des Systems im Hintergrund kontrolliert (Kelley, 1984; Dahlbäck u. a., 1993). Wir untersuchten die phonetische Akkommodation in Bezug auf lokale Prosodie, genauer gesagt die Platzierung des nuklearen Tonakzents in W-Fragen und die diesem Tonakzent folgende finale Intonationskontur, sowie in Bezug auf die deutschen Allophonpaare [ɛː]/[eː] und [ɪç]/[ɪk], die bereits im Shadowing-Experiment untersucht wurden. Die verschiedenen Varianten dieser Merkmale sind im Standarddeutschen akzeptiert.

Um die Interaktion zu motivieren und eine alltagsnahe Situation zu simulieren, wurde den Versuchspersonen das SDS als Tutorin zum Erlernen von Deutsch als Fremdsprache vorgestellt. Der Name der Tutorin war *Mirabella*.

In einer ersten Bedingung bestanden Mirabellas Äußerungen aus natürlicher Sprache, in einer zweiten Bedingung wurden sie durch synthetische Sprache ersetzt. Muttersprachler des Deutschen (natürliche und synthetische Bedingung) und Muttersprachler des Französischen (nur natürliche Bedingung) interagierten mit Mirabella. Da wir davon ausgehen, dass die Faktoren, die die phonetische Akkommodation in HHI auslösen, auch für die Interaktion mit einer virtuellen Person wie Mirabella gelten, unabhängig davon, ob sie mit einer echten menschlichen oder einer synthetischen Stimme spricht, erwarteten wir, dass die Akkommodation in Bezug auf die untersuchten Merkmale in beiden Bedingungen auftritt, aber möglicherweise in unterschiedlichem Ausmaß für die L1- und L2-Sprecher des Deutschen. Zusätzlich analysierten wir den Einfluss der *Big Five* Persönlichkeitsmerkmale der Sprecher auf ihr akkommodierendes Verhalten.

Soweit wir wissen, ist dies die erste Studie, welche die Akkommodation der Frageintonation und der segmentalen Aussprache mit der WOz-Methode untersucht.

Die L1-Deutsch-Sprecher konvergierten zu Mirabella in Bezug auf die modifizierte Frageintonation, nämlich eine ansteigende F_0 -Kontur und die Platzierung des nuklearen Tonakzents auf dem Interrogativpronomen, und in Bezug auf den allophonen Kontrast [ɪç]/[ɪk]. Im Falle des allophonen Kontrastes [ɛː]/[eː] behielten die Versuchspersonen ihr Ausgangsverhalten bei. Die Ergebnisse unterschieden sich nicht zwischen den Versuchsgruppen, die entweder mit der natürlichen oder mit der synthetischen Version von Mirabella kommunizierten. Die L2-Deutsch-Sprecher zeigten ein ähnliches Akkommodationsmuster. Allerdings passten sie sich nicht an den verschobenen nuklearen Tonakzent der W-Fragen an. Auf der Ebene der einzelnen Sprecher fanden wir wiederum erhebliche Variation in Bezug auf Akkommodationsgrad und -richtung.

Die Untersuchung des Einflusses der *Big Five* Persönlichkeitsmerkmale auf das akkommodierende Verhalten zeigte eine Tendenz, dass gewissenhaftere L1-Deutsch-Sprecher häufiger zu Mirabellas Version von ⟨-ig⟩ konvergieren und dass Neurotizismus die Konvergenz zur Frageintonation beeinflusst.

Wir konnten also zeigen, dass phonetische Akkommodation auf der Ebene der lokalen Prosodie und der segmentalen Aussprache bei Benutzern von SDSs unter Verwendung von natürlicher und synthetischer Sprachausgabe auftritt. Außerdem demonstrierten wir, dass dies auch bei Nicht-Muttersprachlern der Zielsprache der Fall ist, was im Kontext des computergestützten Sprachenlernens ausgenutzt werden könnte. Abhängig von der Art des phonetischen Merkmals scheinen jedoch unterschiedliche Persönlichkeitsfaktoren die Konvergenz zu begünstigen.

Das WOz-Experiment bildet [Kapitel 4](#) dieser Dissertation. Das Kapitel basiert auf den folgenden veröffentlichten, von Experten begutachteten Konferenz- und Zeitschriftenartikeln: Gessinger u. a. (2019b), Gessinger u. a. (2019a), Gessinger u. a. (2020) und Gessinger u. a. (2021a).

MENTAL BOUNDARIES Wir übersetzten die BQ-Sh von Rawlings (2001) ins Deutsche und validierten die resultierende Übersetzung mit einer Gruppe von L1-Deutsch-Sprechern verschiedener Altersgruppen und Bildungshintergründe. Wir erhoben auch das NEO Fünf-Faktoren-

Inventar mit den Versuchspersonen des Validierungsprozesses, um zu untersuchen, inwieweit sich mentale Grenzen mit den *Big Five*-Persönlichkeitsmerkmalen überschneiden, da letztere auch im Kontext der phonetischen Akkommodation untersucht werden (z.B. Lewandowski und Jilka, 2019; Yu u. a., 2013). Die strukturelle Validität der deutschen Adaptierung des BQ-Sh (BQ-Sh-G) wurde durch den Vergleich von Werten der zentralen Tendenz, der Variabilität und der internen Konsistenz mit denen des BQ-Sh und der Durchführung einer Maximum-Likelihood-Faktorenanalyse nachgewiesen. Die in der Literatur berichteten Zusammenhänge zwischen dünnen mentalen Grenzen und erhöhten Werten von Neurotizismus und Offenheit, sowie zwischen dicken mentalen Grenzen und Gewissenhaftigkeit, traten auch in den vorliegenden Daten auf.

Unseres Wissens nach ist dies die erste validierte deutsche Adaptierung der BQ-Sh.⁴

Wir wendeten die BQ-Sh-G innerhalb des WOz-Experiments an. Die L1-Deutsch-Versuchspersonen bildeten eine repräsentative Stichprobe in Bezug auf mentale Grenzen. Angesichts der vorangegangenen Analyse der WOz-Daten bezüglich des Einflusses der *Big Five* und der Korrelationen der *Big Five* mit den mentalen Grenzen, nahmen wir an, dass dickere Grenzen die Konvergenz von [ɪç]/[ɪk] begünstigen würden und dünnere Grenzen die Konvergenz der Frageintonation. Während wir Belege für die erste Annahme fanden, konnte die zweite Annahme in unseren Daten nicht bestätigt werden.

Wir stellen damit ein Instrument zur Untersuchung mentaler Grenzen in der Akkommodationsforschung mit L1-Deutsch-Sprechern zur Verfügung und demonstrierten dessen Anwendung.

Die Entwicklung des BQ-Sh-G und der Anwendungsfall mit den WOz-Daten bilden [Kapitel 5](#) dieser Dissertation. Dieses Kapitel ist bisher nicht veröffentlicht worden.

HAUPTBEITRÄGE Dieser Abschnitt fasst die Hauptbeiträge der vorliegenden Dissertation zusammen.

- Shadowing-Experiment
 - Wir zeigen, dass Sprecher bezüglich lokal verankerter phonetischer Merkmale, die in kurze Äußerungen eingebettet sind, konvergieren.
 - Auf der Ebene der wortbasierten spektralen Eigenschaften zeigen wir, dass Konvergenz im Hinblick auf die Verteilung spektraler Energie auch dann stattfindet, wenn die Konvergenz in Bezug auf die zeitliche Struktur separat betrachtet wird.
 - Die beobachteten Konvergenzeffekte treten sowohl bei natürlichen als auch bei verschiedenen Arten von synthetischen Stimmen auf, wobei sie bei letzteren teilweise abgeschwächt sind.
- Wizard-of-Oz-Experiment
 - Wir präsentieren einen interaktiven Austausch mit einem simulierten Sprachdialogsystem namens *Mirabella* in einem computergestützten Sprachlernszenario mit Deutsch als Zielsprache und vergleichen den Effekt von natürlicher und synthetischer Sprache.
 - Wir zeigen, dass Muttersprachler der Zielsprache sowohl im Falle der natürlichen als auch der synthetischen Version von *Mirabella* auf den Ebenen der lokalen Prosodie und der segmentalen Aussprache phonetisch konvergieren.

⁴ Es ist uns bekannt, dass einige Autoren im Bereich der Traumforschung auf eine deutsche Übersetzung der BQ-Langversion verweisen, die vom Institut für Psychologie der Universität Zürich vermutlich Ende der 1990er Jahre übersetzt wurde (e.g., Strauch und Meier, 1999; Funkhouser u. a., 2001; Schredl und Erlacher, 2004). Unseres Wissens nach ist diese Übersetzung jedoch nicht veröffentlicht worden und es liegen keine Informationen über den Adaptierungsprozess vor.

- Für Nicht-Muttersprachler der Zielsprache zeigen wir ebenfalls einen Konvergenzefekt. Dieser Effekt ist im Vergleich zu dem der Muttersprachler für strukturelle phonologische Elemente, die sich radikal von ihrem eigenen muttersprachlichen Muster unterscheiden, abgeschwächt.
- Bezüglich des Einflusses der *Big Five* auf das Akkommodationsverhalten präsentieren wir Hinweise darauf, dass verschiedene Persönlichkeitsfaktoren unterschiedliche Arten von phonetischen Merkmalen beeinflussen.
- Mentale Grenzen
 - Wir präsentieren eine validierte Übersetzung der empirisch abgeleiteten Kurzversion des Boundary Questionnaire als Instrument zur Untersuchung mentaler Grenzen bei L1-Deutsch-Sprechern, z.B. in der Akkommodationsforschung.
 - Wir demonstrieren die Anwendung dieses Instruments für die Teilnehmer des Wizard-of-Oz-Experiments und untersuchen die Vermutung, dass Konvergenz in Bezug auf verschiedene Arten von phonetischen Merkmalen durch dünnere bzw. dickere mentale Grenzen begünstigt werden kann.

PUBLICATIONS

Parts of this dissertation have appeared previously in the following publications:

- Gessinger, I., B. Möbius, B. Andreeva, E. Raveh, and I. Steiner (2019a). “Phonetic accommodation in a Wizard-of-Oz experiment: intonation and segments.” In: *Interspeech*. Graz, pp. 301–305. DOI: [10.21437/Interspeech.2019-2445](https://doi.org/10.21437/Interspeech.2019-2445).
- (2020). “Phonetic accommodation of L2 German speakers to the virtual language learning tutor Mirabella.” In: *Interspeech*. Shanghai, pp. 4118–4122. DOI: [10.21437/Interspeech.2020-2701](https://doi.org/10.21437/Interspeech.2020-2701).
- Gessinger, I., B. Möbius, N. Fakhar, E. Raveh, and I. Steiner (2019b). “A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction.” In: *International Congress of Phonetic Sciences (ICPhS)*. Melbourne, pp. 1475–1479. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1524.pdf.
- Gessinger, I., B. Möbius, S. Le Maguer, E. Raveh, and I. Steiner (2021a). “Accommodation in interaction with a virtual language learning tutor: a Wizard-of-Oz study.” In: *Journal of Phonetics* 86, p. 101029. DOI: [10.1016/j.wocn.2021.101029](https://doi.org/10.1016/j.wocn.2021.101029).
- Gessinger, I., E. Raveh, S. Le Maguer, B. Möbius, and I. Steiner (2017). “Shadowing synthesized speech – segmental analysis of phonetic convergence.” In: *Interspeech*. Stockholm, pp. 3797–3801. DOI: [10.21437/Interspeech.2017-1433](https://doi.org/10.21437/Interspeech.2017-1433).
- Gessinger, I., E. Raveh, J. O’Mahony, I. Steiner, and B. Möbius (2016). “A shadowing experiment with natural and synthetic stimuli.” In: *Phonetik & Phonologie*. Munich, pp. 58–61. DOI: [10.5282/ubm/epub.29405](https://doi.org/10.5282/ubm/epub.29405).
- Gessinger, I., E. Raveh, I. Steiner, and B. Möbius (2021b). “Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing.” In: *Speech Communication* 127, pp. 43–63. DOI: [10.1016/j.specom.2020.12.004](https://doi.org/10.1016/j.specom.2020.12.004).
- Gessinger, I., A. Schweitzer, B. Andreeva, E. Raveh, B. Möbius, and I. Steiner (2018). “Convergence of pitch accents in a shadowing task.” In: *International Conference on Speech Prosody*. Poznań, pp. 225–229. DOI: [10.21437/SpeechProsody.2018-46](https://doi.org/10.21437/SpeechProsody.2018-46).
- Raveh, E., I. Gessinger, S. Le Maguer, B. Möbius, and I. Steiner (2017). “Investigating phonetic convergence in a shadowing experiment with synthetic stimuli.” In: *Conference on Electronic Speech Signal Processing (ESSV)*. Ed. by J. Trouvain, I. Steiner, and B. Möbius. Saarbrücken, pp. 254–261. URL: <http://essv2017.coli.uni-saarland.de/pdfs/Raveh.pdf>.

CONTENTS

1	INTRODUCTION	1
1.1	Shadowing experiment	3
1.2	Wizard-of-Oz experiment	3
1.3	Mental boundaries	4
1.4	Main contributions	5
2	BACKGROUND	7
2.1	Theoretical frameworks	7
2.2	Phonetic features	9
2.3	Experimental settings	10
2.3.1	Shadowing experiments	10
2.3.2	Conversational HCI experiments	11
2.4	Interlocutor characteristics	13
2.5	Voice type	14
2.6	Speaker disposition	14
2.7	Mental boundaries	15
3	SHADOWING EXPERIMENT	19
3.1	Hypotheses and predictions	19
3.2	Material and methods	20
3.2.1	Corpus	20
3.2.1.1	Stimuli	21
3.2.1.2	Participants	23
3.2.2	Analyzed features	24
3.2.2.1	Allophones and schwa epenthesis	24
3.2.2.2	Pitch accent comparison with PaIntE	26
3.2.2.3	Word-level amplitude envelope analysis	28
3.2.3	Further factors	29
3.3	Analysis and results	30
3.3.1	Modeling	30
3.3.2	Segmental pronunciation	31
3.3.3	Pitch accent realization	35
3.3.4	Word-level spectral composition	36
3.3.5	Individual results	38
3.4	Discussion	42
3.4.1	Word ending <-ig>	42
3.4.2	Word-level spectral composition	43
3.4.3	Long vowel <-ä-> and pitch accent realization	44
3.4.4	Word ending <-en>	45
3.4.5	Individual behavior	45
3.4.6	Limitations of difference in distance	46
3.4.7	Natural vs. synthetic speech	48
3.4.8	Model voices	49
3.5	Conclusion	49
4	WIZARD-OF-OZ EXPERIMENT	51
4.1	Hypotheses and predictions	51

4.2	Material and methods	52
4.2.1	Tasks and tested features	53
4.2.1.1	Task 1 — allophonic variation, baseline	53
4.2.1.2	Task 2 — question intonation, baseline	55
4.2.1.3	Task 3 — question intonation, test	56
4.2.1.4	Task 4 — allophonic variation, test	57
4.2.2	Text material	58
4.2.3	Stimuli	59
4.2.4	Participants	60
4.3	Analysis and results	62
4.3.1	Rating of Mirabella	62
4.3.2	Modeling	62
4.3.3	Question intonation	63
4.3.4	Long vowel <-ä->	67
4.3.5	Word ending <-ig>	69
4.3.6	Personality scores	73
4.4	Discussion	76
4.4.1	Reception of Mirabella	77
4.4.2	Accommodation to Mirabella	77
4.4.2.1	Question intonation	78
4.4.2.2	Word ending <-ig>	80
4.4.2.3	Long vowel <-ä->	82
4.4.2.4	Individual behavior	82
4.5	Conclusion	83
5	MENTAL BOUNDARIES	87
5.1	Instruments	87
5.1.1	The Boundary Questionnaire	87
5.1.2	Short versions of the BQ	88
5.1.3	German versions of the BQ	89
5.1.4	Boundaries and the Big Five	90
5.2	Methods	91
5.2.1	Translation	91
5.2.2	Validation	91
5.2.2.1	Participants	91
5.2.2.2	Statistical analysis	92
5.3	Results	92
5.3.1	Structural validity	92
5.3.2	Convergent and discriminant validity	96
5.4	Discussion	97
5.4.1	Structural validity	97
5.4.2	Convergent and discriminant validity	98
5.5	Use case: Mental boundaries and phonetic accommodation	99
5.5.1	Data	99
5.5.2	Analysis and results	101
5.5.3	Discussion	103
5.6	Conclusion	105
6	GENERAL DISCUSSION	107
7	CONCLUSION AND OUTLOOK	113

BIBLIOGRAPHY	115
A SHADOWING: TEXT MATERIAL	127
B WOZ: EXPLAINING UTTERANCES	129
C WOZ: GUIDING UTTERANCES	133
D WOZ: TARGET AND FILLER WORDS	135
E WOZ: QUESTIONS FROM FRAGMENTS AND ANSWERS	137
F WOZ: MAP TASK PREPOSITIONS	139
G WOZ: MAPS	141
H WOZ: INDIVIDUAL DIFFERENCE IN DISTANCE DISTRIBUTIONS	145
I THE GERMAN ADAPTATION OF THE BQ-SH: BQ-SH-G	149

ACRONYMS

AIC	Akaike information criterion
BFI	Big Five Inventory
BNO ^R	Bedürfnis nach Ordnung
BPQ	Boundary Personality Questionnaire
BQ	Boundary Questionnaire
BQ-SH	empirically-derived short version of the Boundary Questionnaire
BQ-SH-G	German adaptation of the BQ-Sh
BQ18	18-item version of the BQ
CALL	computer-assisted language learning
CAT	Communication Accommodation Theory
CEFRL	Common European Framework of Reference for Languages
CH	Childlikeness
DID	difference in distance
DTW	dynamic time warping
GLMM	generalized linear mixed-effects model
HCI	human-computer interaction
HHI	human-human interaction
HMM	Hidden Markov model
HTS	HMM-based Speech Synthesis System
IAM	Interactive Alignment Model
KI	Kindlichkeit
LMM	linear mixed-effects model
LTAS	long-term average spectrum
MLFA	Maximum Likelihood factor analysis
NEO-FFI	NEO Five Factor Inventory
NEO-PI	NEO Personality Inventory
NFO ^R	Need for Order
PAINTE	Parametric Intonation Event

PC ^R	Perceived Competence
Q&A	question-and-answer
SDS	spoken dialog system
SE	Sensitivity
SE	Sensitivität
TR	Trust
UE	Unusual Experiences
UE	Ungewöhnliche Erfahrungen
VE	Vertrauen
VOT	voice onset time
WK ^R	Wahrgenommene Kompetenz
WOZ	Wizard-of-Oz

1 INTRODUCTION

Phonetic accommodation refers to articulatory and vocal adjustments made by a speaker in response to being exposed to speech input from another speaker. As a consequence, the speech of the interlocutors may become more similar or less similar to each other. The former behavior is called *convergence* and the latter *divergence*. This phenomenon has been documented for human spoken interaction (e.g., Pardo, 2006; Levitan and Hirschberg, 2011; Lewandowski, 2012) and contributes to its success and quality (e.g., Lee et al., 2010a; Manson et al., 2013; Borrie et al., 2015).

In addition to an explanatory approach that proposes automatic perception–production integration as a source of convergence (Pickering and Garrod, 2004; Pickering and Garrod, 2013; Pickering and Garrod, 2021), it is assumed that accommodation serves to regulate social distance, with convergence increasing proximity and divergence decreasing it (Giles, 1973; Giles et al., 1991; Shepard et al., 2001; Gallois et al., 2005; Giles, 2016). Social factors that have been found to influence the strength and direction of phonetic accommodation in human-human interaction (HHI) include, for example, the perceived attractiveness and likability of an interlocutor (e.g., Babel et al., 2014; Schweitzer and Lewandowski, 2014; Michalsky and Schoormann, 2017) and the hierarchical relationship between interlocutors (e.g., Gregory and Webster, 1996). For more details on the theoretical frameworks within which accommodation is studied, see [Section 2.1](#).

As a concrete example of a HHI scenario that may benefit from the dynamic process of phonetic accommodation, consider the case of language learning in a classroom setting: The teacher detects incorrect pronunciation on the part of the learner and diverges from it — possibly intentionally. For a continuous feature, e.g., vowel quality, this may imply producing a more extreme version, whereas for a categorical feature, e.g., German word ending ⟨-ig⟩ as [ɪç] or [ɪk], this may mean emphasizing the preferred version. The learner then converges to the teacher’s pronunciation — especially if the teacher is perceived as likable and hierarchically superior. While other factors certainly contribute to the learning process, the latter is a conceivable scenario for accommodation in HHI.

The increasing amount of spoken interactions with computers in our everyday life raises the question whether such dynamic phonetic adaptation is also relevant for human-computer interaction (HCI). One aspect that would imply such relevance is the supposed contribution of accommodation to communicative success and dialog quality, since it is one of the overarching goals in HCI research to make communication with a computer more pleasant for the human user.

Parallel to the above example, a computer-assisted language learning (CALL) context constitutes a situation in which phonetic accommodation, in particular convergence, would be desirable for the user of a spoken dialog system (SDS). CALL offers a low-threshold opportunity to start acquiring a foreign language. There is a wide range of options on the market, from simple, free applications to sophisticated learning programs. On the one hand, it is often argued that a CALL application cannot fully replace a human teacher, especially when it comes to phonetic aspects of communication (e.g., Lee et al., 2014). On the other hand, progress in all areas of the HCI domain, particularly automatic speech recognition and text-to-speech synthesis, enables increasingly successful spoken communication between human learners and virtual teachers. Provided that the speech output of the CALL application is of native-like quality, converging to it would lead to an improvement in the production of the learned language on

the part of the user. Especially the pronunciation of speech segments and the realization of prosodic phenomena such as question intonation, lend themselves as targets for accommodation in this context, as there exist clearly defined standard realizations for these features.

The present work examines phonetic accommodation of human interlocutors in interaction with a computer agent. Prior studies in this setting have focused on global acoustic-prosodic features, such as F_0 , intensity, and speaking rate (e.g., Bell et al., 2003; Oviatt et al., 2004; Staum Casasanto et al., 2010; Gijssels et al., 2016). We expand the existing literature by examining more locally anchored phonetic features. The core of the present thesis is situated in the CALL domain and consists of an interaction with a virtual tutor for learning the German language whose name is *Mirabella*.

First, however, arises the question of what constitutes HCI — more specifically, an interaction with a SDS — from the perspective of human users. We approach the interaction with a computer in two steps:

The first component of spoken HCI is its interface, namely the speech itself. In the case of SDSs, the output usually consists of synthetic speech. Therefore, we investigated whether speakers adopt the phonetic features of interest to a similar extent from synthetic voices as from natural (i.e., human) voices. To this end, we conducted a *speech shadowing* experiment using natural and synthetic speech.

The second, perhaps more decisive, component of HCI is the users' belief that they are in fact interacting with a computer (Branigan et al., 2010). In order to enable a convincing interaction with a computer agent while being able to control the phonetic features of interest, we then conducted a *Wizard-of-Oz (WOz) experiment* simulating a SDS, namely *Mirabella*, again using natural and synthetic speech.

The two experiments are outlined in [Section 1.1](#) and [Section 1.2](#).

An emerging aspect of accommodation research is the exploration of the individual differences between speakers in order to explain the commonly observed variation in accommodating behavior among speakers (e.g., Pardo et al., 2018; Weise et al., 2019).

The potential influencing factors under investigation include phonetic talent in speech production, perception and imitation, as well as cognitive measures such as working memory and attention span (e.g., Lewandowski and Jilka, 2019; Yu et al., 2013). The impact of speaker personality on accommodation is also studied, typically within the *Big Five* personality paradigm (Costa and McCrae, 1992).

As part of the WOz experiment, we also examined the *Big Five* personality factors. In addition, we consider a concept that has not yet been explored in the context of phonetic accommodation, namely *mental boundaries* (Hartmann et al., 1987; Hartmann, 1989; Hartmann, 1991). The permeability of mental boundaries, which can be quantified using the Boundary Questionnaire (BQ), illustrates both the degree of interconnectedness within the mind and between the mind and the surrounding world. Mental boundaries have been suggested as a potential factor influencing the performance in foreign language pronunciation (e.g., Guiora et al., 1972; Ehrman, 1999; Więckowska, 2011; Baran-Łucarz, 2012) — ultimately a manifestation of phonetic accommodation itself — and therefore seem to be a likely candidate for predicting phonetic accommodation.

We present a German adaptation of the empirically-derived short version of the Boundary Questionnaire (BQ-Sh; Rawlings, 2001) in order to enable the consideration of this personality factor in accommodation research involving native speakers of German. As a first use case, we analyzed the WOz data with respect to the influence of mental boundaries on phonetic accommodation.

The adaptation and analysis are outlined in [Section 1.3](#).

1.1 SHADOWING EXPERIMENT

In this experiment, native speakers of German repeated short German sentences immediately after hearing them from a female or a male model speaker. The voices were either natural or synthetic. For the synthetic voices, we used diphone-based synthesis and Hidden Markov model (HMM)-based synthesis. The overall assumption was that the participants would accommodate their speech to that of the model speakers. More specifically, we expected to find convergence towards the model speakers.

The phonetic phenomena under examination include allophonic contrasts and schwa epenthesis as segment-level phenomena and the realization of pitch accents as a phenomenon of local prosody. We also included the word-based temporal structure and distribution of spectral energy as measures of global similarity, which have previously been studied in the context of accommodation as a combined feature only (Lewandowski, 2012; Lewandowski and Jilka, 2019). The allophonic contrasts are [ɛ:] vs. [e:] as a realization of the long vowel ⟨-ä-⟩, e.g., *Mädchen* (girl), and [ɪç] vs. [ɪk] occurring in the word ending ⟨-ig⟩, e.g., *König* (king). Conducting an analysis of accommodation on such a diverse set of features pertaining to different phonetic domains allows for an extensive assessment of the participants' behavior.

To the best of our knowledge, this is the first study that investigates the accommodation of such phenomena when shadowing short utterances and includes both natural and synthetic speech stimuli as accommodation targets.

On the individual level, we found that the participants converged to varying subsets of the examined features, while they maintained their baseline behavior in other cases or, in rare instances, even diverged from the model voices. This confirms that accommodation with respect to one particular phonetic feature does not necessarily predict the behavior with respect to another feature, which was previously documented for acoustic-prosodic features in HHI (e.g., Cohen Priva and Sanker, 2018; Reichel et al., 2018; Weise and Levitan, 2018).

On the group level, the participants of the natural condition converged to all features under examination, however very subtly so for schwa epenthesis. The synthetic voices, while partly reducing the strength of effects found for the natural voices, triggered accommodating behavior as well. The predominant pattern for all voice types was convergence during the interaction followed by divergence after the interaction. Hence, we were able to show that speakers also adopt more locally anchored phonetic features from both natural and synthetic voices — even when they are embedded in short utterances.

The shadowing experiment constitutes Chapter 3 of this thesis. This chapter is based on the following published peer-reviewed conference and journal articles: Gessinger et al. (2016), Raveh et al. (2017a), Gessinger et al. (2017), Gessinger et al. (2018), and Gessinger et al. (2021b).

1.2 WIZARD-OF-OZ EXPERIMENT

We applied the WOz method, in which participants interact with a supposedly intelligent SDS while an experimenter is controlling the output of the system behind the scenes (Kelley, 1984; Dahlbäck et al., 1993). We examined phonetic accommodation with respect to local prosody, more precisely the placement of the nuclear pitch accent in wh-questions and the final intonation contour following this nuclear accent, as well as with respect to the German allophone pairs [ɛ:]/[e:] and [ɪç]/[ɪk] that were already under investigation in the shadowing experiment. The different variants of these features are accepted in Standard German.

To motivate the interaction and simulate a situation similar to that one might encounter in everyday life, the SDS was presented to the participants as a tutoring system for learning German as a foreign language. We named the tutor *Mirabella*.

In a first condition, Mirabella’s utterances consisted of natural speech and in a second condition, they were replaced by synthetic speech. Native speakers of German (natural and synthetic condition) and native speakers of French (only natural condition) interacted with Mirabella. Since we assume that the factors triggering phonetic accommodation in HHI also apply to the interaction with a virtual persona like Mirabella, whether she speaks with a genuine human or a synthetic voice, we expected accommodation with regard to the examined features to occur in both conditions, but possibly to different degrees for the L1 and L2 German speakers. Additionally, we analyzed the influence of the speakers’ *Big Five* personality traits on their accommodating behavior.

To the best of our knowledge, this is the first study examining accommodation to question realization and segmental pronunciation with the WOz method.

The L1 German speakers converged to Mirabella with respect to modified wh-question intonation, i.e., rising F_0 contour and nuclear pitch accent on the interrogative pronoun, and the allophonic contrast [ɪç]/[ɪk]. They did not accommodate to the allophonic contrast [ɛː]/[eː]. The results did not differ between the experimental groups that communicated with either the natural or the synthetic speech version of Mirabella. The L2 German speakers showed a similar pattern of accommodation. However, they did not adapt to the shifted nuclear pitch accent in the wh-questions. On the level of individual speakers, we again found considerable variation with respect to the degree and direction of accommodation.

Testing the influence of the *Big Five* personality traits on the accommodating behavior revealed a tendency for more conscientious L1 German speakers to converge more often to Mirabella’s version of ⟨-ig⟩ and for Neuroticism to influence the convergence to question intonation.

We thus demonstrated that phonetic accommodation at the level of local prosody and segmental pronunciation occurs in users of SDSs using natural and synthetic speech output. We showed that this also occurs for non-native speakers of the target language, which could be exploited in the context of computer-assisted language learning. However, depending on the type of phonetic feature, different personality factors seem to be conducive to convergence.

The WOz experiment constitutes [Chapter 4](#) of this thesis. This chapter is based on the following published peer-reviewed conference and journal articles: Gessinger et al. (2019b), Gessinger et al. (2019a), Gessinger et al. (2020), and Gessinger et al. (2021a).

1.3 MENTAL BOUNDARIES

We translated the BQ-Sh by Rawlings (2001) into German and validated the resulting translation with a group of L1 German speakers of various age groups and educational backgrounds. We also collected NEO Five Factor Inventory (NEO-FFI) data for participants in the validation process to examine the extent to which mental boundaries intersect with the *Big Five* personality traits, since the latter are also investigated in the context of phonetic accommodation (e.g., Lewandowski and Jilka, 2019; Yu et al., 2013). The structural validity of the German adaptation of the BQ-Sh (BQ-Sh-G) was demonstrated through comparing values of central tendency, variability, and internal consistency with those of the BQ-Sh and performing a Maximum Likelihood factor analysis (MLFA). The correlations reported in the literature between thin mental boundaries and elevated levels of Neuroticism and Openness, as well as between thick mental boundaries and Conscientiousness, were reflected in the present data.

To the best of our knowledge, this is the first validated German adaptation of the BQ-Sh.¹

¹ We are aware that some authors in the field of dream research refer to a German translation of the BQ long version, which was translated by the Department of Psychology at the University of Zurich probably at the end of the 1990s (e.g., Strauch and Meier, 1999; Funkhouser et al., 2001; Schredl and Erlacher, 2004). However, to our

We applied the BQ-Sh-G within the WOz experiment. The L1 German participants constituted a representative sample in terms of mental boundaries. The previous analysis of the WOz data using the *Big Five* and the correlations of the latter with mental boundaries suggested that thicker boundaries would favor convergence of [ɪç]/[ɪk] and thinner boundaries would promote convergence of question intonation. While we found evidence for the first assumption, the second assumption could not be confirmed in our data.

We thus provide an instrument for investigating mental boundaries in accommodation research with L1 German speakers and demonstrated its application.

The development of the BQ-Sh-G and the use case involving the WOz data constitute [Chapter 5](#) of this thesis. This chapter has not previously been published.

1.4 MAIN CONTRIBUTIONS

This section summarizes the main contributions of the present thesis.

- Shadowing experiment
 - We demonstrate that speakers converge on locally anchored phonetic features, which are embedded in short utterances.
 - At the level of word-based spectral properties, we show that convergence also happens for the distribution of spectral energy when convergence with respect to the temporal structure is considered separately.
 - The observed convergence effects occur with both natural and different types of synthetic voices, although they are partly attenuated for the latter.
- Wizard-of-Oz experiment
 - We present an interactive exchange with a simulated spoken dialog system called *Mirabella* in a computer-assisted language learning scenario with German as the target language and directly compare the effect of natural and synthetic speech.
 - We demonstrate that native speakers of the target language phonetically converge at the level of local prosody and segmental pronunciation for both the natural and synthetic version of *Mirabella*.
 - For non-native speakers of the target language, we also demonstrate a convergence effect. This effect is attenuated compared to that of the native speakers for structural phonological elements that differ radically from their own native pattern.
 - Regarding the influence of the *Big Five* on accommodation behavior, we present evidence that different personality factors may affect different types of phonetic features.
- Mental boundaries
 - We present a validated translation of the Boundary Questionnaire’s empirically-derived short version as an instrument for investigating mental boundaries with L1 German speakers, e.g., in accommodation research.
 - We demonstrate the application of this instrument for the participants of the Wizard-of-Oz experiment and explore the assumption that convergence with respect to different types of phonetic features can be favored by thinner or thicker mental boundaries, respectively.

knowledge, this translation has not been published and there is no information available about the adaptation process.

2 BACKGROUND

2.1 THEORETICAL FRAMEWORKS

The phenomenon of inter-speaker accommodation is often assessed within the framework of the *Communication Accommodation Theory (CAT)*, a generalized model of communicative interaction (Giles, 1973; Giles et al., 1991; Shepard et al., 2001; Gallois et al., 2005; Giles, 2016). The theory assumes interpersonal conversation to be a dynamic adaptive exchange of verbal and nonverbal behavior. During this exchange, the listener-speaker directs their attention to the speech of the interlocutor and adjusts their own speech as a way of reducing or increasing social distance to the interlocutor and thereby maintaining positive personal and social identities. According to CAT, convergence will therefore occur when social distance should be decreased — as opposed to divergence, which will occur when social distance should be increased. Convergence is motivated by the desire for social approval from the interlocutor or their social group, which positively reinforces the speaker’s identity. In the case of divergence, this positive reinforcement is achieved by setting oneself apart from the interlocutor or their social group. In addition to this *affective* function, CAT also identifies a *cognitive* function of accommodation that results from the speaker’s desire to facilitate comprehension and improve communicative efficiency. This suggests that accommodating behavior is socially motivated and, to some extent, consciously controlled by the speaker.

In line with this social motivation, the direction and extent of phonetic accommodation have been found to depend on factors such as the attitude or the hierarchical relationship towards an interlocutor. For example, it has been shown that an increase in the likability of a conversational partner led to a stronger convergence effect for vowel quality (Schweitzer and Lewandowski, 2014) and fundamental frequency (Michalsky and Schoormann, 2017). However, Schweitzer et al. (2017) report that a decrease in likability promoted both convergence and divergence with respect to pitch accent realization. Results by Gregory and Webster (1996), analyzing long-term average spectra (LTAS), suggest that speakers on the lower end of the hierarchy, or in a less dominant role, converge to the hierarchically higher or more dominant interlocutor.

The *Interactive Alignment Model (IAM)* (Pickering and Garrod, 2004; Pickering and Garrod, 2013; Pickering and Garrod, 2021) represents another point of view, which is reflected in the use of the term *alignment* instead of *accommodation*. Where *accommodation* allows for both converging and diverging behavior, *alignment* necessarily leads to an increased similarity, hence convergence. The model postulates that it is a priming mechanism which leads to alignment between the cognitive representations of the interlocutors during a conversation. The concrete phonetic convergence we observe is thus a manifestation of this low-level alignment. This suggests that accommodating behavior is an automatic process which is triggered subconsciously.

Both theories, as opposing as they might appear at first sight, concurrently agree that convergence is deeply rooted in human communicative behavior. They are not mutually exclusive, as both the social motivation and the automatic process can coexist and may vary in dominance between individuals, which could partially explain the fact that different speakers exhibit different degrees of accommodation. Pickering and Garrod (2021) also point out that alignment within the IAM does not necessarily have to be completely automatic, but can involve a strategic element. However, they argue that maintaining misaligned expressions requires considerable mental effort.

A model of phonetic accommodation combining the *automatic approach* (IAM) and the *social approach* (CAT), as for example suggested by Krauss and Pardo (2004), Babel (2010), and Coles-Harris (2017), is likely to be a better approximation of the actual phenomenon than restricting oneself to either one of the theories.

Assuming such a combined model of phonetic accommodation suggests that convergence represents the unmarked behavior. Divergence would then be expected in cases where a speaker either aims to increase social distance or to counteract extreme behavior of an interlocutor, presumably hoping for them to converge, such as in slowing down a very fast-talking speaker. In these cases, the unmediated tendency to converge may be superseded by a more dominant social motivation to diverge.

The hybrid model of convergence proposed by Lewandowski (2012) goes into further detail and specifies a consciously accessible layer of factors that includes the situational context, the identity and social group membership of the speaker and the interlocutor, as well as the evaluation of the interlocutor, for example in terms of their likability. The influence of these factors on the accommodation process is by and large unconscious. However, it can be consciously deployed if necessary. The very process of accommodation is not consciously accessible and directly determined by factors such as talent — in the present case phonetic talent — along with attention and memory components, as well as personality and other psychological factors, e.g., the need for social approval. An indispensable prerequisite for convergence is the linguistic ability to implement the feature in question. The resulting accommodation effect may under certain circumstances enter conscious perception and serve as feedback for the initially described consciously accessible layer of the model.

This comprehensive set of factors must be kept in mind when assessing accommodating behavior, even if not all factors can be controlled for. Depending on the social complexity of the communicative interaction the automatic processes and the socially motivated components may have different weights (Coles-Harris, 2017). Sections 3.1 and 4.1 discuss their respective relevance for the situational context of the shadowing and Wizard-of-Oz (WOz) experiments at hand. Regarding the evaluation of the interlocutor, we had the participants rate the voices used in the experiments in terms of their likability — in Mirabella’s case, also in terms of her competence (see Sections 3.2.1.1 and 4.3.1). The areas of personality and identity are incorporated into this research through the *Big Five* personality traits and the exploration of mental boundaries (see Sections 4.3.6 and 5.5).

When referring to the interlocutor in the context of spoken interaction, we usually picture this to be another human. What are the predictions of the theoretical frameworks discussed above in the case of human-computer interaction (HCI)?

For the automatic processes it does not matter whether we are communicating with a fellow human or a computer — converging behavior is expected in both cases. The socially motivated components require that the interlocutor is perceived as a social actor, an attribute that we may not intuitively assign to a computer. However, it has been observed that computers can indeed be perceived as social actors and that people exhibit social behavior towards them (Nass et al., 1994), but this does not necessarily apply to every person equally (Lee et al., 2010b). The concept was established as the *Computers are Social Actors* paradigm (Reeves and Nass, 1996; Nass and Moon, 2000). It is reasonable to assume that this status is becoming more established as the development of speech synthesis strives for more naturalness and interactions with SDSs evolve from simple commands to free conversations. The design of a computer agent can further contribute to the degree of personification, for example, by having a name or being represented by an avatar.

It is thus consistent with existing theoretical frameworks for a virtual interlocutor to elicit phonetic accommodation in a human speaker. If the latter believes that convergence is par-

ticularly beneficial for successful communication with a computer, for example because the computer relies on a certain speaking style to understand, the accommodation effect in HCI may be even greater than in communication with a fellow human (Branigan et al., 2010).

2.2 PHONETIC FEATURES

When considering which phonetic features to choose in order to study accommodation, there are many possibilities. One approach is to evaluate accommodation holistically by measuring the perceptual similarity of utterances (e.g., Goldinger, 1998; Namy et al., 2002; Kim et al., 2011; Miller et al., 2013; Babel et al., 2014; Dias and Rosenblum, 2016; Lewandowski and Nygaard, 2018; Clopper and Dossey, 2020). Apart from the perceptual approach, there are global acoustic measures which have been applied to estimate overall accommodation, such as the long-term average spectrum (LTAS; Gregory and Webster, 1996), mel-frequency cepstral coefficients (Delvaux and Soquet, 2007), and amplitude envelopes (Lewandowski, 2012; Lewandowski and Jilka, 2019). A next step in substantiating these holistic findings is to examine the global acoustic-prosodic level, e.g., by measuring accommodation in overall or turn-based fundamental frequency, intensity, or speaking rate (e.g., Coulston et al., 2002; Bell et al., 2003; Ward and Litman, 2007; Levitan and Hirschberg, 2011; Lubold and Pon-Barry, 2014; Michalsky and Schoormann, 2017). Eventually, more local phenomena are targeted, such as vowel quality (e.g., Babel, 2010; Babel, 2012; Nguyen et al., 2012; Dufour and Nguyen, 2013; Lewandowski and Nygaard, 2018; Clopper and Dossey, 2020), voice onset time (VOT; e.g., Fowler et al., 2003; Abrego-Collier et al., 2011; Nielsen, 2011; Yu et al., 2013), pitch accents (Schweitzer et al., 2017), or allophonic variation (e.g., Mitterer and Ernestus, 2008; Honorof et al., 2011; Mitterer and Müsseler, 2013).

While the holistic measures give a more comprehensive impression of accommodation, the acoustic-prosodic and segmental features are more tangible and can be better incorporated into synthetic speech if the goal is, for example, that the computer should also adapt to the human interlocutor.

Prior studies investigating whether humans accommodate to the speech output of spoken dialog systems (SDSs) have, to the best of our knowledge, exclusively examined global acoustic-prosodic features (e.g., Bell et al., 2003; Oviatt et al., 2004; Suzuki and Katagiri, 2007; Staum Casasanto et al., 2010; Gijssels et al., 2016; Raveh et al., 2019). The present thesis focuses on more locally anchored phonetic features, with the shadowing experiment also bridging to global acoustic measures.

In detail, the shadowing experiment examines the following features: On the global level, we use amplitude envelopes to characterize the distribution of spectral energy of individual target words within the short sentences uttered in the shadowing task (see Section 3.2.2.3). On the level of local prosody, we compare pitch accent realization in these sentences by parameterizing their shapes with the Parametric Intonation Event (PaIntE) model (see Section 3.2.2.2). Further, we examine the variation of the German allophones [ɛ:] vs. [e:] as a realization of the long vowel ⟨-ä-⟩ in stressed syllables, e.g., *Bestätigung* (*confirmation*), and [ɪç] vs. [ɪk] as a realization of the word ending ⟨-ig⟩, e.g., *Essig* (*vinegar*), as well as the epenthesis of schwa in a context where schwa is usually elided, namely in the word ending ⟨-en⟩ when preceded by a plosive or a fricative, e.g., *begleiten* (*accompany*; see Section 3.2.2.1).

Amplitude envelopes have been demonstrated to be useful in accounting for phonetic convergence. Lewandowski (2012), for example, showed for a human-human interaction (HHI) corpus of quasi-spontaneous dialogs between non-native and native speakers of English that the amplitude envelopes of tokens of the same word uttered by the interlocutors become more similar over time. We expected this to happen during the present shadowing task as well (see Section 3.3.4).

The analysis of pitch accent realization as parameterized by the PaIntE model is motivated by Schweitzer et al. (2017), who showed that native speakers of German accommodate their realization of pitch accents during spontaneous HHI dialogs in the GERman COversations corpus (Schweitzer et al., 2014; Schweitzer et al., 2015). They diverged when they could see each other and converged when they could not see each other. Based on the design of our experiment, in which the participants do not see the model speakers they are shadowing, we therefore expected convergence to occur with respect to pitch accent realization (see Section 3.3.3).

The segmental phenomena for which accommodation in the form of convergence has been demonstrated to occur include vowel quality, motivating our choice of the vowel contrast [ɛ:] vs. [e:] (e.g., Babel, 2012; Dufour and Nguyen, 2013), as well as the German allophone pair [ɪç] vs. [ɪk] (Mitterer and Müsseler, 2013) in the present study. The pronunciation variation [ɲ] vs. [əŋ] has, to our knowledge, not been studied in a shadowing experiment so far. Although the present shadowing experiment uses longer utterances as stimuli to investigate segment-level phenomena than previous shadowing experiments (see Section 2.3), we expected similar accommodation effects to occur as long as the variations are clearly perceptible (see Section 3.3.2).

The WOz experiment continues the analysis of the allophone pairs [ɪç]/[ɪk] and [ɛ:]/[e:] (see Sections 4.2.1.1 and 4.2.1.4). Furthermore, in terms of local prosody, we examine the placement of the nuclear pitch accent in wh-questions and the final intonational contour following this nuclear accent (see Sections 4.2.1.2 and 4.2.1.3). To our knowledge, this is the first experiment examining accommodation of question intonation in a HCI context. We expected convergence with respect to the investigated features (see Sections 4.3.4, 4.3.5, and 4.3.3).

2.3 EXPERIMENTAL SETTINGS

The experimental settings in which phonetic accommodation has been observed include dynamic, conversational approaches (e.g., Pardo, 2006; Levitan and Hirschberg, 2011; Kim et al., 2011; Lewandowski, 2012; Schweitzer et al., 2017; Michalsky and Schoormann, 2017), and also less interactive tasks, such as consecutive speech shadowing, in which a participant repeats an utterance immediately after hearing it from a model speaker¹ (e.g., Shockley et al., 2004; Babel et al., 2014; Walker and Campbell-Kibler, 2015; Dias and Rosenblum, 2016; Pardo et al., 2017; Lewandowski and Nygaard, 2018; Clopper and Dossey, 2020). The former qualify as fully social scenarios and therefore likely trigger the social motivation for accommodation. A shadowing task, on the other hand, generates a socially impoverished interaction during which the automatic motivation to accommodate might be predominant.

2.3.1 Shadowing experiments

In contrast to previous shadowing experiments investigating accommodation, the participants of the present shadowing experiment repeat full sentences instead of single, often mono- or bisyllabic (non-)words.

We provide a few examples of previous work examining adaptation at the segmental level:

Babel (2012), for example, asked participants to repeat low frequency monosyllabic English words (e.g., *breeze*, *smash*) “as clearly and naturally as possible” (p.180) after a male model speaker and measured the difference in distance in the F1–F2 formant space. While testing various vowels, [æ] and [ɑ] showed the strongest convergence effect. A possible explanation offered by Babel (2012) is the fact that these low vowels vary regionally in North American

¹ To be distinguished from *close speech shadowing* where speech input is repeated while it is still ongoing.

English. This might have resulted in a greater a priori phonetic distance between participants and model speaker, hence more space for the participants to converge to the latter.²

Dufour and Nguyen (2013) used bisyllabic French words ending in /e/ (e.g., *beauté*, *soirée*) or /ɛ/ (e.g., *projet*, *jamaïs*) and measured F1 to test whether speakers of Southern French, who usually produce both endings as [e], converge to a Standard French model speaker, who differentiates [e] and [ɛ]. After hearing a word from the female model speaker, participants were either asked to shadow it (“repeat it as naturally and as clearly as possible”, p.3) or to imitate it (“repeat it by imitating the speaker’s specific pronunciation”, p.3). A convergence effect was found for both groups; however, it was stronger in the imitation group. For the shadowing group, it only occurred for words that had not been used in a pre-test, i.e., words participants heard for the first time during the shadowing task.

In Mitterer and Müsseler (2013) participants repeated bisyllabic German words (e.g., *spielen*, *Stunde*, *fertig*, *Käfig*) and non-words (e.g., *spümen*, *streipen*, *onsig*, *wüssig*) “as quickly as possible” (p.561) after a female model speaker to test the influence of being confronted with different phonetic implementations of the fricative-stop clusters, namely [ʃp]/[ʃt] vs. [sp]/[st], and the word ending ⟨-ig⟩, namely [ɪç] vs. [ɪk]. [ʃp]/[ʃt] and [ɪç] are the Standard German forms. [sp]/[st] are Northern German realizations of the fricative-stop clusters, while [ɪk] is a Southern German realization of the word ending ⟨-ig⟩. However, Mitterer and Müsseler (2013) state that the two variations “differ clearly in their markedness” (p.560), with the fricative-stop cluster variation being undisputedly dialectal and the ⟨-ig⟩ variation having a rather unclear status. Both variations were imitated by the participants, with the more salient fricative-stop clusters showing a stronger effect. Most corrections occurred for the word ending ⟨-ig⟩ from stimulus [ɪç] to participant production [ɪk].

The fact that participants in the present study shadow full sentences moves the task from mere repetition slightly in the direction of conversational interaction. Shadowing short words entails a narrow focus and facilitates attention to phonetic detail, while shadowing longer utterances requires a broader focus and leads to higher cognitive load, as it is the case in fully conversational interaction.

We acknowledge the fact that the accommodating behavior of a speaker can vary between experimental settings, as has been shown by Pardo et al. (2018) comparing perceptual similarity between speakers and model talkers in conversational interaction and speech shadowing. However, since the pronunciation variations are embedded in full sentences in the present study, they are less salient and thus less obvious targets for accommodation. Under these circumstances, although staying within the rather static shadowing paradigm, occurrence of accommodation may be more readily transferable to actual dialog.

2.3.2 Conversational HCI experiments

The body of literature exploring whether humans also accommodate to the speech output of SDSs is growing (e.g., Bell et al., 2003; Oviatt et al., 2004; Staum Casasanto et al., 2010; Gijssels et al., 2016; Raveh et al., 2019). As mentioned above, the phonetic features examined in this context are mainly of global acoustic-prosodic nature. With the exception of Raveh et al. (2019), who studied a commercially available SDS without manipulating its speech output, all of the mentioned HCI studies applied the Woz method to simulate intelligent SDSs. In a Woz setup, users think that they are interacting with an autonomous system, but in reality it is the *wizard*, i.e., the experimenter, who takes the decisions about the system’s responses (Kelley, 1984; Dahlbäck et al., 1993).

² See Section 3.4.6 for a discussion of a possible underlying starting distance bias addressed by MacLeod (2021).

The ability for actual autonomous phonetic accommodation on the part of the computer is not yet developed — there are, however, a few suggestions for possible implementations with respect to specific phonetic features (e.g., Levitan et al., 2016; Raveh et al., 2017b). Developing computers that are themselves capable of phonetic adaptation is complementary to the research on user behavior. Specifically for the application in computer-assisted language learning (CALL), a synergy of the computer detecting erroneous productions by the user, diverging from them to give room for accommodation, and finally the user converging to the computer would probably be an ideal solution.

Until this becomes a reality, the WOz method enables dynamic conversational exchanges between users and a simulated system. Most importantly, this method provides direct control over the speech output of the latter, allowing specific phonetic features to be tested.

We briefly summarize the above-mentioned WOz studies:

Bell et al. (2003), for example, created the embodied graphical agent *Cloddy Hans* who helped the visitors of a museum to solve a puzzle. During the interaction, *Cloddy Hans* either had a slow or a fast speaking rate and his human interlocutors adapted their own speaking rate accordingly, i.e., participants interacting with the fast version spoke faster than those interacting with the slow version. The experiment addressed the issue of hyperarticulation when speaking to a computer. Several instances of misunderstanding on the part of *Cloddy Hans* were deliberately introduced in the dialog, which led to local effects on speaking rate. *Cloddy Hans*' utterances consisted of manipulated natural speech recordings.

Oviatt et al. (2004) explored the behavior of children between 7 and 10 years of age when using the interactive learning platform *I SEE!* to talk to male and female animated marine animals. The animals represented opposite ends of the introvert-extrovert personality spectrum. Their speech was synthesized and varied with respect to intensity, pitch range, utterance duration, pause duration, and response latency. The children adapted the intensity and durational features of their speech to different interlocutors in a bidirectional manner, i.e., in opposite directions when talking to an extrovert or an introvert voice.

In Staum Casasanto et al. (2010), participants entered an immersive virtual reality environment set in a supermarket to talk to the humanoid virtual agent *VIRTUO* about the items in the store. The agent talked with either a fast or a slow speaking rate, which was achieved by manipulating natural speech recordings. Compared to their baseline speaking rate measured before the interaction with *VIRTUO*, the participants of the fast condition spoke significantly faster during the interaction. The participants of the slow condition maintained their baseline speaking rate.

In an extension of the above study, Gijssels et al. (2016) explored whether participants also accommodate on F_0 . In this study, a female agent, *VIRTUA*, was tested as well. Both *VIRTUO* and *VIRTUA* talked with either high or low F_0 , which was again achieved by manipulating natural speech recordings. The participants of the high condition spoke significantly higher during the interaction than the participants of the low condition. A post testing block revealed that this effect did not last beyond the conversation with the system.

These studies demonstrated that humans exhibit accommodating behavior with respect to global acoustic-prosodic features when conversing with virtual interlocutors.³ Combined with the results from HHI research, this supports our assumption that such behavior may also occur with the more locally anchored phonetic features investigated in the present WOz experiment.

³ Another line of research focuses on acoustic-prosodic accommodation on the part of the SDS and its effect on the way the virtual agents are perceived by human users with respect to traits such as social presence, likability, competence, or trustworthiness (e.g., Lubold et al., 2016; Levitan et al., 2016; Gauder et al., 2018; Beňuš et al., 2018; Gálvez et al., 2020).

2.4 INTERLOCUTOR CHARACTERISTICS

It has been shown that the interlocutor significantly influences the degree of phonetic accommodation exhibited by a speaker. We present a selection of influential factors that have been raised in this context:

First of all, the attitude towards the interlocutor in terms of their perceived attractiveness and likability can have an influence on accommodation. Schweitzer and Lewandowski (2014), for example, found a strengthened convergence effect for first and second vowel formants with increasing perceived likability of the interlocutor. In an analysis of measures related to fundamental frequency by Michalsky and Schoormann (2017), likability was predictive of convergence as well, but increasing perceived attractiveness was a stronger predictor. In Schweitzer et al. (2017), however, effects of convergence and divergence with respect to pitch accent realization were more pronounced when speakers disliked their respective interlocutor. These three studies treated conversational interaction. In a shadowing experiment evaluating similarity perceived by listeners, Babel et al. (2014) found that the effect of attractiveness on convergence only applied to female speakers.

Furthermore, the hierarchy between speaker and interlocutor can be relevant for phonetic accommodation. Results by Gregory and Webster (1996) analyzing LTAS in conversational interaction suggest that speakers on the lower end of the hierarchy or in a less dominant role, converge to the hierarchically higher or more dominant interlocutor.

The aspect of social dominance may also play a role for the following factor, namely whether the speaker's sex matches that of the interlocutor (cf. Bilous and Krauss, 1988). Analyses of this factor have yielded varied outcomes. Levitan et al. (2012), for example, found more convergence of acoustic-prosodic features such as fundamental frequency, intensity, voice quality, and speaking rate in the conversational interaction of mixed-sex dyads as opposed to same-sex dyads. Bailly and Martin (2014), on the other hand, observed stronger convergence for same-sex dyads in an analysis of vowel spectra and global convergence as assessed by means of a speaker recognition technique. In a role-neutral conversational map task, eventually, Pardo et al. (2017) found no difference between same-sex and mixed-sex dyads in terms of perceptual similarity.

A final factor, examined by Babel et al. (2014) as well, concerns the typicality of the interlocutor's voice. Typicality was quantified here by means of a speeded identification task to determine the ease of speaker sex classification as female or male. In the evaluation of similarity perceived by listeners, men showed convergence only towards the more atypical voices, whereas women converged to the typical voices as well.

Apart from the obvious differences between settings and features, it should be mentioned that these studies examined various languages situated in different societies whose influence, especially with respect to social factors, needs to be taken into account (German/Germany in Schweitzer and Lewandowski (2014), Michalsky and Schoormann (2017), and Schweitzer et al. (2017), French/France in Bailly and Martin (2014), and English/USA in Gregory and Webster (1996), Levitan et al. (2012), and Babel et al. (2014)).

All of these factors potentially influence the outcome of the present studies.

For instance, the voice typicality factor can be quantified in various other ways than as the ease of female-male-classification. In this thesis, a more prominent source of atypicality is the use of synthetic voices in the experiments. Synthetic voices emulate human voices, but we can assume that they are to some extent not typical of human voices. If there is an effect of atypicality promoting convergence for certain groups of speakers, the synthetic voices may be more successful here — unless there is a threshold of how atypical a voice can be before such an effect is inhibited or even reversed to divergence.

The hierarchical situation in the shadowing experiment is unclear. Does the model speaker who is shadowed by the participant play a dominant role? This may vary according to the perception of the participants. In the case of the WOz experiment, it is possible that the computer agent is perceived as hierarchically inferior due to its machine nature and relative inflexibility. However, it is also possible that the agent is perceived as superior because it provides missing information and, in the case of the L2 German participants, it is perceived as a native speaker of the target language.

At any rate, only the participants can accommodate to the model speakers in the shadowing experiment and to Mirabella in the WOz experiment, since the stimuli in both experiments are prefabricated and played back to the participants during the tasks.

To assess the perceived likability of all voices used in the experiments, we collected respective ratings from the participants. Information about the sex of participant and interlocutor is taken into consideration in the analysis of the data, where applicable.

2.5 VOICE TYPE

Studies comparing the use of natural and synthetic speech in SDSs for tutoring showed that the pre-recorded natural version of a system is sometimes favored by users and can even be more conducive to learning than its synthetic counterpart (e.g., Baylor et al., 2003; Atkinson et al., 2005). Forbes-Riley et al. (2006), in contrast, found almost no influence of a virtual tutor’s voice on learning gain, system usability, or dialog efficiency.

The perception of the virtual interlocutor’s voice is also influenced by whether the agent is graphically represented. In Baylor et al. (2003), students were most motivated when interacting with a graphically animated agent that spoke with a synthetic voice, or with an agent that had a natural human voice and was not graphically animated.

The WOz studies examining phonetic accommodation described in Section 2.3.2 used either manipulated natural speech recordings (Bell et al., 2003; Staum Casasanto et al., 2010; Gijssels et al., 2016) or synthesized speech (Oviatt et al., 2004), and most of them used embodied graphical agents to represent the computer interlocutor, be they humanoid (*Cloddy Hans* in Bell et al., 2003; *VIRTUO/VIRTUA* in Staum Casasanto et al., 2010; Gijssels et al., 2016) or zoomorphic (various marine animals in Oviatt et al., 2004).

In Section 2.4, we noted that the atypicality of a synthetic voice could possibly be conducive to convergence — unless it is too atypical.

In this context, it is of interest to investigate the influence of the voice type on accommodating behavior in a direct comparison of the same SDS, while excluding the possible effect of the virtual interlocutor’s visual appearance by using only their voice for the interaction.

Thomason et al. (2013) compared the accommodation of intensity and F_0 features for students interacting with the ITSPOKE tutoring dialog system (Litman and Silliman, 2004) using either a pre-recorded, i.e., natural, or a synthesized voice. They reported a tendency for F_0 related features to show more convergence in the natural voice condition.

2.6 SPEAKER DISPOSITION

It is commonly observed that different speakers exhibit different degrees of phonetic accommodation (e.g., Pardo et al., 2018; Weise et al., 2019). Exploring the individual differences between speakers causing this variation is becoming a central point of accommodation research.

One factor that may contribute to the individual differences in accommodating behavior is the general speaker disposition, which includes aspects such as innate phonetic talent, personality traits, and cognitive abilities. Only a few studies have investigated these aspects to date.

Lewandowski and Jilka (2019) present an extensive study about the influence of speaker disposition on individual accommodating behavior in the context of nonnative phonetic convergence in conversational interaction, i.e., for native speakers of German speaking English to native speakers of English.

Phonetic talent, as assessed by a series of production, perception and imitation tests (Jilka, 2009a; Jilka, 2009b), was found to significantly influence the degree of phonetic convergence in word-based amplitude envelope match: more talented speakers exhibited higher degrees of convergence.

Concerning the personality measures, Neuroticism and Openness to Experience (assessed by the NEO Five Factor Inventory (NEO-FFI); Costa and McCrae, 1992; Borkenau and Ostendorf, 1993) as well as behavioral inhibition (assessed by the Behavioral Inhibition System questionnaire; Carver and White, 1994; Strobel et al., 2001) had an effect on accommodation, with more neurotic and more open speakers showing more convergence and speakers having less behavioral inhibition converging less.

Finally, with respect to cognitive measures, attention (assessed by the Simon test for mental flexibility; Simon, 1990) influenced accommodation with speakers showing more convergence when they were more successful at suppressing wrong reactions.

Yu et al. (2013) examined the influence of personality and attention (among other factors) in a non-conversational phonetic imitation task focussing on word-initial VOT in English. They found that Openness (assessed by the Big Five Inventory (BFI); John et al., 1991; John et al., 2008) and a strong attention focus (assessed by the Autism-Spectrum Quotient; Baron-Cohen et al., 2001) were positively correlated with the degree of VOT convergence.

These results suggest that it is promising to further investigate the influence of speaker disposition on phonetic accommodation. We collected information about personality traits for the participants of the WOz experiment using the German version of the NEO-FFI (Borkenau and Ostendorf, 1993) and incorporated it in the analysis (see Section 4.3.6).

In addition, we consider a different perspective on personality based on the boundary construct, which we present in the following section.

2.7 MENTAL BOUNDARIES

As a specific component of speaker disposition, we examine the concept of mental boundaries, which has not been considered in accommodation research so far. The concept was introduced as a dimension of human personality and an aspect of the overall organization of the mind by Ernest Hartmann and colleagues to characterize people who suffer from nightmares since they could not be grouped under the umbrella of other personality traits, such as the *Big Five*, for example (Hartmann et al., 1987; Hartmann, 1989; Hartmann, 1991). Mental boundaries refer to both the degree of interconnectedness within the mind and between the mind and the surrounding world. While thick boundaries stand for clear demarcation, thin boundaries represent permeability.

Hartmann (1991) emphasizes that the concept of boundaries in the mind is very broad and includes personality traits that one would not necessarily assume to be related. This is also illustrated by the different types of boundaries considered in the concept, for example: boundaries related to thoughts and feelings, sleep-dream-wake boundaries, boundaries related to memory, boundaries around oneself, boundaries in organizing one's life, boundaries in decision making and action, and others.

Hartmann and colleagues developed the Boundary Questionnaire (BQ) to measure the strength of these mental boundaries, which made the concept of (ego) boundaries, often described in the psychological literature (e.g., Lewin, 1935; Fisher and Cleveland, 1968; Landis, 1970), empirically explorable.

A person with thicker mental boundaries has a clearer sense of separation between, for example, their self and their surroundings; dream and reality; past, present and future; their own versus another group, whereas a person with thinner boundaries tends to merge such concepts. Therefore, the thick-boundary type functions largely unaffected by the environment, while the thin-boundary type tends to absorb the full extent of sensory input (Hartmann et al., 2001; Harrison and Singer, 2013). The boundary construct describes the general nature of a person’s mental boundaries as a value-free dimension of their personality, i.e., without favoring one type over the other.⁴

Apart from this, the concept can also describe intra-individual differences, i.e., a person can behave in a more thin or thick boundary manner depending on the situation (Hartmann et al., 2001). Generally, it was found that women and younger people have thinner mental boundaries than men and older people (Hartmann, 1991; McCrae, 1994). Typical examples of individuals with thinner mental boundaries are arts and music students (Beal, 1998) as well as people suffering from nightmares (Hartmann, 1991; Levin et al., 1991; Cowen and Levin, 1995); thicker mental boundaries can be found, for instance, in salespeople and lawyers, as well as in people with a preference for concreteness and order (Hartmann, 1991; Harrison et al., 2006).

The main field of application of the boundary construct to date is dream research, which is also where it originated from (e.g., Hartmann and Kunzendorf, 2006; Aumann et al., 2012).

We believe that mental boundaries could also have the potential to distinguish individuals with a higher and lower tendency to phonetically accommodate to a conversational partner.

As mentioned above, Hartmann (1991) describes various types of boundaries that are encompassed by the boundary construct. Among these are perceptual boundaries as well as interpersonal boundaries, boundaries of identity, and group boundaries. These seem to be particularly relevant in the context of phonetic accommodation, since we can assume that perceptual receptivity is a prerequisite for the latter and that permeable interpersonal relationships and fluid personal and social identities (i.e., group membership) facilitate accommodation given its social function. For intergroup contexts in particular, the notion of boundary permeability is echoed in CAT, where it is assumed that weak identification with a group and soft intergroup boundaries predispose a speaker to converge to an interlocutor who is part of a different group (e.g., Gallois et al., 2005).

Language learning, especially acquiring the pronunciation of a foreign language, which can be interpreted — at least to some degree — as a manifestation of phonetic accommodation, may again serve as an example here.

Hartmann (1991, p. 221) proposed *en passant* that thinner mental boundaries would be advantageous in learning “to speak a foreign language as native speakers do”. Already prior to that (e.g., Guiora et al., 1972), and then later with reference to Hartmann’s concept (e.g., Więckowska, 2011; Baran-Łucarz, 2012), authors suggested the permeability of ego boundaries as a facilitator for achieving native-like pronunciation in a foreign language.

According to Guiora (1994), pronunciation is anchored in the speaker’s identity. It is central to the language ego, which encompasses all components of self-representation through language and is surrounded by ego boundaries that are more or less flexible. These boundaries, and thus the self-representation, must be temporarily softened in order to speak a foreign language.

⁴ This represents a difference from the term “ego boundaries” as used in the psychoanalytic literature, where solid boundaries are seen as ideal and defective boundaries lead to pathological conditions (e.g., Federn, 1952; Blatt and Ritzler, 1974).

The few studies that have been conducted in this context showed only small effects of enhanced pronunciation skills for thinner boundaries and in some cases had limitations, for example, the indirect assessment of the boundary strength (e.g., Guiora et al., 1972) or an unvalidated test for the pronunciation attainment (e.g., Baran-Łucarz, 2012).

It has been shown that learners generally prefer different learning styles depending on the nature of their mental boundaries. According to Ehrman (1999), learners with thick mental boundaries rely on conscious processes like formal explanation and drilling. They prefer linear learning processes. Thin-boundary learners, on the other hand, rely on strategies of receptivity to outside influence and “make use of [...] native speakers as models with which to identify” (p. 68). They tend to enjoy unexpected learning events.

Another factor in language learning that is associated with boundary permeability is tolerance of ambiguity,⁵ which consists of the following three stages: (1) permitting information to access one’s mind, (2) dealing with contradictions and incomplete information in the mental system, and finally (3) integrating the new information with existing schemata, which is referred to as *accommodation* (Ehrman, 1996). Especially for the first two stages, thinner mental boundaries are considered beneficial, provided that these learners do not become overwhelmed by the abundance of incoming information.

Overall, these considerations and findings suggest that thin, permeable boundaries may be more likely to allow phonetic accommodation than thick, inflexible ones.

The original BQ (Hartmann et al., 1987; Hartmann, 1989; Hartmann, 1991) was developed in English and contains 145 items that operationalize the concept of boundaries at many different levels of behavior and experience, ranging from more abstract interpretations, e.g., “My thoughts blend into one another”, to more literal interpretations, e.g., “I like heavy solid clothing”.

In order to make concepts such as the boundary construct more accessible and easier to use for researchers, it is desirable to develop short versions of the often extensive original questionnaires. Moreover, for the wider dissemination of an instrument, it is required to adapt the questionnaire for other languages, which means translating it, adapting it to the cultural specifics of the target population if necessary, and validating the resulting questionnaire.

We therefore adapt the empirically-derived short version of the Boundary Questionnaire (BQ-Sh; Rawlings, 2001), which contains 46 English items, for German (see Section 5.2) and apply the resulting questionnaire to the data from the present WOz experiment to test whether mental boundaries affect the accommodation behavior of the participants (see Section 5.5).

⁵ The two concepts are so closely connected, in fact, that Ehrman (1996; 1999) suggests assessing tolerance of ambiguity by means of Hartmann’s BQ.

3 SHADOWING EXPERIMENT

In the shadowing experiment, native speakers of German repeat short German sentences immediately after hearing them from a female or a male model speaker. The model speaker voices are either natural or synthetic. For the synthetic voices, two variants are used: diphone-based and Hidden Markov model (HMM)-based synthesis. The investigated phonetic features are allophonic contrasts, schwa epenthesis, the realization of pitch accents, as well as the word-based temporal structure and distribution of spectral energy.

3.1 HYPOTHESES AND PREDICTIONS

We assume that the participants of this experiment generally accommodate their speech to the stimuli during the shadowing task. Specifically, we predict that the participants converge to the stimuli, since convergence has been proposed as the default behavior under the assumption that accommodation is triggered automatically, and because the automatic motivation is presumably dominant in the socially impoverished environment of a shadowing task.

It is unclear to what extent social motivation for accommodation applies under the given circumstances. However, if it does apply — for example, due to mere exposure to a human voice — we have no reason to believe that participants would feel the need to increase social distance to the shadowed speakers and hence phonetically diverge from them, since there is no further interaction between the participants and the model speaker voices beyond the shadowing task itself and the text material used in the latter is uncontroversial, i.e., does not inspire resentment.

The focus of the experiment lies on the question whether participants behave similarly when confronted with either natural or synthetic stimuli. Again, under the assumption that automaticity is the main driving factor of accommodation in a shadowing task, we expect participants to converge to the synthetic stimuli as well. With respect to the social motivation, what was said above holds for both stimulus types: there is no a priori reason to increase social distance to the shadowed voices. However, the fact that the synthetic stimuli are probably recognized as non-human (cf. typicality) by the participants may trigger a feeling of social separation, which may lead to a reduction of the convergence effect and potentially even to divergence.

Although we expect overall accommodation, we predict that there is substantial variation between the participants of the experiment, presumably due to factors mentioned in Sections 2.4 and 2.6 such as their perception of the interlocutor or their own disposition.

This individual variation may, on the one hand, surface as different degrees of accommodation in one phonetic feature; on the other hand, participants may accommodate to different subsets of the phonetic features examined in this experiment, rather than to either all or none of them.

Coming back to the distinction between natural and synthetic stimuli in this context, it is possible that certain phonetic features are difficult to perceive in synthetic speech, and therefore do not lead to accommodating behavior. This may, for example, be the case if a phenomenon is not present in the underlying database or the model built for synthesis. If so, this may concern different features for the two different synthesis methods used in the present experiment.

3.2 MATERIAL AND METHODS

3.2.1 *Corpus*

The present analyses are carried out on a corpus of shadowed speech.¹ The corpus contains 6720 instances of short German sentences (both declaratives and interrogatives) which were uttered by 56 native speakers of German in a shadowing experiment. The experiment included a shadowing task, in which the participants repeated the sentences after hearing them from a female or a male voice, which were either natural or synthetic. The natural stimuli were recorded by a female and a male native speaker of German; two sets of synthetic stimuli were created using diphone- and HMM-based synthesis, both with a female and a male voice (see [Section 3.2.1.1](#)).² Each participant shadowed only one stimulus type, but in both the female and the male version. The participants were not told whether the stimuli they heard were natural or synthetic. The shadowing task was preceded by a baseline production phase and followed by a post production phase in which the participants read the same text material from a screen. Between the baseline production (ca. 4 min) and the shadowing task (ca. 6 min), the participants played a game on a tablet that involved no linguistic input or output, which we refer to as the visual task (ca. 7 min). The post production (ca. 4 min) immediately followed the shadowing task. The entire experimental procedure including informed consent, instructions, a final questionnaire (see [Section 3.2.1.2](#)), and the remuneration took about 45 min.

While the baseline production serves to determine the participants' preference with respect to the pronunciation variants — or the baseline values of the other examined features — and the shadowing task tests the accommodation towards the model voices, the post production allows to evaluate whether the accommodation effect is fully or partially sustained after the shadowing task or the participants return to the baseline level of the respective feature. The visual task was incorporated to weaken the participants' mental representation of their own baseline productions before continuing with the shadowing task.

The text material presented to the participants consists of 15 target and 15 filler sentences (see [Appendix A](#)). Every target sentence contains one of three segments for which two prototypical pronunciation variants are expected to occur in native speakers of German (see [Section 3.2.2.1](#)):

- (1) ⟨-ä-⟩ as [ɛ:] or [e:] — e.g., in: Die Bestätigung ist für Tanja.

the confirmation is for Tanja

- (2) ⟨-ig⟩ as [ɪç] or [ɪk] — e.g., in: Kommt Essig in den Salat?

does go vinegar into the salad

- (3) ⟨-en⟩ as [ɲ] or [ən] — e.g., in: Sie begleiten dich zur Taufe.

they accompany you to the baptism

All natural and synthetic target stimuli exist in both versions, with the exception of the female HMM [ɛ:] targets, which were indeed more open than the [e:] targets, but not undisputably distinguishable due to technical reasons (see [Section 3.2.1.1](#)).

For the purpose of this study, the three variations are initially regarded as binary contrasts: [ɛ:] vs. [e:], [ɪç] vs. [ɪk], and [ɲ] vs. [ən]. During the baseline phase, the participants' productions

¹ The corpus was annotated using the WebMAUS services (Kisler et al., 2017). Manual corrections were carried out where necessary for the analyses.

² I would like to thank Eran Raveh and Sébastien Le Maguer for generating the diphone and HMM stimuli, respectively.

are auditorily identified by the experimenters as belonging to one of the two categories. Each of the three variations appears in five target sentences. Since some speakers use both variants of a pair interchangeably, forms with a minimum of three out of five possible occurrences are considered to be the preferred variant of a speaker.

During the shadowing task, the stimuli were selected so that the participants heard the opposite of the pronunciation variant they had uttered predominantly during the baseline production, hence their dispreferred variant, for most of the items. This provided them with the opportunity to accommodate phonetically. For some of the items, namely the ones the participants produced with their dispreferred variant during the baseline production, they would hear their preferred version during the shadowing task.

The filler sentences are comparable in length, but do not contain any of the target features listed above, for example:

- Die Glühbirne ist leider kaputt.
the lightbulb is unfortunately broken
- Habt ihr das rote Auto erkannt?
did you the red car recognize

Apart from the explicit manipulation on the segmental level, it can be assumed that all stimuli — targets as well as fillers — naturally differed from the versions the participants uttered in the baseline production on various levels (e.g., speaking rate, intonation pattern, rhythm, segmental pronunciation), giving additional opportunities to accommodate.

The participants of the experiment (see [Section 3.2.1.2](#)) were recorded in a sound-attenuated booth using a stationary cardioid microphone. The instructions for the experiment were given in written form on a screen in front of the participants. To avoid priming of convergence through the instructions (cf. Dufour and Nguyen, 2013), words such as “repeat” and “imitate” were not used. For the baseline and post productions the pertinent part of the instructions read as follows (English translation): “We will now record 30 short sentences with you. Please speak completely normally.”; for the shadowing task the instructions were: “We will now record another 60 short sentences with you. This time, you will not read the sentences, but hear them. Please speak completely normally again.” We cannot exclude the possibility that some participants might still have interpreted the task as an imitation task. Given the length of the stimuli, which is unusual for a shadowing experiment in accommodation research, the target for imitation would still be rather broad and it would not be obvious which specific features were to be imitated.

To allow the participants to become familiar with the task, a small number of test sentences were provided at the beginning of the experiment, which were not included in the later analysis.

During the shadowing task, 60 stimuli in two blocks of 30 stimuli each from a female and a male voice were played back to the participants over headphones, with half of the participants hearing the female voice first, the other half hearing the male voice first. Within the blocks, the stimuli were semi-randomized for balanced distribution of the targets over the two sets.

[Figure 1](#) illustrates the flow of the data collection process.³

3.2.1.1 *Stimuli*

For the natural stimuli, two native speakers of German (female, 25 years old; male, 23 years old) were recorded in a sound-attenuated booth using a stationary cardioid microphone. The 30 target and filler sentences were presented on a computer screen and the speakers were instructed

³ I would like to thank Eran Raveh for co-designing and co-executing the experimentation process.

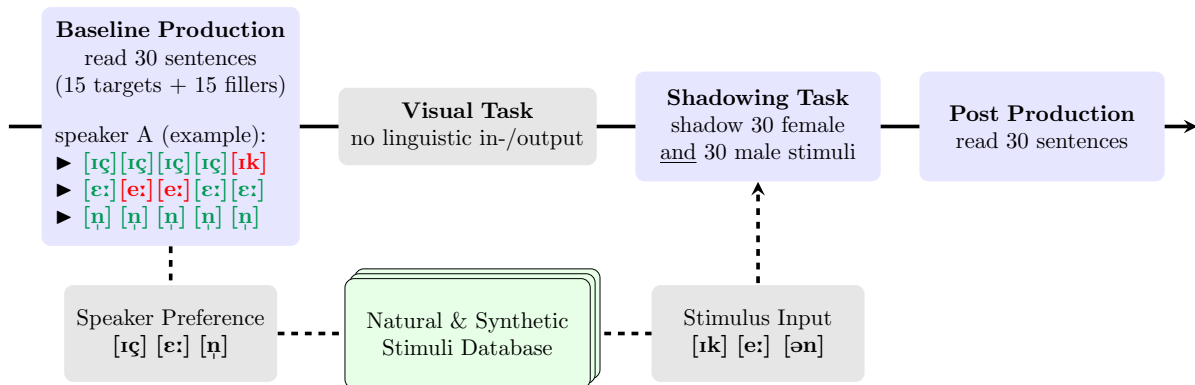


Figure 1: Overview of the data collection process. The stimuli presented during the shadowing task are selected from the database depending on the *speaker preference*, i.e. the participant’s preferred pronunciation variants in the baseline production. The stimuli containing the dispreferred variants are passed to the participants as *stimulus input*. Note that for **some items**, the participants still hear their preferred variant during the shadowing task; for **most items**, however, they hear their dispreferred variant.

to speak naturally, as if in conversation with someone. Subsequently, the 15 target sentences were presented again. The three pronunciation variations were explained to the speakers and they were asked to distinctly produce the two corresponding variants for every target sentence. The best tokens in terms of target feature production and overall clarity were selected.

The first set of synthetic stimuli was created using diphone-based synthesis with MBROLA (Dutoit et al., 1996). One female and one male voice were used to match the sex of the natural speakers. For the realization of the segmental variations, different phonetic transcriptions of the target sentences were provided to the system, one for each of the pronunciation variants. To control for potential differences in prosody and information structure between the natural and synthetic stimuli, the F_0 contours and segment durations of the natural stimuli were specified as parameters to the synthesis system. This resulted in diphone-based stimuli with the same F_0 contours and segment durations as the natural stimuli.

The second set of synthetic stimuli was created using the HMM-based Speech Synthesis System (HTS, version 2.3; Zen and Toda, 2005) with the BITS unit selection corpus (Ellbogen et al., 2004). Again, one female and one male voice were used and the F_0 contours and segment durations of the natural stimuli were imposed on the synthetic stimuli.

The resulting 270 stimuli ($45 \text{ stimuli} \times 3 \text{ types} \times 2 \text{ sexes}$) were stored in a database for use in the experiment.

To assess the perceived quality of the stimuli, scores of naturalness and likability were collected from the participants directly after the experiment. They rated only those female and male voices they had heard during the experiment on 8-point scales from *1 – very unnatural* to *8 – very natural* and from *1 – not likable* to *8 – likable*. We used 8-point scales to provide enough room for differentiation between the six voices. Since we can assume that the participants interpreted the unlabeled steps between the endpoints as equidistant intervals, we can consider this an approximation of an interval scale and calculate the mean as a measure of the central tendency.

The naturalness of the natural stimuli was judged with a mean score of 6.2 ($SD = 1.2$) for the female voice and 5.5 ($SD = 1.6$) for the male voice. Thus, even the natural stimuli were not evaluated as perfectly natural. This may be partly due to the central tendency bias, which disfavors extreme responses on such rating scales. But it also suggests that the participants’ concept of a *very natural* sounding voice is not necessarily fulfilled by a natural voice. The diphone stimuli received mean naturalness scores of 2.6 ($SD = 1.8$) for the female voice and

3.5 ($SD = 1.9$) for the male voice and were thus perceived as least natural. The HMM stimuli, finally, were rated with a mean naturalness of 4.0 ($SD = 2.4$) for the female voice and 4.3 ($SD = 2.8$) for the male voice and thus showed the greatest variance in ratings, indicating that the participants were less in agreement about their degree of naturalness. We can conclude that the synthetic voices were perceived as less natural than the natural voices. This was the case although they were evaluated separately by different listeners. A direct comparison in a joint evaluation would likely reinforce this difference. However, for the present study only the assessment of the stimuli that the participants actually heard is of relevance.

The likability of the six voices was rated somewhat more uniformly by the participants. The natural stimuli received mean likability scores of 5.5 ($SD = 1.4$) for the female voice and 5.1 ($SD = 2$) for the male voice; the diphone stimuli were rated with a mean likability of 3.8 ($SD = 1.3$) for the female voice and 4.6 ($SD = 1.6$) for the male voice; the HMM stimuli scored a mean likability of 5.1 ($SD = 2.1$) for the female voice and 5.6 ($SD = 1.9$) for the male voice. Thus while the likability of the natural and HMM voices was rated almost the same, slightly on the positive side of the scale, the female diphone voice was rated both as the most unnatural and least likable, followed by the male diphone voice.

We selected two synthesis methods which made it possible to control the output on the level of individual segments. This was done directly, by changing the desired target diphone, in the case of diphone synthesis, and indirectly, by training the voices with the pronunciation variants, in the case of HMM synthesis.

Since HMM synthesis uses machine learning techniques, the degree of flexibility depends on the corpus used for training. Due to an imbalanced number of occurrences of the target sounds [ɛ:] ($n = 282$) and [e:] ($n = 1457$) in the corpus underlying the female HMM voice, it was not possible, with the synthesis process applied here, to produce female HMM stimuli containing the target allophone [ɛ:] that were clearly distinguishable from those containing the target allophone [e:]. Therefore, we decided to let the participants of the HMM group shadow both male and female stimuli to keep the experimental flow identical to the natural and diphone group, but only included their productions shadowing the male HMM [ɛ:] stimuli in the present analysis. Note that this merely concerns the participants with a baseline preference for the allophone [e:] ($n = 7$, see [Table 1](#)).

One long-standing point of criticism toward diphone synthesis is the large number of concatenation points, which is detrimental to the perceived naturalness (Olive et al., 1998; Taylor, 2009). Discontinuities at concatenation points result in discontinuities of spectral trajectories that may be audible. Diphone systems use several techniques, including a careful construction of the diphone inventory, to reduce the discontinuities. Our diphone stimuli are generally rather smooth, but it is still possible that audible glitches affect the ability of the stimuli to trigger accommodation.

For these reasons, the present experiment used stimuli generated by HMM synthesis, which are generally smooth but can sound buzzy, and by diphone synthesis, which can be directly controlled but may contain discontinuities.

3.2.1.2 *Participants*

The participants were recruited on the Saarland University campus and paid for taking part in the experiment. 50 participants were students and six had non-academic jobs. All 56 participants were native speakers of German and 11 spoke more than one native language (e.g., Turkish, French, Vietnamese, Dutch). All had learned at least one, and the majority more than two, foreign languages. The most frequent foreign languages were English ($n = 55$), French ($n = 44$), and Spanish ($n = 30$). Multilingualism may be a favorable basis for phonetic accommodation,

as it entails a certain amount of experience in switching between different pronunciation settings. It is likely that this basis would be more pronounced in native bilinguals and speakers who have achieved native-like pronunciation in a foreign language, as they either have more experience with different pronunciation settings or have been more successful in switching between them than speakers who have maintained a strong accent of their native language in acquired foreign languages. However, the participants in the present study were not selected according to these criteria and we are therefore not in a position to systematically investigate the effect of multilingualism on phonetic accommodation.

The participants came from ten different German states and Austria with roughly 60% from central regions and 20% from northern and southern regions, respectively. The regional origin of the speakers will mainly be reflected in the baseline productions of the allophonic contrasts [ɛ:]/[e:] and [ɪç]/[ɪk], as these are regionally distributed (see [Section 3.2.2.1](#)). We do not expect the regional origin to influence the accommodating behavior and do not investigate this further.

In a questionnaire completed after the experiment, which asked the participants to assess their general communicative behavior, 80% answered affirmatively to the question whether they change the way they speak depending on their respective interlocutor; 50% believed they would converge to an interlocutor of the same dialectal background; only 15% claimed they would do the same with an interlocutor of a different dialectal background; 16% said that they intentionally imitate the pronunciation of interlocutors.

These numbers, although they may not agree with the actual behavior of the participants, show that there is a certain awareness of the phenomenon of accommodation to an interlocutor in spoken communication. The readiness to accommodate seems to be higher when the accommodation target is more familiar (e.g., own vs. different dialect). A small number of participants perceives convergence to an interlocutor even as an intentional, active process. We will assess whether participants who indicated that they would converge to dialects of other regions or intentionally imitate interlocutors exhibit particular patterns of accommodation.

Each participant of the present study was presented with only one of the three stimulus types — natural, diphone, or HMM. This resulted in the following three experimental groups: the natural group with 21 participants (17 female, 4 male; mean age 26.6 years; age range 19 to 34 years), the diphone group with 18 participants (14 female, 4 male; mean age 26.2; age range 19 to 50 years), and the HMM group with 17 participants (13 female, 4 male; mean age 26.8 years; 18 to 51 years). The between-subjects design was chosen to avoid learning and transfer effects over conditions that might have occurred if the same participants had been exposed to all three stimulus types.

3.2.2 Analyzed features

3.2.2.1 Allophones and schwa epenthesis

We examine whether participants accommodate on the level of segmental pronunciation with respect to the three types of pronunciation variation that were explicitly manipulated during the shadowing task. These pronunciation variations are commonly found among native speakers of German.

The realization of the long vowel ⟨-ä-⟩ in stressed syllables as [ɛ:] or [ɛ:],⁴ and the realization of the word ending ⟨-ig⟩ as [ɪç] or [ɪk], vary regionally, occurring roughly in the North and South of the German-speaking region of Europe, respectively (Kleiner, 2011).⁵ The Standard

⁴ This contrast also occurs word-initially, but we only take word-medial occurrences into account in this study.

⁵ Note that for Austria [e:] is more common in the East, whereas [ɛ:] is typically encountered in the West (Dudenredaktion, 2015; Kleiner, 2011).

German variants of each pair are [ɛ:] (predominant in the South) and [ɪç] (predominant in the North; cf. Dudenredaktion, 2015). However, it has been shown that the respective non-standard forms, [e:] and [ɪk], are not perceived as strong dialectal markers by native listeners of German. According to Kiesewalter (2019), the realization of ⟨-ä-⟩ as [e:] subjectively corresponds to the standard, and the realization of ⟨-ig⟩ as [ɪk] is perceived as only slightly dialectal.

Elision or epenthesis of [ə] in the word ending ⟨-en⟩ when preceded by a plosive or a fricative varies mainly based on speaking style. In Standard German, schwa is elided in this position. An epenthetic schwa in this context, despite occurring in certain German dialects, is primarily produced when speaking particularly slowly and clearly. It is often perceived as hyperarticulation, especially when the quality is additionally shifted towards [e] or [ɛ]. Since humans have been shown to apply hyperarticulation when conversing with computers (e.g., Burnham et al., 2010), it may be the case that participants are more likely to pick up this trait from a synthetic voice than from a natural one.

Although speakers have their preferred variants in the contexts given in this study, [ɛ:], [e:], [ɪk], [ɲ], and [ən] are all part of the basic phonetic inventory of native speakers of German and used by all speakers in other contexts. Only [ɪç] is an exception here, since many speakers realize [ç] as [ʃ] or [ç]. In our analysis, the latter are evaluated as phonetically different members of the underlying fricative class and included in the [ɪç] category. Ultimately, every participant has the necessary means to accommodate with respect to the pronunciation variations examined in the present study.

The degree of accommodation for the three pronunciation variations was quantified as follows:

The vowel quality [ɛ:] vs. [e:] was evaluated as a continuum in the F1–F2 formant space. Automatic annotations (WebMAUS, Kisler et al., 2017) of all target vowel segments were manually corrected by a trained phonetician. The first and second formants of each target vowel were measured at the temporal midpoint in all productions as well as in the stimuli using Praat’s (Boersma and Weenink, 2017) Burg algorithm. In contrast to a preliminary analysis in Gessinger et al. (2017), where the mean of all model speaker vowels (female and male combined) was defined as the overall convergence target, we now took a more fine-grained approach by calculating the Euclidean distance between each of the speakers’ productions and the corresponding vowel of the model speaker they were shadowing in the respective instance, as

$$dist = \sqrt{(F1_{participant} - F1_{model})^2 + (F2_{participant} - F2_{model})^2}$$

Figure 2 illustrates the utterance pairings for which the Euclidean distance was calculated. For the baseline and post productions, the Euclidean distance was calculated twice per speaker production: once in comparison to the female model speaker vowels and once in comparison to the male model speaker vowels. For the shadowing productions, only the Euclidean distance to the stimulus shadowed in the respective instance was calculated. This resulted in six comparisons per speaker and item.

A decrease of Euclidean distance in the F1–F2 formant space indicates convergence of vowel quality to the model speakers; conversely, an increase indicates divergence.

For further analysis, the difference in Euclidean distance (DID_{vowel}) was calculated between baseline and shadowing (bs), baseline and post (bp), and shadowing and post (sp) productions. DID_{vowel} is positive in the case of convergence and negative in the case of divergence.

The variation [ɪç] vs. [ɪk] was evaluated as a binary contrast. All target segments were manually annotated by a trained phonetician as belonging to the fricative or plosive class of the contrast by correcting automatic annotations. As mentioned above, some speakers produced instances of both categories in the baseline phase. In those cases, participants heard their preferred variant for some of the items in the shadowing phase (see Figure 1). The present analysis

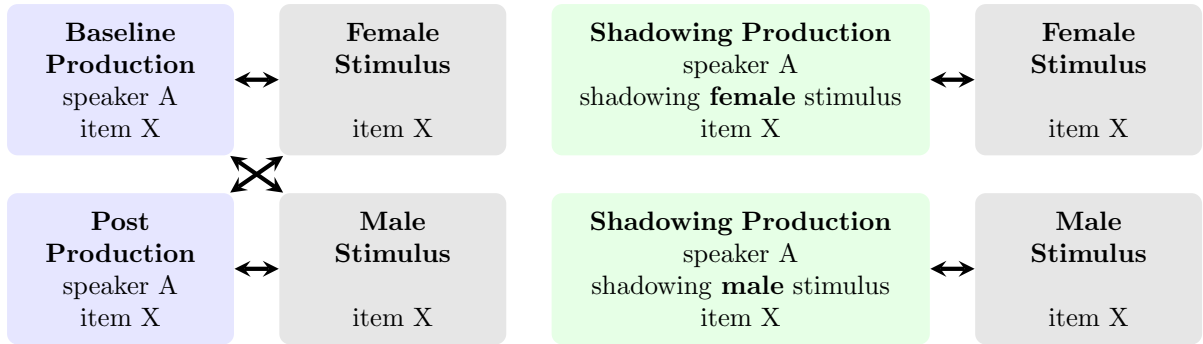


Figure 2: Utterance pairings for the analyses of vowel quality, pitch accent realization, and amplitude envelopes. Baseline and post productions are compared twice, i.e., to the corresponding female and male stimuli. Shadowing productions are compared to the stimulus shadowed in the respective instance. This results in six comparisons per speaker and item.

of the data accounts for this fact by comparing each participant production to the variant they heard from the model speakers and determining whether these are the same or different variants of the binary contrast. A significant increase of *same*-cases indicates convergence of the pronunciation variant to the model speakers, and a decrease indicates divergence.⁶

The presence or absence of [ə] in the word ending (-en) was determined by measuring the duration of potential schwa segments between the preceding consonant (here [d], [x], [t], [ç], or [f]; see Appendix A) and the final nasal, which were determined by manual correction of automatic annotations as performed by a trained phonetician. A duration of 30 ms was established as a minimum threshold to count the segment in question as a schwa. This decision is supported by the fact that all unambiguous schwas occurring in the stimuli were at least 30 ms long. As in the case of [ɪç] vs. [ɪk], we were taking all speaker productions into account and counted *same* (as model) vs. *different* (from model) cases. A significant increase of *same*-cases indicates convergence of the pronunciation variant to the model speakers, while a decrease indicates divergence.

3.2.2.2 Pitch accent comparison with PaIntE

In German, post-lexical accentuation is achieved by increasing intensity and length, as well as producing full instead of reduced vowel qualities. If such stressed units are further accompanied by pitch movement, they are called pitch accents (Möbius, 1993). A nuclear pitch accent is the last pitch accent in a prosodic phrase and may, in the text material of the present study, coincide with the last syllable of an utterance or occur in non-final position. Prenuclear pitch accents are all pitch accents occurring before the nuclear pitch accent in a prosodic phrase. To characterize and compare the pitch accents phonetically in the present study, we use the Parametric Intonation Event (PaIntE) model (Möhler, 1998; Möhler and Conkie, 1998; Schweitzer et al., in press).

The Parametric Intonation Event (PaIntE) model approximates the F_0 contour of intonation events with the sum of a rising and a falling sigmoid as shown in Figure 3. Each parametrization takes the syllable carrying the intonation event σ^* , as well as one preceding and one following syllable σ as the basis for the analysis. The length of each syllable is normalized to 1; the three syllables thus fit into the range of -1 to 2 .

⁶ Note that a preliminary analysis of [ɪç] vs. [ɪk] in this corpus only counted cases of convergence vs. cases of non-convergence from baseline to shadowing phase and excluded those instances where the same variant was already produced in the baseline phase (Gessinger et al., 2017).

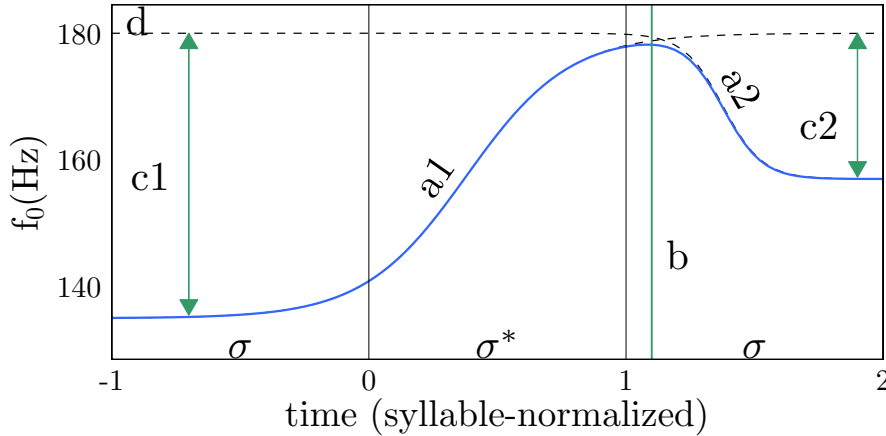


Figure 3: Parameterization of an intonation event on the syllable σ^* by the PaIntE model. The F_0 contour is approximated with the sum of a rising and a falling sigmoid function. The approximation is characterized by six parameters: $a1$, $a2$, b (peak alignment), $c1$, $c2$, and d (peak height). Figure adapted from Möhler and Conkie (1998) and Schweitzer et al. (2017).

The model function is characterized by six parameters: $c1$ and $a1$ represent the height and slope of the rising sigmoid, respectively; $c2$ and $a2$ provide the same information for the falling sigmoid. The parameters d and b describe the absolute height and the relative syllable alignment of the F_0 peak, respectively.

If the F_0 contour cannot be fitted with two sigmoids, only a single sigmoid is applied (either rising or falling, see the dashed lines in Figure 3), leaving one set of c and a parameters unspecified. If a single sigmoid is not a good fit either, PaIntE only provides the mean F_0 value as the d parameter, leaving all other parameters unspecified.

To extract the PaIntE parameters, F_0 is tracked using the `get_f0` function of the Entropic Signal Processing System (ESPS; Talkin, 1995). The resulting raw contour is smoothed by the `smooth_f0` algorithm authored by Gregor Möhler (March 2000). It uses the `smooth_phrase` algorithm from the Edinburgh Speech Tools Library⁷ (King et al., 1999).

To determine the target syllables for the present study, prenuclear and nuclear pitch accents of all stimuli used in the shadowing task were manually annotated by a trained phonetician. Since the F_0 contours and segment durations from the natural stimuli were imposed on the two types of synthetic stimuli during their generation, it is expected that the same pitch accent locations are found in all three stimulus sets — natural, diphone, and HMM. This was true for the vast majority of the utterances. In very few cases, additional pitch accents occurred in the synthetic stimuli. These were taken into account in the analysis. Overall, the distribution of accent types found in the stimuli was 59% prenuclear and 41% nuclear pitch accents. Of the nuclear pitch accents, 34% coincided with the last syllable of an utterance and 66% occurred in non-final position.

The PaIntE parameters were extracted for every syllable of all utterances after manual correction of the automatically determined syllable boundaries by a trained phonetician. Then, the data was cleaned by executing the following steps: Cases in which the pitch accent shape was estimated by the mean F_0 alone, leaving all other parameters unspecified, were excluded from the analysis (approx. 6% of the data). Furthermore, cases in which one of the six (two sigmoids fitted) or four (one sigmoid fitted) parameter values fell into the 1st or 99th percentile for that parameter in one speaker, were excluded as well (approx. 10% of the data) to remove potential

⁷ http://festvox.org/docs/speech_tools-2.4.0

measurement errors while keeping atypical yet plausible values in the data. Such atypical values are expected when a speaker accommodates to an interlocutor.

To subsequently calculate the Euclidean distance between 6-dimensional PaIntE parameter vectors, the c (height) and a (slope) parameters were set to 0 wherever they were unspecified. Remember that this is the case when only a single sigmoid was fitted. Then, all PaIntE parameters were standardized to speaker specific z-scores to eliminate differences linked to the speaker sex and to give all parameters the same weight in the distance analysis.

Finally, the Euclidean distance between the 6-dimensional PaIntE parameter vectors \vec{p} of a participant and \vec{m} of a model voice was calculated for the same syllable as

$$d(\vec{p}, \vec{m}) = \sqrt{\sum_{i=1}^6 (\text{participant}_i - \text{model}_i)^2}$$

This was done for the pairings detailed in [Figure 2](#), as described above in detail for the vowel quality analysis. A decrease of Euclidean distance indicates convergence of pitch accent realization to the model speakers; conversely, an increase indicates divergence.

For further analysis, we reduced the data set to the target syllables defined above, i.e., the syllables carrying a prenuclear or nuclear pitch accent in the stimuli, and calculated the difference in Euclidean distance ($\text{DID}_{\text{PaIntE}}$) between baseline and shadowing (bs), baseline and post (bp), as well as shadowing and post (sp) productions. $\text{DID}_{\text{PaIntE}}$ is positive in the case of convergence and negative in the case of divergence.

3.2.2.3 *Word-level amplitude envelope analysis*

Contrary to the features discussed so far, amplitude envelopes represent the speech signal globally by the distribution of spectral energy across time and do not single out specific areas of interest from the signal (Wade et al., 2010).

In the present study, the amplitude envelope analysis is carried out on one word per utterance. In the target utterances, this is the word containing the segmental manipulation, whereas in the filler utterances, a regular content word was selected. For the target utterances, the analysis of word-level spectral composition is therefore related to the assessment of segmental pronunciation. See [A](#) for an overview of the words in question and their location in the original target and filler sentences. It is possible that the spectral composition assimilates to a greater degree in utterances for which the stimulus explicitly encourages, or makes room for, accommodation, hence the target utterances.

The word boundaries were manually corrected in automatic annotations by a trained phonetician. For the analysis, the acoustic signal of a word was separated into four logarithmically spaced frequency bands between 80 Hz and 7 800 Hz in MATLAB (version R2017a). An amplitude envelope was calculated for each resulting band using the linear Hilbert transform. The band-separated amplitude envelopes were then compared to their corresponding counterpart as detailed in [Figure 2](#).

Subsequently, each pairing of amplitude envelopes was transformed to have equal length while taking spectral characteristics into account, by performing dynamic time warping (DTW) with the Speech Signal Processing Toolkit⁸ (version 3.7). This resulted in the first similarity measure, i.e., the cost of the DTW operation, which is lower for more similar signals.

The resulting time-warped amplitude envelopes were then compared by cross-correlation. This resulted in the second similarity measure, i.e., the match value, which is the maximum

⁸ <http://sp-tk.sourceforge.net>

Table 1: Number of participants preferring the respective pronunciation variant as identified during the baseline phase. Participants in parentheses were excluded from the analysis of the corresponding feature.

condition	[ɛ:]	vs.	[e:]	[ɪç]	vs.	[ɪk]	[ɲ]	vs.	[ən]
natural	11		10	12		9	21		–
diphone	14		4	9		9	17		(1)
HMM	10		7	6		11	16		(1)

value of the cross-correlation transformed onto a scale from zero to one with 1 indicating maximal similarity, i.e., identity.⁹

As it was done for the PaIntE analysis, the DTW cost and match value data sets were cleaned by excluding values that fell into the 1st or 99th percentile for the respective parameter in one speaker (approx. 6% of the data in both data sets).

For further analysis, we calculated the difference in distance for both similarity measures, DID_{DTW} and DID_{match} , between baseline and shadowing (bs), baseline and post (bp), and shadowing and post (sp) productions. DID_{DTW} is negative in the case of convergence and positive in the case of divergence, whereas DID_{match} is positive in the case of convergence and negative in the case of divergence.

3.2.3 Further factors

Apart from the influence of the experimental phase itself, namely baseline production, shadowing task or post production, there are further factors which might influence the measured variables and need to be accounted for in the analyses. These factors, discussed below, are either given by the design of the experiment or motivated by theoretical considerations (see [Chapter 2](#)).

SPEAKER PREFERENCE For each variation of segmental pronunciation examined in this study, the participants’ preferred variant was identified during the baseline phase. Refer to [Table 1](#) for an overview of the preference groups. Since only two out of 56 participants had a preference to produce [ən] in the baseline phase, we excluded these participants from the analysis of the schwa epenthesis. For the other two variations of segmental pronunciation, there are two preference groups: [ɛ:] or [e:] and [ɪç] or [ɪk]. It is possible that the readiness to produce the respective other variant depends on the speaker’s preference group. Especially since one of the variants is considered Standard German for the respective variation in the given context (see [Section 3.2.2.1](#)), there might be a bias in favor of producing this more prestigious variant. The factor PREFERENCE is included in the analysis of the allophonic contrasts (see [Section 3.3.2](#)).

SPEAKER ATTITUDE At the end of the experiment, the participants were asked which variant of each pronunciation variation they believe to produce themselves and what they think of the respective other variant.¹⁰ The majority of the participants reported a positive attitude towards the variants they do *not* believe to produce themselves — 80% for [ɛ:]/[e:], 70% for [ɪç]/[ɪk], and 72% for [ɲ]/[ən]. This includes ratings such as “also ok”, “better”, and “Standard German”. Only a minority of participants showed a negative attitude towards the other versions

⁹ The scripts to extract and cross-correlate the amplitude envelopes are taken from Lewandowski (2012).

¹⁰ Note that approximately 30% of the participants for each of the three features misjudged which variant they (predominantly) produce themselves. This is in line with the assumption stated in Mitterer and Müsseler (2013) with regard to [ɪç] vs. [ɪk] that speakers are often not consciously aware of which variant they use.

such as “wrong”, “weird”, and “sounds artificial”. It seems plausible that a positive attitude towards a pronunciation variant might entail a higher probability of converging to it, whereas the production of variants carrying a negative connotation might be inhibited. The factor ATTITUDE with the two levels *positive* and *negative* is included in the analyses of all three pronunciation variations (see Section 3.3.2).

PAIRING: SAME-SEX VS. MIXED-SEX In the present study, each speaker shadowed a female and a male model voice. As discussed in Section 2.4, this factor has yielded different outcomes in prior analyses, some suggesting that more accommodation occurs in same-sex, others in mixed-sex pairings. The analysis includes the factor PAIRING with the two levels *same-sex* and *mixed-sex*, where applicable, namely for *vowel quality* (see Section 3.3.2), *pitch accent realization* (see Section 3.3.3), as well as *DTW cost* and *match value* (see Section 3.3.4).

SENTENCE TYPE The analyses of pitch accent realization and word-level spectral composition are performed on both target and filler sentences. While the analysis of pitch accent realization has no particular link to the pronunciation variations in the target sentences, the analysis of spectral composition is based on the words containing these segmental variants. For the measures associated with spectral composition, i.e., the *DTW cost* and the *match value*, it can therefore be assumed that the distance between participant and model speaker baseline productions is greater for the target sentences than for the filler sentences. This additional space may enhance the accommodation effect. The factor SENTENCE with the two levels *filler* and *target* is therefore included in the analyses of the *DTW cost* and the *match value* (see Section 3.3.4).

ACCENT TYPE In the analysis of pitch accent realization, an additional factor comes into play that is motivated by prosodic theory, namely the accent type. We distinguish between *prenuclear* and *nuclear* pitch accents as the two levels of the factor ACCENT. The latter are known to be perceptually more salient (Jagdfeld and Baumann, 2011). We therefore expect a stronger accommodation effect for nuclear than for prenuclear pitch accents (see Section 3.3.3).

3.3 ANALYSIS AND RESULTS

3.3.1 Modeling

The dependent variables (see Section 3.2.2) are analyzed using linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs) formulated with the lme4 package (1.1-18-1; Bates et al., 2015) and evaluated with the lmerTest package (3.0-1; Kuznetsova et al., 2017) in R (3.5.1; RCore Team, 2018).

To strike a compromise between accuracy and complexity, model selection is carried out bottom-up, starting with a model which only includes the random factor intercepts for SUBJECT and ITEM. Then, theoretically relevant fixed factors (sum coded) and interactions as given by the design of the experiment or as motivated by the predictions made in Section 3.2.3 are added to the model. Random slopes for SUBJECT and/or ITEM are added for every effect where there is more than one observation for each unique combination of SUBJECT/ITEM and treatment level. Random slopes are only removed to simplify the model in cases of convergence errors or to allow a non-singular fit. The influence on the model fit is assessed by means of the Akaike information criterion (AIC), which estimates the relative quality of a statistical model for a given data set by taking into account the likelihood function and the number of estimated parameters (Akaike, 1973). A factor is kept in the model if the AIC value decreases by at least

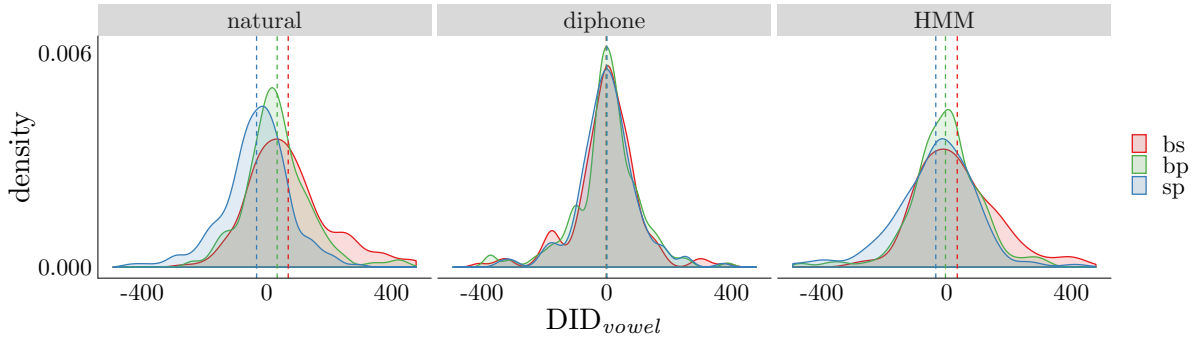


Figure 4: Distributions of DID_{vowel} for the pronunciation variation [ɛ:] vs. [e:] in the three experimental groups. Comparisons are made between **base–shadow**, **base–post**, as well as **shadow–post** phases. The dashed lines indicate the distribution means.

two points as compared to the model without the factor in question. Modeling is concluded by visual inspection of the residuals’ normality and homoscedasticity. Factors kept in the model are being considered significant predictors of the respective dependent variable at $\alpha = 0.05$.

For the analyses taking a difference in distance (DID) measure as dependent variable (DID_{vowel} , DID_{PaIntE} , DID_{DTW} , DID_{match}), the information about the experimental phase is included in the dependent measure, since the DID measures are calculated as comparisons of the experimental phases: baseline and shadowing (bs), baseline and post (bp), as well as shadowing and post (sp). It is therefore the model intercept that provides insight about accommodating behavior. The intercept is considered to significantly differ from zero at $\alpha = 0.05$.

In comparison, the analyses of the binary contrasts [ɪç] vs. [ɪk] and [ɲ] vs. [əŋ] take PHASE as a fixed factor into the model to assess accommodation. As for the DID measures, all experimental phases are compared to each other.

Comparing all experimental phases to each other allows to assess whether participants accommodate to the model speakers during the shadowing task (baseline vs. shadowing), whether the respective effect is sustained or reverted in the post phase (shadowing vs. post), and whether participants reach their baseline level again in the post phase (baseline vs. post).

3.3.2 Segmental pronunciation

LONG VOWEL <-Ä-> The distributions of DID_{vowel} measured for the vowel realizations are shown in Figure 4. A positive DID_{vowel} indicates convergence to the model speakers, a negative DID_{vowel} divergence, and a DID_{vowel} close to zero maintenance of the vowel quality.

Recall that the analysis of the seven participants constituting the HMM group with a baseline preference for [ɛ:] only includes their productions shadowing the male HMM [ɛ:] stimuli (see Section 3.2.1.1).

Note also that the baseline productions of the two preference groups [ɛ:] and [e:] were located at opposite ends of the F1–F2 space and their shadowing productions were expected to move towards each other, i.e., towards the model speaker vowels of the other variant. However, this difference in direction is canceled out in the calculation of the Euclidean distance. The two preference groups can therefore be jointly analyzed.

LMMs with DID_{vowel} as the dependent variable were fitted for each stimulus type and phase comparison data set separately, resulting in nine models. The factors PREFERENCE, PAIRING, and ATTITUDE were tested following the method in Section 3.3.1. Including the random factor intercepts for ITEM resulted in a singular fit for eight out of the nine models. Therefore, we only

Table 2: Results for variation [ɛ:] vs. [e:] — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{vowel} as estimated in separate models for the three different stimulus types and the three phase comparisons.

natural	base–shadow	base–post	shadow–post
intercept	69.94*** (14.89)	32.88 (17.53)	−33.67** (11.44)
PREFERENCE	33.79* (14.89)		
observations	210	210	209
diphone			
intercept	−1.68 (8.72)	−1.49 (9.93)	0.44 (7.24)
observations	177	179	178
HMM			
intercept	32.64* (13.61)	−2.12 (12.38)	−31.23 (15.86)
observations	134	135	133

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

included SUBJECT as a random factor in all models. Table 2 shows the parameter estimates for the nine final models.

In the natural data set, mean DID_{vowel} is significantly positive for the base–shadow comparison, indicating convergence to the model speakers during the shadowing task, significantly negative for the shadow–post comparison, indicating divergence from the model speakers after the shadowing task, and not significantly different from zero for the base–post comparison, indicating that the participants reached their baseline level again in the post phase. Additionally, the convergence effect in the shadowing task is stronger for participants with baseline preference [e:], as indicated by the significant effect of PREFERENCE.

No effect was found for the diphone data set; participants do not seem to have accommodated to the diphone model speaker vowels.

In the HMM data set, we found a significant convergence effect in the shadowing task, but no significant divergence effect in the post phase. The diverging movement from shadowing task to post phase is, however, so substantial that participants ended up close to their baseline level again, as shown by the non-significant base–post phase comparison.

The factors PAIRING and ATTITUDE did not account for variance in the data.

WORD ENDING <-IG-> The percentages of cases in which participant and model speakers realized the *same* or a *different* variant of the segmental pronunciation variation [ɪç] vs. [ɪk] are shown in Figure 5. In all three data sets, the number of same variants increases by about 30 % from the baseline phase to the shadowing phase, and decreases again in the post phase, yet to different degrees.¹¹

¹¹ Note that the numbers given in the text are descriptive values, whereas the GLMM result tables contain model estimates.

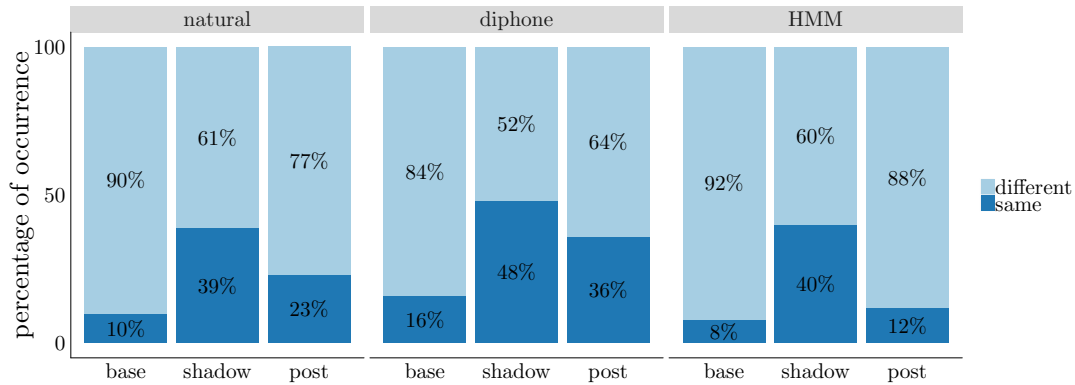


Figure 5: Results for variation [ɪç] vs. [ɪk]. Cases where speaker and model realize the same variant are indicated in dark blue; cases where they realize a different variant are indicated in light blue.

GLMMs with *IDENTITY* (*same* or *different*) as the dependent variable were fitted for each stimulus type data set separately, always comparing two experimental phases at a time, resulting in nine models. The outcome *same* is coded as success in the models.

The factors *PHASE*, *PREFERENCE*, and *ATTITUDE* were tested following the method described in Section 3.3.1. Table 3 shows the parameter estimates for the nine final models. Note that these are binomial models and the coefficients are hence in logit-space. If a logit-coefficient is positive, the effect of the corresponding predictor on the response variable is positive as well, and vice versa.

The increase of same variants in the shadowing task is significant for both the natural data set (10% to 39%) and the diphone data set (16% to 48%). Moreover, in both data sets the number of same variants decreases again in the post phase, although, not all the way to the baseline level (natural: 39% to 23%; diphone: 48% to 36%).

For the HMM data set, the increase of same variants in the shadowing task (8% to 40%) does not reach significance in the statistical model. However, the decrease of same variants in the post phase is significant and reaches the baseline level (40% to 12%). The latter is shown by the fact that *PHASE* did not account for variance in the data set and was therefore not included in the HMM base–post model.

The factors *PREFERENCE* and *ATTITUDE* did not show any significant effect on *IDENTITY*, although the former factor did improve overall fit in various models.

WORD ENDING <-EN> The percentages of cases in which participant and model speakers realized the *same* or a *different* variant of the segmental pronunciation variation [ɪ] vs. [əɪ] are shown in Figure 6. In 85% to 95% of the cases over all experimental phases of all three data sets, participants produced a different variant than the model speakers. The statistical analysis was carried out as described for the [ɪç] vs. [ɪk] variation above, without testing the factor *PREFERENCE*, however, since all analyzed speakers preferred [ɪ] in the baseline phase. Table 4 shows the parameter estimates for the nine final models.

Only in the case of the natural data set did participants produce significantly more [əɪ] (i.e., same variants) during the shadowing task, compared to the baseline phase (5% to 15%). The amount of same variants does not decrease significantly from the shadowing task to the post phase (15% to 10%) and there is no significant difference between the baseline and the post phase.

For both synthetic data sets, the factor *PHASE* did not remain in the final models; participants do not seem to have accommodated to the synthetic model speakers with respect to this feature.

The factor *ATTITUDE* did not influence *IDENTITY*.

Table 3: Results for variation [ɹ̥] vs. [ɹk] — parameter estimates (coefficients with standard errors in parenthesis) of the effects on identity as estimated in separate models for the three different stimulus types and the three phase comparisons.

natural	base-shadow	base-post	shadow-post
intercept	−1.75*** (0.47)	−3.01*** (0.79)	−1.47* (0.6)
PHASE	−0.93** (0.3)	−0.7 ** (0.25)	0.75* (0.3)
PREFERENCE	0.05 (0.57)	−0.07 (0.64)	
observations	315	210	315
diphone			
intercept	−1.43** (0.54)	−1.54*** (0.42)	−0.44 (0.57)
PHASE	−1.33*** (0.24)	−0.7 *** (0.21)	0.49** (0.18)
PREFERENCE	−0.37 (0.62)	0.01 (0.39)	−0.52 (0.67)
observations	270	180	270
HMM			
intercept	−1.9 ** (0.62)	−3.64*** (1.03)	−2.87** (0.98)
PHASE	−0.84 (0.49)		1.34*** (0.3)
PREFERENCE			−0.52 (0.88)
PHASE:PREF			−0.5 (0.29)
observations	254	170	254

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

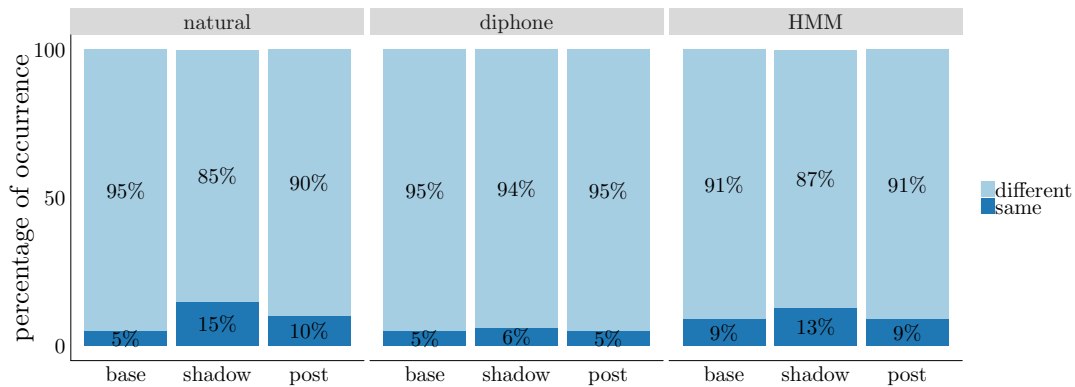


Figure 6: Results for variation [ɹ̥] vs. [ɹn]. Cases where speaker and model realize the same variant are indicated in dark blue; cases where they realize a different variant are indicated in light blue. Since all of the participants heard the model variant [ɹn], the percentage indicating *same*-cases coincides with the percentage of [ɹn] occurrences.

Table 4: Results for variation [ŋ] vs. [ən] — parameter estimates (coefficients with standard errors in parenthesis) of the effects on identity as estimated in separate models for the three different stimulus types and the three phase comparisons.

natural	base–shadow	base–post	shadow–post
intercept	−3.27*** (0.63)	−10.11 (5.46)	−2.76*** (0.63)
PHASE	−0.79** (0.28)	−1.04 (0.56)	−0.01 (0.27)
observations	315	210	315
diphone			
intercept	−4.79*** (1.28)	−10.98* (4.71)	−4.79*** (1.28)
observations	255	170	255
HMM			
intercept	−3.13*** (0.73)	−2.74*** (0.6)	−3.89*** (1.11)
observations	239	159	238

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

3.3.3 Pitch accent realization

The distributions of DID_{PaIntE} measured for the pitch accent realizations is shown in Figure 7. A positive DID_{PaIntE} indicates convergence to the model speakers, a negative DID_{PaIntE} divergence, and a DID_{PaIntE} close to zero maintenance of the pitch accent realization. As mentioned in Section 3.2.3, we distinguish prenuclear and nuclear pitch accents.

LMMs with DID_{PaIntE} as the dependent variable were fitted for each stimulus type and phase comparison data set separately, resulting in nine models. The factors PAIRING and ACCENT were tested following the method in Section 3.3.1. Table 5 shows the parameter estimates for the nine final models.

In the natural data set the participants converged to the model speakers during the shadowing task and diverged again in the post phase, reaching the baseline level.

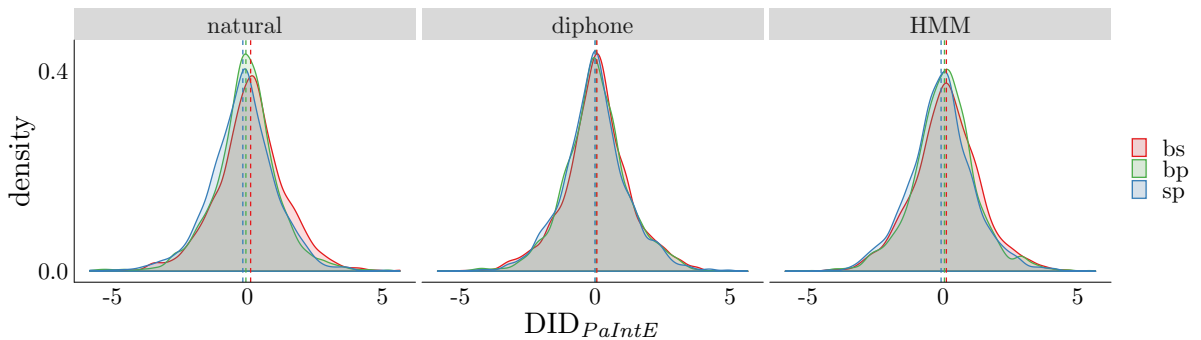


Figure 7: Distributions of DID_{PaIntE} for the comparison of pitch accent realizations in the three experimental groups. Comparisons are made between **base–shadow**, **base–post**, as well as **shadow–post** phases. The dashed lines indicate the distribution means.

Table 5: Results for the PaIntE analysis of the pitch accent realization — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{PaIntE} as estimated in separate models for the three different stimulus types and the three phase comparisons.

natural	base–shadow	base–post	shadow–post
intercept	0.11* (0.04)	−0.04 (0.06)	−0.17*** (0.04)
ACCENT		−0.04 (0.05)	
observations	2 136	2 065	2 031
diphone			
intercept	0.06 (0.04)	0.03 (0.05)	−0.01 (0.04)
ACCENT	0.01 (0.03)		
PAIRING	0.02 (0.03)		
observations	1 687	1 653	1 643
HMM			
intercept	0.12* (0.05)	0.05 (0.06)	−0.8 (0.05)
observations	1 637	1 599	1 580

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

In the diphone group, no accommodation is observed.

For the HMM group participants converged towards the model speakers during the shadowing task, and, although there is no significant divergence effect in the post phase, became indistinguishably close to the baseline level in the post phase.

The factors PAIRING and ACCENT did not show any significant effect on the intercept, although they improved overall fit in two models.

3.3.4 Word-level spectral composition

DTW COST The distributions of DID_{DTW} resulting from the DTW cost analysis is shown in [Figure 8](#). A positive DID_{DTW} indicates convergence to the model speakers, a negative DID_{DTW} divergence, and a DID_{DTW} close to zero maintenance of the temporal structure of the target words.

LMMs with DID_{DTW} as the dependent variable were fitted for each stimulus type and phase comparison data set separately, resulting in nine models. As expected, the effects are very small, since we are comparing the same word spoken by different speakers and the room for variation, on the temporal as well as the spectral level, is therefore quite limited. The factors PAIRING and SENTENCE were tested following the method in [Section 3.3.1](#). [Table 6](#) shows the parameter estimates for the nine final models.

In the natural and diphone data sets, participants converged to the model speakers in the shadowing task and diverged during the post phase, reaching the baseline level.

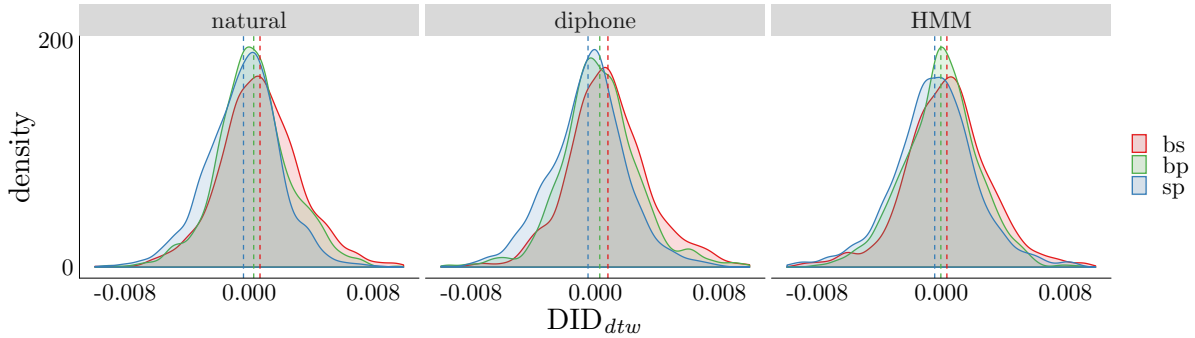


Figure 8: Distributions of DID_{DTW} for the DTW cost analysis in the three experimental groups. Comparisons are made between **base-shadow**, **base-post**, as well as **shadow-post** phases. The dashed lines indicate the distribution means.

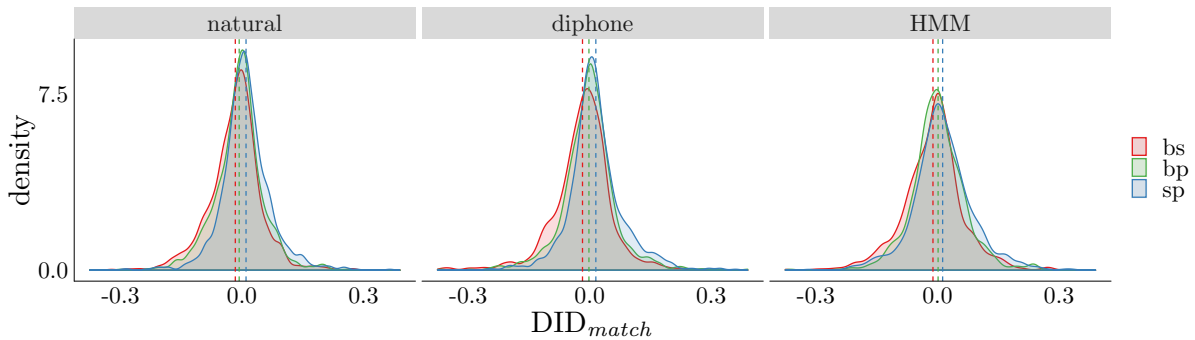


Figure 9: Distributions of DID_{match} for the amplitude envelope analysis in the three experimental groups. Comparisons are made between **base-shadow**, **base-post**, as well as **shadow-post** phases. The dashed lines indicate the distribution means.

For the HMM data set, convergence during the shadowing task is not significant; however, there is substantial movement away from the model speakers during the post phase and, eventually, no difference between baseline and post phase. Additionally, the sentence type accounts for variability in the case of the HMM data set: the diverging movement from shadowing task to post phase is stronger for the target sentences than for the filler sentences. Furthermore, the HMM base-post model suggests that — although there is no significant difference between baseline and post phase for the entire data set — the filler sentences are relatively closer to the model speakers in the post phase, compared to the baseline phase, while the target sentences are relatively farther away from the model speakers.

The factor PAIRING did not account for variance in the data.

MATCH VALUE The distribution of DID_{match} resulting from cross-correlating the time-warped amplitude envelopes is shown in Figure 9. Contrary to the other DID measures, a **negative** DID_{match} indicates convergence to the model speakers and a **positive** DID_{match} divergence from the model speakers with respect to the spectral composition of the target words. As before, a DID_{match} close to zero indicates maintenance of the baseline behavior.

Recall that the match value itself is bounded between 0 and 1 and can therefore be interpreted as probability, with 1 indicating maximal similarity, i.e., identity. The distribution of the match value is skewed towards 1, since we are comparing the same word spoken by different speakers. Using DID_{match} as a dependent variable resolved these issues and we could still fit LMMs for each stimulus type and phase comparison data set separately, which resulted in nine models.

Table 6: Results for the DTW analysis of the amplitude envelopes — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{DTW} as estimated in separate models for the three different stimulus types and the three phase comparisons.

natural	base–shadow	base–post	shadow–post
intercept	$6.95 \times 10^{-4***}$ (1.59×10^{-4})	2.81×10^{-4} (1.42×10^{-4})	$-3.78 \times 10^{-4**}$ (1.36×10^{-4})
SENTENCE	-0.01×10^{-4} (1.5×10^{-4})		
observations	1 136	1 139	1 137
diphone			
intercept	$8.28 \times 10^{-4***}$ (1.95×10^{-4})	2.9×10^{-4} (2.05×10^{-4})	$-4.86 \times 10^{-4*}$ (1.9×10^{-4})
observations	966	965	960
HMM			
intercept	3.40×10^{-4} (2.27×10^{-4})	-0.18×10^{-4} (1.68×10^{-4})	$-4.06 \times 10^{-4**}$ (1.32×10^{-4})
SENTENCE	-2.0×10^{-4} (1.39×10^{-4})	$2.03 \times 10^{-4*}$ (0.92×10^{-4})	$3.93 \times 10^{-4**}$ (1.24×10^{-4})
observations	908	909	909

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The factors PAIRING and SENTENCE were tested following the method in [Section 3.3.1](#). [Table 7](#) shows the parameter estimates for the nine final models.

As for the DTW analysis, participants shadowing natural and diphone stimuli converged to the model speakers during the shadowing task and diverged again in the post phase. However, the natural group did not reach the baseline level in the post phase, but stayed in between baseline and shadowing levels. The diphone group reached the baseline level in the post phase. Additionally, for the diphone group, the convergence effect in the shadowing task was influenced by the pairing of participants: the effect is stronger in mixed-sex than in same-sex pairings.

For the HMM group, the accommodating effect from baseline to shadowing phase again does not reach significance. There is, however, a significant movement away from the model speakers in the post phase, reaching the baseline level. As for DID_{DTW} , the sentence type accounts for variability in the HMM data set, with the target sentences showing a stronger divergence effect from shadowing task to post phase and reaching values farther from the model speakers in the post phase, compared to the baseline phase.

3.3.5 Individual results

To go beyond the analysis of accommodation on the group level, we assessed the performance of the individual participants with respect to the six features discussed above, focusing on the comparison of the baseline phase to the shadowing task.

For the DID measures (DID_{vowel} , DID_{PaIntE} , DID_{DTW} , DID_{match}), we conducted Wilcoxon signed-rank tests to determine whether each individual participant converged to or diverged from the model speakers (i.e., significant difference of their individual DID measure distribution

Table 7: Results for the match value analysis of the amplitude envelopes — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{match} as estimated in separate models for the three different stimulus types and the three phase comparisons.

natural	base-shadow	base-post	shadow-post
intercept	−0.017*** (0.003)	−0.007* (0.003)	0.010* (0.004)
PAIRING	0.003 (0.002)		
observations	1 138	1 142	1 139
diphone			
intercept	−0.019*** (0.005)	−0.002 (0.006)	0.016** (0.006)
SENTENCE	0.007 (0.005)		
PAIRING	0.005* (0.002)		
observations	964	971	965
HMM			
intercept	−0.011 (0.006)	0.001 (0.005)	0.011** (0.004)
SENTENCE	0.004 (0.005)	−0.006* (0.003)	−0.010* (0.004)
observations	907	910	911

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

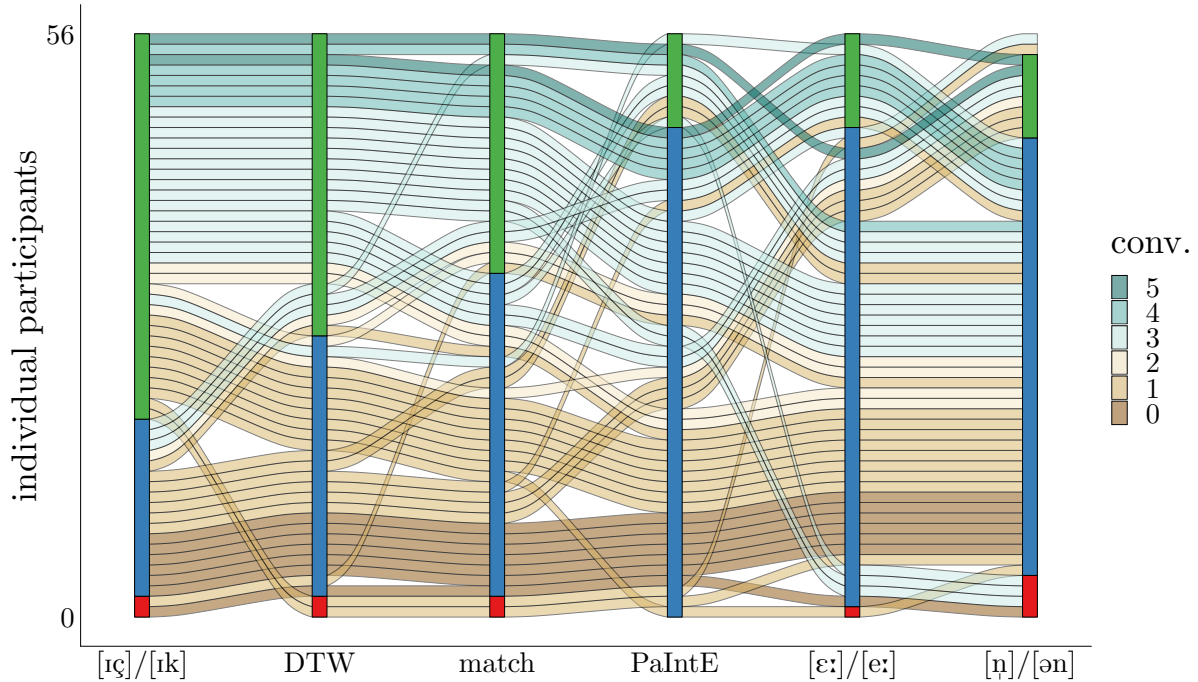


Figure 10: Accommodating behavior of the 56 participants for the comparison of baseline phase and shadowing task. Each vertical bar stands for one examined feature; the colors of the sections indicate whether the corresponding participant shows **convergence**, **maintenance**, or **divergence** for the respective feature. Two participants were excluded from the analysis of schwa epenthesis as they were the only participants producing schwa as a baseline preference. Hence they are not included in the rightmost vertical bar. Each horizontal line stands for one individual participant; the colors of the lines indicate with respect to how many features each participant converged to the model speakers (see legend).

from 0 at $\alpha = 0.05$), or whether they maintained the distance to the model speakers (i.e., no significant difference of their individual DID measure distribution from zero).

The degree of accommodation for the two binary contrasts [ɪç] vs. [ɪk] and [ŋ] vs. [əŋ] was assessed as the percentage of possible category changes. When determining the number of possible instances of accommodation, cases in which a participant already produced the same variant as the model speakers during the baseline phase were taken into consideration. The degree of accommodation for the binary contrasts was classified at the following thresholds so that single occurrences of convergence or divergence were still considered as maintaining behavior

- **convergence**: increase of same variants $\geq 20\%$
- **maintenance**: increase of same or different variants $< 20\%$
- **divergence**: increase of different variants $\geq 20\%$

Figure 10 shows a summary of the individual results. The six features under examination are ordered by decreasing number of individual participants converging significantly to them. Most participants converge with respect to the binary contrast [ɪç] vs. [ɪk] ($n = 37$, 66.1%), followed by the two measures related to the amplitude envelopes, DTW cost ($n = 29$, 51.8%) and match value ($n = 23$, 44.1%). The pitch accent realization assessed by the PaIntE model as well as the binary contrast [ɛː] vs. [eː] trigger convergence in 9 participants (16.1%), respectively, and the binary contrast [ŋ] vs. [əŋ] in 8 participants (14.8%).

Table 8 breaks these numbers down for the three experimental groups: natural, diphone, and HMM. Conducting 2×3 two-tailed Fisher’s exact tests for the distribution of participants

Table 8: Percentage of participants converging to the model speakers with respect to the respective feature in the three experimental groups, as well as in the entire participant group.

feature	natural %	diphone %	HMM %	total %
	$n = 21$ (for [ɲ]/[əɲ]: $n = 21$)	$n = 18$ $n = 17$	$n = 17$ $n = 16$	$n = 56$ $n = 54$
[ɾç]/[ɪk]	61.9	72.2	64.7	66.1
DTW cost	52.4	61.1	41.2	51.8
match value	47.6	50.0	23.5	41.1
PaIntE	23.8	5.6	17.6	16.1
[ɛ:] [e:]	28.6	11.1	5.9	16.1
[ɲ]/[əɲ]	19.0	5.9	18.8	14.8

Table 9: Distribution of participants converging to the model speakers with respect to a different number of features (0 to 6) in the three experimental groups, as well as over all 56 participants. Dominant groups are highlighted in gray.

no. of features	natural %	diphone %	HMM %	total %
	$n = 21$	$n = 18$	$n = 17$	$n = 56$
0	9.5	11.1	17.6	12.5
1	28.7	33.3	35.3	32.1
2	9.5	5.6	11.8	8.9
3	33.3	38.9	29.4	33.9
4	9.5	11.1	5.9	8.9
5	9.5	–	–	3.6
6	–	–	–	–

converging or non-converging (i.e., maintaining their behavior or diverging) over these three experimental groups, did not yield a significant result for any of the features. This suggests that, in every experimental group and for every feature, a similar proportion of participants converged to the model speakers.

Figure 10 further illustrates the number of features with respect to which each individual participant converges out of the possible 6 under examination. No participant actually converged to all 6 features and only two participants converged to 5 features. Five participants converged to 4 and 2 features, respectively. The majority of the participants accumulated at 3 features ($n = 19$) and 1 feature ($n = 18$). A total of seven participants did not converge at all.

Table 9 details how these different degrees of convergence are distributed over the three experimental groups: natural, diphone, HMM. Conducting a 6×3 two-tailed Fisher’s exact test for the distribution of participants converging to the model speakers for 0 to 5 features over these three experimental groups, did not yield a significant result. This suggests that, in every experimental group, a similar proportion of participants showed convergence to the model speakers with respect to the same number of features.

Some cases of individual divergence from the model voices were found as well, i.e., 4 cases for [ɲ]/[əɲ], 2 cases for [ɾç]/[ɪk], DTW cost and match value, respectively, and one case for [ɛ:]|[e:]. No individual divergence was found for the pitch accent comparison with PaIntE.

We identified the participants who stated in the questionnaire after the experiment that they converge to dialects of other regions ($n = 8$) or intentionally imitate the pronunciation of interlocutors ($n = 9$; see Section 3.2.1.2). Only two participants appeared in both groups.

For the first group, we could assume that they would specifically pick up the two regionally distributed features, [ɪç]/[ɪk] and [ɛ:]/[e:]. However, only four of the eight speakers converged with respect to [ɪç]/[ɪk] and none with respect to [ɛ:]/[e:], while one speaker from this group even diverged with respect to [ɪç]/[ɪk]. This does not indicate a particular inclination for convergence to regional features. In terms of overall convergence, the members of this group were not particularly successful either: they converged to a maximum of 3 features.

The second group, namely the speakers who claimed to intentionally imitate the pronunciation of interlocutors, also did not include any of the speakers converging to more than 3 features. With respect to [ɪç]/[ɪk] and DTW cost, five of the nine speakers converged, respectively; two speakers each picked up the schwa from the model voices and converged with respect to the amplitude envelope match; only one speaker converged to the pitch accent realization and none with respect to [ɛ:]/[e:]. Divergence was not found in this group. These results do not notably reflect a possible effect of intentional imitation.

Of the two speakers who claimed to converge to dialects of other regions and to intentionally imitate interlocutors, one converged to 3 features and the other to none.

3.4 DISCUSSION

The goal of the present study was to investigate phonetic accommodation of human interlocutors in a shadowing task with a specific focus on the accommodation effect evoked by synthetic stimuli. Diphone- and HMM-based synthetic stimuli, as well as natural stimuli, were used in the process. The language under investigation in this study is German. The shadowing task was carried out by native speakers of German.

To get a broader picture of phonetic accommodation in the experimental data, we examined features pertaining to different phonetic domains, i.e., variation of segment-level phenomena as well as variation with respect to pitch accent realization (local prosody) and word-based global similarity (temporal structure and distribution of spectral energy). The segment-level phenomena under investigation are allophonic variation of [ɪç]/[ɪk] and [ɛ:]/[e:], as well as schwa epenthesis. To make the systematic investigation of accommodation with respect to these features possible, the stimuli for the shadowing task were chosen depending on the participants' baseline productions: the participants were presented with the opposite of their preferred variants.

Analyses were carried out at the group level and for individual participants. Since the experimental procedure comprised three phases — baseline production, shadowing task, and post production — we drew three comparisons for each group data set, namely baseline vs. shadowing, shadowing vs. post, and baseline vs. post. For the individual behavior, we focused on the comparison of baseline phase and shadowing task. Combining the results of these analyses provides an overview of the phonetic accommodation in the present shadowing corpus.

3.4.1 *Word ending <-ig>*

The allophonic variation [ɪç] vs. [ɪk] was most successful in triggering convergence when looking at the individual results, with two thirds of all participants converging to the model speakers during the shadowing task. This may be due to the relative salience of this feature, even though it was embedded in a larger utterance, and to the fact that participants can presumably access the binary variation between fricative and plosive more easily than other, more gradual, changes.

At the group level, we found the same pattern for the natural and the diphone group: convergence to the model speakers during shadowing, divergence in the post phase, although not

entirely falling back to the baseline level, but rather sustaining the convergence effect in attenuated form.

In the HMM group, although the relative increase of *same* forms is equal to the diphone group, the convergence effect is not significant in the statistical model. The group does, however, show a divergence effect from shadowing task to post phase and reaches the baseline level in the latter.

These results, in combination with the fact that a similar proportion of participants converges to the model speakers in all three groups, shows that [ɪç] vs. [ɪk] is a rather successful target for convergence in native speakers of German for natural as well as synthetic stimuli.

Neither the preference for one or the other variant in the baseline phase, nor the attitude towards the dispreferred variant being positive or negative, had an impact on the accommodation for this feature. The standard variant [ɪç] does not seem to be an easier target for convergence. This may have to do with the fact that the participants were in many cases not certain which of the variants is the standard, or at least did not have a negative attitude towards the variant they believed not to produce.

3.4.2 *Word-level spectral composition*

The second and third most frequent cases of convergence for individual speakers were found in the measures pertaining to the word-based global similarity.

The first measure, i.e., the cost of the dynamic time warping (DTW) process, emphasizes changes in the temporal domain while taking the spectral domain into account. It shows that more than half of the participants converged to the model speakers with respect to word-based timing.

The second measure, i.e., the match value, resulted from cross-correlating the time-warped amplitude envelopes. The analysis of the match value shows that almost half of the participants converged to the model speakers with respect to word-based distribution of spectral energy alone, i.e., excluding timing.

Figure 10 shows that these two groups widely overlap, with 21 participants converging with respect to both features, eight participants converging only with respect to DTW cost, and two only with respect to the match value. This was expected, since these measures are closely related. However, taking both measures into account disentangles the contributions of timing and energy distribution across spectral bands to the accommodation effect.

On the group level, both measures behave similarly. There is a pattern of convergence to the model speakers in the shadowing task and divergence in the post phase reaching the baseline level, which occurred in the natural and diphone group for both measures. The only exception to this pattern is that the match value does not reach the baseline level in the natural group, meaning that the convergence effect was partially sustained in this case.

As before, the HMM group behaved differently. There is no significant convergence effect in the shadowing task for either of the two measures. However, as for the other two groups, we found a significant divergence effect in the post phase and the HMM group did reach the baseline level in the post phase.

It may have been the case that the target sentences, in particular, drive the accommodation for DTW cost and match value, since they specifically offer room for convergence in the form of the dispreferred segmental variants. This presupposes, of course, that participants accommodate with respect to the offered variants. Overall, such an influence of the sentence type did not manifest itself, especially not in the actual shadowing phase. Only in the HMM group, the sentence type emerged as a significant predictor: the divergence effect in the post phase was stronger for the target sentences resulting in post productions which were relatively farther from

the model speakers, compared to the baseline phase. Whether this behavior is indeed causally related to the dispreferred segmental variants in the target stimuli remains unclear.

The participant–model pairing only surfaced as a significant predictor in the match value analysis of the diphone group: convergence during the shadowing task was slightly stronger in mixed-sex pairings. Remember that prior studies on phonetic accommodation found both cases of more convergence in mixed-sex and same-sex pairings. Our results, although showing one incident of increased convergence in mixed-sex pairings, do not make a strong case in favor of speakers converging more to a model talker of the opposite sex.

3.4.3 Long vowel <ä-> and pitch accent realization

The number of individual participants converging drops drastically for the allophonic variation [ɛ:] vs. [e:] and the pitch accent realization, to only 16%, respectively. Figure 10 shows that among these, there is only one participant who converged with respect to both features.

On the group level, the observed patterns for both measures are again very similar to each other and also distinct from the patterns observed for other measures. The natural group converged to the model speakers, with respect to both vowel production and pitch accent realization, and diverged again in the post phase reaching the baseline level.

The HMM group showed convergence in the shadowing task, too. However, the divergence effect in the post phase was not significant and the group still reached the baseline level.

The diphone group, finally, did not show any accommodation with respect to vowel quality and pitch accents.

The speaker preference had a significant influence on the vowel production in the natural group: the convergence effect in the shadowing task was stronger for those participants whose baseline preference was [ɛ:]. One possible explanation may be that [ɛ:] is an easier target for convergence, since it is the standard German form and therefore more prestigious. However, although still being considered the prescriptive norm, [ɛ:] is generally used less frequently by native speakers of German in Germany. Therefore, it may also be the case that [ɛ:] is more salient to hearers and therefore picked up from the speech input more easily. Remember that atypicality has been shown to promote accommodation for some speakers (Babel et al., 2014). As for the allophonic contrast [ɪç] vs. [ɪk], the attitude towards the variant participants did not believe to produce themselves did not influence accommodation for the vowel quality. Recall that this attitude was predominantly positive, only 20% of participants had a negative attitude towards the dispreferred variant in the case of [ɛ:] vs. [e:].

For DID_{PaIntE} as a measure of similarity in pitch accent realization, the accent type was tested as an additional factor. The expected effect of higher perceptual salience of nuclear as opposed to prenuclear pitch accents did not appear.

Eventually, the participant–model pairing did not surface as a significant predictor in the analysis of vowel quality and pitch accent realization. This means that it did not make a difference for the accommodating behavior whether participant and model were of the same or different sex.

Note that the Euclidean distance between the 6-dimensional PaIntE vectors that underlies the DID_{PaIntE} measure is a rather coarse estimation of similarity in pitch accent realization. The relative contribution of the individual PaIntE parameters to the accommodation effect is subject to further analysis.

3.4.4 *Word ending <-en>*

Epenthesis of schwa in the word ending ⟨-en⟩ was least successful in triggering accommodating behavior.

On the individual level, there are still about 15% of participants who converge to the model speakers with respect to schwa epenthesis. However, taking the entire group into account, the only significant convergence effect emerged in the shadowing task of the natural group. Contrary to every other feature, even in the natural group there was no significant divergence effect in the post phase and the baseline level was still reached, which suggests a rather weak effect.

For both synthetic groups, no accommodation was observed, nor did the attitude towards schwa epenthesis play a role in the statistical models. As stated initially, producing a schwa in the word ending ⟨-en⟩ is rather unusual. This statement was confirmed by the fact that the vast majority of the participants (54 of 56) preferred schwa elision in the baseline production and hence shadowed [ən] stimuli. Recall that only these 54 participants were subsequently analyzed. It was claimed above that such atypicality might promote accommodation. We can assume, however, that there are limits to how atypical such a variant may be to still be considered a target for accommodation. It may have been the case that an unusual variant such as [ən] would be more likely picked up from a synthetic than a natural voice, since hyperarticulation occurs in human-computer interaction (HCI; Burnham et al., 2010). However, synthetic voices alone do not make HCI. The present shadowing scenario lacks the need to be understood by the interlocutor, which is an important layer to HCI and spoken interaction in general. Therefore, even schwa might be picked up in a more conversational scenario, which would presumably also trigger the speaker’s belief that converging to the computer leads to greater communicative success (cf. Branigan et al., 2010).

3.4.5 *Individual behavior*

Concerning the individual accommodation behavior of the participants in the shadowing task, we found mainly convergence and maintenance, as well as some cases of individual divergence. Note that we are taking a categorical approach and do not further distinguish degrees of convergence or divergence here. The participants varied regarding the number of tested features they accommodated to. This supports our initial assumption that we would find considerable variation between the participants, which manifests itself in the form of accommodation to different subsets of the features. The two top convergers — both from the natural condition — accommodated to five out of the six examined features; for both of them it was the schwa epenthesis which they did not pick up from the model speakers.

The self-assessment of a few participants stating that they converge to dialects of other regions or consciously imitate the pronunciation of their interlocutors was not confirmed by the data.

Recall that the regionally distributed features were deliberately chosen not to be strong dialectal markers. In order to trigger the convergence to a dialect that the participants were referring to, more salient dialectal features may be required.

For a speaker to be able to intentionally imitate their conversational partners, the salience of the features in question plays a role, as does their selective realizability. In the present study, the allophonic contrasts and the schwa epenthesis lend themselves as targets for such intentional imitation. The other features, namely the pitch accent realization, the temporal structure and the distribution of spectral energy, seem to be less easily imitated intentionally, but rather a result of a more holistic high-level adjustment. This should be examined in a further study, in which participants are explicitly asked to imitate the stimuli.

It is not unexpected that the participants’ self-assessment of phonetic accommodation is often inaccurate. An adaptation at the phonetic level is certainly more difficult for speakers to evaluate and quantify than, for example, an adaptation at the lexical level, where the use of certain words is easier to capture.

Another factor that may influence individual differences in accommodating behavior is the general speaker disposition, which includes aspects such as innate phonetic talent, personality traits, and cognitive abilities. Yu et al. (2013) observed, for example, that Openness to Experience and a strong attention focus were positively correlated with the degree of word-initial voice onset time (VOT) convergence during a non-conversational phonetic imitation task in English.

Lewandowski and Jilka (2019) examined accommodation of word-based amplitude envelope match in dialogs between non-native and native speakers of English. They found a higher degree of convergence among phonetically talented, more neurotic and more open speakers, as well as among speakers with higher attention scores. Convergence was found to be negatively correlated with behavioral inhibition.

This factor was not included in the present study and deserves further investigation.

3.4.6 *Limitations of difference in distance*

Cohen Priva and Sanker (2019) point out potential limitations of the DID measure to account for convergence in corpora of spoken interaction and, particularly, for the attempt to establish individual differences with respect to accommodating behavior. Their three main concerns are: firstly, in an extreme case of over-convergence, the DID measure might not reflect the convergence that has taken place, but suggests maintaining behavior; secondly, convergence might be underestimated for small initial distances between participant and model speaker; and lastly, the baseline measures might not be representative of the speaker’s usual behavior and therefore convergence might partly be an effect of becoming closer to the latter independent of the interlocutor’s influence.

Although Cohen Priva and Sanker (2019) examined a very different set of features from the one used in the present study, namely median and range of fundamental frequency, speaking rate, as well as the ratio of two types of filled pauses, and mention that their findings may be less problematic for other features, their concerns should be discussed with respect to their implications for the present study.

For the DTW cost (DID_{DTW}) and the match value (DID_{match}), the concern regarding over-convergence does not hold, since identity is an upper boundary to similarity inherent to these measures. This is not the case for the vowel quality measure (DID_{vowel}). It needs to be considered that, contrary to the one-dimensional features examined in Cohen Priva and Sanker (2019), vowel quality is a two-dimensional feature here, which makes the definition of over-convergence difficult. However, the space to move is somewhat bounded by neighboring vowel categories. Given that we systematically maximize the baseline difference between speaker and model and minimize contextual variability (see discussion below), we assume that cases of over-convergence to the extent that they will be mistaken for maintenance are unlikely to occur. For the comparison of pitch accent realizations within the six dimensions of the PaIntE model (DID_{PaIntE}), the definition of over-convergence becomes even more difficult and would have to be established for individual dimensions. The dimensions themselves differ with respect to their linguistic interpretability and presumably their relative contribution to the perception of pitch accents. Specifically, this relative contribution would have to be examined further to establish what over-convergence really means in the realm of pitch accent realization. A certain limitation for over-convergence seems to be given by the plausible and well-formed pitch accent shapes.

Regarding the concerns about variance in initial distance to the model speakers, the features examined in the present study are very different from each other. While participants are expected to exhibit small initial distances to the model speakers for the DTW cost and the match value, since we compare the same lexical items, the design of the study maximizes initial distances with respect to the vowel quality for all participants by presenting them with instances of their dispreferred variant. In the case of the allophonic variation, maximizing this distance is possible without leaving the range of normal human performance, and therefore without jeopardizing the ecological validity of the findings. The initial distances in PaIntE parameters are mainly guided by the sentence structure and an assumed default placement of pitch accents. If the initial distances vary mainly by feature and are rather balanced between speakers for the same feature, the concern of potential underestimation of convergence would be less of a problem for the analysis of the individual behavior of different participants, but more so for the different features as a whole. However, the small initial distances for the DTW cost and the match value do not exhibit the same problem as small initial distances in speaking rate, for example, since there is very little expected variability of these features as opposed to a feature like speaking rate.

In accommodation research, it is always a point of concern whether the selected baseline is representative of the speaker’s usual behavior. The shadowing paradigm entails a switch of elicitation technique — in the present case from reading text to repeating speech, which is a certain limitation. In the specific shadowing experiment at hand, there may be a further effect of first exposure — in the baseline phase — versus repetition — in the shadowing task and the post phase. However, this repetition, or in other words the stability of the linguistic context throughout the experiment, also enhances the relative representativeness of the baseline productions: Although a lot of variation is possible within a vowel category, the variation occurring in our data is limited due to the comparison of identical vowel contexts (i.e., lexical items) in all three phases of the experiment; the same is true for the word-based measures and pitch accent realizations, which are themselves embedded and tested in the same sentences throughout the experiment. Moreover, allophonic variation, pitch accent realization, and word-based intensity distribution of targets embedded in short utterances are less likely affected by extreme baseline values than measures stemming from targets read and shadowed in isolation. These features also seem less prone to task-induced variation as opposed to features such as the range of fundamental frequency or speaking rate, which are likely to change over the course of an interaction as a result of familiarization with the task at hand.

MacLeod (2021) adds to the discussion about the limitations of the DID measure by exploring accommodation with respect to word duration in a speech shadowing corpus. In particular, she sheds light on the fact that findings suggesting that greater starting distance between interlocutors leads to more convergence (e.g., Babel, 2012; Walker and Campbell-Kibler, 2015; Kim and Clayards, 2019; Clopper and Dossey, 2020) may be due to a bias introduced by the way DID is calculated. However, in case the model speaker production lies outside the participants’ baseline production range, which explicitly increases the starting distance for all of them — as for the vowel quality in the present experiment —, she assumes that the bias is alleviated or even eliminated.

As an alternative to the DID measure, Cohen Priva and Sanker (2019) and MacLeod (2021) suggest *linear combination*, where a LMM is used to predict the participants’ values of a phonetic feature during an interaction — e.g., while shadowing — by means of the participant’s baseline values — capturing consistency — and the model speaker values — capturing accommodation. This approach overcomes some of the problems of the DID measure but, according to (MacLeod, 2021), also has some limitations of its own. Relevant to the present work is the fact that Euclidean distances between participant and model, which we use for the vowel quality and

pitch accent analyses, cannot be processed in such a model. Similarly, the measures of word-level amplitude envelope analysis inherently compare participant and model and thus cannot be split in the manner required by the linear combination approach.

While we certainly must keep in mind the potential limitations of the DID measure, we hope to have shown that for certain features they do not apply or apply only partially, and alternative approaches are not feasible in every case. It is safe to say that the concerns have to be evaluated separately for each feature used to examine accommodation.

3.4.7 *Natural vs. synthetic speech*

Coming back to the focus of the present study, namely the question whether participants behave similarly when confronted with either natural or synthetic stimuli, we can summarize that the participants of the natural condition have accommodated during the shadowing task in the expected direction, i.e., towards the model speakers, on all tested features. Remember, however, that the effect was weak for schwa epenthesis, which supports the assumption that speakers accommodate less to unusual features. Furthermore, with the exception of schwa epenthesis, the participants of the natural condition always diverged significantly from the model speakers in the post phase, partly reaching the baseline level (vowel quality, pitch accent realization, and DTW cost), partly showing a sustained convergence effect (allophonic variation [ɪç] vs. [ɪk] and match value).

The participants of the two synthetic conditions did not show an accommodation effect for schwa epenthesis. The two other cases for which no accommodation was found, are the vowel quality and pitch accent realization measures for the participants of the diphone condition. However, for the remaining features — allophonic variation [ɪç] vs. [ɪk], DTW cost, and match value — the participants of the diphone condition behaved similarly to those of the natural condition.

The participants of the HMM condition, finally, never showed the complete pattern of significant convergence in the shadowing task, complemented by significant divergence in the post phase reaching the baseline level. However, they always showed substantial movement within the overall constellation of the three phase comparisons carried out in the present study, which suggests that this general pattern — even if in a weaker form — is underlying the HMM data as well. That is, we either found convergence in the shadowing task and no significant divergence in the post phase while still reaching the baseline level (vowel quality and pitch accent realization), or no significant convergence in the shadowing task, yet divergence in the post phase, again reaching the baseline level (allophonic variation [ɪç] vs. [ɪk], DTW cost, and match value).

For the HMM voices, our initial assumption that certain phonetic features might not be clearly distinct in the synthetic stimuli proved true: with the synthesis process applied here, it was not possible to produce female HMM stimuli with a clearly distinguishable target allophone [ɛ:]. The seven participants of the HMM condition with a baseline preference for [e:] therefore heard a lower total number of clear [ɛ:] target allophones, namely only from the male model voice, which could be a disadvantage for the emergence of an accommodation effect. Nevertheless, we found overall convergence of vowel quality for the entire HMM group, in contrast to the diphone group, in which all participants heard clear target allophones from both model voices, but still no overall convergence occurred.

In summary, we observe the same behavior in the diphone group as in the natural group with respect to several features and no accommodation for other features. For the HMM group, we observe a similar underlying pattern as for the natural group, but in some individual phase comparisons the effect is not up to par with that of the latter. Technical differences between the

synthesis methods may have contributed to the differences in performance. However, neither of the two synthesis qualities made accommodation impossible.

3.4.8 *Model voices*

One aspect which needs to be taken into consideration is that the six model voices employed in the present study differ with respect to stimulus type (natural, diphone, and HMM) and sex (female and male), but of course exhibit a variety of other characteristics that may affect the degree of accommodation to them, for example their perceived naturalness and likability (see [Section 3.2.1.1](#)).

The participants of the natural condition gave higher ratings of naturalness to the voices they shadowed than the participants of the HMM condition. The diphone voices were rated as sounding least natural by the participants of the respective condition. This supports our initial assumption that the participants would recognize the synthetic voices as non-human. We had further speculated that this could trigger a feeling of social separation in the participants, which may lead to a reduction of the convergence effect or even to divergence. It may be the case that this factor indeed contributed to the overall weaker effects of the synthetic stimuli. However, the diphone stimuli that were rated as most unnatural sounding showed effects of similar strength as the natural stimuli for some of the examined features and it is unclear why the social component should only influence such a subset.

In terms of likability, the natural voices were rated on a par with the HMM voices, while the diphone voices again received the lowest ratings. Thus the diphone voices, on the one hand, set themselves apart from the two other voices by their lower naturalness and likability, but still triggered considerable accommodation effects for a subset of the examined features. The HMM voices, on the other hand, although being as likable as the natural voices, did not trigger the same strength of accommodation for most examined features.

Such differences need to be explored further by testing various voices of each stimulus type. However, the present experiment showed that synthetic voices, while partly reducing the strength of effects, do trigger accommodating behavior. As for the natural voices, convergence during interaction followed by divergence after the interaction is the predominant pattern.

3.5 CONCLUSION

The present shadowing experiment used natural and two types of synthetic voices (diphone- and HMM-based) to test whether native speakers of German accommodate to these voices when repeating short German sentences after them. The use of short sentences as target utterances provided a controlled context while still keeping a broad focus. The examined features pertain to different phonetic domains allowing for an extensive assessment of the participants' behavior: allophonic variation ([ɛ:] vs. [e:], [ɪç] vs. [ɪk]), schwa epenthesis, realization of pitch accents (PaIntE parameters), as well as word-based temporal structure (DTW cost) and distribution of spectral energy (match value). We predicted accommodation in the form of convergence to occur with respect to these features.

The results of the individual accommodation behavior analysis need to be interpreted with caution due to potential limitations of the difference in distance (DID) measures. Concerning the predicted individual variation, we found that the participants converged to varying subsets of 0 to 5 out of the six examined features, with the most individual convergers for [ɪç] vs. [ɪk], followed by DTW cost and match value, and least for the PaIntE parameters, [ɛ:] vs. [e:], and the schwa epenthesis, in that order. Very few cases of divergence were found for all features

but the pitch accent realization for which no such cases occurred. Although almost half of the participants individually converged to at least three out of six features, this demonstrates that accommodation with respect to one particular feature does not necessarily predict the behavior with respect to another feature.

Describing accommodating behavior more broadly for different speaker groups is a step towards modeling the given individual variation for the HCI context in order to gain a better understanding of the user or even to implement such behavior in the computer.

On the group level, the participants of the natural condition converged to all features under examination, however very subtly so for schwa epenthesis. The participants of the diphone condition behaved similarly to the natural group with respect to several features ([ɪç] vs. [ɪk], DTW cost, and match value) or did not show any accommodation for other features. For the participants of the HMM condition, the effects were less clear overall. A significant convergence effect in the shadowing task only emerged for [ɛ:] vs. [e:] and the PaIntE parameters. However, taking into account the post production, we conclude that the same pattern of convergence in the shadowing task and divergence after the shadowing task observed in the natural group for all features but schwa epenthesis, is underlying the HMM group, too.

The present experiment showed that German native speakers converge to various features ranging from segmental variation and local prosody to the word-based temporal structure and distribution of spectral energy when shadowing short sentences from natural voices. For segment-level features, like the ones we examined, accommodation had previously only been investigated in shorter, mono- or bisyllabic utterances (Babel, 2012; Dufour and Nguyen, 2013; Mitterer and Müsseler, 2013). We could show that such features are also picked up from longer utterances. The analysis of pitch accent realizations differed from an earlier approach investigating conversational speech (Schweitzer et al., 2017) in that it included the accent type. The assumption that nuclear pitch accents might cause a greater convergence effect due to their higher perceptual salience was not confirmed. An earlier approach to investigate the accommodation of the word-based distribution of spectral energy in conversational speech (Lewandowski, 2012; Lewandowski and Jilka, 2019) was expanded in this study to include the aspect of temporal structure, showing a convergence effect for the distribution of energy over spectral bands, even when convergence with respect to timing is already accounted for.

As the participants in the present experiment shadowed both a female and a male voice, we examined whether they showed a higher degree of accommodation to a model talker of the same or the opposite sex. However, no strong tendency could be observed, since only one case of increased convergence in mixed-sex pairs was found.

Regarding the comparison of natural and synthetic model speakers in speech shadowing, synthetic voices were found to induce accommodating behavior as well, but partly reduce the strength of effects found for the natural voices. One difference between the synthetic voices used in this study was that the diphone voices were perceived as generally more unnatural and unlikable than the HMM voices, which could be a source for different accommodating behavior towards them. The predominant pattern of accommodation for all voice types, however, was convergence during the interaction, followed by divergence after the interaction. We conclude that phonetic accommodation does occur in human-computer interaction involving synthetic speech, but for the phonetic features and model voices examined here, to a lesser extent overall than in human-human interaction.

4 WIZARD-OF-OZ EXPERIMENT

In the Wizard-of-Oz (WOz) experiment, native and non-native speakers of German interact with the supposedly intelligent spoken dialog system (SDS) *Mirabella* while the experimenter is controlling the output of the system behind the scenes. *Mirabella* is presented as a tutoring system for learning German as a foreign language. Her voice is either natural or synthetic — the latter is generated with Hidden Markov model (HMM)-based synthesis. The investigated phonetic features are allophonic contrasts and the intonation of wh-questions. The influence of the participants' *Big Five* personality traits on their accommodating behavior is considered.

4.1 HYPOTHESES AND PREDICTIONS

The virtual language learning tutor *Mirabella* was designed to lead a friendly conversation, i.e., she explains the tasks at hand to the participants, asks whether everything was understood, praises and encourages the participants, and does not exhibit extreme behavior that would provoke counteraction. Therefore, we have no reason to believe that the participants would show divergence in conversation with *Mirabella*, for example in order to increase the social distance to her. We expect mainly converging behavior on the part of the participants. To assess the impression that the participants have of *Mirabella*, we collect simple scores for her perceived likability and competence, as well as her intelligibility and response time after the experiment (see [Section 4.3.1](#)).

The first part of the present study compares two voice types, i.e., a natural and a synthetic voice, in their ability to trigger accommodating behavior in users of a SDS. As discussed above, it was shown in WOz experiments using embodied graphical agents that both voice types can individually lead to phonetic accommodation of global acoustic-prosodic features (e.g., Bell et al., 2003; Oviatt et al., 2004; Staum Casasanto et al., 2010; Gijssels et al., 2016). We expect that accommodation also occurs for the more locally anchored phonetic features investigated in the present study. See [Section 4.2.1](#) for more details on the tested features and specific predictions.

In the case of *Mirabella*, the two voice types are directly compared using the same SDS and a possible effect of the virtual interlocutor's visual appearance is excluded, as she communicates only through her voice. We expect both versions of *Mirabella* to trigger accommodating behavior in the participants. The natural version may be at an advantage, since it has been shown that natural voices are often preferred in tutoring settings, specifically so when there is no accompanying graphical representation of the virtual interlocutor (Baylor et al., 2003; Atkinson et al., 2005). Moreover, the natural version may be more readily perceived as a social actor, which according to the Communication Accommodation Theory (CAT) would promote accommodation. In addition, our own prior shadowing experiment has shown a stronger accommodation effect for natural voices compared to different synthetic voices (see [Chapter 3](#)).

However, the synthetic version of *Mirabella* may have an advantage in that it sounds rather atypical, which has been shown to increase convergence for some speakers (Babel et al., 2014). Furthermore, the synthetic version may be perceived as more machine-like and therefore more likely to benefit from convergence (Branigan et al., 2010).

Although the experiment is situated in a language learning context, the participants of this first part of the study are native speakers of German. We therefore essentially investigate L1–L1 communication.

The second part of the present study extends the experiment to L1–L2 communication by having a group of non-native speakers of German, i.e., native speakers of French, interact with the natural version of Mirabella. See [Section 4.2.4](#) for more information about the participants.

In both contexts, the question remains open, whether Mirabella is actually perceived as a “native speaker” of German by the participants. It is conceivable that a SDS which does not possess complete linguistic flexibility is not regarded as a fully competent speaker of the language in question and that, with respect to accommodation, similar mechanisms apply as in dialogs with non-native speakers (see [Costa et al. \(2008\)](#) for an overview). Specifically for the native speaker group, the belief in the limited linguistic competence of the addressee may, for example, lead to a higher degree of adaptation on the part of the participants. In contrast, native speakers are likely to be confident in their own pronunciation and may perceive the SDS as hierarchically inferior to them — two aspects that contradict a strong tendency towards convergence (see [Gregory and Webster \(1996\)](#) for hierarchy).

For the non-native speakers, it is conceivable that they show more adaptation than the native speakers because they are less confident in their own pronunciation and Mirabella, if perceived as a native speaker of the target language, is hierarchically superior to them. On the other hand, it is possible that the non-native speakers have greater difficulty in perceiving the phonetic detail in Mirabella’s speech and implementing it in their production, and therefore accommodate less.

Apart from the general expectation to find convergence to Mirabella at the group level, we predict that the individual participants will differ considerably in their behavior, as was the case in previous studies (e.g., [Pardo et al., 2018](#); [Chapter 3](#)). To further investigate a possible source of this variation, we include the *Big Five* personality traits in the analysis (see [Section 4.3.6](#)). Openness and Neuroticism have been suggested to promote convergence in the context of phonetic accommodation ([Yu et al., 2013](#); [Lewandowski and Jilka, 2019](#)). We therefore expect a possible influence of these factors on our data.

4.2 MATERIAL AND METHODS

The WOz experiment is presented to the participants as an interaction with an application for learning the German language. This resembles a realistic use case as it simulates a scenario from the growing field of computer-assisted language learning (CALL). The interaction is presented to be about “learning German”; the topic of pronunciation is not mentioned at any point. For the native speaker group, the experiment is disguised as a test run of the application before it is deployed to learners of German. The same is true for the non-native speakers, except that they are actually part of Mirabella’s supposed target audience. Both contexts motivate the situation for the participants and shift the focus from the participants being tested themselves to the system being under scrutiny.

The system introduces itself as a female tutor for German as a foreign language called *Mirabella*. During the experiment, the participants only interact with Mirabella’s voice; she is not represented by an embodied virtual agent. All utterances available to the *wizard*, i.e., the experimenter, to choose from during the experiment were either pre-recorded by a native speaker of German or pre-synthesized (see [Section 4.2.3](#)). These stimuli are manually played back to the participants by the experimenter, while the participants believe to interact with a fully automatic SDS which understands their speech input and reacts accordingly.

During the interaction with Mirabella, the participants are seated in front of a monitor in a sound-attenuated booth and recorded with a sampling rate of 48 kHz using a stationary cardioid microphone. Mirabella’s utterances are played to the participants over headphones. The recordings are followed by a questionnaire about the participants themselves and their opinion about Mirabella, as well as the German version of the NEO Five Factor Inventory

(NEO-FFI; Borkenau and Ostendorf, 2007) for the native speakers of German and the French version of the Big Five Inventory (BFI; Plaisant et al., 2005; Plaisant et al., 2010) for the native speakers of French to collect information about their personality traits.

4.2.1 Tasks and tested features

The interaction with Mirabella consists of four tasks and lasted about 30 minutes for the native speakers and 40 minutes for the non-native speakers (including short breaks after tasks 1 and 3). Mirabella explains the tasks to the participants and takes part in them. The interaction is supported by visualization of the tasks on a screen. The features tested for accommodating behavior are the intonation of constituent questions such as “Wo hat sich der Hase versteckt?” (*Where did the rabbit hide?*) and the variation of the German allophone pairs [ɛ:] vs. [e:] as a realization of the long vowel ⟨-ä-⟩¹ in stressed syllables, e.g., Käse (*cheese*), and [ɪç] vs. [ɪk] as a realization of the word ending ⟨-ig⟩², e.g., Honig (*honey*).

The first two tasks familiarize the participants with the system and the text material occurring in the experiment and elicit baseline productions of the target utterances.

The two tasks testing for accommodation are a *question-and-answer (Q&A) game* of two rounds (task 3), in which the participants and Mirabella take turns asking each other questions about the location of the animals on the screen, and a *map task* of four rounds (task 4), in which the participants have to describe their way to a destination while asking Mirabella about the hidden objects they encounter.

4.2.1.1 Task 1 — allophonic variation, baseline

This task ensures that the participants know all 71 German words (24 targets — 12 per allophonic contrast — and 47 fillers) they need to recognize during the experiment (see Appendix D) and reveals which versions of [ɛ:] vs. [e:] and [ɪç] vs. [ɪk] they produce naturally.

The set of words contains 35 nouns, which are presented to the participants as pictures, and 36 adjectives, which are presented in their English translations for the native speakers of German and in their French translation for the native speakers of French. The participants name the pictures and translate the English/French adjectives to German by pronouncing them in the carrier sentence “Das Wort ⟨item⟩ kenne ich.” (*I know the word ⟨item⟩*.)

In case they do not recognize an item, they state: “Das Wort kenne ich nicht.” (*I do not know the word.*)

In the event that an item is not recognized (correctly), the participants are provided with the initial letter of the word in question and the opportunity to try again (see Figure 11). If they fail a second time, the word is presented in written form and needs to be read out loud to move on with the task. That way, while avoiding to present the written form as long as possible, all items are uttered by every participant.

Among the L1 German speakers, the items were correctly recognized at the first attempt in 88 % of all cases. In 9 % of the cases the initial letter was provided and in 3 % of the cases, of which just over half were target items (56 %), the word was eventually read. The L2 German speakers recognized the items correctly at first sight in 75 % of all cases. The first letter was needed to correctly name 11 % of the items and in 14 % of the cases the word was read. With 52 %, the amount of target items in the group of *read* words was similar to that of the L1 German speakers.

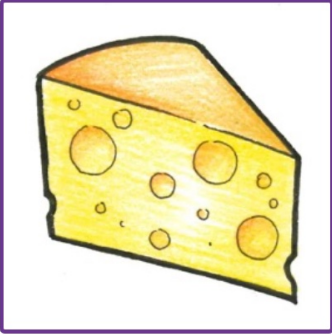
¹ In one of twelve target items the vowel occurs word initially; in two target items the graphematic form is ⟨äh⟩. For simplification, we are referring to all of these with ⟨-ä-⟩.

² In many cases this ending constitutes a morphemic suffix.

Aufgabe 1: Kennst du das?

Probier's nochmal!

K _____



Nicole: „Das Wort _____ kenne ich.“
oder
„Das Wort kenne ich nicht.“

Figure 11: Task 1 — Do you know that? Picture naming and translation task to familiarize the participants with the text material and elicit baseline productions of the target items. Here: second attempt to name the *cheese* picture. The first letter of the target word *Käse* is provided as a hint. The blue box contains the carrier sentence. *Probier's nochmal* (try again)!

In this first task, Mirabella accepts allophonic variation in order to avoid that the participants change their pronunciation simply because they were not understood. But she only accepts the expected target words, i.e. no synonyms, in order to be perceived as a non-human interlocutor who does not have the full range of human linguistic flexibility.

The individual realizations of ⟨-ä-⟩ and ⟨-ig⟩ are auditorily categorized as [ɛ:] or [e:] and [ɪç] or [ɪk], respectively, by the experimenter. The categorization has to be performed in real-time and on the basis of the phonetician's auditory impression in order to ensure a smooth and seamless interaction with Mirabella for the participant. The validity of these online annotations is evaluated in [Section 4.3.4](#) for [ɛ:]/[e:] and in [Section 4.3.5](#) for [ɪç]/[ɪk]. Note that we consider fricative variants such as [ʃ] or [ç] as part of the [ɪç] category.

The occurrence of the allophones under examination varies regionally throughout the German-speaking region of Europe. The codified Standard German variants of each pair are [ɛ:] (predominant in the South) and [ɪç] (predominant in the North; Dudenredaktion, 2015; Kleiner, 2011).³ However, Kiewewalter (2019) has shown that the respective non-standard forms are perceived as subjectively corresponding to the standard (for [e:]; predominant in the North and Eastern Austria) or only slightly dialectal (for [ɪk]; predominant in the South) by native listeners of German. Therefore, we do not expect dialectality to influence accommodating behavior for these features.

While it is possible for a speaker to use both forms interchangeably, we expected the participants of the present study to have a preference for one of the two forms. The preference was determined for each participant as the majority variant produced for the 12 items per allophonic

³ Often, the opposite is thought to be the case by speakers, since the written form of the word ending ⟨-ig⟩ hints towards [ɪk] being the standard and there is a tendency of long, stressed ⟨-ä-⟩ merging to [e:] across the German-speaking regions.

Aufgabe 2: Stelle mir Fragen!

Du hast sicher schon gemerkt, dass ich dich verstehe...

Ich kann dir aber auch antworten! 🗨️

Probiere es mal aus! Formuliere in beliebiger Reihenfolge fünf Fragen mit den angegebenen Wörtern. Ich gebe dir jeweils die Antwort.

wann – Italien – den Euro – eingeführt haben

was – die Hauptstadt – von Lettland – sein

wo – die Brüder Grimm – geboren sein

wer – die erste Frau – im Weltall – sein

wie viele Tage – der August – haben

Figure 12: Task 2 – Ask me questions! Participants formulate five wh-questions in random order from the given fragments. They are answered by Mirabella. This familiarizes them with Mirabella’s voice and elicits baseline question intonation patterns. Green questions have already been asked.

variation.⁴ It was stored in the system and retrieved in task 4 to test for accommodation to the respective non-preferred variant.

4.2.1.2 Task 2 — question intonation, baseline

The participants formulate five wh-questions in random order whose components are given as fragments, e.g., wer – die erste Frau – im Weltall – sein (*who – the first woman – in space – be*; see Figure 12). Mirabella talks for the first time when she answers these questions.

This task familiarizes the participants with Mirabella’s voice and reveals the intonation they usually apply when producing constituent questions. See Appendix E for the expected questions and the corresponding answers given by Mirabella.

Slight variations to the expected questions are accepted by the experimenter to show a certain flexibility. In the case of bigger deviations or disfluencies, Mirabella encourages the participants to try again (see Appendix C, 18, 21, 23, 26). This behavior in combination with utterances such as “Lass mich überlegen...” (*Let me think...*) or “Sehr gute Frage!” (*Great question!*) interspersed in the dialog, aims to reinforce the impression of talking to a non-human yet social interlocutor.

The general unmarked expectation is for German wh-questions to be produced with falling intonation. Rising intonation is mainly applied in the case of echo questions, i.e., when the answer was not understood and the question is uttered again (cf. Möbius, 1993; Grice and Baumann, 2002; Wochner et al., 2015).

⁴ In the event of a tie, the Standard German variant was set as the speaker preference. This was the case only once among the participants of the present study.



Figure 13: Task 3 — Where did the animals hide? Q&A game testing accommodation to the intonation of constituent questions. Here: second round of the game; both players have asked each other seven questions so far; the animals that Mirabella has already asked for are marked by green frames, those that the participant has asked for by blue frames; it is Mirabella’s turn.

Also French *wh*-questions are usually realized with a final F_0 fall, but rising contours are possible as well. As in German, such rising contours mainly occur in echo questions (Di Cristo, 1998; Delais-Roussarie et al., 2015).

We therefore expected to find mainly falling intonation contours for the questions asked in this task from both the L1 German and L1 French speaker groups.

4.2.1.3 Task 3 — question intonation, test

In this task Mirabella and the participants take turns asking (Q) and answering (A) each other about ten animals hiding in ten houses (see Figure 13), in the following form:

Q: Wo hat sich **<the animal>** versteckt?

Where did <the animal> hide?

A: **<the animal>** hat sich in Haus Nummer **<number>** versteckt.

<the animal> hid in house number <number>.

The order in which Mirabella and the user ask for the animals on the screen is free. The task includes two rounds of 20 turns, with Mirabella and the participants each asking and answering 10 questions per round.⁵ The realization of questions on the part of the system differs between round one and round two with respect to pitch accent placement and intonation, giving room for accommodation. In round one (R1), Mirabella produces all questions with a nuclear pitch accent on the *<animal>* followed by a final F_0 fall, whereas in round two (R2), all questions are

⁵ We did not include explicit filler material in this task, e.g., questions with different intonation contours, since we assume that accommodation requires a certain amount of repetition. The answers uttered between the questions serve as filler material for the questions themselves, in that they have a different intonation contour, thus providing a certain amount of variety and distraction.

produced with a nuclear pitch accent on the interrogative pronoun *wo* (*where*) followed by a final high F_0 rise – here illustrated using the example “*Where did the lion hide?*”:

R1: Wo hat sich der **L**öwe versteckt? ↘

R2: **W**o hat sich der Löwe versteckt? ↗

The latter version constitutes the typical shape of an echo question asking for information that was already given, but not understood. Such echo questions are unlikely to occur naturally in the context of the Q&A exchange at hand, since the answers do not necessarily have to be understood by the participants: the correct pictures are always visually marked on the screen as well.

In the second round of the game, all animals stay paired with the same house numbers as before, however the arrangement of the houses on the screen differs from that of the first round. Therefore, it is unexpected, yet not pragmatically wrong, to ask for the location of the animals in the form of an echo question.

For both the native and non-native speaker groups, we expected to find falling intonation contours for the first round of the Q&A and a substantial increase of rising contours from the first to the second round of the Q&A. Additionally, we expected the nuclear pitch accent to be shifted from the $\langle animal \rangle$ in the first round to the interrogative pronoun *wo* in the second round.

4.2.1.4 Task 4 — allophonic variation, test

In this map task the participants describe the path from leaving a house until reaching a destination on the map while walking past different objects (see Figure 14). To that end, they are using the prepositions given on the right side of the screen (see Appendix F for details). Additionally, the participants describe the object in question with the adjective given next to it at every step. This results in two-part utterances of the following type:

- Ich gehe um die **S**äge herum. Die **S**äge ist schwer.
I walk around the saw. The saw is heavy.
▶ bold target contains the [ɛ:] vs. [e:] contrast
- Ich gehe an dem Pferd vorbei. Das Pferd ist **mutig**.
I walk past the horse. The horse is brave.
▶ bold target contains the [ɪç] vs. [ɪk] contrast

Some of the objects (**O**) and adjectives (**A**) are hidden behind boxes. The participants ask Mirabella about these items: “Mirabella, was ist hinter der $\langle color \rangle$ Box?” (*Mirabella, what is behind the $\langle color \rangle$ box?*)

The information about the participants’ preference with respect to the [ɪç] vs. [ɪk] and [ɛ:] vs. [e:] contrasts is automatically retrieved from the results of task 1 before the map task. Mirabella then uses the non-preferred variants when providing the requested information:

- O:** Hinter der $\langle color \rangle$ Box ist $\langle the\ object \rangle$.
Behind the $\langle color \rangle$ box is $\langle the\ object \rangle$.
- A:** Das Wort hinter der $\langle color \rangle$ Box ist $\langle adjective \rangle$.
The word behind the $\langle color \rangle$ box is $\langle adjective \rangle$.

Given this information, the participants can formulate the required two-part utterance. Subsequently, the hidden item is revealed.

Aufgabe 4: Wie kommst du zum Ziel?

um...herum

aus...heraus

in...hinein

an...vorbei

durch...hindurch

Nicole: „Ich gehe... Der/die/das ist...“

Figure 14: Task 4 – How do you reach the destination? Map task testing accommodation to allophonic variation. Here: The participant has made her way through the map up until the position marked by the yellow frame. She will ask Mirabella for the item behind the yellow box, use the preposition *um...herum* to say that she goes *around* the item, and use the given adjective *müde* to further describe the item as *tired*.

The task consists of four maps with nine object-adjective pairs each and contains a total of 12 occurrences per allophonic contrast (see [Appendix G](#) for an overview of all maps). Each map contains:

- three pairs including an [ɪç] vs. [ɪk] target
e.g., **Honig** (*honey*) – süß (*sweet*); Baum (*tree*) – schattig (*shady*)
- three pairs including an [ɛ:] vs. [e:] target
e.g., Mädchen (*girl*) – schlau (*smart*); Bus (*bus*) – verspätet (*delayed*)
- three filler pairs not including a target⁶
e.g., Haus (*house*) – leer (*empty*); Autos (*cars*) – laut (*loud*)

If the target item is an object, it occurs twice in the two-part utterance (e.g., *Honig*, *Mädchen*; see *Säge* in the example above); if the target item is an adjective, it occurs only once, in the second part of the utterance (e.g., *schattig*, *verspätet*; see *mutig* in the example above).

For both the native and non-native speaker groups, we expected to find a substantial increase of the non-preferred variant for the [ɪç] vs. [ɪk] contrast and a substantial shift in the F1–F2 space in the direction of the non-preferred variant for the [ɛ:] vs. [e:] contrast during the map task as compared to the baseline task.

4.2.2 Text material

The text material used in the experiment pertains to two different categories. The first category contains structural utterances, which are either used to explain the tasks or to guide the

⁶ The [ɪç]/[ɪk] items additionally serve as fillers for the [ɛ:]/[e:] items and vice versa.

conversation. While the explaining utterances are presented at the beginning of a new task and follow a chronological order that is the same for all participants (see [Appendix B](#)), the guiding utterances are available to the experimenter at any time during the experiment and may be used to react to the participants' behavior if needed (see [Appendix C](#)).

The second category contains utterances which are part of the actual tasks testing for phonetic accommodation, either as target or filler material. More details about these utterances were given above, together with the explanations of the individual tasks in [Section 4.2.1](#).

Since the experiment is designed as an application for learning the German language, the text material used in the experiment was chosen to be accessible to advanced learners of German. This constrains the selection of possible target items substantially.

4.2.3 *Stimuli*

The first set of Mirabella's utterances was pre-recorded by a female native speaker of German (aged 26 years). The recordings were carried out with a sampling rate of 48 kHz using a stationary cardioid microphone in a sound-attenuated booth. The speaker was instructed to speak in a friendly tone, basing her performance on experience with the usual tone of commercial language assistance systems. She produced the target stimuli in their different forms. The best versions in terms of target feature clarity were selected for use in the experiment.

The second set of utterances consists of synthesized speech.⁷ As the idea of the present study is to extend the analysis presented in [Chapter 3](#) we rely on the same paradigm with an updated process. This updated process uses three main toolkits: MaryTTS (Le Maguer et al., 2018) as the front-end, HMM-based Speech Synthesis System (HTS; Zen and Toda, 2005) to achieve the modeling, and the vocoder WORLD (Morise et al., 2016) to render the signal from the acoustic parameters generated by HTS.

The HTS models were trained using the BITS corpus (Ellbogen et al., 2004). We used the samples recorded by speaker *spk1*, which in total correspond to about 3 h of speech sampled at 48 kHz. The provided alignment was discarded, as our voice building pipeline (Steiner and Le Maguer, 2018) already includes an automatic alignment step.

In order to achieve German based synthesis, we defined a feature set derived from the one proposed for English (Tokuda et al., 2002). The major modification is the adaptation of the phonetic part for German. This adaptation corresponds to the extension of the phonetic alphabet and the addition of corresponding questions in the question file.

Within the synthesis pipeline we imposed three main parameters. On the one hand, we modified the front-end decision by inducing the allophonic contrasts [ɪç]/[ik] and [ɛ:]/[e:]. This enabled HTS to produce the different variants in the map task stimuli. On the other hand, we extracted the segment durations and F_0 contours from the natural stimuli and applied these values in the synthesis process – the durations at the phone level and the fundamental frequency at the frame level. By imposing these parameters, it was possible to generate the variations of prosodic structure in the synthetic utterances of the Q&A game.

Imposing the duration at the phone level is straightforward as this option is directly implemented in HTS. To impose F_0 , we had the choice between two main solutions: using the voicing prediction from the system or creating a new voicing prediction using the generated spectral information in combination with a simple neural network. After informal subjective evaluation, we concluded that using the voicing information predicted by HTS leads to a more consistent quality and is less likely to introduce artifacts. Applying this voicing mask when imposing

⁷ I would like to thank Sébastien Le Maguer for generating the synthetic stimuli.

the fundamental frequency avoided mismatches between F_0 and the harmonic structure of the spectrum.

Both versions of Mirabella thus use the natural source signal, but they differ with respect to the filter applied to the latter: the human vocal tract for the natural stimuli and HTS for the synthetic stimuli.

HTS produces speech with a degraded voice quality, which is often described as buzzy or muffled (Zen et al., 2009). We can therefore assume that the synthetic version of Mirabella is clearly perceived as non-human by the participants, whereas in the case of Mirabella’s natural version, the impression of talking to a computer is mainly caused by the interaction itself. The process of imposing the natural segment durations and F_0 contours during synthesis, however, ensured that the synthetic version of Mirabella was still as similar as possible to the natural version in its perceived personality, insofar as the latter is conveyed through prosody (e.g., Smith et al., 1975; Apple et al., 1979; Nass and Lee, 2001; Trouvain et al., 2006). This is relevant since the perceived personality of the interlocutor can influence the accommodating behavior towards them (e.g., Yu et al., 2013; Lewandowski and Jilka, 2019).

4.2.4 Participants

The participants were recruited from Saarland University and other educational institutions in Saarbrücken. They were paid for taking part in the experiment.

L1 GERMAN SPEAKERS This group consisted of 42 native speakers of German. Four of them spoke more than one native language: English ($n = 2$), Polish ($n = 1$), and Greek ($n = 1$). All had learned at least one foreign language, the majority two or more. The most frequent foreign languages were English ($n = 42$), French ($n = 31$), and Spanish ($n = 16$). Thirty-nine participants were students and three had non-academic jobs. The participants came from eleven German states with 61% from central regions⁸, 22% from southern regions⁹, and 17% from northern regions¹⁰.

Each participant was presented with only one of the two stimulus types — natural or HMM. This resulted in two experimental groups: the *L1 natural group* with 20 participants (16 female, 4 male; mean age 25.8 years; age range 18 to 55 years) and the *L1 synthetic group* with 22 participants (15 female, 7 male; mean age 23.7 years; 18 to 32 years).

L2 GERMAN SPEAKERS This group consisted of 11 native speakers of French (5 female, 6 male; mean age 25.2 years; age range 16 to 53 years). All participants indicated French as their sole or dominant native language. Two participants indicated a second native language: one Portuguese and the other a Bamileke language. The participants spoke 2 to 4 foreign languages. Besides German, the most common foreign languages were English ($n = 11$) and Spanish ($n = 5$). Their self-assessed command of German ranged from *B2: upper intermediate* ($n = 2$) to *C1: advanced* ($n = 9$) according to the Common European Framework of Reference for Languages (CEFR). The participants were students ($n = 8$) or employees ($n = 3$) of educational institutions in Saarbrücken and came from different regions of France ($n = 10$)¹¹ and Cameroon ($n = 1$).

⁸ Saarland ($n = 16$), Rheinland-Pfalz ($n = 4$), Hessen ($n = 3$), and Berlin ($n = 2$).

⁹ Baden-Württemberg ($n = 7$) and Bayern ($n = 2$).

¹⁰ Niedersachsen ($n = 2$), Nordrhein-Westfalen ($n = 2$), Hamburg ($n = 1$), Bremen ($n = 1$), and Sachsen-Anhalt ($n = 1$).

¹¹ Île-de-France ($n = 4$), Grand-Est ($n = 2$), Normandie ($n = 2$), Auvergne-Rhône-Alpes ($n = 1$), and Provence-Alpes-Côte d’Azur ($n = 1$).

Table 10: Percentage of agreement with statements concerning the general communicative behavior in the three experimental groups.

Statement	L1 natural %	L1 synthetic %	L2 natural %
	<i>n</i> = 20	<i>n</i> = 22	<i>n</i> = 11
1 Depending on who I talk to, my way of speaking changes.	90	100	73
2 If someone speaks a dialect from my region, I adapt to it.	80	59	55
3 If someone speaks a dialect from another region, I adapt to it.	25	27	9
4 I deliberately imitate the pronunciation of interlocutors.	15	18	27
5 My way of speaking almost never changes.	–	–	45

All participants were presented with the natural stimuli, therefore henceforth: *L2 natural group*.

COMMUNICATIVE BEHAVIOR (SELF-ASSESSED) Table 10 shows the results of a questionnaire completed after the experiment, which asked the participants to assess their general communicative behavior. The results are reported separately for the L1 natural, L1 synthetic, and L2 natural groups.

Most speakers answered affirmatively to the question whether they change their way of speaking depending on their respective interlocutor — however, considerably less in the L2 group than in the two L1 groups (see Statement 1). A substantial part of the participants also believed they would converge to an interlocutor of the same dialectal background — here the L1 synthetic and the L2 natural groups exhibit lower numbers than the L1 natural group (see Statement 2). About a quarter of the participants in both L1 groups claimed they would do the same with an interlocutor of a different dialectal background — the vast majority of the L2 group believed that this statement does not apply to them (see Statement 3). However, about a quarter of L2 speakers believed that they intentionally imitate the pronunciation of interlocutors — this opinion was less prevalent in the two L1 groups (see Statement 4). Eventually, almost half of the L2 speakers also agreed with the statement that they almost never change the way they speak — in the two L1 groups no one expressed this opinion (see Statement 5).

These numbers, although they may not agree with the actual behavior of the participants, show that there is a certain awareness of the phenomenon of accommodation to an interlocutor in spoken communication. The readiness to accommodate seems to be higher when the accommodation target is more familiar (e.g., own vs. different dialect). A small number of participants perceives convergence to an interlocutor even as an intentional, active process.

Overall, the opinions in the two L1 groups are similar, while the L2 group shows a different pattern. Particularly striking is the lower readiness of the L2 group to adapt to speakers of other dialects and the agreement with the statement that their way of speaking almost never changes. Given our data, we cannot determine whether these patterns indeed differ systematically, possibly for cultural reasons, between native speakers of German and French. If such systematic differences exist, they could influence actual accommodation behavior: an attitudinal pattern like that of the L1 German speakers seems more conducive to accommodation.

4.3 ANALYSIS AND RESULTS

4.3.1 *Rating of Mirabella*

After the experiment, the participants rated Mirabella on 5-point scales with regard to her likability (*unpleasant to very likable*), competence (*incompetent to very competent*), intelligibility (*bad to very good*), and response time (*too slow to too fast*). Since we can assume that the participants interpreted the unlabeled steps between the endpoints as equidistant intervals, we can consider this an approximation of an interval scale and calculate the mean as a measure of the central tendency.

Among the L1 German speakers, the ratings of the two versions of Mirabella differed most for intelligibility, with the synthetic version (mean = 3.9, $SD = 0.8$) being less intelligible than the natural version (mean = 5, $SD = 0.2$). In addition, the synthetic version of Mirabella was judged to be less likable (synthetic: mean = 3.8, $SD = 1$; natural: mean = 4.5, $SD = 0.6$), but only slightly less competent (synthetic: mean = 4, $SD = 0.9$; natural: mean = 4.3, $SD = 0.4$).

Mirabella's response time, i.e., the response time of the experimenter, was considered equally appropriate in both cases (synthetic: mean = 2.9, $SD = 0.9$; natural: mean = 2.9, $SD = 0.6$).

The L1 French speakers rated the natural version of Mirabella they had heard very similarly to the L1 German speakers, with a perfect score for intelligibility (mean = 5, $SD = 0$) and a high score for likability (mean = 4.5, $SD = 0.5$). Mirabella's competence was rated somewhat higher by the L1 French speakers (mean = 4.7, $SD = 0.5$) compared to the L1 German speakers.

The L1 French speakers also considered Mirabella's reaction time to be appropriate (mean = 3.2, $SD = 0.8$).

4.3.2 *Modeling*

The dependent variables are analyzed using linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs) formulated with the lme4 package (1.1-21; Bates et al., 2015) and evaluated with the lmerTest package (3.1-0; Kuznetsova et al., 2017) in RStudio (1.1.463; RStudio Team, 2016) with R (3.5.2; RCore Team, 2018).

To strike a compromise between accuracy and complexity, model selection is carried out bottom-up, starting with a model which only includes the random factor intercepts for SUBJECT and ITEM. Then, theoretically relevant fixed factors (sum coded) and interactions as given by the design of the experiment are added to the model. Random slopes for SUBJECT and/or ITEM are added for every effect where there is more than one observation for each unique combination of SUBJECT/ITEM and treatment level. Random slopes are only removed to simplify the model in cases of convergence errors or to allow a non-singular fit. The influence on the model fit is assessed by means of the Akaike information criterion (AIC), which estimates the relative quality of a statistical model for a given data set by taking into account the likelihood function and the number of estimated parameters (Akaike, 1973). A factor is kept in the model if the model fit improves significantly and the AIC value decreases by at least two points as compared to the model without the factor in question. Factors kept in the model are being considered significant predictors of the respective dependent variable at $\alpha = 0.05$.

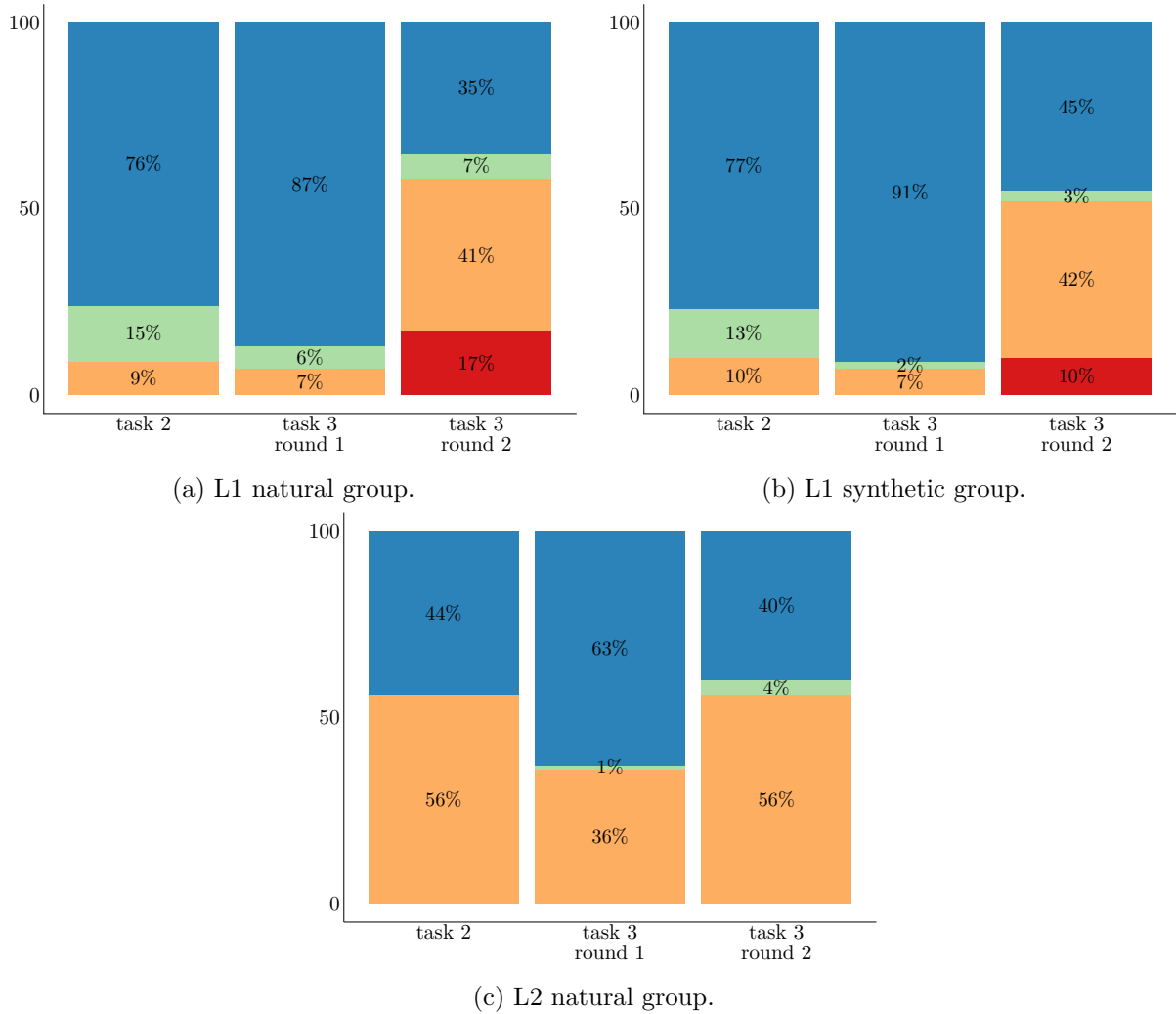


Figure 15: Percentages of questions realized with **falling**, **falling-rising**, **rising(a)**, or **rising(w)** intonation contour during the baseline production (task 2) and the two rounds of the question-and-answer game (task 3).

4.3.3 Question intonation

The intonation contours of the 1378 questions uttered in tasks 2 and 3 (L1 natural: $n = 526$, L1 synthetic: $n = 568$, L2 natural: $n = 284$)¹² were perceptually classified by two trained phoneticians, taking the position of the nuclear pitch accent into account.¹³ Three contour types were found in the data: *falling*, *falling-rising*, and *rising* (cf. Grice and Baumann, 2002). The latter occurs in two variants: first, as *rising(a)* contours with a nuclear pitch accent on the respective *animal* in task 3 or an equivalent word in focus in task 2, and second, as *rising(w)* contours with a nuclear pitch accent on the interrogative pronoun *wo*. Figure 15 shows the results of the evaluation for the three experimental groups.

L1 GERMAN SPEAKERS The results of the native speakers of German are given in Figures 15a (natural group) and 15b (synthetic group). In task 2, where the participants formulate wh-questions from given fragments, the *falling* contours are predominant in both groups (natural:

¹² Theoretically expected number of data points: (5 base questions + 2 × 10 animal questions) × number of participants. Small deviations due to repetitions.

¹³ I would like to thank Bistra Andreeva for help with the intonation analysis.

76 %, synthetic: 77 %), but *falling-rising* (natural: 15 %, synthetic: 13 %) and *rising(a)* (natural: 9 %, synthetic: 10 %) contours are produced as well.

In the first round of task 3, where Mirabella produces exclusively *falling* contours, the predominance of *falling* contours on the part of the participants becomes more pronounced in both groups (natural: 87 %, synthetic: 91 %), yet *falling-rising* (natural: 6 %, synthetic: 2 %) and *rising(a)* (natural and synthetic: 7 %) contours still occur.

In the second round of task 3, where Mirabella produces exclusively *rising(w)* contours, the amount of *rising(a)* contours increases in both groups (natural: 41 %, synthetic: 42 %) and *rising(w)* contours emerge in both groups as well (natural: 17 %, synthetic: 10 %). While the amount of *falling-rising* contours stays about the same in both groups (natural: 7 %, synthetic: 3 %), the number of *falling* contours is considerably smaller in the second round of task 3 (natural: 35 %, synthetic: 45 %).

The increase of rising contours (this includes *falling-rising*, *rising(a)*, and *rising(w)* contours) from round 1 to round 2 of task 3 per experimental group was evaluated by fitting GLMMs with the binary response *falling/rising* as dependent variable and testing the factors TASK (round1/round2) and SPEAKER SEX (female/male) following the method described in Section 4.3.2. Note that these are binomial models and the coefficients are hence in logit-space. If a logit-coefficient is positive, the effect of the corresponding predictor on the response variable is positive as well, and vice versa.

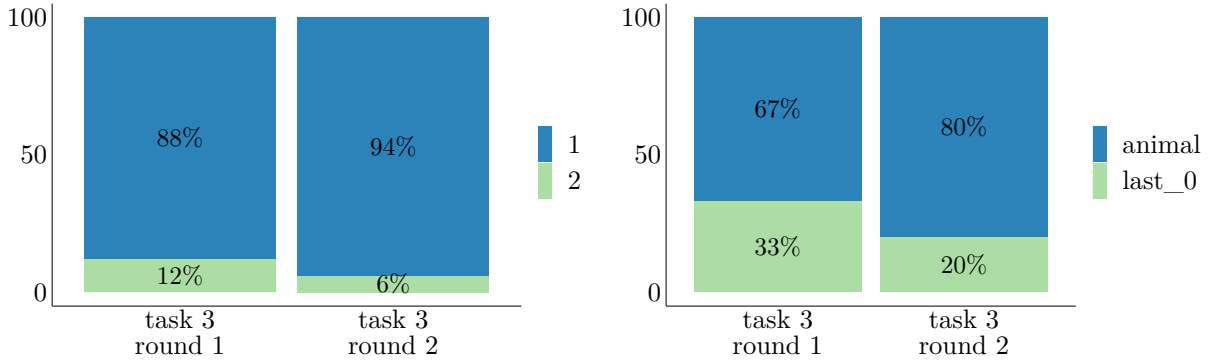
The model of the natural data set did not converge when random intercepts for ITEM, i.e., the different animals, were included, therefore the models for both experimental groups were fitted including random intercepts only for USER. The factor TASK is a significant predictor of the dependent variable in the natural group (Estimate (log-odds) = -4.87 , SE = 1.24, $z = -3.94$, $p < 0.001$) and the synthetic group (Estimate (log-odds) = -2.73 , SE = 0.8, $z = -3.44$, $p < 0.001$) indicating an increase of rising contours in round 2 of task 3. The models include random slopes for TASK by USER to account for the individual reactions of the participants. The factor SPEAKER SEX did not improve the fit of the models and was therefore not included.

L2 GERMAN SPEAKERS Figure 15c shows the results of the non-native speakers of German. While *falling* contours were predominant in the baseline productions of the native speakers, the non-native speakers produced 56 % *rising* and only 44 % *falling* contours in the same task. Like the L1 German groups, the French speakers produced more *falling* contours (63 %) when interacting with Mirabella in the first round of the Q&A. However, they still produced a substantial amount of *rising* (36 %) and some *falling-rising* (1 %) contours, as well. In the second round of the Q&A, where Mirabella produced *rising(w)* contours with a nuclear pitch accent on the interrogative word, the amount of *rising* (56 %) and *falling-rising* (4 %) contours in the French group increased again. Unlike the German speakers, however, the French speakers did not produce any *rising(w)* contours.

As for the native speakers above, the increase of rising contours (this includes *falling-rising* and *rising(a)* contours) from the first to the second round of the Q&A was evaluated by fitting a GLMM to the data of the non-native speakers.

The model includes random intercepts for USER and ITEM, as well as by-user random slopes for TASK. Although including the factor TASK (round1/round2) improved the fit of the model, it was not a significant predictor of the contour type (Estimate (log-odds) = -2.16 , SE = 2.57, $z = -0.84$, $p = 0.4$).¹⁴ Again, the factor SPEAKER SEX did not improve the fit of the model and was therefore not included.

¹⁴ For consistency with the L1 German models, we fitted a model without random intercepts for ITEM, as well. This affected the model estimates only slightly: Estimate (log-odds) = -2.27 , SE = 2.16, $z = -1.05$, $p = 0.3$.



(a) Percentages of questions realized with a single intonational phrase (1) or two separate intonational phrases (2). (b) Percentages of questions realized with a nuclear pitch accent on the **animal** or the ultima of the phrase-final word (**last_0**).

Figure 16: Special intonational patterns in the L2 natural group.

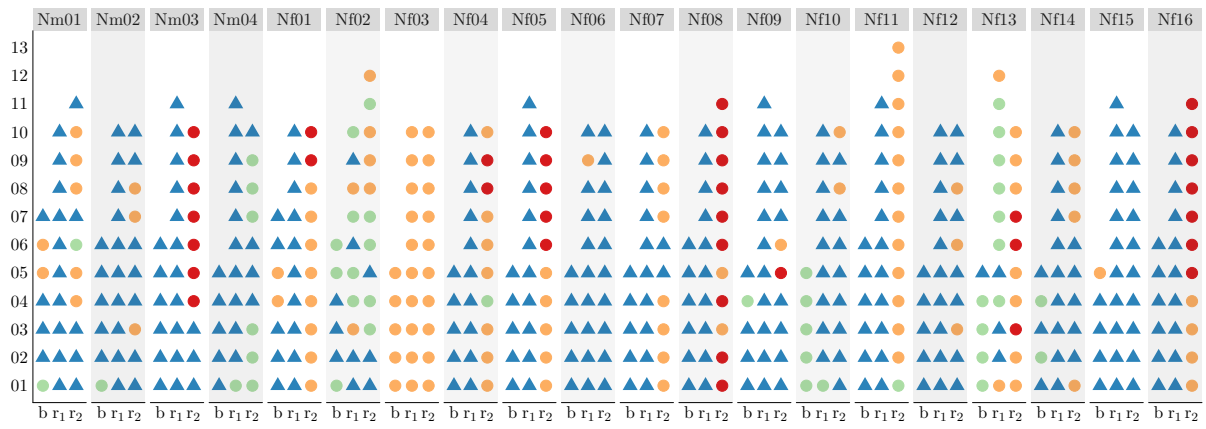
Two further points are noteworthy in the non-native speaker data. First, the questions were not always produced as one single intonational phrase, but some participants had a tendency to produce the final part of the question separately: “Wo hat sich das Pferd | versteckt?” (Where did the horse | hide?) This occurred in 12% of all questions in the first Q&A round, but only 6% in the second Q&A round (see Figure 16a). Second, the nuclear pitch accent was not always realized on the respective animal, but sometimes on the ultima of *versteckt* — which coincides with the lexical stress in German. While this can partly be an effect of the unusual phrasing mentioned above, it occurred more frequently, namely in 33% of all questions in the first Q&A round and 20% in the second Q&A round (see Figure 16b). Whereas the decrease in cases of unusual phrasing was not significant in a GLMM with random intercepts for USER and ITEM (Estimate (log-odds) = 0.46, SE = 0.28, $z = 1.66$, $p = 0.1$), the increase of nuclear pitch accents on the respective animal in the second round of the Q&A was significant in an equivalent model (Estimate (log-odds) = 0.51, SE = 0.19, $z = 2.68$, $p < 0.01$).

INDIVIDUAL BEHAVIOR Figure 17 shows the individual question realizations in chronological order by each speaker of the three experimental groups. Note that some speakers never deviate from their preferred question intonation, e.g., speakers *Sm03* and *Ff04* always produce the expected *falling* pattern, while speakers *Nf03* and *Fm04* only utter *rising(a)* questions. In contrast, *Nm03*, *Nf05*, *Sm02*, and *Sf09*, are examples of speakers who have a clear preference to produce the *falling* pattern, but ultimately converge to the *rising(w)* pattern produced by Mirabella, either directly or via instances of *rising(a)*.

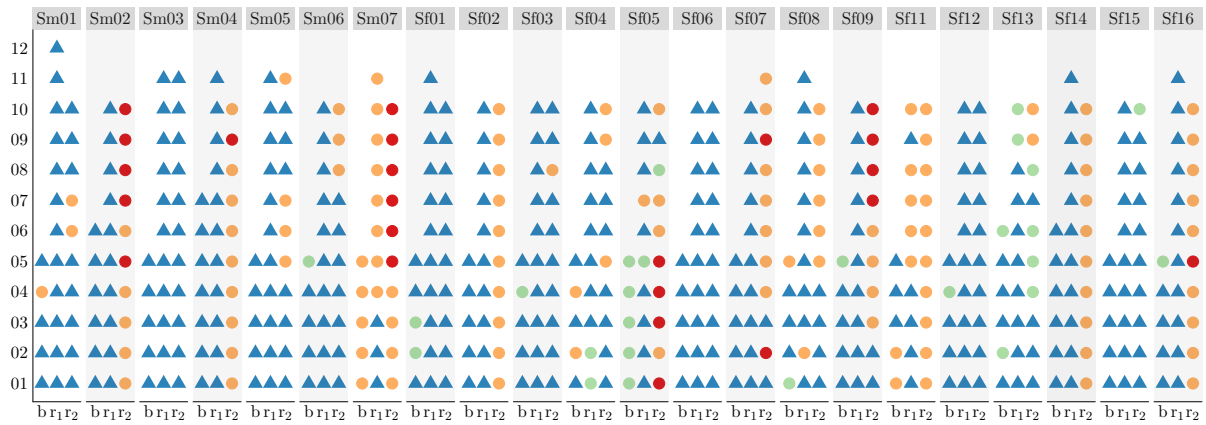
To evaluate the accommodating behavior on the individual level we classified all participants according to the following thresholds, comparing the number of *rising(a)* or *rising(w)* occurrences in round 2 to round 1:

- increase of ≥ 5 \rightarrow substantial convergence
- increase of ≥ 2 \rightarrow moderate convergence
- in-/decrease of 1 \rightarrow maintenance
- decrease of ≥ 2 \rightarrow moderate divergence
- decrease of ≥ 5 \rightarrow substantial divergence

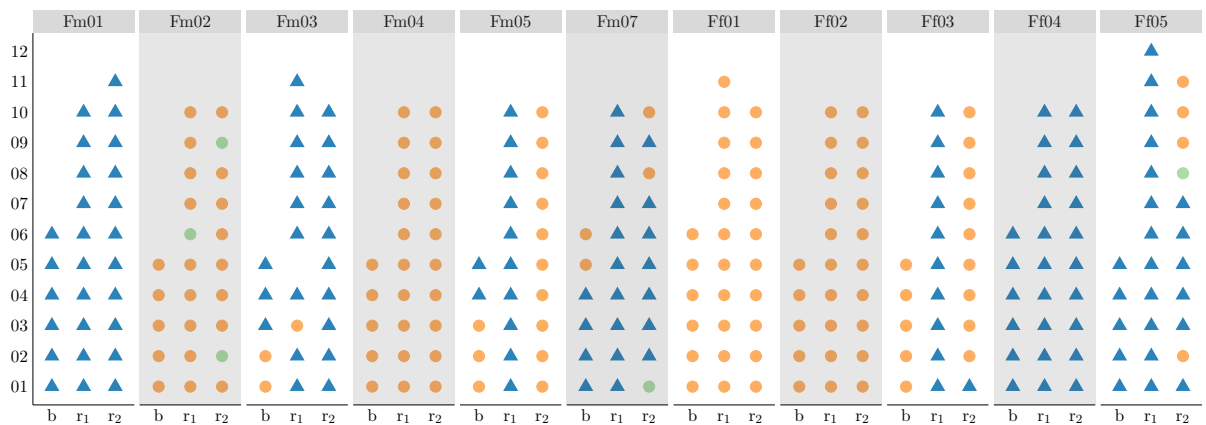
According to these criteria, 23 participants show substantial convergence (L1 natural: 11, L1 synthetic: 10, L2 natural: 2), moderate convergence is found in 13 participants (L1 natural: 5, L1 synthetic: 6, L2 natural: 2), and 17 participants do not change their question intonation (L1 natural: 4, L1 synthetic: 6, L2 natural: 7). Divergence on the individual level was not found.



(a) L1 natural group.



(b) L1 synthetic group.



(c) L2 natural group.

Figure 17: Individual question realizations with **falling** \blacktriangle , **falling-rising** \bullet , **rising(a)** \circ , or **rising(w)** \bullet intonation contour in their order of occurrence during the baseline production (b), as well as during round 1 (r_1) and round 2 (r_2) of the question-and-answer game.

4.3.4 Long vowel <-ä->

As auditorily determined by the experimenter during the baseline task, 25 of the 42 L1 German speakers participating in the present experiment had a preference for [ɛ:] (20 female, 5 male) and 17 speakers had a preference for [e:] (11 female, 6 male). In the L2 German group, eight speakers preferred [ɛ:] (4 female, 4 male) and three speakers preferred [e:] (1 female, 2 male).

In order to validate the online annotations, all baseline [ɛ:]/[e:] targets were annotated again by the original annotator, i.e., the experimenter, and an additional phonetically trained annotator without time pressure and with the option to visualize the spectrogram. The inter-rater agreement between these two offline annotations (Cohen’s kappa = 0.91) and the intra-rater agreement between the online and offline annotations of the experimenter (Cohen’s kappa = 0.88) were both found to be almost perfect. Among the cases in which inter- or intra-rater agreement was not given, a maximum of four — usually only one or two — clustered on a single participant. This means that the ambiguous cases never affected the choice of Mirabella’s variant. Although the auditory classification of vowel quality in a binary way poses a certain challenge in the experimental procedure, because ambiguous forms can be difficult to assign to a category, we conclude from this validation that the participants’ preference with respect to [ɛ:]/[e:] was determined correctly.

For all 1705 realizations of long, stressed <-ä-> uttered by the participants in tasks 1 (L1 natural: $n = 247$, L1 synthetic: $n = 264$, L2 natural: $n = 129$) and task 4 (L1 natural: $n = 391$, L1 synthetic: $n = 431$, L2 natural: $n = 219$) as well as by Mirabella ($n = 12$ per natural and synthetic version), the first and second formants were measured at the temporal midpoint of the vowel using Praat’s Burg algorithm (Boersma and Weenink, 2019).

In a second step, the Euclidean distance ($dist$) in the F1–F2 space between each participant realization (U) and the respective realization by Mirabella (M) was calculated for the baseline task (Equation 1) and the map task (Equation 2), e.g., for <-ä-> in Käse (*cheese*):

- participant’s base production vs. Mirabella’s production
- participant’s map production vs. Mirabella’s production

Finally, the difference in Euclidean distance ($dDist$) between the baseline task and the map task was calculated (Equation 3), resulting in data sets of 403 values for the L1 natural group, 431 values for the L1 synthetic group, and 214 values for the L2 natural group.¹⁵

$$dist(b) = \sqrt{(U_{baseF1} - M_{F1})^2 + (U_{baseF2} - M_{F2})^2} \quad (1)$$

$$dist(m) = \sqrt{(U_{mapF1} - M_{F1})^2 + (U_{mapF2} - M_{F2})^2} \quad (2)$$

$$dDist = dist(b) - dist(m) \quad (3)$$

Difference in Euclidean distance has the following potential outcomes:

- $dDist > 0$, if the participants shift their productions in the direction of Mirabella (convergence);
- $dDist = 0$, if the participants do not shift their productions in the F1–F2 space (maintenance);
- $dDist < 0$, if the participants shift their productions away from Mirabella (divergence).

The difference in Euclidean distance measure contains the information about the experimental task, since it is calculated as a comparison of the baseline and map task. It is therefore the model

¹⁵ Theoretically expected number of data points: 20 map items [i.e., 2×8 nouns + 4 adj.] compared with their base counterpart \times number of participants. Small deviations due to missing values and repetitions.

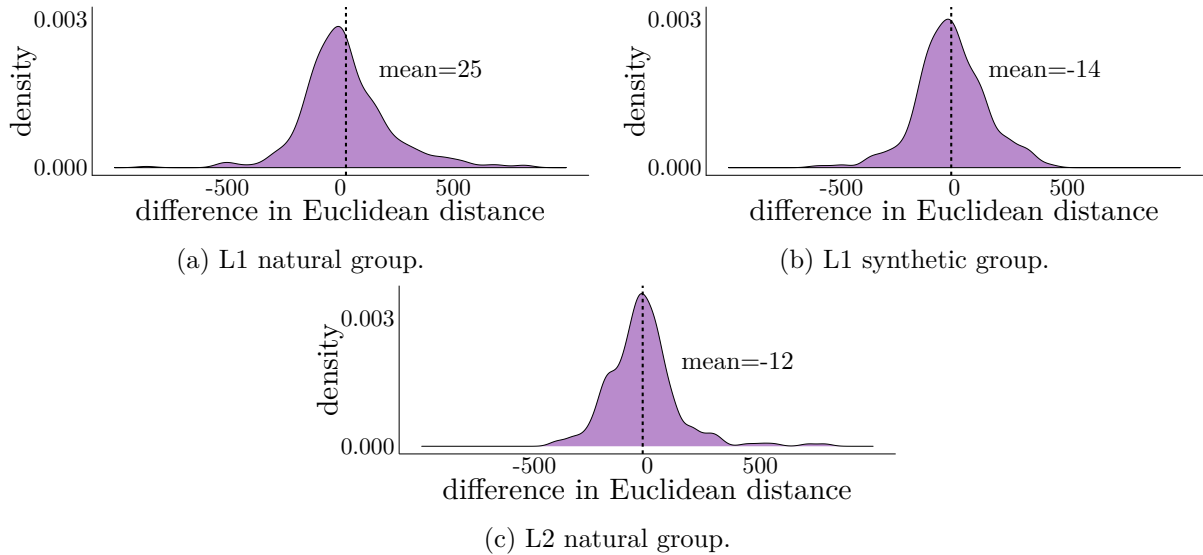


Figure 18: Difference in Euclidean distance in the F1–F2 space (in Hz) between participant realizations of ⟨-ä-⟩ and the respective realizations by Mirabella in the baseline compared to the map task. Positive values indicate convergence, negative values divergence. The distribution means are shown by the dashed lines. They do not differ significantly from zero for either of the groups.

intercept that provides insight about accommodating behavior. The intercept is considered to significantly differ from zero at $\alpha = 0.05$.

Figure 18 shows the distributions of $dDist$ for the three experimental groups. The distribution of the L1 natural group has a mean of 25 which is positive and therefore suggests convergence; the distribution mean of the L1 synthetic group (−14) and the L2 natural group (−12) are both negative and therefore suggest divergence.

However, fitting LMMs with $dDist$ as dependent variable and testing the factors SPEAKER SEX (female/male) and PREFERENCE ([ɛ:]/[e:]) following the method described in Section 4.3.2, revealed that the means do not differ significantly from zero for the L1 natural group (Estimate = 26.32, SE = 24.54, df = 20.36, $t = 1.07$, $p = 0.3$), the L1 synthetic group (Estimate = −19.31, SE = 18.4, df = 24.46, $t = -1.05$, $p = 0.3$), as well as the L2 natural group (Estimate = −23.93, SE = 24.83, df = 12.99, $t = -0.96$, $p = 0.35$). These models include random intercepts for USER and ITEM, i.e., the target words. The factor PREFERENCE was a significant predictor only in the model of the L1 synthetic group, indicating that the participants with a baseline preference for [e:] have a stronger tendency to diverge than the participants preferring [ɛ:], whose group intercept is slightly above zero (Estimate = 37.06, SE = 14.38, df = 19.55, $t = 2.6$, $p < 0.05$). The factor SPEAKER SEX did not improve the fit of the models and was therefore not included.

Figure 19 shows the individual productions of ⟨-ä-⟩ by each speaker in the three experimental groups relative to the vowels they heard from Mirabella. To evaluate the accommodating behavior on an individual level, two complementary tests were carried out per participant. First, a kernel density based global two-sample comparison test for 2-dimensional data was performed to determine whether the set of baseline vowels differed significantly from the set of map task vowels ($\alpha = 0.05$). Second, a two-sided one-sample Wilcoxon signed-rank test evaluated whether the individual $dDist$ distribution differed significantly from zero ($\alpha = 0.05$). If both tests reach significance, we consider the individual participant to accommodate to Mirabella, since their map task productions are substantially farther from their original baseline distribution while being substantially closer to (convergence) or farther from (divergence) Mirabella’s vowels. This approach suggests three cases of convergence with respect to vowel quality (*Nm02*, *Nf04*, and

Nf15) and five cases of divergence (*Nm01*, *Nm03*, *Sm01*, *Sm05*, and *Sf12*). None of these cases belong to the L2 speaker group.

Since the tests were performed for each participant individually, we have to consider adjusting the p-values to control the false discovery rate using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). For all participants mentioned above, the adjusted p-values of the Wilcoxon signed-rank test remain below 0.05. However, only for *Nf15*, *Sm01*, *Sm05*, and *Sf12*, the same is true for the kernel density based comparison, as well. While keeping this limitation in mind, we still consider all eight speakers to show accommodating behavior with regard to [ɛ:]/[e:]. See [Appendix H](#) for more detailed individual results.

4.3.5 Word ending <-ig>

The 1374 realizations of the word ending <-ig> uttered in tasks 1 and 4 (L1 natural: $n = 518$, L1 synthetic: $n = 570$, L2 natural: $n = 286$)¹⁶ were auditorily and visually classified as belonging to the fricative or plosive category by the author of the study and an additional phonetically trained annotator. The fricative category included variants of [ɪç] such as [ɪʃ]. The two resulting annotations of the baseline items were compared with the online annotation performed by the author of the study during the experiment on a purely auditory basis and under time pressure. The two offline annotations did not differ from each other and the online and offline annotations of the author differed in a single instance only. We conclude from this validation that the participants’ preference with respect to [ɪç]/[ɪk] was determined correctly.

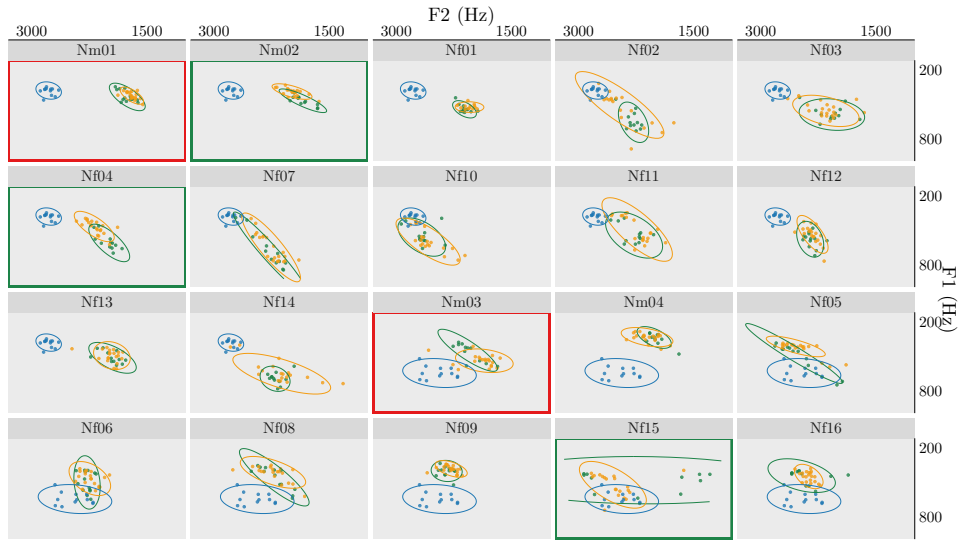
Since speakers are not always consistent in using only one variant during the baseline task, preference reflects the majority variant produced during task 1. Of the 42 L1 German speakers participating in the present experiment, 17 had a preference for the [ɪç] variant (13 female, 4 male) and 25 for the [ɪk] variant (18 female, 7 male). In the L2 German group, 9 speakers preferred [ɪç] (4 female, 5 male) and 2 speakers preferred [ɪk] (1 female, 1 male). Individual realizations were further classified as being the *same* as or a *different* variant than the one produced by Mirabella. [Figure 20](#) shows the results of the [ɪç] vs. [ɪk] evaluation for the three experimental groups.

L1 GERMAN SPEAKERS The results of the native speakers of German are given in [Figures 20a](#) (natural group) and [20b](#) (synthetic group). The clear majority of all baseline instances is produced with a *different* variant of the target contrast than the one the participants hear from Mirabella in the map task (natural: 90%, synthetic: 83%). This is expected, since the variant used by Mirabella is selected to be the opposite of each participant’s preference. In the remaining cases (natural: 10%, synthetic: 17%), the participants uttered the non-preferred variant in the baseline task, hence the *same* variant as Mirabella.

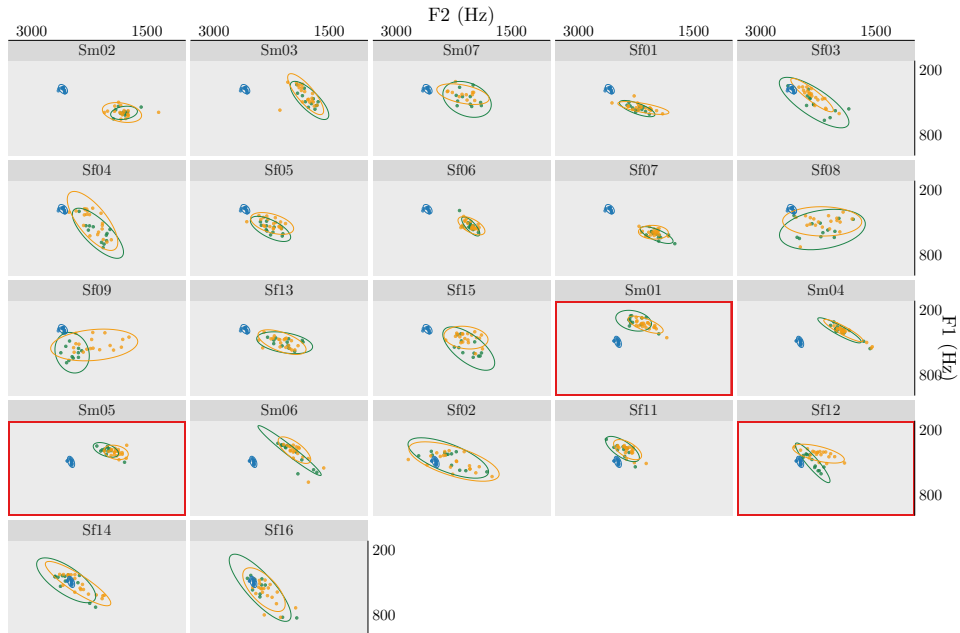
While the participants in the natural group are split equally between those preferring [ɪç] and those preferring [ɪk] and in each of the subgroups the *different* variant of the target contrast is produced in 90% of the baseline instances, the synthetic group contains more participants preferring [ɪk] (68%), and within this subgroup only 78% of the baseline instances are of the *different* type (compared to 94% for the [ɪç]-preference subgroup). This means that there is more variation in the baseline productions of the synthetic [ɪk]-preference subgroup than in the three other subgroups.

In the map task, the amount of non-preferred variants uttered by the participants increases by 27% to a total of 37% in the natural group and by 31% to a total of 48% in the synthetic

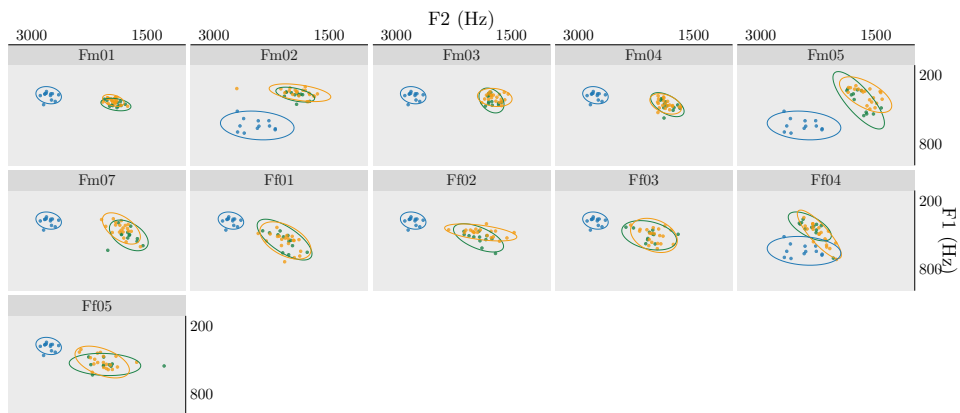
¹⁶ Theoretically expected number of data points: (12 base items + 14 map items [i.e., 2×2 nouns + 10 adj.]) \times number of participants. Small deviations due to missing values.



(a) L1 natural group.



(b) L1 synthetic group.



(c) L2 natural group.

Figure 19: Individual participant realizations of $\langle -\ddot{a}- \rangle$ in the F1–F2 space (in Hz) from the **baseline task** and the **map task**, relative to the vowels the participants heard from **Mirabella**. The ellipses indicate the 95% confidence interval. Framed participants were found to **converge** to or **diverge** from Mirabella.

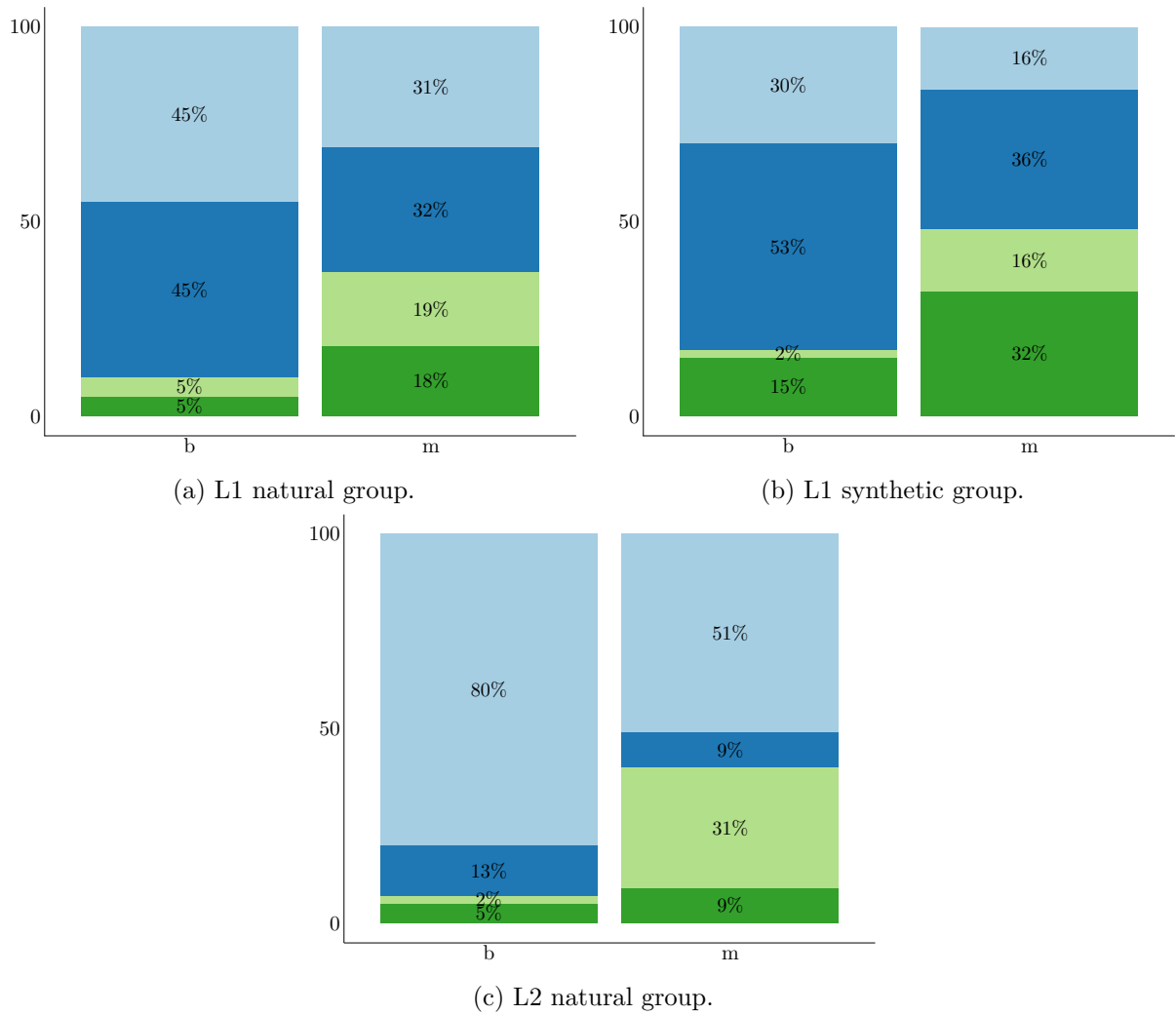


Figure 20: Percentages of the word ending <-ig> realized as the *same* variant (green tones) or a *different* variant (blue tones) as Mirabella in the baseline production (b) and the map task (m) split by participants whose baseline preference is [ɪç] (light tones) or [ɪk] (dark tones).

group. In the natural data, the occurrences of *same* variants quadruple for both subgroups ([ɪç]: 19%, [ɪk]: 18%). In the synthetic data, the [ɪç] and [ɪk] subgroups contribute to the increase to different proportions: There are eightfold as many *same* variants in the [ɪç] subgroup (16%) while the occurrences only double (32%) in the [ɪk] subgroup.

The increase of non-preferred variants per experimental group was evaluated by fitting a GLMMs with the binary response *different/same* as dependent variable and testing the factors TASK (base/map), SPEAKER SEX (female/male), and PREFERENCE ([ɪk]/[ɪç]) following the method described in Section 4.3.2.

The models did not converge when random intercepts for ITEM, i.e., the target words, were included, therefore the models were fitted only including random intercepts for USER. The factor TASK is a significant predictor of the dependent variable in the natural group (Estimate (log-odds) = -0.91, SE = 0.44, $z = -2.05$, $p < 0.05$) and the synthetic group (Estimate (log-odds) = -0.73, SE = 0.23, $z = -3.23$, $p < 0.01$) indicating an increase of *same* variants of the target contrast in the map task. The models include random slopes for TASK by USER to account for the individual reactions of the participants. The factor PREFERENCE is a significant predictor in the model of the synthetic group establishing the above made observation that the group of participants preferring [ɪk] is larger (Estimate (log-odds) = 0.69, SE = 0.28, $z = 2.47$, $p < 0.05$). However, there is no significant interaction of TASK and PREFERENCE. The factor SPEAKER SEX did not improve the fit of the models and was therefore not included.

L2 GERMAN SPEAKERS **Figure 20c** shows the results of the non-native speakers of German. In 93% of all baseline task instances, the French speakers produced a *different* variant of the target contrast than they heard from Mirabella in the map task. The remaining 7% are cases where the participants uttered the dispreferred variant in the baseline task, hence the *same* variant as Mirabella.

In the map task, the amount of dispreferred variants uttered by the non-native speakers increased by 33% to a total of 40%.

The baseline distribution and the accommodative effect in the map task is nearly identical to the L1 natural group, with the only difference that the majority of the French speakers (82%) had a baseline preference for [ɪç], while the German speakers were equally distributed between the two preference groups.

As for the native speakers above, the increase of dispreferred variants was evaluated by fitting a GLMM to the data of the non-native speakers. The model includes random intercepts for USER and ITEM,¹⁷ as well as by-user random slopes for TASK. Both the factor TASK (Estimate (log-odds) = -1.02, SE = 0.36, $z = -2.80$, $p < 0.01$) and the factor PREFERENCE (Estimate (log-odds) = -1.45, SE = 0.49, $z = -2.94$, $p < 0.01$) were significant predictors for *different/same*. This indicates an increase of *same* target contrast variants in the map task and it establishes that the group of participants preferring [ɪç] is larger. Again, there is no significant TASK/PREFERENCE-interaction and the factor SPEAKER SEX did not improve the fit of the model.

INDIVIDUAL BEHAVIOR **Figure 21** shows the individual realizations of the word ending <-ig> in chronological order by each speaker of the three experimental groups. Note that some speakers never deviate from their preferred allophonic variant, e.g., speakers *Nm02*, *Sf03*, and *Fm01* always produce the fricative variant, while speakers *Nf05* and *Nf06* only produce the plosive variant. In contrast, *Nf07*, *Sf14*, and *Fm02* are examples of speakers who have a clear

¹⁷ For consistency with the L1 German models, we fitted a model without random intercepts for ITEM, as well. However, it yielded a singular fit.

preference for one variant in the baseline task, but converge almost entirely to Mirabella during the map task.

To evaluate the accommodating behavior on the individual level we classified all participants according to the following thresholds, comparing the number of *same* instances in task 4 to task 1:

- increase of ≥ 7 \rightarrow substantial convergence
- increase of ≥ 2 \rightarrow moderate convergence
- in-/decrease of 1 \rightarrow maintenance
- decrease of ≥ 2 \rightarrow moderate divergence
- decrease of ≥ 7 \rightarrow substantial divergence

According to these criteria, 17 participants show substantial convergence (L1 natural: 6, L1 synthetic: 7, L2 natural: 4), moderate convergence is found in 18 participants (L1 natural: 5, L1 synthetic: 9, L2 natural: 4), 14 participants do not increase nor decrease the number of *same* instances (L1 natural: 8, L1 synthetic: 3, L2 natural: 3), and 4 participants moderately diverge from Mirabella (L1 natural: 1, L1 synthetic: 3). Substantial divergence on the individual level was not found.

4.3.6 Personality scores

To explore the influence of different personality traits on the accommodation occurring in the present study, we collected personality scores of all participants using the German version of the NEO-FFI (Borkenau and Ostendorf, 2007) for the native speakers of German and the French version of the BFI (Plaisant et al., 2005; Plaisant et al., 2010) for the native speakers of French.

These self-description questionnaires measure the *Big Five* personality traits, i.e., Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. The NEO-FFI uses a total of 60 items (12 items per trait) and takes approximately 10 minutes to complete. The BFI contains 45 items (8 to 10 per trait) and takes about 5 minutes to complete. The questionnaires were administered after the experiment.

To create a larger database for this analysis, we merge the data of the L1 natural and L1 synthetic groups. For the L2 speakers, we look at individual cases.

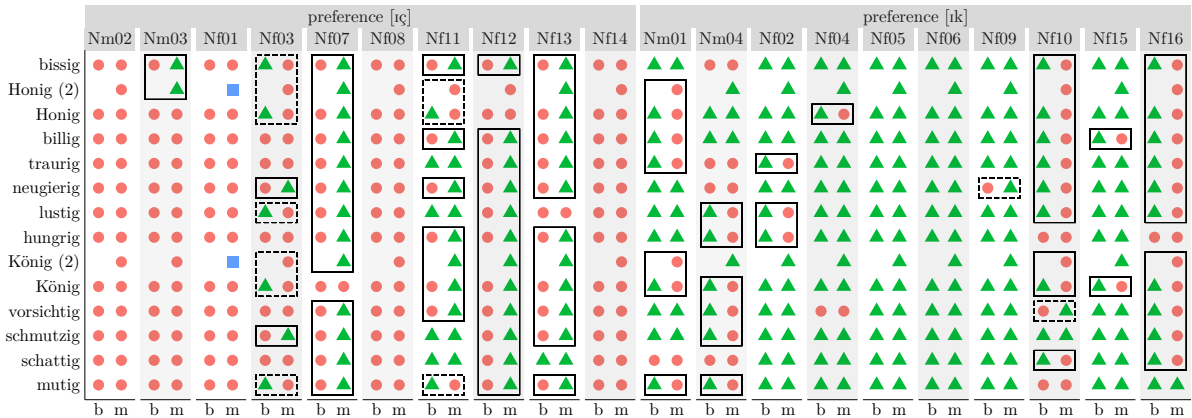
L1 GERMAN SPEAKERS Raw values were calculated for each personality trait and converted into standard T-values according to the guidelines provided by NEO-FFI. These standard values take the sex and age of the participants into account.

For each personality trait, we selected the 35% of all participants (combined natural and synthetic group) with the lowest values and the 35% with the highest values. This resulted in balanced subsets of 29 to 30 participants. For each of the five subsets and three phonetic features we fitted the statistical models described above again, always including random intercepts for USER and ITEM.

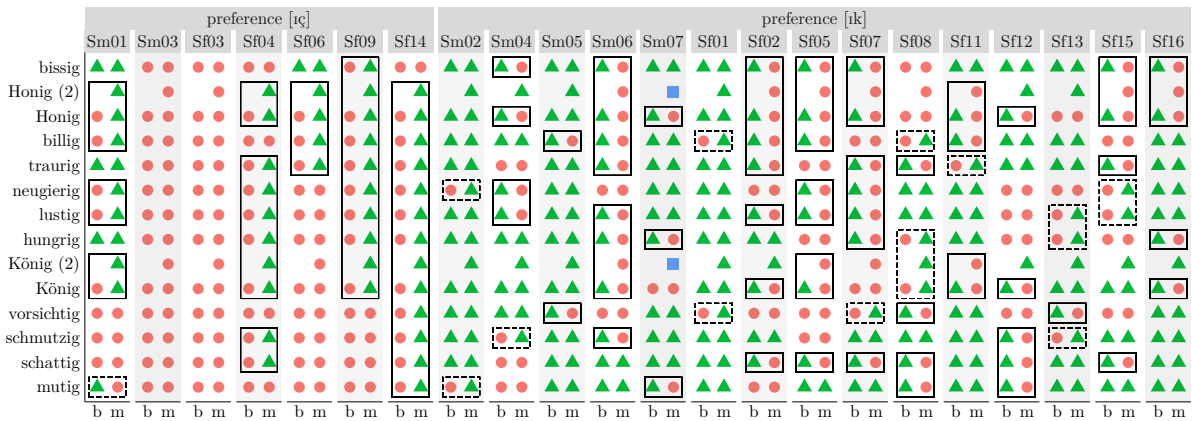
We tested the factors TASK (where applicable, i.e., for question intonation and word ending <-ig>), PREFERENCE (where applicable, i.e., for word ending <-ig> and long vowel <-ä->), and PERSONALITY TRAIT (high/low).

For one subset a significant effect of PERSONALITY TRAIT emerged:

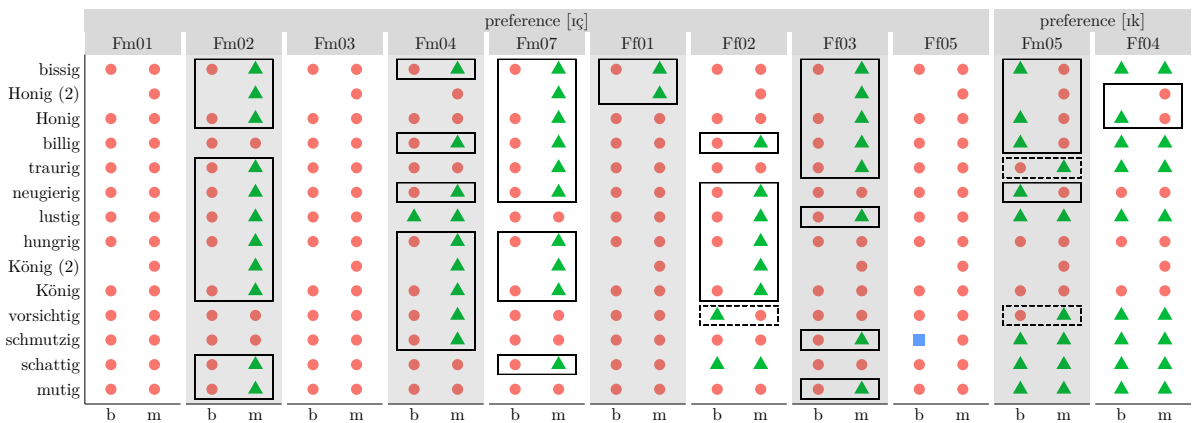
In the group of participants with very high or very low values for Neuroticism, question intonation was influenced by TASK, i.e., round 1 or round 2 of the question-and-answer game, (Estimate (log-odds) = -4.76 , SE = 1.14, $z = -4.02$, $p < 0.001$) and there was a significant interaction of TASK and PERSONALITY TRAIT (Estimate (log-odds) = -1.27 , SE = 0.57, $z = -2.22$, $p < 0.05$), indicating that participants who scored high values for Neuroticism were more



(a) L1 natural group.

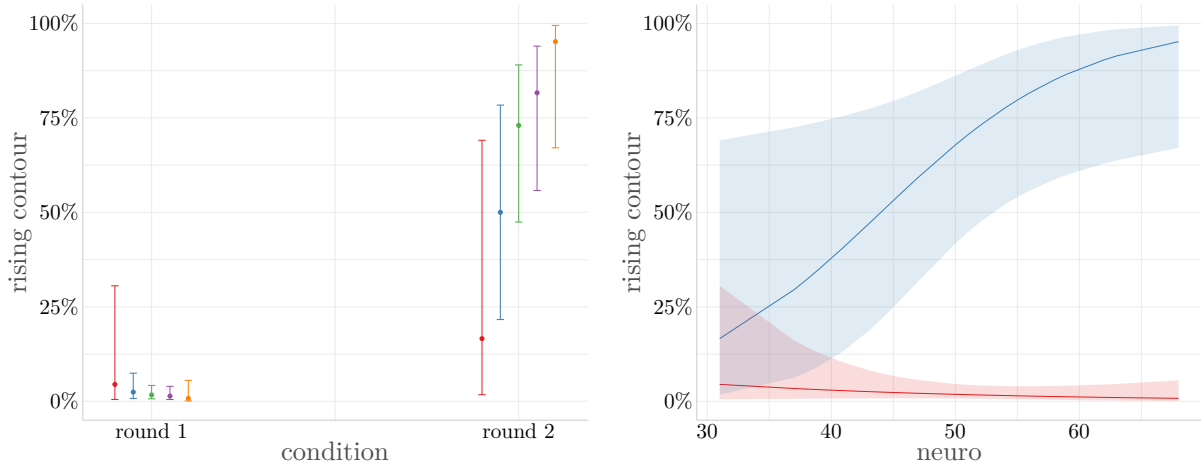


(b) L1 synthetic group.



(c) L2 natural group.

Figure 21: Individual results for the realization of the word ending <-ig> as [ɪç] ● or [ɪk] ▲ in the baseline production (b) and the map task (m). Target words are given in the order of occurrence in the map task, starting with *mutig*. Solid boxes show cases of convergence, dashed boxes cases of divergence. The ■ indicates missing values. Participants are grouped by their baseline preference for [ɪç] or [ɪk].



(a) Percentage of *rising* contours in the two conditions for different levels of *Neuroticism*. (b) Percentage of *rising* contours over *Neuroticism* for the two conditions.

Figure 22: Influence of Neuroticism on question intonation for the L1 German speakers. Figure (a) shows predictions for the **minimum value** (31; low), **lower quartile** (44), **median quartile** (52), **upper quartile** (56), and the **maximum value** (68; high) of Neuroticism per experimental condition. Figure (b) shows continuous predictions for **round 1** and **round 2** of the Q&A. The figures display the 95 % confidence interval.

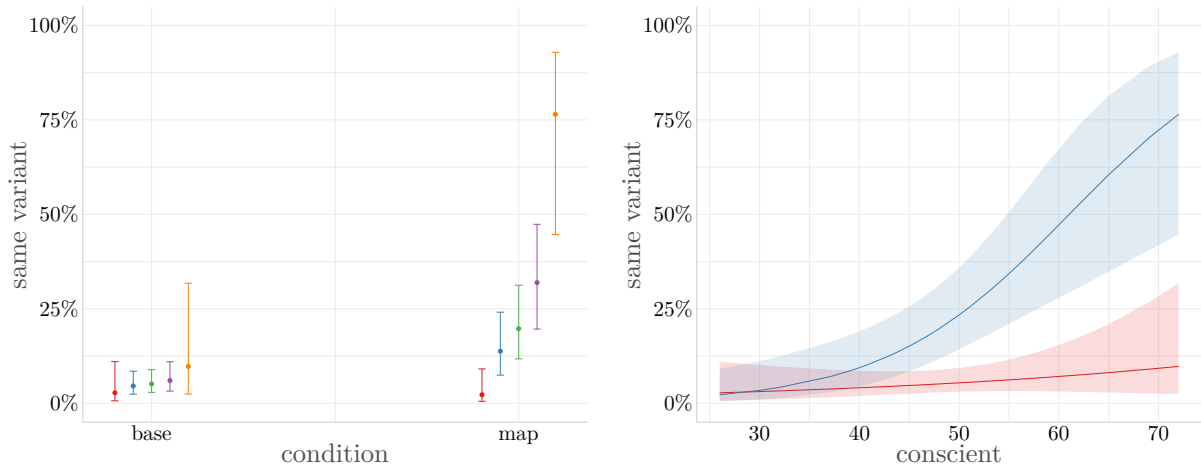
likely to produce questions with rising intonation in round 2 of the game and therefore more likely to converge to Mirabella. The model includes random slopes for TASK by USER.

The reported p-values are not adjusted for the fact that the general hypothesis of whether personality traits influence accommodation with respect to a particular phonetic feature was tested for five subsets of the same data set, which increases the probability of a false positive result. If we adjust the p-values to control the false discovery rate using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), TASK is still a significant predictor of the intonation contour ($p < 0.001$). However, with $p = 0.1$ only a trend remains of the interaction with PERSONALITY TRAIT.

We extended this rather coarse analysis, which used only the top and bottom 35 % of the speakers for each personality trait, with a more fine-grained analysis, which used the entire data set and included the personality traits as continuous factors (as above, in separate models).

The resulting model for the influence of Neuroticism on question intonation is illustrated in Figure 22. It confirms the finding that more neurotic participants were more likely to converge to Mirabella’s rising contours, with a significant interaction of TASK and PERSONALITY TRAIT (Estimate (log-odds) = -0.08 , SE = 0.22, $z = -3.86$, $p < 0.001$) that also persists when the p-values are adjusted to control the false discovery rate ($p < 0.001$). TASK is not a significant predictor in this model and the latter only converged when the random slopes for TASK by USER were omitted.

The analysis further revealed that including Conscientiousness as a continuous factor significantly improved the fit of the respective model for the realization of the word ending <ig>, see Figure 23. This suggests that more conscientious participants were more likely to converge to Mirabella’s realization of <ig>. The model showed a main effect of PERSONALITY TRAIT (Estimate (log-odds) = 0.07, SE = 0.03, $z = 2.30$, $p < 0.05$), but the interaction of TASK and PERSONALITY TRAIT turned out not to be significant (Estimate (log-odds) = -0.04 , SE = 0.02, $z = -1.85$, $p = 0.06$). Both TASK and PREFERENCE improved the model fit as well, but were also not significant predictors. The model includes random slopes for TASK by USER.



(a) Percentage of *same* variants in the two conditions for different levels of *Conscientiousness*. (b) Percentage of *same* variants over *Conscientiousness* for the two conditions.

Figure 23: Influence of Conscientiousness on word ending ⟨-ig⟩ for the L1 German speakers. Figure (a) shows predictions for the **minimum value** (26; low), **lower quartile** (44), **median quartile** (48), **upper quartile** (54), and the **maximum value** (72; high) of Conscientiousness per experimental condition. Figure (b) shows continuous predictions for **baseline production** and **map task**. The figures display the 95 % confidence interval.

L2 GERMAN SPEAKERS Because of the small number of speakers in the L2 natural group, we only take a look at individual cases. More specifically, we examine the speakers with the 35 % highest and the 35 % lowest scores for both Neuroticism and Conscientiousness, as well as their results with respect to intonation accommodation (for Neuroticism) and realization of the word ending ⟨-ig⟩ (for Conscientiousness).

Only one of the four speakers who converged on question intonation (substantial convergence) also shows particularly high scores for Neuroticism. Two of the four speakers (1 × substantial and 1 × moderate convergence) are in fact from the group with particularly low Neuroticism scores.

Three of the eight speakers who converged to Mirabella’s realization of ⟨-ig⟩ (2 × substantial and 1 × moderate convergence) also show particularly high scores for Conscientiousness. Also in this case, two speakers (1 × substantial and 1 × moderate convergence) are from the group with particularly low scores.

4.4 DISCUSSION

We conducted a WOz experiment with 42 native speakers of German and 11 native speakers of French to investigate phonetic accommodation by human interlocutors in an human-computer interaction (HCI) context. The participants of the experiment solved four tasks in interaction with the virtual language learning tutor *Mirabella*, who was created for this purpose. The participants were confronted with *Mirabella* using either natural or synthetic speech. The latter was generated using MaryTTS, HTS, and WORLD (see Section 4.2.3). The prosodic parameters segment duration and fundamental frequency were extracted from the natural stimuli and imposed on the synthetic ones. Due to this combined process, *Mirabella*’s synthetic voice was clearly identifiable as non-natural while the stimuli still exhibited a natural prosody.

4.4.1 *Reception of Mirabella*

After the experiment, the participants rated Mirabella with regard to her likability, competence, and intelligibility. Among the L1 German speakers, the natural Mirabella version was rated as being more intelligible, more likable, and somewhat more competent than the synthetic Mirabella version. However, both versions of Mirabella achieved high scores on all three 5-point scales with mean values well above 3 in each case.

The L1 French speakers also understood the natural version of Mirabella perfectly well and considered her to be very likable. They rated Mirabella's competence even higher than the L1 natural group did. This is consistent with the assumption that, as learners of German, the L1 French speakers perceive themselves to be hierarchically inferior to the "native speaker" Mirabella.

Mirabella's response time, i.e., the response time of the experimenter, was evaluated as well. It was considered equally appropriate in all three experimental groups. This is plausible, since the experimenter was always the same person.

Overall, the ratings of the synthetic Mirabella version showed more variability, which means that the participants were less in agreement in her case. It is possible that for some participants the attitude towards the non-natural sounding synthetic voice interferes with the evaluation of the different qualities, while other participants are able to abstract from this impression and evaluate Mirabella independently of it. This could lead to the wider range of scores we observe. For the L2 natural group, the reaction time score showed similar variability. It stands to reason that, especially for language learners, the evaluation of reaction time depends on listening comprehension skills and therefore varies considerably. Future work could investigate the influence of these evaluations on the accommodating behavior in detail.

As part of the questionnaire administered after the experiment, the participants could also express their thoughts and assumptions about the experiment. None of the participants raised any doubt that Mirabella functioned fully automatically, neither in the questionnaire nor through informal comments. On the contrary, they referred to their experience in a way that suggests they believed that they were interacting with a computer, which is a key component of HCI (Branigan et al., 2010). Furthermore, speakers expressed no suspicion that Mirabella was testing particular pronunciation-related phenomena. In the L1 German group, a frequently expressed assumption about the purpose of the study was to evaluate the dialog system in terms of how well it understands different participants and how quickly it responds to speech input. The interaction was perceived in many cases as a training for Mirabella with the presumed goal of improving HCI. One participant described the system as being child-friendly and suggested that it could be used in schools. In the L2 German group, a recurring assumption was that the test was about how well artificial intelligence can understand non-native speakers, which indicates that Mirabella was indeed perceived as an intelligent system.

4.4.2 *Accommodation to Mirabella*

We tested accommodation with respect to the intonation of constituent questions in a Q&A game, and the variation of the German allophone pairs [ɛ:] vs. [e:] as a realization of the long vowel ⟨-ä-⟩ in stressed syllables, e.g., *Käse* (*cheese*), and [ɪç] vs. [ɪk] as a realization of the word ending ⟨-ig⟩, e.g., *Honig* (*honey*), in a map task.

Both the Q&A game and the map task are of a rather repetitive nature. However, they are structured to reflect a possible interaction of human speakers, they enabled an engaging, dynamic and meaningful exchange between the participants and Mirabella, and it is conceivable that they could occur in a real-life learning context, especially in CALL.

In the Q&A game, the participants took turns with Mirabella asking about the location of animals on the screen and providing the requested information. The distribution of roles was therefore relatively equal in this task. The questions and answers always followed the same pattern. Therefore, the participants had hardly any difficulty in formulating them. However, since the order in which the questions were asked was not predetermined and Mirabella's questions also had to be answered correctly in order to continue playing, we assume that the participants' attention was not particularly focused on the phonetic realization of the questions.

In the map task, Mirabella held more of a leading role because she had knowledge about the hidden information, i.e., the target words, before the participant did. She provided the information in full sentence contexts and the participants had to include it in a two-part utterance. To construct an utterance, the participants had to select a suitable preposition and formulate grammatically correct sentences. This seemed to be difficult at times — even for the participants who were native speakers of German —, but always resulted in acceptable utterances for the purpose of the current study. In any event, the participants' attention had to be divided between different domains and we assume that pronunciation did not stand out as an obvious target.

4.4.2.1 Question intonation

L1 GERMAN SPEAKERS As expected for native speakers of German, all participants produced predominantly *falling* intonation contours when formulating constituent questions from given fragments.

When interacting with Mirabella in the first round of the Q&A game, where she produced her questions with a nuclear pitch accent on the $\langle animal \rangle$ followed by a final F_0 fall, this predominance was reinforced in both experimental groups. The small amount of *falling-rising* contours and *rising(a)* contours that occurred in these two tasks, could either be idiosyncratic behavior — speaker *Nf03*, for example, produced exclusively *rising(a)* contours — or an expression of insecurity or politeness — such feelings are likely to weaken in the course of the interaction, e.g., because Mirabella's behavior confirms that the task is being carried out correctly. Therefore, it is unlikely that an increase in rising contours at later points in the interaction is attributable to insecurity or politeness.

The crucial change happened in the second round of the Q&A game, where Mirabella produced all questions with a nuclear pitch accent on the interrogative pronoun *wo* (where) followed by a final high F_0 rise (*rising(w)*). This behavior led to a significant increase of rising contours (this includes *falling-rising*, *rising(a)*, and *rising(w)* contours) on the part of the participants in both experimental groups. This increase can mainly be attributed to a change in intonation contour while keeping the nuclear pitch accent on the $\langle animal \rangle$. However, in a smaller number of cases, participants also shifted the nuclear pitch accent to the interrogative pronoun. This suggests that the participants were primarily receptive to the overall rising contour. It seems sensible to ask to what extent convergence can take place without giving the impression to mock the interlocutor. The question intonation in the present study may well be a case in which full convergence, i.e., a rising contour *with* a shifted pitch accent, seems to go one step too far for many participants. Since a rising contour *without* a shifted pitch accent results in a more acceptable form than a shifted pitch accent with a *falling* contour — no such cases occurred in our data —, we can observe this clear two-step convergence hierarchy.

L2 GERMAN SPEAKERS The L1 French speakers had no particular preference for *falling* or *rising* contours in their baseline productions. The reasons for this are unclear, since a preference for falling contours was expected among the L1 French speakers as well. As described above

for the L1 German group, the rising contours could be an expression of insecurity. It is fair to assume that such insecurity may be more pronounced and persistent in non-native speakers of the target language. A comparison with the participants' performance in their native language on the same task would be a valuable future extension to this analysis.

As it was the case for the L1 German speakers, the number of *falling* contours increased as Mirabella produced *falling* contours in the first Q&A round, and the number of *rising* contours increased as she produced *rising* contours in the second Q&A round. However, the increase from first to second round was not significant for the L1 French speakers. This could be a result of the unexpectedly high occurrence of rising contours in the baseline task and the first round of the Q&A, as this gave the L1 French group less room to adapt. We suspect that a feature that is common in the speaker's own speech may, on the one hand, be an easier target for accommodation because it is mastered anyway and can be easily implemented, but on the other hand, the very lack of room to the interlocutor may also prevent moving *even further* towards them and becoming *too similar*. The decrease of rising contours in round 1 of the Q&A and the increase in round 2 on the part of the L1 French speakers are, in our opinion, most certainly a consequence of the interaction with Mirabella and, as such, meaningful illustrations of accommodation.

Mirabella's shift of the nuclear pitch accent to the question word *wo* (where) was never adopted by the L1 French speakers. It can be assumed that the metrical pattern of French, which uses relatively small accentual phrases and has an obligatory phrase-final accent (Di Cristo, 1998), contradicts the realization of an initial nuclear pitch accent considerably and therefore even advanced learners (here: CEFRL B2/C1) do not adopt this pattern. In this case a limit of accommodation may have been reached. To emphasize the question word and still follow the native pattern, French would favor a syntactic variation in combination with a *rising* contour, namely: "L'animal se cache où?" (*The animal is hiding where?*; Delais-Roussarie et al., 2015).

Two particular patterns emerged among L1 French speakers. In the first round of the Q&A, an unusual two-part phrasing of the questions occurred: "Wo hat sich die Kuh | versteckt?" (Where did the cow | hide?), and the nuclear pitch accent was often placed on the ultima of *versteckt* (hidden) instead of the ⟨*animal*⟩. In the second Q&A round, the questions were more often produced as a single intonation phrase and the nuclear pitch accent was significantly more often placed on the ⟨*animal*⟩.

This may be interpreted either as a reduction of insecurity in the interaction with Mirabella on the part of the L1 French speakers, or as accommodation of their own native pattern to that of the foreign language.

PRAGMATIC CONTEXT In the following we would like to return briefly to the influence of the pragmatic context on the task at hand. We have already mentioned that the echo questions that Mirabella produces in round two do not contradict the context. But they are also not expected to occur, since the change in the pragmatic context is not very obvious. In round two, the animals are arranged differently on the screen than in round one and we can assume that this is consciously perceived by the participants. Furthermore, the animals are still paired with the same house number, which justifies an echo question, but in our opinion is probably not consciously noticed by the participants.

Could it still be accommodation to the changed pragmatic context instead of to Mirabella's speech output that we observe in our data? The majority of the participants adopted the rising intonation, but did not shift the pitch accent to the interrogative pronoun in round two. However, there is no pragmatic motivation for this, because only the shift of the nuclear pitch

accent (in combination with rising intonation) changes the function of the question to suit the changed pragmatic context.

In conclusion, we do not believe that changing the pragmatic context alone would trigger the observed amount of questions with rising intonation, nor do these questions fit functionally to the changed pragmatic context. We therefore assume that the observed change in question intonation by the participants constitutes accommodation to Mirabella.

PERSONALITY SCORES In a separate analysis of the joint L1 natural and L1 synthetic data, including personality scores collected with the German version of the NEO-FFI (Borkenau and Ostendorf, 2007), Neuroticism emerged as a significant predictor of accommodation to question intonation, with more neurotic participants converging more to Mirabella. In a first analysis comparing only the participants with the highest and lowest scores for Neuroticism, the effect did not hold when applying the Benjamini–Hochberg correction to account for multiple comparisons. However, a more fine-grained analysis that included Neuroticism as a continuous factor demonstrated its influence on question intonation even after adjusting the p-values. The finding is in line with Lewandowski and Jilka (2019), where more neurotic speakers show more convergence with respect to word-based amplitude envelope match. We would like to discuss a possible explanation for the occurrence of such an effect regarding question intonation. A high degree of Neuroticism is synonymous with emotional instability. People with a high level of Neuroticism are more likely to state that they are easily out of balance, more insecure and nervous, and less able to control their needs (Borkenau and Ostendorf, 2007). Lewandowski and Jilka (2019) relate the degree of Neuroticism to the need of social approval and suggest that under the CAT perspective (Giles, 1973; Giles et al., 1991; Shepard et al., 2001), this might imply that neurotic people have a tendency to converge in an attempt to avoid distress. However, the degree of Neuroticism was not predictive of the other features tested in the present study. A possible difference between the question intonation and the allophonic contrasts is that deviating from the expected way of formulating questions may have more potential to cause communicative distress than using another allophonic variant.

Note that we conducted an analysis of isolated personality traits, while traits may also interact with each other in influencing accommodating behavior.

Among the native speakers of French, we did not observe any particular connection between a high level of Neuroticism, as determined by the French version of the BFI (Plaisant et al., 2005; Plaisant et al., 2010), and accommodation with respect to question intonation. However, we have to take into account the small group size, in light of which this result should not be given too much weight.

We consider it reasonable to assume that personality traits such as Neuroticism — or Openness as found by Yu et al. (2013) — facilitate accommodating behavior. However, since they are certainly not the only determinants of this behavior, they do not necessarily allow precise predictions for individual speakers with respect to specific phonetic features.

A limitation of the present analysis that should be considered is that the participants were not selected on the basis of their personality. It is possible that existing effects would be more pronounced in an experimental group selected on the basis of an extreme expression (particularly high/low) of the personality traits in question.

4.4.2.2 *Word ending <-ig>*

L1 GERMAN SPEAKERS With respect to the [ɪç]/[ik] contrast, we found a significant convergence effect during the map task for both experimental groups, L1 natural and L1 synthetic. This effect did not depend on the baseline preference of the speakers, which was equally dis-

tributed between both variants in the L1 natural group and skewed towards [ɪk] in the L1 synthetic group. Although [ɪç] is codified Standard German and [ɪk] a Southern German variant, which might imply that the former is more prestigious and therefore able to trigger more convergence, an effect of baseline preference was not expected, since Kiesewalter (2019) showed that [ɪk] is perceived as being close to the standard by native listeners of German. Further evidence for the ambiguous status of the [ɪç]/[ɪk] contrast comes from the participants of the present study: In the post-experiment questionnaire, almost 40 % of the participants misjudged which variant of the contrast they predominantly produce themselves.¹⁸ When asked for their opinion about the respective other variant, the vast majority judged it as acceptable. Only five participants had a negative opinion about the variant they did not produce themselves, e.g., “wrong” or “weird”. For speakers *Nm02*, *Nf01*, and *Nf14* this was [ɪk] and they did indeed not produce a single instance of it. For speakers *Sf02* and *Nf09* the disliked variant was [ɪç]. *Nf09* did produce one [ɪç] in the baseline task and then diverged to only producing [ɪk] during the map task, while *Sf02* produced three instances of [ɪç] in the baseline task and then even showed substantial convergence to Mirabella during the map task. This suggests that the attitude of a speaker towards the feature in question might influence their accommodating behavior, but does not fully predict it.

L2 GERMAN SPEAKERS The L1 French speakers also exhibited a significant increase in dispreferred variants when communicating with Mirabella in the map task, which means that they converged to her. As for the L1 German speakers, the effect did not depend on their baseline preference. The majority of the L1 French speakers had a baseline preference for the fricative variant, which may be due to the fact that, as non-native speakers, they had learned German through formal instruction and [ɪç] is the codified Standard German variant. Just 20 % of the L1 French speakers misjudged their own baseline preference, and all speakers had a positive attitude towards the variant they thought not to produce themselves. Only participant *Fm07* indicated that he had “never heard” the plosive variant before. Since this statement was made after the experiment and Mirabella had produced [ɪk] in his case, this is evidently not true, but could probably be replaced with “never consciously perceived”. Moreover, *Fm07* exhibited considerable convergence to Mirabella’s plosive productions, showing that he had perceived [ɪk] at least subconsciously.

PERSONALITY SCORES The above discussed analysis of the influence of personality traits revealed a tendency for more conscientious L1 German speakers to converge more often to Mirabella’s version of ⟨-ig⟩. However, this effect was not significant in the final statistical model. What could be a possible reason for Conscientiousness to affect accommodation in the case of the word ending ⟨-ig⟩? The personality trait Conscientiousness is characterized by diligent, efficient, and orderly behavior, a desire to perform well in assignments, great discipline, and a preference for planning (Borkenau and Ostendorf, 2007). Therefore, it is conceivable that Conscientiousness may promote convergence for a binary feature such as the [ɪç]/[ɪk] contrast, whose categorical nature enables choosing the “correct” variant, i.e., that of the interlocutor, and thus show a “better” performance in the task. Moreover, the fact that there is a certain awareness — however vague — that in the case of [ɪç]/[ɪk] one of the two variants pertains to codified Standard German may make a conscientious speaker even more likely to converge.

For the L1 French speakers, we found no relationship between Conscientiousness and accommodation with respect to the realization of the word ending ⟨-ig⟩.

¹⁸ This is consistent with the assumption made in Mitterer and Müsseler (2013) that speakers are often unaware which variant of [ɪç]/[ɪk] they use.

Overall, we conclude that Conscientiousness may influence accommodation, which to our knowledge has not been found in the previous literature. However, the underlying motivation for accommodation of a conscientious speaker seems to be different from that of a neurotic speaker, and therefore a different set of phonetic features is likely to be affected.

4.4.2.3 Long vowel <-ä->

The analysis of the [ɛ:]/[e:] contrast by measuring the difference in Euclidean distance in the F1–F2 space did not reveal an accommodation effect for either of the three experimental groups, L1 natural, L1 synthetic, and L2 natural. Only a stronger divergence tendency among the L1 German speakers with a baseline preference for [e:] in the synthetic group was predicted by the statistical model. The absence of substantial accommodating behavior on the group level was not expected, since the participants of the previous shadowing experiment converged with respect to the [ɛ:]/[e:] contrast when shadowing natural stimuli, but also, to a smaller extent, when shadowing HMM-based stimuli (see [Chapter 3](#)). However, formulating a new utterance entails a higher cognitive load than repeating a given utterance. Therefore, the attention to phonetic detail at the level needed to capture the fine-grained differences in vowel quality may not have been available to the participants of the present study. In addition, it is possible that the gradual change in vowel quality is generally more difficult for speakers to access and control than the binary variation between fricative and plosive in the case of the [ɪç]/[ik] contrast or the different forms of question intonation. The present analysis of the vowel quality is also stricter than that of the [ɪç]/[ik] contrast, since the distribution is considered as a whole and individual cases of accommodation are thus not taken into account. Other ways of evaluating the [ɛ:]/[e:] contrast, e.g., as a categorical change between [ɛ:] and [e:], may provide more insight and would be a valuable future extension to the present results.

4.4.2.4 Individual behavior

To get an impression of the individual accommodating behavior within the two experimental groups, we determined for each participant whether they converged, diverged, or maintained their preference for the three analyzed features. The accommodation to the question intonation and the [ɪç]/[ik] contrast was further classified as being moderate or substantial. For the [ɛ:]/[e:] contrast the individual result reflects a combination of a significant shift away from the participants' own baseline vowel productions and towards/away from Mirabella.¹⁹ [Figure 24](#) shows the resulting individual accommodating behavior of the 42 native speakers of German and the 11 native speakers of French.

The majority of substantial convergence cases are found for the question intonation in both L1 groups and for the [ɪç]/[ik] contrast in the L2 group. Overall convergence (moderate and substantial) in the L1 natural group is led by question intonation as well ($n = 16$), followed by the [ɪç]/[ik] contrast ($n = 11$), and even three individual cases of vowel convergence, i.e., for speakers *Nm02*, *Nf04* and *Nf15*. In the L1 synthetic group, question intonation and [ɪç]/[ik] contrast are on par (both: $n = 16$) and no individual cases of vowel convergence occurred. For the L2 natural group, finally, [ɪç]/[ik] convergence was most prevalent ($n = 8$), followed by question intonation ($n = 4$). Again, no individual cases of [ɛ:]/[e:] convergence were observed. Occasional divergence is found for the [ɪç]/[ik] contrast (speakers *Nf03*, *Sm02*, *Sf01*, and *Sf13*) and the [ɛ:]/[e:] contrast (speakers *Nm01*, *Nm03*, *Sm01*, *Sm05*, and *Sf12*).

¹⁹ As pointed out in [Section 4.3.4](#), the individual results for the [ɛ:]/[e:] contrast are based on unadjusted p-values. When adjusting the p-values to control the false discovery rate, one case of convergence remains in the natural group and three cases of divergence in the synthetic group.

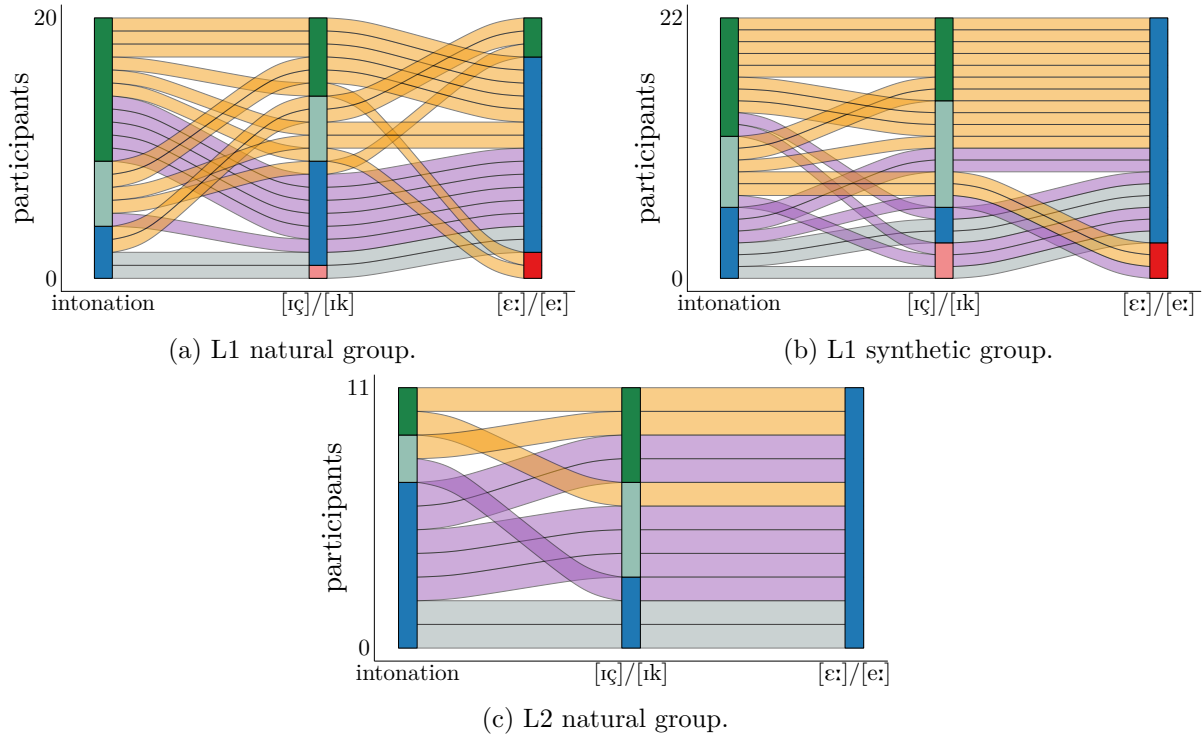


Figure 24: Individual accommodation behavior of the participants on the three examined features. The colors code **substantial convergence**, **moderate convergence**, **maintenance**, **moderate divergence**, and **substantial divergence**. Some participants converge with respect to **two features**, some only for **one feature**, and some do **not converge** at all.

According to these measures, 60% of the L1 German participants converged to two out of the three tested features (natural: $n = 12$, synthetic: $n = 13$) and 28% to one feature only (both: $n = 6$). Very few participants did not converge at all (natural: $n = 2$, synthetic: $n = 3$). Among the L2 German participants, only 27% converged to three features, 55% to one feature, and 18% to none at all. This confirms that accommodating behavior with respect to one phonetic feature does not necessarily predict the behavior with respect to another feature, which was previously documented for acoustic-prosodic features in human-human interaction (HHI; e.g., Sanker, 2015; Cohen Priva and Sanker, 2018; Reichel et al., 2018; Weise and Levitan, 2018).

In the questionnaire administered after the experiment, very few participants — only L1 German speakers — stated that they had consciously perceived some of the tested features. Three participants pointed out that Mirabella produced ⟨-ig⟩ differently than they expected. Among them were *Nm02* and *Nm14*, who showed no accommodation to Mirabella with respect to this feature, and *Nm01*, who converged substantially. Two other participants commented on the varying question intonation, namely *Sm03*, who did not change their own intonation at all, and *Nm03*, who adopted both the rising intonation and the shifted pitch accent from Mirabella at the fourth trial. This illustrates on a small scale that the conscious perception of a phonetic change neither necessarily leads to nor prevents accommodation. The extent to which the other participants consciously reflected on pronunciation characteristics of Mirabella cannot be further evaluated.

4.5 CONCLUSION

In summary, the participants of the present study accommodated their phonetic productions to the speech of a virtual language learning tutor with respect to two out of three tested features,

i.e., question intonation and the allophonic contrast [ɪç] vs. [ɪk]. This shows that accommodating behavior in users of a SDS is indeed triggered by locally anchored phonetic features. Also in line with our predictions (see Section 4.1), the accommodation occurred in the form of convergence. This was the expected behavior under both the assumption that alignment between interlocutors is an automatic process (cf. *Interactive Alignment Model (IAM)*; Pickering and Garrod, 2004; Pickering and Garrod, 2013) and the assumption that we aim to decrease social distance to an interlocutor by converging to them (cf. *CAT*; Giles, 1973; Giles et al., 1991; Shepard et al., 2001), since the participants considered Mirabella to be likable, showing that they had a positive attitude towards her. The participants did not accommodate to the allophonic contrast [ɛ:] vs. [e:], which in turn demonstrates that phonetic convergence does not necessarily occur for all features.

The absence of accommodating behavior at the group level, as in the present case of [ɛ:]/[e:], may be related to the salience of the feature in question: if it is not recognized as a potential target for accommodation (consciously or subconsciously), it cannot lead to convergence.

Considering the motivation to reduce social distance or to facilitate communication with an interlocutor through convergence, it is possible that different phonetic features contribute to these goals to varying degrees and speakers may implement accommodation accordingly. The intonation of a question, for example, has a direct influence on eliciting the correct answer and may also bear social meaning. Both the binary [ɪç]/[ɪk] contrast and the gradual [ɛ:]/[e:] contrast may influence comprehensibility, especially for interlocutors with a strong preference for one variant. Due to its residual perceived dialectality, [ɪç]/[ɪk] may bear a certain degree of social significance, as well. In the case of [ɛ:]/[e:], which is in the process of merging to [e:] across the German-speaking regions, such social significance may not — or no longer — apply.

As expected, we found considerable variation with respect to the degree and direction of accommodation on the level of individual speakers. It has already been suggested that a model of phonetic accommodation that combines the *automatic approach* (IAM) and the *social approach* (CAT) is influenced by additional factors (e.g., Lewandowski, 2012). For example, various aspects of the speaker disposition may be associated with individual differences in accommodating behavior (Yu et al., 2013; Lewandowski and Jilka, 2019). We tested the influence of the *Big Five* personality traits on the accommodating behavior in our data, which revealed, for the L1 German speakers, an influence of Neuroticism on the convergence to question intonation, as well as a tendency for Conscientiousness to affect convergence with respect to the [ɪç]/[ɪk] contrast. We assume that the underlying motivation for accommodation may vary between these personality types. While the more neurotic speakers want to avoid communicative distress by adopting intonational features, the more conscientious speakers aim to perform well in the assignment and adopt the potentially “correct” version of the binary feature. Openness, which had previously also been shown to positively correlate with convergence, did not appear as a predictor in our data.

In keeping with our predictions, the overall results did not differ between the experimental groups (both L1 German speakers) that communicated with either the natural or the synthetic speech version of Mirabella. Mirabella’s synthetic voice was clearly identifiable as non-natural, which did not prevent nor promote accommodating behavior. It remains unclear to what extent the presumed advantages that the different voice types hold (see Section 4.1) have worked in their favor, e.g., *natural voice*: potentially more straight-forwardly perceived as social actor, therefore more accommodation according to CAT; *synthetic voice*: potentially perceived as more machine-like and more likely to benefit from convergence.

Initially, we had hypothesized that accommodation might be weakened for the native speakers of German because, first, they are most likely confident in their own pronunciation and, second, they interact with a virtual language learning tutor for German whom they are likely to per-

ceive as hierarchically inferior to them. Both aspects argue against a strong convergence effect. Nevertheless, such an effect was observed. It could be that our second assumption counteracted the latter, namely that the participants probably did not perceive the SDS as fully linguistically flexible and therefore assumed (consciously or subconsciously) that it could likely benefit from convergence.

For all three examined features, the non-native speakers of German behaved similarly to the native speakers, but exhibited specific patterns — especially in the case of the question intonation, where they never shifted the nuclear pitch accent to phrase-initial position. That the L1 French speakers showed accommodation to Mirabella in form of convergence is consistent with the assumption that convergence occurs in the direction of the hierarchically superior interlocutor: As a “native speaker” of the target language, Mirabella has a model function and furthermore provides the participants with information on how to solve the tasks. Another initial assumption was that the non-native speakers might have greater difficulty in perceiving the phonetic detail in Mirabella’s speech and implementing it in their production. For the phrase-initial pitch accent placement, this may have been the case, as this pattern deviates strongly from the native pattern of L1 French speakers. However, the group of non-native participants in this study consisted of very proficient speakers of German. The described difficulties would certainly affect beginners to a greater extent.

We conclude that phonetic accommodation on the level of local prosody and segmental pronunciation occurs in users of SDSs. This may be exploited, for example, in computer-assisted language learning applications in a way that is beneficial for many users of such systems. Non-native speakers interacting with a virtual language learning tutor show a similar degree of accommodating behavior towards the latter as do native speakers, which enables incidental inductive learning, i.e., automatic learning by generalizing from examples without intending to do so (Williams, 2009). However, structural phonological elements of the target language that deviate too radically from the native pattern seem to require more explicit training.

5 MENTAL BOUNDARIES

We provide an overview of the structure and psychometric properties of the Boundary Questionnaire and its short versions, the situation with respect to German BQ versions, as well as the relationship of mental boundaries and the *Big Five* personality traits. We present the adaptation process of the empirically-derived short version of the Boundary Questionnaire for German, which consists of the translation and the validation with native speakers of German. This process also considers the *Big Five*. As a first use case, we applied the resulting German adaptation in the Wizard-of-Oz experiment to investigate the influence of mental boundaries on the accommodating behavior.

5.1 INSTRUMENTS

5.1.1 *The Boundary Questionnaire*

In the original English 145-item version of the Boundary Questionnaire (BQ; Hartmann et al., 1987; Hartmann, 1989; Hartmann, 1991), respondents express their attitude towards items like “I feel unsure of who I am at times” or “I think children need strict discipline” on a five-point scale from 0 (*not at all true of me*) to 4 (*very true of me*). To control for the tendency of respondents to agree with statements, 58 of the 145 items, including the second example above, are reverse scored. The overall boundary score *SumBound* is determined by summing the points; higher total numbers represent thinner boundaries. Seven items were excluded from the calculation of *SumBound* due to non-significant or negative correlations with the latter. However, they remained in the questionnaire. *SumBound* BQ is therefore based on 138 items.

For a sample of 866 subjects (53% female, 47% male; $M_{\text{age}} = 33$ years; $SD_{\text{age}} = 15$ years; students, patients, and research participants) who took the BQ and yielded a mean *SumBound* value of 273 with a standard deviation of 52 (female: $M = 288$, $SD = 51$; male: $M = 263$, $SD = 49$), Hartmann (1991) and Harrison et al. (2006) report a very high internal consistency with Cronbach’s $\alpha = 0.93$ (Cronbach, 1951). This suggests that some items may be redundant (Tsang et al., 2017) and the questionnaire could be shortened.

Conceptually, the 145 items of the BQ are divided into twelve topic areas, such as *Sleep, wake, dream*, e.g., “When I awake in the morning, I am not sure whether I am really awake for a few minutes”, *Unusual experiences*, e.g., “I have had déjà vu experiences”, or *Thoughts, feeling, mood*, e.g., “At times I feel happy and sad all at once”.

To determine the empirical internal structure of the 138 items that are included in *SumBound*, Harrison et al. (2006) performed a principal component analysis (Pearson, 1901; Hotelling, 1933) on the sample data described above, followed by Varimax rotation (Kaiser, 1958) of thirteen factors as suggested by the scree-plot of eigenvalues (Cattell, 1966). All items with a loading of 0.25 or above in the resulting factor matrix were considered as belonging to a factor. This resulted in twelve easily interpretable factors and one non-interpretable factor, namely factor XIII. According to Harrison et al. (2006, p. 371), factors I through XII pertain to three conceptual groups: “experiential barriers within the psyche”, e.g., *Primary Process Thinking* and *Overinvolvement*, “boundaries between internal and external events”, e.g., *Openness* and *Flexibility*, and “opinions about the world, and styles of organizing them”, e.g., *Belief in*

Impenetrable Inter-group Boundaries and Identification with Children. The internal consistency of the factors is given by theta coefficients (Armor, 1973) ranging from 0.92 (I) to 0.56 (XII).

The validity of the instrument was assessed by examining subjects for whom particularly thin (art students and nightmare sufferers) or particularly thick (naval officers) mental boundaries were assumed according to theory. These predictions were confirmed in the data, with the thin group scoring higher *SumBound* values ($M = 336$) than the thick group ($M = 248$). The BQ has thus been shown to be a valid and reliable instrument for measuring the strength of mental boundaries.

5.1.2 Short versions of the BQ

There are two English short versions of the BQ, which could be used as a basis for an adaptation to German. In the following, we present both versions and explain our choice.

Kunzendorf et al. (1997) propose an 18-item version of the BQ (BQ18) for which they selected 18 of the 138 original items according to the following three criteria: the items' face validity, i.e., "the degree to which test respondents view [the items] as relevant to the context in which the test is being administered" (Weiner and Craighead, 2010, pp. 637), their correlation with BQ's *SumBound*, and their distribution over BQ's topical categories. They refer to an unpublished study with 856 subjects that showed positive correlations of *SumBound* BQ and each of the 18 individual BQ18 items ($M_r = 0.36$, $SD_r = 0.09$), as well as a substantial correlation with *SumBound* BQ18 ($r = 0.87$). Thus the 18-item version of the BQ seems to capture the boundary construct very well.

Rawlings (2001) developed an empirically-derived short version of the Boundary Questionnaire (BQ-Sh) from the data of 300 early-stage psychology students (74 % female, 26 % male; $M_{\text{age}} = 19$ years; age range 17 to 56 years) who completed the original BQ in exchange for course credit. Factor extraction was carried out using Maximum Likelihood factor analysis (MLFA; Lawley and Maxwell, 1962). To determine the number of factors for the transformation of the factor matrix, both the parallel analysis method (Horn, 1965), which suggested seven factors, and the scree-plot of eigenvalues, which suggested six factors, were considered. Seven factors were then transformed applying the Promax method (Hendrickson and White, 1964) and interpreted as I: *Unusual Experiences* (UE), II: *Need for Order* (NfO^r), III: *Childlikeness* (Ch), IV: *Perceived Competence* (PC^r), V: *Trust* (Tr), VI: *Sensitivity* (Se), and VII: *Mysticism*.¹

In order to reduce the size of the questionnaire, items with a very low loading were removed, as well as items with a cross-loading above 0.3 or equal to the target loading. Some additional items were excluded after reliability analyses of the subscales derived in this process. The 46 remaining items were examined by means of a further factor analysis. Promax transformations of five (suggested by parallel analysis) and six (suggested by scree-plot) factors were carried out. The emerging factors were similar to those of the initial analysis and occurred in the following order: UE (I), NfO^r (II), Tr (III), PC^r (IV), Ch (V), and Se (VI). Although Factor VI included only two items, Rawlings (2001) decided to retain this subscale since "it was shown to have reasonable 'reliability', to be conceptually meaningful, and to represent aspects of the construct not covered by the other factors" (pp. 135). In contrast, Factor III remained in the BQ-Sh, but was not included in the calculation of *SumBound* for conceptual reasons.

Mean *SumBound* of the BQ-Sh in the total sample was 78.5 with a standard deviation of 15.4 (female: $M = 79.09$, $SD = 16.04$; male: $M = 76.87$, $SD = 13.42$). Cronbach's alpha

¹ In the present work, the NfO^r and PC^r scales are marked with a superscript "r" for *reversed* to emphasize that the concepts they measure are negatively correlated with *SumBound* and the entire scales are therefore inversely included in its calculation. As a result, higher scores on the NfO^r scale denote a *lower* need for order, and higher scores on the PC^r scale denote *lower* perceived competence.

for *SumBound* in the total sample was 0.74, indicating adequate internal consistency (Tsang et al., 2017), and for the subscales it was UE: 0.8, NfO^r: 0.79, Tr: 0.7, PC^r: 0.65, Ch: 0.69, and Se: 0.69. Both the scores of the full scale and the subscales were approximately normally distributed. Overall, there were no substantial correlations between the five subscales included in the calculation of *SumBound* BQ-Sh, with $|r| < 0.3$ in all cases. According to Rawlings (2001), this demonstrates the complexity of the boundary construct and the fact that individuals do not necessarily have weak boundaries with respect to all aspects identified. However, the correlations did suggest a clustering of the factors NfO^r and PC^r on the one hand and the factors UE, Ch and Se on the other. This finding was substantiated by a further factor analysis which resulted in two uncorrelated factors ($r = 0.07$) with loadings of 0.73 for UE, 0.62 for Ch and 0.67 for Se on the first factor as well as 0.81 for NfO^r and 0.76 for PC^r on the second factor. According to Rawlings (2001), these factors could reflect two broader personality dimensions which are hierarchically located between *SumBound* BQ-Sh and its five subscales. Finally, *SumBound* BQ had positive correlations with each of the BQ-Sh subscales² ($M_r = 0.45$, $SD_r = 0.17$) and further correlated substantially with *SumBound* BQ-Sh ($r = 0.88$), which demonstrates that the short version captures the boundary construct very well.

While the BQ18 clearly stands out for its brevity, the BQ-Sh with its effectively 40 items — when excluding the six items of the Tr subscale — is a very concise short version too. Only eight items appear in both short versions. In the sample of Rawlings (2001) the internal consistency was lower for the BQ18 ($\alpha = 0.66$) than for the BQ-Sh ($\alpha = 0.74$) and the two short versions correlated with $r = 0.77$. Since the BQ-Sh is based on a solid empirical foundation and introduces five fairly independent subscales that could contribute to further insights into the concept of mental boundaries, we used the BQ-Sh as a basis for the German adaptation.

5.1.3 German versions of the BQ

As mentioned in Chapter 2, we are aware that some authors in the field of dream research refer to a German translation of the BQ long version, which was translated by the Department of Psychology at the University of Zurich, probably at the end of the 1990s (e.g., Strauch and Meier, 1999; Funkhouser et al., 2001; Schredl, 2004). However, to our knowledge, this translation has not been published and there is no information available about the adaptation process.

We present a sample of mean *SumBound* values reported in the studies referring to the translation in order to assess whether they behave similarly as in the English-speaking population. For a student sample of 123 participants (68% female, $M_{\text{age}} = 28$ years; 32% male, $M_{\text{age}} = 30$ years) who completed the German translation, Strauch and Meier (1999) report a mean *SumBound* value of 302 ($SD = 38$). Funkhouser et al. (2001) report that in a sample of 61 Swiss participants over the age of 60 (69% female, $M_{\text{age}} = 71.7$ years, $SD_{\text{age}} = 5.6$; 31% male, $M_{\text{age}} = 72.2$ years, $SD_{\text{age}} = 5.4$), the mean *SumBound* value was 240 ($SD = 40$), and when retested after 26 weeks, it was 234 ($SD = 42$). This suggests that also in a German-speaking population, older subjects have thicker mental boundaries than younger subjects do.

Funkhouser et al. (2001) further report a significantly higher mean *SumBound* value for female subjects than for male subjects which is in line with findings for the English BQ long version (Hartmann, 1991) — however, this difference was significant only for the first test (female: $M = 260$, $SD = 36$; male: $M = 233$, $SD = 49$), but not for the retest.

Schredl (2004) reports a mean *SumBound* value of 290 ($SD = 43$) for a sample of 444 participants (85% female, 15% male; $M_{\text{age}} = 23.5$ years, $SD_{\text{age}} = 5.7$; mainly psychology students) and indicates an internal consistency of $r = 0.92$ for the total scale in this sample. This cor-

² The excluded Tr subscale had a low negative correlation with *SumBound* BQ ($r = -0.18$).

responds to the internal consistency of the English BQ long version — under the assumption that r denotes Cronbach's alpha.

Some authors reduce the aforementioned translation of the BQ long version to the BQ18 short version for their dream-related analyses (e.g. Schredl and Engelhardt, 2001; Funkhouser et al., 2008; Aumann et al., 2012). Since we did not use this version for our adaptation, we will not go into further detail about these studies.

The BQ-Sh is not mentioned in the context of the studies referring to the translation and to our knowledge there is no German adaptation available for this instrument.

As for the German translation of the BQ, it has not been published and is thus not accessible. Moreover, there is no information available about its adaptation process.

We therefore believe that a systematic German adaptation of the BQ-Sh, consisting of the translation and validation of the instrument, can offer an important contribution to future research on mental boundaries.

5.1.4 *Boundaries and the Big Five*

The boundary concept is to some extent related to the *Big Five* personality factors Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness.

For instance, McCrae (1994) compared the BQ results of 124 subjects (57 % female, 44 % male; age range 26 to 91 years) with results in the NEO Personality Inventory (NEO-PI; McCrae and Costa, 1985), which assesses the *Big Five*. The BQ was administered to the respondents seven years after the completion of the NEO-PI. He found significant correlations of *SumBound* with Neuroticism ($r = 0.32$), Extraversion ($r = 0.27$), and Openness ($r = 0.66$). Furthermore, in a joint factor analysis of various personality and cognitive ability measures for a subset of 85 respondents resulting in six factors reflecting the *Big Five* and general intelligence, the BQ *SumBound* score had a loading above 0.4 in absolute magnitude on the factors representing Neuroticism (0.44) and Openness (0.63). This indicates a connection between thin mental boundaries and these personality dimensions, which is promising with regard to our assumption that thinner boundaries are more likely to enable phonetic accommodation, as the few findings available in this context so far suggest that Openness and Neuroticism are predictive of such adaption (Yu et al., 2013; Lewandowski and Jilka, 2019).

Hartmann et al. (2001) refer to the remarkably high correlation of *SumBound* and the NEO-PI factor Openness to Experience. They observe that the Openness concept of the NEO-PI assesses only such qualities of Openness that are perceived positively, e.g., enjoying abstract ideas and speculations about the nature of the universe. In contrast, the Openness concept of the BQ also considers the more negative aspects of Openness, such as feeling overwhelmed or vulnerable, or becoming over-involved. They therefore consider the boundary construct as a more comprehensive approach to Openness that extends beyond the concept covered by the *Big Five*.

Schredl (2004) examined the relationship of the *Big Five*, assessed with the German version of the Revised NEO Personality Inventory (Ostendorf and Angleitner, 2004), and *SumBound*, assessed with the abovementioned German translation of the BQ, in a sample of 444 subjects (see Section 5.1.3) and found significant correlations for Neuroticism ($r = 0.33$), Openness ($r = 0.54$), Agreeableness ($r = 0.18$), and Conscientiousness ($r = -0.35$).

In the interest of completeness, we would also like to highlight the work of Schredl et al. (2009), who developed a new Boundary Personality Questionnaire (BPQ) that measures boundary strength while systematically excluding aspects of Neuroticism. The BPQ contains 20 items, some of which are taken from the original BQ and some of which are newly constructed, and is provided in English and German. For a sample of 59 psychology students (72 % female,

14% male, 14% unspecified; $M_{\text{age}} = 20.7$ years, $SD_{\text{age}} = 2.6$), Schredl et al. (2009) report the following significant correlations with *SumBound* BPQ: $r = 0.49$ for Openness and $r = -0.5$ for Conscientiousness.

We included the *Big Five* in the validation of the BQ-Sh adaptation presented here to further explore this relationship in our data (see Section 5.2.2).

5.2 METHODS

The adaptation of the BQ-Sh for German was prepared taking into consideration the *ITC Guidelines for Translating and Adapting Tests* (International Test Commission, 2018).

5.2.1 Translation

We used a double forward translation and reconciliation procedure in which the author of this work and a professional translator independently translated the 46 items of the BQ-Sh from English into German and then discussed and reconciled any discrepancies between the two translations. The result was assessed by an expert in the area of differential psychology and psychodiagnostics. All three translators involved in the process were native speakers of German and familiar with the target culture, namely German-speaking regions in Central Europe.

A small number of structural changes were made to the questionnaire. BQ-Sh item 13 “*I have dreams, daydreams, nightmares in which my body or someone else’s body is being stabbed, injured, or torn apart.*” was judged to be difficult to understand and disturbing. It was therefore excluded. Item 90 “*East is East and West is West, and never the twain shall meet. (Kipling)*” was excluded as well, because as a quotation it stood out from the set of items and its face validity was expected to be low, since the metaphorical statement obscured the actual meaning of the item. In the case of item 79 “*I cannot imagine living with or marrying a person of another race.*” and item 105 “*There are no sharp dividing lines between normal people, people with problems, and people who are considered psychotic or crazy.*”, the negation was considered to complicate the response and was therefore removed from the items. Consequently, item 79 is now reverse-scored (i.e., 0 = 4; 1 = 3; 3 = 1; 4 = 0), whereas item 105 is no longer reverse-scored. Item 33 “*Children and adults have a lot in common. They should give themselves a chance to be together without any strict roles.*” and item 108 “*I am a down-to-earth, no-nonsense kind of person.*” were found to contain two statements and were therefore split into two separate items. The result of this process is the German adaptation of the BQ-Sh (BQ-Sh-G), which contains 46 items (see Appendix I). Like the BQ-Sh, we kept the Tr subscale in the questionnaire for use in future research, although it is not included in the calculation of *SumBound*. The BQ-Sh-G can thus also be administered in a more compact 40-item version that does not include the Tr subscale.

5.2.2 Validation

5.2.2.1 Participants

A total of 341 native speakers of German completed both the 46-item BQ-Sh-G and the 60-item German version of the *NEO Five Factor Inventory* (NEO-FFI; Borkenau and Ostendorf, 2007). The questionnaires were made available online using the *PsyToolkit* platform (Stoet, 2010; Stoet, 2017) and could thus be individually answered by the participants. The participants were recruited from the author’s extended circle of acquaintances. They were not paid, but took part in a raffle of vouchers if they were interested. Six participants incorrectly answered a control

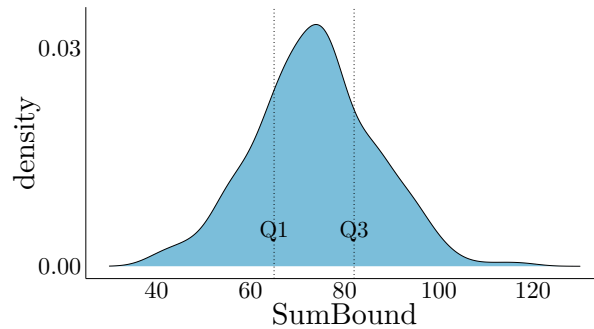


Figure 25: Distribution of the BQ-Sh-G total score (*SumBound*) in the sample. The lines indicate the 25 % (Q1) and 75 % (Q3) quantiles.

question built into the questionnaires and were excluded from the analysis. This left 335 (64 % female; 36 % male) participants for the validation of the BQ-Sh-G, who took a median time of 16 minutes to complete both questionnaires. The female participants had a mean age of 33 years (SD = 12 years; age range 17 to 71 years; skewness = 1.25; kurtosis = 0.71). The mean age of the male participants was 37 years (SD = 14 years; age range 17 to 79 years; skewness = 1.05; kurtosis = 0.32). Thirty-two participants were native speakers of another language besides German (e.g., Turkish, Kurdish, Albanian, Croatian, Greek, Russian, French, Italian, etc.). The educational level of the participants varied. The largest groups were those with a master’s degree ($n = 132$), a high school diploma ($n = 71$), a doctorate ($n = 53$), or a bachelor’s degree ($n = 48$). The remaining 31 participants held other degrees or were still attending school. The participants had professions or were students in a wide range of fields, such as business and administration, services, education and science, language studies, art and music, information technology, health and medicine, law, and journalism.

5.2.2.2 *Statistical analysis*

The statistical analysis is organized in two sections.

First, we explore the structure and internal consistency of the collected BQ-Sh-G data (Section 5.3.1) by giving means, standard deviations, and alpha coefficients for the total score and the subscales. A correlation matrix is provided as well. The respective BQ-Sh values from Rawlings (2001) are given for comparison. This is followed by a Maximum Likelihood factor analysis (MLFA) of the BQ-Sh-G data with Promax transformation of six factors — as suggested by the number of subscales in the BQ-Sh.

Second, we assess the relationship with the collected NEO-FFI data (Section 5.3.2) using a heterotrait–heteromethod correlation matrix of the BQ-Sh-G total score and subscales with the NEO-FFI subscales. Alpha coefficients are also reported for the latter to estimate their internal consistency.

5.3 RESULTS

5.3.1 *Structural validity*

Figure 25 gives an overview of how the total *SumBound* score is distributed among the participants. Theoretically achievable are scores from 0 to 160 (40 items \times 0 to 4 points). The *SumBound* scores in the total sample range from 39 to 118, with 50 % of them falling between 65 (Q1) and 82 (Q3) points.

Table 11: Means (M), standard deviations (SD) and alpha coefficients of *SumBound* and the subscales Unusual Experiences (UE), Need for Order (NfO^r), Trust (Tr), Perceived Competence (PC^r), Childlikeness (Ch), and Sensitivity (Se) for the BQ-Sh-G (top) and the BQ-Sh (bottom; from Rawlings, 2001). Sample size (n_s) and number of items per scale (n_i) are indicated.

	Sum Bound	UE	NfO ^r	Tr	PC ^r	Ch	Se
BQ-Sh-G	$n_i = 46$	$n_i = 11$	$n_i = 11$	$n_i = 6$	$n_i = 10$	$n_i = 6$	$n_i = 2$
$n_s = 335$							
M	73.61	13.94	23.22	11.33	16.73	15.19	4.54
SD	12.91	7.27	6.01	3.76	4.72	3.38	1.69
alpha	0.77	0.79	0.79	0.71	0.63	0.68	0.64
BQ-Sh	$n_i = 46$	$n_i = 12$	$n_i = 12$	$n_i = 6$	$n_i = 9$	$n_i = 5$	$n_i = 2$
$n_s = 300$							
M	78.5	17.09	26.19	13.15	16.98	12.71	5.54
SD	15.4	8.68	7.16	4.5	5.33	3.29	1.82
alpha	0.74	0.8	0.79	0.7	0.65	0.69	0.69

Table 11 gives the mean values with standard deviations as well as alpha coefficients of *SumBound* and the six subscales for the German BQ-Sh-G and the English BQ-Sh (Rawlings, 2001; see Section 5.1.2).

Mean *SumBound* of the BQ-Sh-G in the total sample is 73.61 with a standard deviation of 12.91. The female participants show a slightly higher mean *SumBound* ($M = 74.41$, $SD = 12.83$) than the male participants ($M = 72.21$, $SD = 12.99$). However, an unpaired two-samples t-test showed that this difference is not significant ($t(333) = 1.5$, $p = 0.13$).

The participants' age and *SumBound* show a very weak yet significant negative correlation with a Spearman's rank correlation coefficient of -0.18 ($p < 0.001$).

SumBound and most subscales show adequate internal consistency with Cronbach's alpha close to or above 0.7 (range: 0.68 to 0.79); only the coefficients of the PC^r and Se subscales are somewhat lower with $\alpha = 0.63$ and 0.64, respectively. Recall, however, that the Se subscale contains only two items.

Figure 26 shows correlation matrices for the BQ-Sh-G and the BQ-Sh. For both versions, the total *SumBound* score correlates strongly with the UE, NfO^r, and PC^r subscales (BQ-Sh-G: 0.65, 0.59, and 0.64; BQ-Sh: 0.74, 0.59, and 0.59, respectively) and moderately with the Ch subscale (BQ-Sh-G: 0.39; BQ-Sh: 0.37), whereas the correlations with the Tr and Se subscales are weak or statistically non-significant (BQ-Sh-G: 0.25 and 0.22; BQ-Sh: non-significant and 0.2, respectively).

The correlations between the individual subscales were mostly weak or statistically non-significant; only the subscales NfO^r and PC^r showed a moderate correlation with $r = 0.42$.

Table 12 shows the results of a MLFA with Promax transformation investigating the internal structure of the BQ-Sh-G. Based on the number of subscales resulting from the analysis in Rawlings (2001), we extracted six factors explaining 31% of the total variance. Most items have substantial loadings on their respective scales ($M = 0.5$).

Among the exceptions are items NfO^r 22 and PC^r 5, whose highest loadings are below 0.3. The same applies to the two items of the Se scale (8 and 18), which, in addition, do not constitute a factor of their own, but are associated with the UE scale. For the items of the PC^r scale a special pattern emerges, they form three clusters: items 5, 9, 11, 12, and 34 on the one hand,

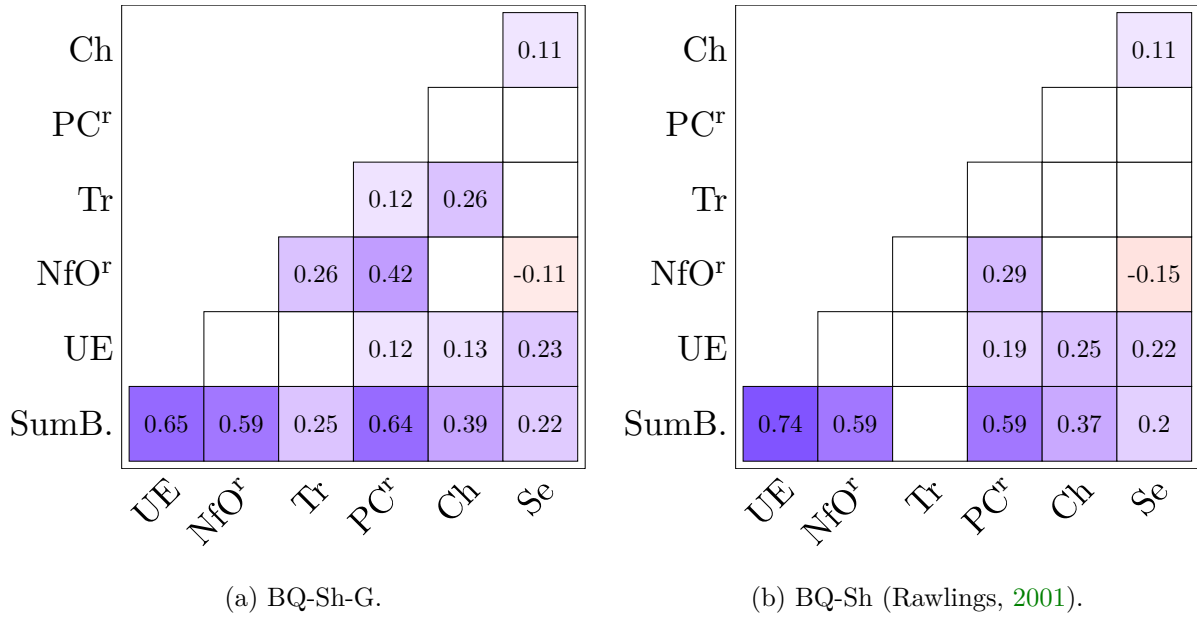


Figure 26: Correlation matrices of the BQ short versions' total scores (*SumBound*) and their subscales Unusual Experiences (UE), Need for Order (NfO^r), Trust (Tr), Perceived Competence (PC^r), Childlikeness (Ch), and Sensitivity (Se). Given correlations are significant with $\alpha < 0.05$ or smaller.

and items 4, 17, and 45 on the other hand, occupy two unassigned factors (we will refer to them as PC^r I and PC^r II, respectively), while items 32 and 41 are associated with the NfO^r scale.

Few cross-loadings with an absolute value above 0.2 occur, only three of which are even greater than 0.3. Item UE 19 loads highest on factor PC^r II (0.43), but still shows a substantial loading on the factor associated with the UE scale (0.38). Item PC^r 45 loads 0.37 on factor PC^r I and 0.51 on factor PC^r II. Finally, item PC^r 32 has both a loading of 0.43 on the factor associated with the NfO^r scale and a cross-loading of 0.31 on the PC^r II factor.

Table 12: MLFA of the BQ-Sh-G data with Promax transformation of six factors (F1–F6). Number of factors suggested by number of subscales resulting from analysis in Rawlings (2001). All loadings with an absolute value ≥ 0.2 are reported. The highest absolute loading for each item is highlighted in gray (if ≥ 0.3) or by a box (if < 0.3).

Scale	Item	F1	F2	F3	F4	F5	F6
NfO ^r	2	0.56					
	7	0.6					
	13	0.49				0.26	
	15	0.48				0.28	
	22	0.24					
	23	0.55					
	25	0.46					
	26	0.56					
	30	0.41					
	40	0.56					0.22

Scale	Item	F1	F2	F3	F4	F5	F6
	46	0.54					
UE	16		0.53				
	21		0.45				
	24		0.62				
	28		0.58				
	35		0.63		-0.26		
	36		0.5				
	38		0.56				
	39		0.33				-0.22
	43		0.44				0.21
	44		0.51				
	19	-0.21	0.38				0.43
Ch	6			0.43			
	10			0.32			
	14			0.56			
	20			0.77			
	27	0.28		0.31			
	42			0.94			
Tr	1				0.53		
	3	0.21			0.44		
	29				0.62		
	31				0.5		
	33		-0.22		0.52		
	37		0.27		0.53		
PC^r	5	0.21				0.28	
	9					0.55	
	11					0.54	
	12					0.5	
	34					0.31	
	4						0.32
	17						0.41
	45					0.37	0.51
	32	0.43					0.31
	41	0.45					
Se	8		0.27		-0.27		
	18		0.26	0.22			

Se	0.52				
Ch		0.29	0.17	0.17	
PC ^r	0.17	-0.14	0.13		-0.66
Tr	-0.17	0.45	0.2	0.26	
NfO ^r			0.29	0.16	-0.29
UE	0.41		0.21	-0.15	-0.16
SumB.	0.32		0.35		-0.47
	Neuro	Extra	Open	Agree	Consc.

Figure 27: Correlation matrix (heterotrait–heteromethod) of the BQ-Sh-G total score (*SumBound*) and subscales Unusual Experiences (UE), Need for Order (NfO^r), Trust (Tr), Perceived Competence (PC^r), Childlikeness (Ch), Sensitivity (Se) with the NEO-FFI subscales Neuroticism, Extraversion, Openness to Experience, Agreeableness, Conscientiousness. Given correlations are significant with $\alpha < 0.05$ or smaller.

5.3.2 Convergent and discriminant validity

Figure 27 shows a heterotrait–heteromethod correlation matrix of the BQ-Sh-G and the NEO-FFI data. The Neuroticism and Openness to Experience subscales of the NEO-FFI correlate positively with the BQ-Sh-G *SumBound* score: 0.32 and 0.35, respectively; the Conscientiousness subscale shows a negative correlation of -0.47 with *SumBound*. Some moderate to strong correlations between the NEO-FFI and BQ-Sh-G subscales emerge as well: Neuroticism with UE (0.41) and Se (0.52); Extraversion with Tr (0.45); Conscientiousness with PC^r (-0.66). There are also several weak correlations: Extraversion with Ch (0.29); Openness to Experience with Tr (0.2), NfO^r (0.29), and UE (0.21); Agreeableness with Tr (0.26); Conscientiousness with NfO^r (-0.29). In addition, a number of very weak, yet significant correlations appear.

The internal consistency of the five NEO-FFI subscales in our data is adequate, with alpha coefficients of 0.88 for Neuroticism, 0.82 for Extraversion, 0.69 for Openness to Experience, 0.76 for Agreeableness, and 0.85 for Conscientiousness.

5.4 DISCUSSION

5.4.1 *Structural validity*

The total *SumBound* score and the subscales of the German BQ-Sh-G exhibited means and standard deviations similar to those of the English BQ-Sh in Rawlings (2001). They also showed adequate internal consistency, as was the case for the BQ-Sh.

Contrary to the expectation that women would have thinner mental boundaries than men, there was no difference in boundary strength between female and male participants in our data. Participant age, on the other hand, significantly influenced the latter, with older participants exhibiting slightly thicker mental boundaries than younger participants ($r = -0.18$).

All six subscales of the BQ-Sh-G had significant positive correlations with *SumBound*, equal in strength to those of the BQ-Sh. The only exception was the Tr subscale, which was not included in the calculation of *SumBound* and also did not correlate with it in the BQ-Sh. In our data, however, the Tr subscale did not only correlate with *SumBound* — although it was still excluded from its calculation —, but also showed some statistically significant, weak, positive correlations with the NfO^r, PC^r, and Ch subscales.

While the correlations between the individual subscales of the BQ-Sh in Rawlings (2001) were all weak ($|r| < 0.3$) or statistically non-significant, we found a moderate correlation between the subscales NfO^r and PC^r in our data ($r = 0.42$). This is in line with the comparatively tighter connection between these two subscales suggested by Rawlings (2001).³

A MLFA with Promax transformation of six factors, as suggested by the number of factors in the BQ-Sh, yielded a good fit for most items with substantial factor loadings on their associated BQ-Sh-G subscale. In the following, we will discuss the content of the few items whose highest loading was below 0.3.

Item NfO^r 22 was met with some reservations during the translation process: “I cannot imagine living with or marrying a person of another race”. Since the term “race” is not suitable in this context in the target culture (namely German-speaking regions in Central Europe), the translation “ethnische Gruppe” (ethnic group) was chosen. Additionally, however, it can be assumed that the concept of interracial marriage holds a particular significance in the US-American culture that is not similarly present in the target culture of this adaptation, and item NfO^r 22 therefore probably does not carry the same weight in the German translation.

The other items with lower loadings are item PC^r 5 (“I keep my desk and worktable neat and well organized”), item Se 8 (“I am easily hurt”), and item Se 18 (“I am a very sensitive person”). In our opinion, the content of these items does not exhibit any particularities that would be expected to negatively influence the fit.

Despite their slightly lower loadings, we do not see any reason to exclude the discussed items from the questionnaire.

For the PC^r subscale, the MLFA suggested three clusters: PC^r I (items 5, 9, 11, 12, and 34), PC^r II (items 4, 17, and 45) and items 32 and 41, which were associated with the NfO^r subscale. A closer look at the content of these items suggests that PC^r I represents a more concrete form of Perceived Competence, e.g., “I keep my desk and worktable neat and well organized” and PC^r II with topics such as psychotherapy, memory of the past and sense of time, a more abstract form. Items 32 (“There are sharp dividing lines between normal people, people with problems, and people who are considered psychotic or crazy”) and 41 (“I know exactly what parts of the town I live in are safe and what parts are unsafe”) could be easily integrated into the Need for Order scale. Note, however, that item 32 shows a strong connection to PC^r II as well. We will

³ In Rawlings (2001), the correlation of NfO^r and PC^r is the highest between-subscale correlation as well. However, with $r = 0.29$, it is not moderate yet.

continue to refer to the original subscale PC^r as a single entity. Future work with the BQ-Sh-G may investigate whether the clustering suggested here solidifies.

Furthermore, the MLFA did not assign the two items of the Se subscale to a factor of their own, but associated them with the UE subscale. However, their content (see Se 8 and 18 above) does not integrate well with the Unusual Experiences scale. We will continue to refer to Se as a separate subscale.

Overall, we were able to demonstrate the structural validity for our German adaptation of the BQ-Sh to a satisfactory degree.

5.4.2 *Convergent and discriminant validity*

To further explore the relationship of mental boundaries and the *Big Five* discussed in [Section 5.1.4](#), we collected information regarding the latter using the NEO-FFI.

McCrae (1994) reported correlations of the original BQ's *SumBound* with Neuroticism and Openness to Experience: 0.32 and 0.66, respectively. He further substantiated the connection between these personality dimensions and thin mental boundaries in a factor analysis. In our data, the BQ-Sh-G *SumBound* correlated with Neuroticism (0.32) and Openness to Experience (0.35), as well. While McCrae (1994) additionally found a correlation with Extraversion (0.27) — which was, however, not reflected in the subsequent factor analysis —, in our data, Conscientiousness correlated negatively with *SumBound* (−0.47). The latter is in agreement with the results of Schredl (2004), who examined a German-speaking population with a translation of the BQ long version and found a correlation of −0.35 with Conscientiousness, as well as correlations of 0.33 with Neuroticism and 0.54 with Openness. In addition, there was a very weak correlation with Agreeableness ($r = 0.18$), which in turn does not occur in our data.

Two observations are noteworthy: First, in the BQ long versions, the correlation of *sumBound* and Openness is stronger than in the case of the BQ-Sh short version. We suppose that in the item reduction process for constructing the BQ-Sh, which was concerned with reliably capturing the concept of mental boundaries, the connection to Openness was weakened — as it was intentionally done with Neuroticism in the case of the BPQ (Schredl et al., 2009).

Second, in the German-speaking population, a negative correlation of *SumBound* and Conscientiousness occurs for both the BQ and the BQ-Sh that was not present in the English-speaking population. We see no reason why the two target populations should behave differently with respect to this relationship. Future work may establish whether this finding holds.

We observe that the correlation of Neuroticism and *SumBound* in our data is mainly attributable to the Unusual Experiences and Sensitivity subscales. The correlation of Openness to Experience is distributed relatively evenly across all subscales, except Se. The negative correlation of Conscientiousness is weakly influenced by the Need for Order subscale and strongly by the Perceived Competence subscale.

With respect to the convergent validity, there seems to be a certain relationship between the thickness of mental boundaries and the *Big Five* personality dimensions, which manifests itself repeatedly: thinner boundaries tend to be associated with increased Neuroticism and Openness to Experience. This finding supports our hypothesis that thinner boundaries facilitate phonetic accommodation, since previous evidence from the literature suggests that Openness and Neuroticism favor such accommodation (Yu et al., 2013; Lewandowski and Jilka, 2019).

Correlations with Extraversion or Agreeableness emerge only occasionally — if so, they are weak and unstable. In our data these two personality factors did not correlate with *SumBound*.

Regarding the relationship with Conscientiousness, the situation is more ambiguous. To our knowledge, a negative correlation with the latter has so far only been found for a German-speaking population — and accordingly also occurred in our data.

The observed correlations with the subscales of the NEO-FFI are only weak to moderate, which demonstrates that the BQ-Sh-G possesses discriminant validity, too. This is also reflected in the comment by Hartmann et al. (2001) on the relationship between *SumBound* and Openness to Experience, mentioned above, stating that the Openness concept of the BQ encompasses negative aspects as well and is thus a more comprehensive approach to Openness that goes beyond the concept captured by the *Big Five*.

Overall, we were able to demonstrate convergent and discriminant validity between our German adaptation of the BQ-Sh and the *Big Five* to a satisfactory degree.⁴

5.5 USE CASE: MENTAL BOUNDARIES AND PHONETIC ACCOMMODATION

To start exploring whether mental boundaries can be used to predict phonetic accommodation, we examine their effect on the *Mirabella* data (see Chapter 4).

Initially, we hypothesized that thinner mental boundaries would favor adaptation to an interlocutor because they stand for permeability, place less emphasis on group membership and separation from others, and allow for the absorption of external input (Hartmann et al., 2001; Harrison and Singer, 2013). This assumption was supported by the fact that a correlation of thin boundaries and the *Big Five* personality traits Neuroticism and Openness has been demonstrated (McCrae, 1994; Hartmann et al., 2001), and the same personality traits have also been found to facilitate phonetic accommodation — in the few studies conducted on this subject (Yu et al., 2013; Lewandowski and Jilka, 2019).

When validating our German adaptation of the Boundary Questionnaire’s empirically-derived short version, we observed a correlation of thin boundaries with Neuroticism and Openness as well. However, we additionally found a correlation of thick boundaries with Conscientiousness, which was even slightly stronger (see Figure 27).

We also tested the influence of the *Big Five* on the *Mirabella* data and found that more neurotic speakers were more likely to converge to their interlocutor’s question intonation and a tendency for more conscientious speakers being more likely to adopt the variant of the allophonic contrast [ɪç]/[ik] from their interlocutor. This could be due to the different reasons for accommodation that these personality types entail, e.g., avoiding stress vs. performing well (see Sections 4.4.2.1 and 4.4.2.2), which may promote convergence with respect to different types of phonetic features.

Following these findings, we adjust our expectations to assume that thinner mental boundaries will favor the adaptation to question intonation and thicker mental boundaries will lead to the [ɪç]/[ik] contrast being adopted more readily.

5.5.1 Data

The 42 native speakers of German who participated in the WOz experiment with *Mirabella* (combined natural and synthetic group) are a subset of the 335 participants whose data were used to validate the BQ-Sh-G. We will summarize the boundary data for this subset hereafter to determine whether it constitutes a representative sample.

The 11 native speakers of French who interacted with *Mirabella* also completed the BQ-Sh-G — for lack of a French version. Thanks to their excellent command of German, they were able

⁴ For the purpose of this study, we examined convergent and discriminant validity only with respect to the *Big Five* personality traits, which are also relevant in the context of phonetic accommodation. For further validation of the BQ-Sh-G, future studies should include other established concepts that have possible intersections with the boundary concept, such as the Minnesota Multiphasic Personality Inventory (Hathaway and McKinley, 1943) or the Myers–Briggs Type Indicator (Briggs-Myers and Myers, 1995) — cf. Hartmann et al., 2001.

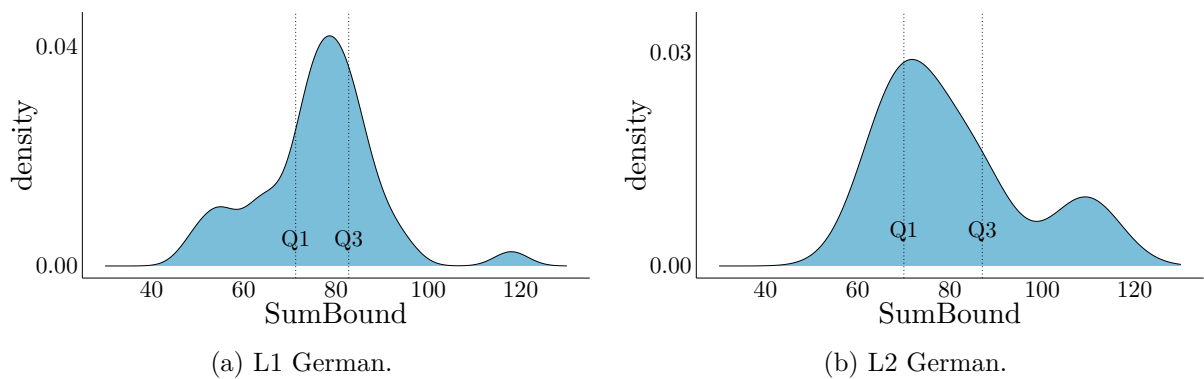


Figure 28: Distribution of the BQ-Sh-G total score (*SumBound*) in the two speaker groups. The lines indicate the 25 % (Q1) and 75 % (Q3) quantiles.

to do so without difficulty. We will summarize their boundary data as well, keeping in mind that the questionnaire was not validated for this group.

Figure 28 gives an overview of how the total *SumBound* score is distributed among the L1 and L2 German participants. In the L1 German sample, the scores range from 49 to 118, with 50 % of them falling between 71 (Q1) and 83 (Q3). The L2 German speakers have *SumBound* scores ranging from 63 to 112, half of which are within 70 (Q1) and 87 (Q3) points.

This demonstrates a limitation of the present analysis, which we already mentioned in relation to the *Big Five*: Since the participants were not selected for their boundary composition, we have fewer cases with an extreme expression (particularly thick/thin) and more participants exhibiting medium scores.

Table 13 gives the mean values with standard deviations as well as alpha coefficients of the total *SumBound* score and the six BQ-Sh-G subscales for the L1 and L2 German speakers.

The mean *SumBound* in the L1 German sample of 75.95 (SD = 12.82) is similar to that of the validation sample (M = 73.61, SD = 12.91). As with the latter, the female L1 German speakers have a slightly higher mean *SumBound* (M = 76.1, SD = 13.72) than the male L1 German speakers (M = 75.55, SD = 10.41). Yet again, an unpaired two-samples t-test showed that this difference is not significant ($t(40) = 0.12, p = 0.9$).

Since the participants' age is relatively homogeneous in this sample, we do not test for an effect of age on *sumBound*.

SumBound and most subscales show adequate internal consistency with Cronbach's alpha close to or above 0.7 (range: 0.69 to 0.79); only the coefficient of the PC^r subscale is somewhat lower with $\alpha = 0.58$. In the validation sample, the PC^r subscale had the lowest alpha coefficient as well (0.63).

The L2 German group shows a higher mean *SumBound* and more variation (M = 80.91, SD = 16.14). The difference in mean *SumBound* between female and male participants is not significant (female: M = 86, SD = 22.27; male: M = 76.67, SD = 8.76; $t(9) = 0.95, p = 0.37$). *SumBound* and the subscales UE and Se exhibit strong internal consistency in the L2 German group with alpha coefficients above 0.8. The NfO^r scale falls in the lowest acceptable range with $\alpha = 0.61$. The PC^r, Tr, and Ch scales exhibit very poor internal consistency with alpha coefficients of 0.36, 0.14, and 0.03, respectively.

Figure 29 shows BQ-Sh-G correlation matrices for the L1 and L2 German participants. As in the validation sample, *SumBound* correlates strongly with the UE, NfO^r, and PC^r subscales (L1 German: 0.66, 0.66, and 0.73; L2 German: 0.75, 0.72, and 0.73, respectively). For the German L1 group, the Ch scale also shows a weak correlation of 0.32 with *SumBound*, which is consistent with the finding in the validation sample (0.39). In contrast, the L2 German group

Table 13: Means (M), standard deviations (SD) and alpha coefficients of the BQ-Sh-G total *SumBound* score and its subscales Unusual Experiences (UE), Need for Order (NfO^r), Trust (Tr), Perceived Competence (PC^r), Childlikeness (Ch), and Sensitivity (Se) for the L1 German speakers (top) and the L2 German speakers (bottom). Sample size (n_s) and number of items per scale (n_i) are indicated.

	Sum Bound	UE	NfO ^r	Tr	PC ^r	Ch	Se
L1 German	$n_i = 46$	$n_i = 11$	$n_i = 11$	$n_i = 6$	$n_i = 10$	$n_i = 6$	$n_i = 2$
$n_s = 42$							
M	75.95	14.57	23.67	10.86	18.12	15.67	3.93
SD	12.82	6.86	5.88	3.83	4.43	3.27	1.69
alpha	0.78	0.76	0.79	0.72	0.58	0.69	0.7
L2 German							
$n_s = 11$							
M	80.91	18.82	22.36	9.82	18.09	16.64	5.00
SD	16.14	9.22	4.92	2.79	4.44	2.29	2.65
alpha	0.81	0.84	0.61	0.14	0.36	0.03	0.86

exhibits an unexpectedly strong correlation of *SumBound* with the Se scale (0.64). For the German L1 group, the correlations between the individual subscales are all weak or statistically non-significant. The moderate correlation between the NfO^r and PC^r subscales of 0.42 in the validation sample compares to 0.33 in the L1 German data and an unexpectedly high 0.74 in the L2 German data — the only significant correlation between subscales in that group.

Figure 30 shows heterotrait–heteromethod correlation matrices of the BQ-Sh-G and the *Big Five* data for both speaker groups. The *Big Five* were assessed with the German NEO-FFI (Borkenau and Ostendorf, 2007) in the case of the L1 German speakers and the French Big Five Inventory (BFI; Plaisant et al., 2005; Plaisant et al., 2010) was used for the L2 German, i.e., L1 French speakers.

As in the validation sample, in the German L1 group *SumBound* also correlates positively with Neuroticism (0.39) and Openness to Experience (0.4), and negatively with Conscientiousness (−0.54). The L2 German group shows more extreme correlations overall. For *SumBound* in particular, they are $r = 0.81$ with Neuroticism and $r = -0.87$ with Conscientiousness — none, however, with Openness.

5.5.2 Analysis and results

Similarly to the analysis regarding the influence of the *Big Five* on accommodation (see Section 4.3.6), we use the entire L1 German data set (combined natural and synthetic group) to refit the statistical models for the three phonetic features under investigation: question intonation, word ending ⟨-ig⟩, and long vowel ⟨-ä-⟩, while including BOUNDARY STRENGTH, i.e., *SumBound*, as a continuous factor.

Figure 31 visualizes the three resulting models. Neither for question intonation (31a and b), nor for vowel quality (31e) did the inclusion of the factor BOUNDARY STRENGTH improve the fit of the model.

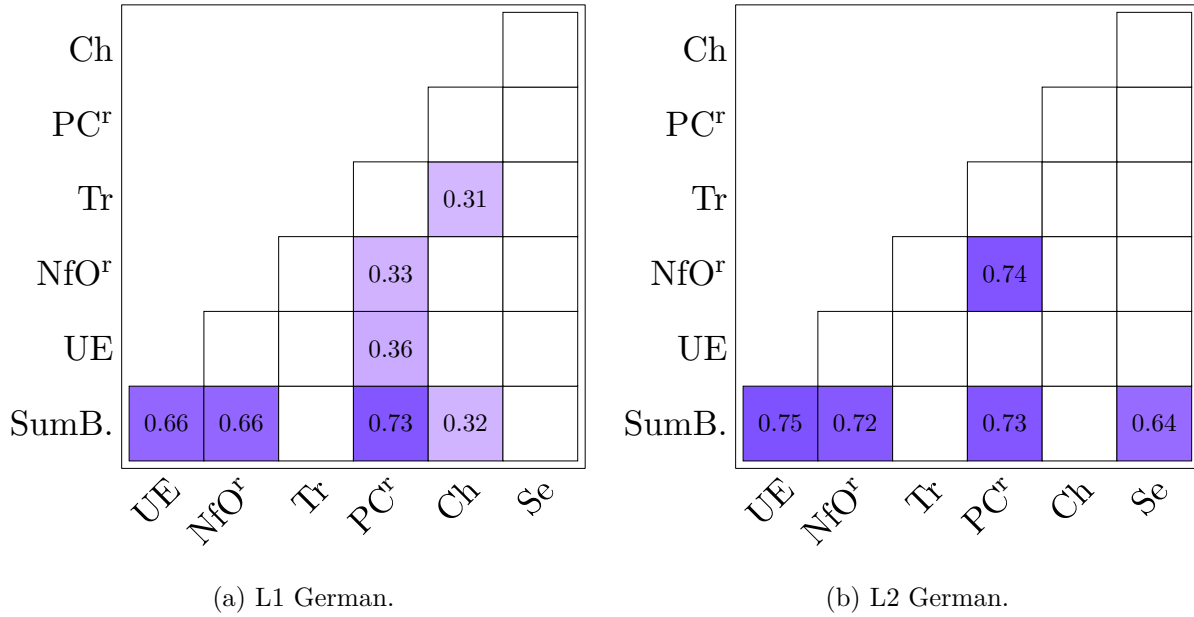


Figure 29: Correlation matrices of *SumBound* and the subscales Unusual Experiences (UE), Need for Order (NfO^r), Trust (Tr), Perceived Competence (PC^r), Childlikeness (Ch), and Sensitivity (Se). Given correlations are significant with $\alpha < 0.05$ or smaller.

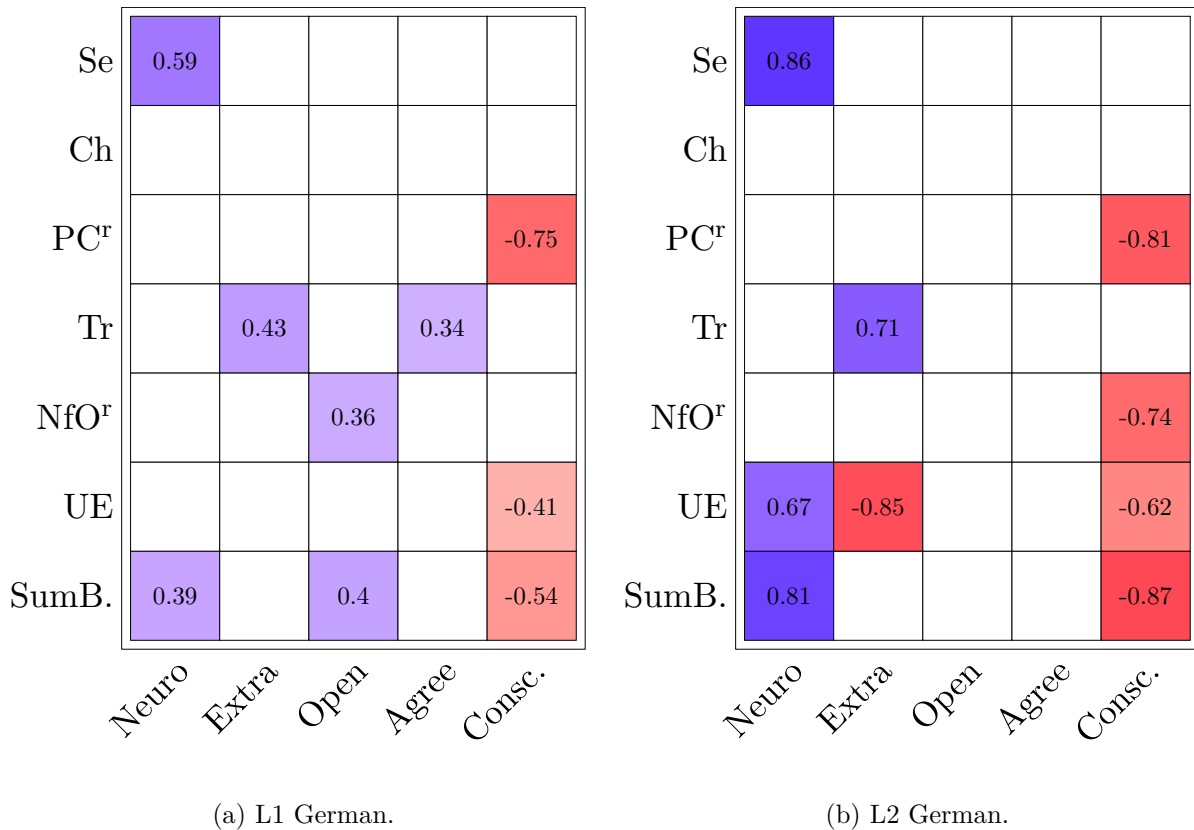


Figure 30: Correlation matrices (heterotrait-heteromethod) of the BQ-Sh-G total score (*SumBound*) and subscales Unusual Experiences (UE), Need for Order (NfO^r), Trust (Tr), Perceived Competence (PC^r), Childlikeness (Ch), Sensitivity (Se) with the *Big Five* Neuroticism, Extraversion, Openness to Experience, Agreeableness, Conscientiousness. Given correlations are significant with $\alpha < 0.05$ or smaller.

In the case of the [ɪç]/[ɪk] contrast, however, including the factor BOUNDARY STRENGTH did improve the fit of the model, suggesting that participants with thicker mental boundaries were more likely to converge to Mirabella’s realization of ⟨-ig⟩ (31c and d). Yet, the interaction with TASK was not significant (Estimate (log-odds) = 0.03, SE = 0.02, $z = 1.92$, $p = 0.06$). The model showed main effects of TASK (Estimate (log-odds) = -3.07 , SE = 1.16, $z = -2.66$, $p < 0.01$) — demonstrating the increase of *same* variants in the map task — and BOUNDARY STRENGTH (Estimate (log-odds) = -0.06 , SE = 0.02, $z = -2.55$, $p < 0.05$) — indicating an a priori higher occurrence of *same* variants in participants with thicker mental boundaries. The model includes random intercepts for USER and ITEM, as well as random slopes for TASK by USER.

Given the small number of speakers in the L2 German group, we consider only individual cases. That is, we examine the performance in terms of intonation accommodation and realization of the word ending ⟨-ig⟩ for the speakers with the thinnest and thickest mental boundaries in the group (35 % in each case).

Of the four speakers who converged on question intonation, one has relatively thin boundaries (substantial convergence) and another relatively thick boundaries (moderate convergence). Among the eight speakers who converged to Mirabella’s realization of ⟨-ig⟩, two speakers have relatively thin and two have relatively thick boundaries (1 × substantial and 1 × moderate convergence, respectively).

5.5.3 Discussion

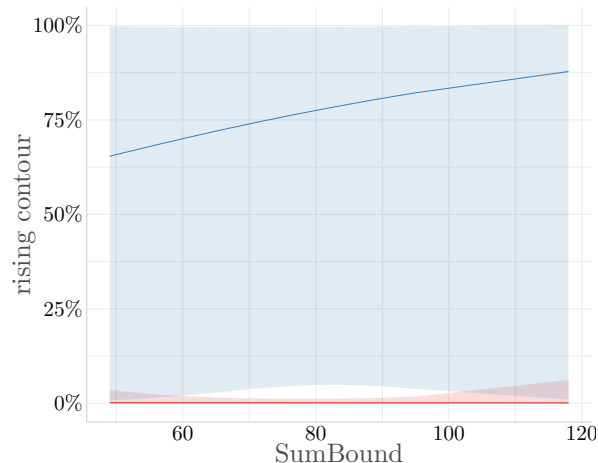
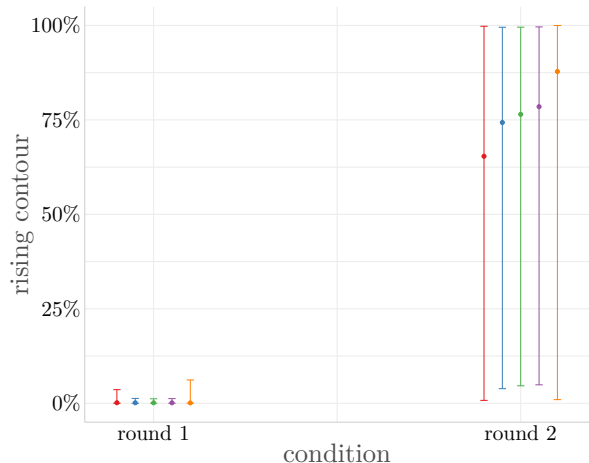
We demonstrated that the L1 German speakers participating in the WOz experiment can be considered a representative sample in terms of mental boundaries, since key characteristics such as values of central tendency, variability, and internal consistency, as well as correlations within the boundary data and between the boundary data and the *Big Five* were consistent with the validation sample. The L2 German speakers exhibit some dissimilarities from the validation sample, in that they have higher boundary values, more extreme correlations overall, and problems with the internal consistency of some subscales. This could be due to the small sample size, but also to the fact that the BQ-Sh-G is simply not intended nor validated for them — which underlines the importance of adapting and validating such questionnaires for different target populations. However, some key characteristics of the boundary data are present for the L2 German speakers as well. For example, as expected, there is a positive correlation of *SumBound* with Neuroticism and a negative correlation with Conscientiousness.

Concerning the influence of mental boundaries on phonetic accommodation, our prediction that thicker boundaries would favor convergence with respect to the binary [ɪç]/[ɪk] contrast was corroborated for the L1 German speakers: Including the interaction of the boundary strength with the task in the statistical model yielded a better representation of the data. However, the interaction was not significant in the final model.

Regarding the question intonation, we expected that thinner boundaries would promote convergence, but we found no evidence for this in our data. We observed only a slight tendency of the data in the predicted direction, which cannot be reasonably interpreted because of the large variation.

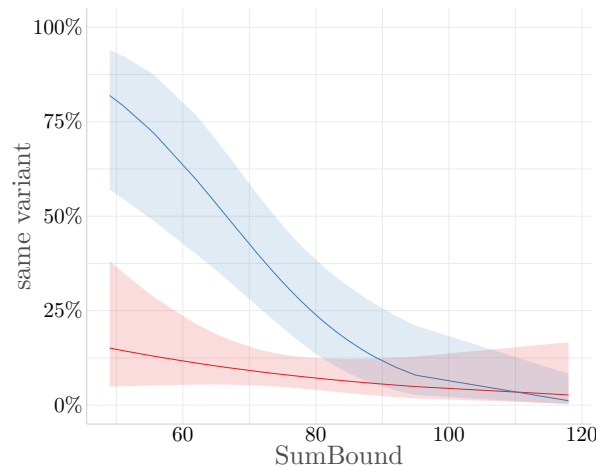
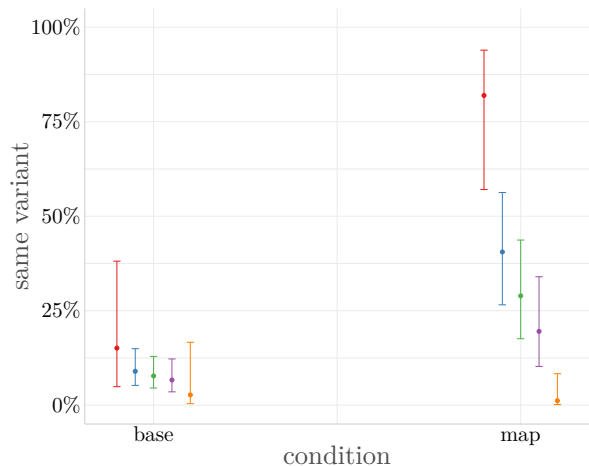
For completeness, we also tested the vowel quality data where no accommodation effect had occurred at the group level — boundary strength did not inform this case further.

For the L2 German speakers, we found no connection of mental boundaries and phonetic accommodation when considering the participants with the thickest and thinnest boundaries. Note that this was a small sample and the boundary questionnaire was not validated for this group.



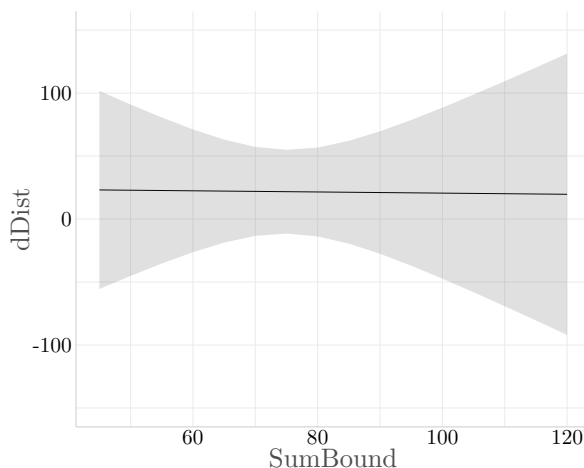
(a) Question intonation: percentage of *rising* contours in the two conditions for different levels of *SumBound*.

(b) Question intonation: percentage of *rising* contours over *SumBound* for the two conditions.



(c) Word ending <-ig>: percentage of *same* variants in the two conditions for different levels of *SumBound*.

(d) Word ending <-ig>: percentage of *same* variants over *SumBound* for the two conditions.



(e) Long vowel <-ä-): difference in Euclidean distance (*dDist*) — which compares the conditions *base* and *map* — over *SumBound*.

Figure 31: Influence of boundary strength on the examined phonetic features for the L1 German speakers. Figures (a) and (c) show predictions for the **minimum value** (49; thick), **lower quartile** (71), **median quartile** (77), **upper quartile** (83), and the **maximum value** (118; thin) of *SumBound* per experimental condition. Figures (b), (d), and (e) show continuous predictions — the first two separately for **condition 1** (round 1/*base*) and **condition 2** (round 2/*map*). Each figure displays the 95% confidence interval.

We would like to reiterate that the participants of the WOz study were not selected based on their boundary composition. Testing a group of participants in which extreme expressions of mental boundaries (particularly thick/thin) also occur in sufficient numbers would be a relevant extension for the present work. We assume that existing effects would be more evident in such a group.

In the case of mental boundaries, we are considering a very broad concept that combines many personality dimensions. While it covers Neuroticism and Openness at one end of the continuum and Conscientiousness at the other to some extent, it cannot be regarded as a substitute for the separate evaluation of the *Big Five*. In the present case, the latter have in fact exhibited higher predictive power for convergence with respect to the specific phonetic features (see [Section 4.3.6](#)). We evaluate the small amount of information provided by the boundary construct for [ɪç]/[ɪk] as shining through of the *Big Five* factor Conscientiousness due to its correlation with the latter.

Is there, however, additional explanatory power for accommodating behavior in the boundary construct? The present analysis represents but a first use case for the BQ-Sh-G in accommodation research. We hypothesize that boundary strength may predict a more general disposition toward accommodation that is possibly eclipsed when considering individual phonetic phenomena.

In [Section 2.7](#), we mentioned the assumption made in the literature that thinner mental boundaries may be conducive to the acquisition of native-like pronunciation in a foreign language (e.g., Guiora et al., 1972; Więckowska, 2011; Baran-Łucarz, 2012). For this scenario, it seems reasonable to consider a holistic view of native-like pronunciation, since performance with respect to an isolated phonetic feature certainly contributes to the overall impression, but does not necessarily determine it.

While we observe considerable variability regarding accommodation behavior with respect to various phonetic features in the literature as well as in our own data, we nevertheless assume that there are speakers who have a greater tendency to converge to their conversational partner on a holistic level than others. This is where the broad concept of mental boundaries could complement the previously assumed influencing factors for phonetic accommodation. In this case, we would again assume that thinner mental boundaries are more likely to facilitate the adaptation to an interlocutor.

Future studies of accommodation behavior — especially at a more holistic level — should consider boundary strength as a possible influencing factor to shed further light on its predictive potential.

5.6 CONCLUSION

We introduced the German adaptation of the Boundary Questionnaire’s (BQ; Hartmann, 1991) empirically-derived short version (BQ-Sh; Rawlings, 2001). The English BQ-Sh was professionally translated into German, a small number of structural changes were made where deemed necessary, and the resulting BQ-Sh-G was validated with a sample of 335 participants.

We demonstrated the structural validity of the German adaptation. Our data exhibited values of central tendency, variability, and internal consistency similar to those of the BQ-Sh. A tendency for thicker mental boundaries in older participants was observed ($r = -0.18$). Female and male participants did not differ with respect to boundary strength. When extracting six factors from the BQ-Sh-G data — as suggested by the number of subscales in the BQ-Sh — most items showed substantial loadings on their associated subscale.

Furthermore, we explored the relationship of the BQ-Sh-G and the *Big Five* personality traits. We found positive correlations of the total *SumBound* score with Neuroticism ($r = 0.32$) and Openness to Experience ($r = 0.35$), as well as a negative correlation with Conscientiousness

($r = -0.47$), which is in line with previous literature. We conclude that convergent validity with the *Big Five* is observed where it was expected. However, in light of the relatively weak correlations, the discriminant validity of the BQ-Sh-G is given as well.

The BQ-Sh-G was developed for the purpose of application in phonetic research, but is not limited to it. We encourage to administer the BQ-Sh-G in a more compact 40-item version that does not include the Trust subscale, since following Rawlings (2001) the latter is not included in the calculation of the total score *SumBound*.

As a first use case, we presented the application of the BQ-Sh-G in the context of the WOz experiment to predict phonetic accommodation.

We found a tendency for thicker mental boundaries to promote convergence to the computer agent's version of the allophonic contrast [ɪç]/[ɪk] for L1 German speakers. However, this turned out not to be significant in the statistical model and, moreover, is presumably due to the correlation of boundary strength with the *Big Five* factor Conscientiousness.

Boundary strength did not inform accommodation to question intonation and vowel quality.

We believe that the predictive potential of the boundary construct is more applicable to a holistic approach to phonetic accommodation, with thinner mental boundaries indicating a greater disposition to accommodate. This deserves to be investigated in future research.

6 GENERAL DISCUSSION

The overarching goal of this dissertation was to investigate the accommodation behavior of human interlocutors in interaction with a computer agent, while focusing on locally anchored phonetic phenomena. We approached this goal in three main steps. First, the shadowing experiment provided a detailed analysis of accommodation to natural and synthetic voices in terms of various phonetic features, highlighting the interface of spoken human-computer interaction (HCI), namely the speech itself. Second, the Wizard-of-Oz (WOz) experiment featured a dynamic interaction with the virtual language learning tutor Mirabella that incorporated the users' belief that they are actually communicating with a computer, which we believe to be a core component of HCI. Finally, the chapter about mental boundaries paved the way for further in-depth investigation of individual differences between speakers in accommodation behavior.

SHADOWING EXPERIMENT In the shadowing experiment, our main questions were whether the locally anchored phonetic features under investigation would be adopted from longer utterances, and whether this would be true to the same extent for natural as for synthetic voices.

The findings on accommodation behavior from shadowing experiments using single, often mono- or bisyllabic (non-)words as stimuli (e.g., Babel, 2012; Dufour and Nguyen, 2013; Mitterer and Müsseler, 2013) are informative with respect to the adoptability of the phonetic feature in question, but, in our opinion, unlikely to be predictive of the behavior in dynamic conversational interaction. In the case of such short utterances, the shadower's attention is directed to a well-circumscribed acoustic event, the phonetic details of which may be fully captured and then reproduced more easily than in the case of longer utterances, which require a broader focus and induce higher cognitive load. For example, the finding of Pardo et al. (2018) that there is only a moderate relationship of accommodation behavior assessed by perceptual similarity between speech shadowing and conversational interaction, is based on a comparison of mono- and bisyllabic shadowed utterances with utterances from a conversational map task. It could be worthwhile to include longer shadowed utterances in such a comparison, since we think that their results may be more easily transferable to actual dialog.

In our experiment, we used short German sentences as stimuli and demonstrated that speakers converge on locally anchored phonetic features, i.e., segment-level variation as well as variation of local prosody, even when these are embedded in such longer utterances and are therefore less salient targets for accommodation. With the word-based temporal structure and distribution of spectral energy, we also examined the level of global similarity in our data. We demonstrated that convergence still occurs for the distribution of spectral energy when convergence with respect to the temporal structure is already accounted for. These two aspects have previously been studied in the context of accommodation as a combined feature only (Lewandowski, 2012; Lewandowski and Jilka, 2019).

While the higher-level features of word-based global similarity were successful triggers for convergence overall, the locally anchored features elicited varying degrees of convergence in our data, which we assume to be due to several contributing factors including accessibility and plausibility. First, the accessibility of the factor, both in terms of perception (i.e., salience) and production (i.e., selective realizability), certainly matters, such that binary features (e.g., [ɾ]/[ɪk]) may be more successful triggers for convergence than gradual features that are more difficult to isolate (e.g., [ɛː]/[eː], pitch accents). In addition, the plausibility of the variant plays

a role, so that more unusual variants may generally be adopted to a lesser extent (e.g., [ŋ]/[ən]). Furthermore, speakers differ greatly in their individual accommodation behavior.

As mentioned above, the second question this experiment examined was whether the observed effects would occur for both natural and synthetic voices. But how do we imagine a synthetic voice in the context of modern HCI? With commercial systems using state-of-the-art speech synthesis methods based on deep learning (e.g., van den Oord et al., 2016; Wang et al., 2017; Shen et al., 2018) we are used to a quality of synthetic speech that is hardly distinguishable from natural speech. However, the end-to-end nature of this approach to speech synthesis, does not allow for selective modification of the speech signal at the segmental level, as necessary for the purpose of the present study. Thus, there are two possibilities: On the one hand, experiments can involve natural voices. When presented in an HCI context, it is reasonable to assume that study participants will accept them as natural-sounding synthetic voices, since they are used to such quality from commercial systems. On the other hand, synthetic voices can be used that are based on synthesis methods permitting manipulation at the segmental level. These may then sound more synthetic than study participants are used to nowadays.

We opted for both ways by examining natural voices as well as diphone- and Hidden Markov model (HMM)-based synthetic voices in the shadowing experiment. In this particular experiment, the HCI context is established only through the participants' perception of the voices. While the natural voices were perceived as being natural, the synthetic voices were actually perceived as being synthetic due to their quality, which turned a limitation of the diphone- and HMM-based methods into a desirable attribute for this study.

Overall, the observed convergence effects occurred for both natural and synthetic voices, although they were partly attenuated for the latter.

Since the synthetic voices contained not only the segmental variations but also various deviations due to synthesis artifacts, it would have been conceivable that the participants would distance themselves from the phonetic form they heard and rather reproduce the content of what they heard in their usual way of speaking. Instead, it seems to be the case that accommodation is not prevented even in the presence of synthetic-sounding voices, but listener-speakers can still extract appropriate accommodation targets.

We believe that state-of-the-art deep-learning-based synthetic voices — were they capable to produce the targeted segmental variations — would elicit an accommodation effect equivalent to that of the natural voices, since the two are almost indistinguishable to the listener, as mentioned above.

WIZARD-OF-OZ EXPERIMENT The main question of the WOz experiment was whether the participants would adopt locally anchored phonetic features from the simulated virtual language learning tutor Mirabella in an interactive spoken exchange with her. Mirabella was designed as a female agent in order to match a real-life interaction with a spoken dialog system (SDS) as closely as possible, since most commercial systems feature a female voice exclusively, or at least use it as a default.

The results of the shadowing experiment motivated the use and direct comparison of natural as well as synthetic speech in the subsequent WOz experiment. We hence created two versions of Mirabella. The version featuring natural speech represents a SDS of a quality that users would expect nowadays. In the WOz experiment, as far as we can tell, the participants perceived the natural voice as a natural-sounding synthetic voice, since it was presented in the corresponding context. The second version of Mirabella used HMM-based synthetic speech, which, we assume, led to a reinforcement of the perceived HCI-context on the part of the participants.

The wizard, i.e., the experimenter, had a limited number of pre-produced stimuli at her disposal for interacting with the participants, and some of them were used several times within

a conversation, e.g., “Ok?”, “Versuch’s nochmal!” (*Try again!*), or “Sehr gut!” (*Very good!*). The fact that such recurring utterances of the same content did not differ at all in their phonetic form enhanced the impression of communicating with a computer agent, both for the synthetic and the natural version of Mirabella. Thus, again, a limitation of the employed method — here WOz — translated into a desirable effect for the study.

The experiment was embedded in a computer-assisted language learning (CALL) scenario and although Mirabella sometimes had more information than the participants, e.g., she knew the hidden items in the map task, she presented herself less as a teacher and more as a cooperative peer. While she sometimes did not understand a statement, asked participants to repeat it, or provided a hint on how to proceed, she did not actually correct them. The participants engaged in the interaction and conveyed the impression of seeing Mirabella as a social actor.

In general, the participants used only the utterances required for the tasks. However, there were also instances of witty remarks (Example 1) or expressions of politeness (Example 2) by participants (P) towards Mirabella (M):

- (1) P: Wo hat sich der Hase versteckt?
Where did the rabbit hide?
 M: Ich bin dran!
It’s my turn!
 P: **Ja dann hau mal raus, Mirabella!**
Well then, shoot, Mirabella!
 M: Wo hat sich der Löwe versteckt?
Where did the lion hide?
 P: [lacht] Der Löwe hat sich in Haus Nummer 9 versteckt.
[laughs] The lion hid in house number 9.
- (2) M: Es hat mir viel Spaß gemacht, mit dir zu arbeiten!
I had a lot of fun working with you!
 P: **Ebenso!**
Likewise!
 M: Vielen Dank, dass du teilgenommen hast!
Thank you so much for participating!
 P: **Gern geschehen!**
You’re welcome!
 M: [überlappend] Du kannst jetzt die Kopfhörer absetzen und die Kabine verlassen. Bis bald!
[overlapping] You may now take off the headphones and leave the booth. See you soon!
 P: **Tschau!**
Bye!

In conclusion, we successfully induced a dynamic exchange between the participants and the simulated SDS Mirabella.

For native speakers of the target language German, we were able to demonstrate a very similar degree of phonetic convergence to both the natural and synthetic versions of Mirabella at the level of local prosody and segmental pronunciation. This shows that accommodating behavior in users of a SDS is indeed triggered by locally anchored phonetic features, even if the users are continuously reminded of the interlocutor’s machine-nature by a synthetic-sounding voice.

Again, the different phonetic features yielded different effects. While the gradual variation in vowel quality caused accommodation only for a few individual speakers, the binary allo-

phonic variation [ɪç]/[ɪk] and the locally anchored prosodic features were successful triggers for convergence at the group level.

We related some of the speaker-specific differences in accommodation behavior present in the WOz data to the *Big Five* personality factors Neuroticism and Conscientiousness, with the former affecting question intonation and the latter showing a tendency of affecting the binary allophonic variation [ɪç]/[ɪk]. To date, there are not enough studies investigating this aspect in order to assess whether indeed different personality factors influence different types of phonetic features with respect to accommodation. It is, however, quite conceivable that, in addition to a general disposition to adapt to an interlocutor, which could well be favored by a certain personality structure, the specific communicative function of a phonetic feature within the interaction also plays a role. For different personality types and their inherent needs in spoken interaction, phonetic features they assume to be conducive to these respective needs might be particularly “worthwhile” accommodation targets. Thus, a more neurotic speaker to whom social approval is important might feel that convergence on intonational features has the potential to create such approval, whereas a more conscientious speaker might find that convergence on a binary segmental feature allows them to demonstrate that they are performing a task well. If such tendencies should indeed apply, we assume that this selection takes place largely at an unconscious level.

For non-native speakers of the target language German, we also demonstrated a convergence effect. Regarding the allophonic variation [ɪç]/[ɪk], the effect was as strong as that of the native speakers, and regarding the vowel quality, as with the native speakers, it did not occur at the group level. In the case of the question intonation, however, we observed that the non-native speakers initially tended to deviate further from the expected standard realization than the native speakers, i.e., in terms of phrasing and placement of the nuclear pitch accent. With regard to the components of question intonation that we examined, the convergence effect then turned out to be attenuated compared to that of the native speakers, i.e., no significant increase in rising intonation contours and no shift in the nuclear pitch accent. This could be due to the fact that structural phonological elements that deviate strongly from the own native pattern — here the phrase-initial position of the nuclear pitch accent — are less likely to be adopted through accommodation. In addition, it could be that by adjusting the initially more deviant structure, i.e., “normalizing” the phrasing and shifting the nuclear pitch accent to the default position in German, accommodation has already occurred to the extent that can be expected in a single step, and that a longer interaction would be needed for further adaptation.

MENTAL BOUNDARIES We hypothesize that there is a connection between individuals who have a general tendency to converge to their interlocutor in conversation and those who acquire a native-like pronunciation when learning a foreign language. While phonetic talent plays a crucial role in this context (e.g., Lewandowski and Jilka, 2019), it is reasonable to assume that such talent will only come to full fruition in communicative interaction when paired with a favoring personality structure. How the latter is composed, however, remains to be understood (e.g., Hu and Reiterer, 2009; Reiterer, 2019).

One personality dimension under consideration is the strength of mental boundaries, indicating the permeability of an individual’s mind both within itself and towards its environment, which is assessed with the Boundary Questionnaire (BQ, Hartmann, 1991). The influence of boundary strength on foreign language attainment has been sparsely studied to date (e.g., Guiora et al., 1972; Baran-Łucarz, 2012), and the connection with accommodation not at all. We suppose that the boundary construct, which is deemed informative in the context of foreign language learning, has the potential to enrich accommodation research as well.

We present a validated German translation of the empirically-derived short version of the Boundary Questionnaire (BQ-Sh) as an instrument for investigating mental boundaries with L1 German speakers. It may be applied in accommodation research, but is not limited to it.

We demonstrated the application of this instrument for the participants of the WOz experiment in order to explore the assumption that convergence with respect to different types of phonetic features can be favored by thinner or thicker mental boundaries, respectively.

Since the boundary construct is partially correlated with the *Big Five* personality traits, effects reflecting the prior personality analysis with the *Big Five* were expected. This was corroborated in a tendency of thicker mental boundaries to influence accommodation to the allophonic contrast [ɪç]/[ɪk], since Conscientiousness correlated negatively with boundary strength. Thinner mental boundaries, however, did not impact the accommodation to question intonation, although Neuroticism correlated positively with boundary strength.

The mental boundaries constitute a very broad personality dimension that may reveal tendencies in accommodation to locally anchored phonetic features depending on their type, but whose predictive power will probably be more apparent in a holistic approach to phonetic accommodation.

Recalling the hybrid model of convergence by Lewandowski (2012) mentioned at the beginning of this work, it is important to state again that personality is but one of many factors influencing the accommodation process. Therefore, in future research, the aspect of mental boundaries should ideally be addressed in conjunction with other factors, such as phonetic talent, but also attentional and memory components.

FURTHER CONSIDERATIONS The question arises how speakers who are perceived as strong convergers from a holistic perspective would behave with respect to the locally anchored phonetic features we examined. In our data, we observed a large variation with respect to these features and no speaker ever converged on all of them. It is probable that even a strong converger does not accommodate with respect to all possible features, but possesses a certain talent for (subconsciously) selecting features that are appropriate and contribute to the perception of convergence. This selection of features then certainly spans over all possible phonetic levels, from local phenomena to the acoustic-prosodic level, all the way to global similarity.

A targeted study of speakers perceived as strong convergers, either by self-assessment or as assessed by their environment, would certainly be an insightful addition to the present work, which investigated accommodation in the average user of an SDS.

For the HCI context, the finding that users of SDSs adopt locally anchored phonetic phenomena from computer agents is relevant from two points of view. From a human perspective, such accommodation can be exploited in a CALL context, as it has the potential to lead to incidental inductive learning at the levels of segmental pronunciation and local prosody. From the perspective of the computer agent, it must be acknowledged that the phonetic realization at precisely these levels is relevant to, perceived, and possibly even adopted by the user. Thus, it should be a goal that high quality synthetic voices can be purposefully manipulated at the level of locally anchored features to eventually enable phonetically responsive SDSs (Raveh et al., 2017b) — as has been proposed for global acoustic-prosodic features (Levitan et al., 2016).

To what extent converging behavior should then be implemented in computer agents is a different matter. There is already some evidence that accommodation at the level of acoustic-prosodic features by the computer agent can influence how socially present, likable, competent, or trustworthy users consider it to be (e.g., Lubold et al., 2016; Levitan et al., 2016; Gauder et al., 2018; Beňuš et al., 2018). However, since there is so much individual variation regarding phonetic accommodation in humans, what would be an appropriate behavior for a computer agent? The goal would certainly be to model “successful” accommodation behavior. What the

latter might look like and whether this also depends on the accommodation behavior of the respective interlocutor — which we assume to be the case — requires further investigation.

In the context of CALL, native-like pronunciation forms a clear target for accommodation, which in turn provides the required parameters for implementing useful accommodation behavior in an SDS. A phonetically responsive virtual tutor — as proposed at the beginning of this dissertation for human-human interaction — should be able to detect incorrect pronunciation on the part of the learner and diverge from it, in order to prompt the learner to converge to the tutor.

7 CONCLUSION AND OUTLOOK

With the present work we are contributing to a better understanding of phonetic accommodation in human-computer interaction (HCI). While the focus of previous studies in this context has mostly been on global acoustic-prosodic phenomena, we demonstrate the relevance of locally anchored phenomena, especially segmental allophonic variation and local prosodic features. Speakers identify them as accommodation targets and converge to them when repeating short sentences after model voices and when dynamically interacting with a computer agent, even in cases where the interlocutor voices sound distinctly synthetic and therefore highlight the HCI context. Consequently, modern synthetic voices should be manipulable at this level of phonetic detail in order to implement accommodation behavior similar to that of humans in the computer agents as well.

The observed phonetic convergence does not only occur in native speakers, but also in non-native speakers of the target language — German in the present case. However, for non-native speakers, greater a priori phonetic distance from the interlocutor, as well as constraints arising from dominant patterns in the own native language, may attenuate the degree of accommodation for equal interaction length compared to native speakers. It could still be exploited in the context of computer-assisted language learning (CALL) — again under the assumption that the synthetic voices used in CALL applications are able to realize the relevant phonetic forms and to reflect native-like variation in pronunciation.

As in previous studies, we observed individual differences between speakers in their accommodation behavior and considered personality in particular as a possible predictor for this variation. Based on our findings regarding the influence of the *Big Five* personality traits, we hypothesize that accommodation with respect to different types of phonetic features may serve the needs that different personality types have in spoken interaction. This merits further investigation.

As far as the broad personality concept of mental boundaries is concerned, the findings presented in this dissertation constitute a starting point. By making a validated German adaptation of the empirically-derived short version of the Boundary Questionnaire available, we provide the means for further research into their influence on phonetic accommodation in native speakers of German. We hypothesize that their predictive power applies primarily to a general disposition favoring phonetic accommodation, which is more likely to be captured by a holistic approach than by examining isolated locally anchored features.

In the course of our investigations, additional related questions presented themselves that may be explored in the future. It may, for example, be worthwhile to complement the current results with an explicit imitation task to better discriminate between the effects of intentional imitation and those of dynamic convergence in interaction. We assume that features which are not identified as targets for imitation within the stimuli are unlikely to become targets for accommodation.

The different voice types used in our experiments (i.e., natural, diphone, HMM) were rated differently on attributes such as naturalness, likability, and competence by the experimental subjects, which may influence their accommodation behavior. However, several voices per type that differ systematically in these attributes should be examined in order to draw conclusions about this influence that can be clearly distinguished from the influence of the type itself.

Moreover, speakers could be pre-selected based on commonalities in their personality structure — extreme expressions of the *Big Five* and/or mental boundaries — to shed further light on

their presumed influence on phonetic accommodation. Ideally, personality structure and other aspects of speaker disposition could eventually be used to derive user profiles that would predict common patterns in accommodation behavior. Such profiles could then serve as orientation for the implementation of accommodation behavior in spoken dialog systems.

Finally, the perception of phonetic accommodation should be further investigated. For example, the question arises whether there is a relationship between one's own accommodation type during production and one's perception of accommodation in an interlocutor, be they another human or a computer.

We are looking forward to future findings of research on phonetic accommodation in the HCI context!

BIBLIOGRAPHY

- Abrego-Collier, C., J. Grove, M. Sonderegger, and C. L. Alan (2011). “Effects of speaker evaluation on phonetic convergence.” In: *International Congress of Phonetic Sciences (ICPhS)*. Hong Kong, pp. 192–195. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Abrego-Collier/Abrego-Collier.pdf>.
- Akaike, H. (1973). “Information theory and an extension of the maximum likelihood principle.” In: *International Symposium on Information Theory*, pp. 267–281.
- Apple, W., L. A. Streeter, and R. M. Krauss (1979). “Effects of pitch and speech rate on personal attributions.” In: *Journal of Personality and Social Psychology* 37.5, p. 715. DOI: [10.1037/0022-3514.37.5.715](https://doi.org/10.1037/0022-3514.37.5.715).
- Armor, D. J. (1973). “Theta reliability and factor scaling.” In: *Sociological Methodology* 5, pp. 17–50. DOI: [10.2307/270831](https://doi.org/10.2307/270831).
- Atkinson, R. K., R. E. Mayer, and M. M. Merrill (2005). “Fostering social agency in multimedia learning: Examining the impact of an animated agent’s voice.” In: *Contemporary Educational Psychology* 30.1, pp. 117–139. DOI: [10.1016/j.cedpsych.2004.07.001](https://doi.org/10.1016/j.cedpsych.2004.07.001).
- Aumann, C., O. Lahl, and R. Pietrowsky (2012). “Relationship between dream structure, boundary structure and the Big Five personality dimensions.” In: *Dreaming* 22.2, pp. 124–135. DOI: [10.1037/a0028977](https://doi.org/10.1037/a0028977).
- Babel, M. (2010). “Dialect divergence and convergence in New Zealand English.” In: *Language in Society* 39, pp. 437–456. DOI: [10.1017/S0047404510000400](https://doi.org/10.1017/S0047404510000400).
- (2012). “Evidence for phonetic and social selectivity in spontaneous phonetic imitation.” In: *Journal of Phonetics* 40, pp. 177–189. DOI: [10.1016/j.wocn.2011.09.001](https://doi.org/10.1016/j.wocn.2011.09.001).
- Babel, M., G. McGuire, S. Walters, and A. Nicholls (2014). “Novelty and social preference in phonetic accommodation.” In: *Laboratory Phonology* 5.1, pp. 123–150. DOI: [10.1515/lp-2014-0006](https://doi.org/10.1515/lp-2014-0006).
- Bailly, G. and A. Martin (2014). “Assessing objective characterizations of phonetic convergence.” In: *Interspeech*. Singapore, pp. 2011–2015. URL: https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_2011.pdf.
- Baran-Lucarz, M. (2012). “Ego boundaries and attainments in FL pronunciation.” In: *Studies in Second Language Learning and Teaching* 2.1, pp. 45–66. DOI: [10.14746/ss11t.2012.2.1.3](https://doi.org/10.14746/ss11t.2012.2.1.3).
- Baron-Cohen, S., S. Wheelwright, R. Skinner, J. Martin, and E. Clubley (2001). “The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians.” In: *Journal of Autism and Developmental Disorders* 31.1, pp. 5–17. DOI: [10.1023/A:1005653411471](https://doi.org/10.1023/A:1005653411471).
- Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). “Fitting linear mixed-effects models using lme4.” In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: [10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01).
- Baylor, A., J. Ryu, and E. Shen (2003). “The effects of pedagogical agent voice and animation on learning, motivation and perceived persona.” In: *EdMedia + Innovative Learning*. Honolulu, Hawaii, pp. 452–458.
- Beal, S. W. (1998). “The boundary characteristics of artists.” PhD thesis. Boston University.
- Bell, L., J. Gustafson, and M. Heldner (2003). “Prosodic adaptation in human-computer interaction.” In: *International Congress of Phonetic Sciences (ICPhS)*. Barcelona, pp. 2453–2456.

- URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_2453.pdf.
- Benjamini, Y. and Y. Hochberg (1995). “Controlling the false discovery rate: A practical and powerful approach to multiple testing.” In: *Journal of the Royal Statistical Society* 57.1, pp. 289–300. DOI: [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).
- Beňuš, Š., M. Trnka, E. Kuric, L. Marták, A. Gravano, J. Hirschberg, and R. Levitan (2018). “Prosodic entrainment and trust in human-computer interaction.” In: *International Conference on Speech Prosody*. Poznań, pp. 220–224. DOI: [10.21437/SpeechProsody.2018-45](https://doi.org/10.21437/SpeechProsody.2018-45).
- Bilous, F. R. and R. M. Krauss (1988). “Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads.” In: *Language & Communication* 8, pp. 183–194. DOI: [10.1016/0271-5309\(88\)90016-x](https://doi.org/10.1016/0271-5309(88)90016-x).
- Blatt, S. J. and B. A. Ritzler (1974). “Thought disorder and boundary disturbances in psychosis.” In: *Journal of Consulting and Clinical Psychology* 42.3, pp. 370–381. DOI: [10.1037/h0036688](https://doi.org/10.1037/h0036688).
- Boersma, P. and D. Weenink (2017). *Praat: doing phonetics by computer [computer program]*. Version 6.0.25, retrieved 11 February 2017 from <http://www.praat.org/>.
- (2019). *Praat: doing phonetics by computer [computer program]*. Version 6.1.06, retrieved 8 November 2019 from <http://www.praat.org/>.
- Borkenau, P. and F. Ostendorf (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen: Hogrefe.
- (2007). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (Vol. 2. neu normierte und vollständig überarb. Auflage)*. Göttingen: Hogrefe.
- Borrie, S. A., N. Lubold, and H. Pon-Barry (2015). “Disordered speech disrupts conversational entrainment: a study of acoustic-prosodic entrainment and communicative success in populations with communication challenges.” In: *Frontiers in Psychology* 6.1187. DOI: [10.3389/fpsyg.2015.01187](https://doi.org/10.3389/fpsyg.2015.01187).
- Branigan, H. P., M. J. Pickering, J. Pearson, and J. F. McLean (2010). “Linguistic alignment between people and computers.” In: *Journal of Pragmatics* 42.9, pp. 2355–2368. DOI: [10.1016/j.pragma.2009.12.012](https://doi.org/10.1016/j.pragma.2009.12.012).
- Briggs-Myers, I. and P. B. Myers (1995). *Gifts differing: Understanding personality type*. 2nd ed. Mountain View, CA: Davies-Black. ISBN: 978-0891060741.
- Burnham, D., S. Jeffry, and L. Rice (2010). “‘D-o-e-s-Not-C-o-m-p-u-t-e’: vowel hyperarticulation in speech to an auditory-visual avatar.” In: *Auditory-Visual Speech Processing (AVSP)*. Hakone. URL: https://www.isca-speech.org/archive/avsp10/papers/av10_P18.pdf.
- Carver, C. S. and T. L. White (1994). “Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales.” In: *Journal of Personality and Social Psychology* 67.2, p. 319. DOI: [10.1037/0022-3514.67.2.319](https://doi.org/10.1037/0022-3514.67.2.319).
- Cattell, R. B. (1966). “The scree test for the number of factors.” In: *Multivariate behavioral research* 1.2, pp. 245–276. DOI: [10.1207/s15327906mbr0102_10](https://doi.org/10.1207/s15327906mbr0102_10).
- Clopper, C. G. and E. Dossey (2020). “Phonetic convergence to Southern American English: acoustics and perception.” In: *The Journal of the Acoustical Society of America* 147.1, pp. 671–683. DOI: [10.1121/10.0000555](https://doi.org/10.1121/10.0000555).
- Cohen Priva, U. and C. Sanker (2018). “Distinct behaviors in convergence across measures.” In: *Annual Meeting of the Cognitive Science Society (CogSci)*. Austin, TX, pp. 1515–1520.
- (2019). “Limitations of difference-in-difference for measuring convergence.” In: *Laboratory Phonology* 10.1, p. 15. DOI: [10.5334/labphon.200](https://doi.org/10.5334/labphon.200).
- Coles-Harris, E. H. (2017). “Perspectives on the motivations for phonetic convergence.” In: *Language and Linguistics Compass* 11.12. DOI: [10.1111/lnc3.12268](https://doi.org/10.1111/lnc3.12268).

- Costa, A., M. J. Pickering, and A. Sorace (2008). “Alignment in second language dialogue.” In: *Language and Cognitive Processes* 23.4, pp. 528–556. DOI: [10.1080/01690960801920545](https://doi.org/10.1080/01690960801920545).
- Costa, P. T. and R. R. McCrae (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): professional manual*. Odessa, FL: Psychological Assessment Resources.
- Coulston, R., S. Oviatt, and C. Darves (2002). “Amplitude convergence in children’s conversational speech with animated personas.” In: *International Conference on Spoken Language Processing (ICSLP)*. Denver, pp. 2689–2692. URL: https://www.isca-speech.org/archive/archive_papers/icslp_2002/i02_2689.pdf.
- Cowen, D. and R. Levin (1995). “The use of the Hartmann Boundary Questionnaire with an adolescent population.” In: *Dreaming* 5.2, pp. 105–114. DOI: [10.1037/h0094428](https://doi.org/10.1037/h0094428).
- Cronbach, L. J. (1951). “Coefficient alpha and the internal structure of tests.” In: *Psychometrika* 16.3, pp. 297–334. DOI: [10.1007/bf02310555](https://doi.org/10.1007/bf02310555).
- Dahlbäck, N., A. Jönsson, and L. Ahrenberg (1993). “Wizard of Oz studies – why and how.” In: *Knowledge-based systems* 6.4, pp. 258–266. DOI: [10.1016/0950-7051\(93\)90017-N](https://doi.org/10.1016/0950-7051(93)90017-N).
- Delais-Roussarie, E., B. Post, M. Avanzi, C. Buthke, A. Di Cristo, I. Feldhausen, S.-A. Jun, P. Martin, T. Meisenburg, A. Rialland, R. Sichel-Bazin, and H.-Y. Yoo (2015). “Intonational phonology of French: Developing a ToBI system for French.” In: *Intonation in Romance*. Ed. by S. Frota and P. Prieto. Oxford University Press, pp. 63–100.
- Delvaux, V. and A. Soquet (2007). “Inducing imitative phonetic variation in the laboratory.” In: *International Congress of Phonetic Sciences (ICPhS)*. Saarbrücken, pp. 369–372. URL: <http://www.icphs2007.de/conference/Papers/1318/1318.pdf>.
- Di Cristo, A. (1998). “Intonation in French.” In: *Intonation systems: A survey of twenty languages*. Ed. by D. Hirst and A. Di Cristo. Cambridge University Press. Chap. 11, pp. 195–218. ISBN: 978-0521395502.
- Dias, J. W. and L. D. Rosenblum (2016). “Visibility of speech articulation enhances auditory phonetic convergence.” In: *Attention, Perception, & Psychophysics* 78.1, pp. 317–333. DOI: [10.3758/s13414-015-0982-6](https://doi.org/10.3758/s13414-015-0982-6).
- Dudenredaktion (2015). *Duden - Das Aussprachewörterbuch: Betonung und Aussprache von über 132.000 Wörtern und Namen*. Vol. 6. Duden - Deutsche Sprache in 12 Bänden. Mannheim: Bibliographisches Institut GmbH.
- Dufour, S. and N. Nguyen (2013). “How much imitation is there in a shadowing task?” In: *Frontiers in Psychology* 4.346. DOI: [10.3389/fpsyg.2013.00346](https://doi.org/10.3389/fpsyg.2013.00346).
- Dutoit, T., V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken (1996). “The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes.” In: *International Conference on Spoken Language Processing (ICSLP)*. Vol. 3. Philadelphia, PA, pp. 1393–1396. DOI: [10.1109/icslp.1996.607874](https://doi.org/10.1109/icslp.1996.607874).
- Ehrman, M. E. (1996). *Understanding second language learning difficulties*. Thousand Oaks, CA: SAGE Publications. DOI: [10.4135/9781452243436](https://doi.org/10.4135/9781452243436).
- (1999). “Ego boundaries and tolerance of ambiguity in second language learning.” In: *Affect in language learning*. Cambridge: Cambridge University Press, pp. 68–86. ISBN: 978-0521659635.
- Ellbogen, T., F. Schiel, and A. Steffen (2004). “The BITS speech synthesis corpus for German.” In: *International Conference on Language Resources and Evaluation (LREC)*. Lisbon, pp. 2091–2094. URL: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/72.pdf>.
- Federn, P. (1952). *Ego psychology and the psychoses*. New York: Basic Books.
- Fisher, S. and S. E. Cleveland (1968). *Body image and personality*. 2nd ed. New York: Dover Publications.

- Forbes-Riley, K., D. J. Litman, S. Silliman, and J. R. Tetreault (2006). “Comparing synthesized versus pre-recorded tutor speech in an intelligent tutoring spoken dialogue system.” In: *Florida Artificial Intelligence Research Society Conference (FLAIRS)*. Melbourne Beach, Florida, pp. 509–514.
- Fowler, C. A., J. M. Brown, L. Sabadini, and J. Weihing (2003). “Rapid access to speech gestures in perception: evidence from choice and simple response time tasks.” In: *Journal of Memory and Language* 49.3, pp. 396–413. DOI: [10.1016/S0749-596X\(03\)00072-X](https://doi.org/10.1016/S0749-596X(03)00072-X).
- Funkhouser, A. T., O. Würmle, K. Carnes, P. Locher, F. Ramseyer, and M. Bahro (2008). “Boundary Questionnaire results and dream recall among persons going through retirement.” In: *International Journal of Dream Research* 1.2, pp. 34–38. DOI: [10.11588/ijodr.2008.2.78](https://doi.org/10.11588/ijodr.2008.2.78).
- Funkhouser, A. T., O. Würmle, C. M. Cornu, and M. Bahro (2001). “Boundary questionnaire results in the mentally healthy elderly.” In: *Dreaming* 11.2, pp. 83–88. DOI: [10.1023/a:1009432520849](https://doi.org/10.1023/a:1009432520849).
- Gallois, C., T. Ogay, and H. Giles (2005). “Communication Accommodation Theory: a look back and a look ahead.” In: *Theorizing about intercultural communication*. Ed. by W. B. Gudykunst. Thousand Oaks, CA: SAGE Publications, pp. 121–148.
- Gálvez, R. H., A. Gravano, Š. Beňuš, R. Levitan, M. Trnka, and J. Hirschberg (2020). “An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars.” In: *Speech Communication* 124, pp. 46–67. DOI: [10.1016/j.specom.2020.07.007](https://doi.org/10.1016/j.specom.2020.07.007).
- Gauder, L., M. Reartes, R. H. Gálvez, Š. Beňuš, and A. Gravano (2018). “Testing the effects of acoustic/prosodic entrainment on user behavior at the dialog-act level.” In: *International Conference on Speech Prosody*. Poznań, pp. 374–378. DOI: [10.21437/SpeechProsody.2018-76](https://doi.org/10.21437/SpeechProsody.2018-76).
- Gessinger, I., B. Möbius, B. Andreeva, E. Raveh, and I. Steiner (2019a). “Phonetic accommodation in a Wizard-of-Oz experiment: intonation and segments.” In: *Interspeech*. Graz, pp. 301–305. DOI: [10.21437/Interspeech.2019-2445](https://doi.org/10.21437/Interspeech.2019-2445).
- (2020). “Phonetic accommodation of L2 German speakers to the virtual language learning tutor Mirabella.” In: *Interspeech*. Shanghai, pp. 4118–4122. DOI: [10.21437/Interspeech.2020-2701](https://doi.org/10.21437/Interspeech.2020-2701).
- Gessinger, I., B. Möbius, N. Fakhar, E. Raveh, and I. Steiner (2019b). “A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction.” In: *International Congress of Phonetic Sciences (ICPhS)*. Melbourne, pp. 1475–1479. URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1524.pdf.
- Gessinger, I., B. Möbius, S. Le Maguer, E. Raveh, and I. Steiner (2021a). “Accommodation in interaction with a virtual language learning tutor: a Wizard-of-Oz study.” In: *Journal of Phonetics* 86, p. 101029. DOI: [10.1016/j.wocn.2021.101029](https://doi.org/10.1016/j.wocn.2021.101029).
- Gessinger, I., E. Raveh, S. Le Maguer, B. Möbius, and I. Steiner (2017). “Shadowing synthesized speech – segmental analysis of phonetic convergence.” In: *Interspeech*. Stockholm, pp. 3797–3801. DOI: [10.21437/Interspeech.2017-1433](https://doi.org/10.21437/Interspeech.2017-1433).
- Gessinger, I., E. Raveh, J. O’Mahony, I. Steiner, and B. Möbius (2016). “A shadowing experiment with natural and synthetic stimuli.” In: *Phonetik & Phonologie*. Munich, pp. 58–61. DOI: [10.5282/ubm/epub.29405](https://doi.org/10.5282/ubm/epub.29405).
- Gessinger, I., E. Raveh, I. Steiner, and B. Möbius (2021b). “Phonetic accommodation to natural and synthetic voices: behavior of groups and individuals in speech shadowing.” In: *Speech Communication* 127, pp. 43–63. DOI: [10.1016/j.specom.2020.12.004](https://doi.org/10.1016/j.specom.2020.12.004).

- Gessinger, I., A. Schweitzer, B. Andreeva, E. Raveh, B. Möbius, and I. Steiner (2018). “Convergence of pitch accents in a shadowing task.” In: *International Conference on Speech Prosody*. Poznań, pp. 225–229. DOI: [10.21437/SpeechProsody.2018-46](https://doi.org/10.21437/SpeechProsody.2018-46).
- Gijssels, T., L. Staum Casasanto, K. Jasmin, P. Hagoort, and D. Casasanto (2016). “Speech accommodation without priming: The case of pitch.” In: *Discourse Processes* 53.4, pp. 233–251. DOI: [10.1080/0163853x.2015.1023965](https://doi.org/10.1080/0163853x.2015.1023965).
- Giles, H. (1973). “Accent mobility: a model and some data.” In: *Anthropological Linguistics*, pp. 87–105.
- ed. (2016). *Communication accommodation theory: Negotiating personal and social identities in context*. Cambridge, UK: Cambridge University Press. ISBN: 978-1107105829.
- Giles, H., N. Coupland, and J. Coupland (1991). “Accommodation theory: communication, context, and consequence.” In: *Contexts of Accommodation: Developments in Applied Sociolinguistics*. Ed. by H. Giles, J. Coupland, and N. Coupland. Cambridge University Press, pp. 1–68. DOI: [10.1017/cbo9780511663673.001](https://doi.org/10.1017/cbo9780511663673.001).
- Goldinger, S. D. (1998). “Echoes of echoes? An episodic theory of lexical access.” In: *Psychological Review* 105.2, pp. 251–279. DOI: [10.1037/0033-295X.105.2.251](https://doi.org/10.1037/0033-295X.105.2.251).
- Gregory, S. W. and S. Webster (1996). “A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions.” In: *Journal of personality and social psychology* 70.6, pp. 1231–1240. DOI: [10.1037/0022-3514.70.6.1231](https://doi.org/10.1037/0022-3514.70.6.1231).
- Grice, M. and S. Baumann (2002). “Deutsche Intonation und GToBl.” In: *Linguistische Berichte*, pp. 267–298.
- Guiora, A. Z. (1994). “The two faces of language ego.” In: *Psychologica Belgica* 34.2/3, pp. 83–97.
- Guiora, A. Z., B. Beit-Hallahmi, R. C. L. Brannon, C. Y. Dull, and T. Scovel (1972). “The effects of experimentally induced changes in ego states on pronunciation ability in a second language: an exploratory study.” In: *Comprehensive Psychiatry* 13.5, pp. 421–428. DOI: [10.1016/0010-440X\(72\)90083-1](https://doi.org/10.1016/0010-440X(72)90083-1).
- Harrison, A. and J. Singer (2013). “Boundaries in the mind: historical context and current research using the boundary questionnaire.” In: *Imagination, cognition and personality* 33.1, pp. 205–215. DOI: [10.2190/IC.33.1-2.h](https://doi.org/10.2190/IC.33.1-2.h).
- Harrison, R. H., E. Hartmann, and J. Bevis (2006). “The Boundary Questionnaire: Its preliminary reliability and validity.” In: *Imagination, Cognition and Personality* 25.4, pp. 355–382. DOI: [10.2190/8120-6340-T808-7001](https://doi.org/10.2190/8120-6340-T808-7001).
- Hartmann, E., R. Harrison, J. Bevis, I. Hurwitz, A. Holevas, and H. Dawani (1987). “The Boundary Questionnaire: A measure of thin and thick boundaries derived from work with nightmare sufferers.” In: *Sleep Research* 16, p. 274.
- Hartmann, E. (1989). “Boundaries of dreams, boundaries of dreamers: thin and thick boundaries as a new personality measure.” In: *Psychiatric Journal of the University of Ottawa* 14.4, pp. 557–560.
- (1991). *Boundaries in the mind: a new psychology of personality*. New York, NY: Basic Books.
- Hartmann, E., R. Harrison, and M. Zborowski (2001). “Boundaries in the mind: past research and future directions.” In: *North American Journal of Psychology* 3.3, pp. 347–368.
- Hartmann, E. and R. G. Kunzendorf (2006). “Boundaries and dreams.” In: *Imagination, Cognition and Personality* 26.1, pp. 101–115. DOI: [10.2190/HK76-038K-407M-8670](https://doi.org/10.2190/HK76-038K-407M-8670).
- Hathaway, S. R. and J. C. McKinley (1943). *The Minnesota Multiphasic Personality Inventory*. Minneapolis, MN: University of Minnesota Press.

- Hendrickson, A. E. and P. O. White (1964). “Promax: A quick method for rotation to oblique simple structure.” In: *British journal of statistical psychology* 17.1, pp. 65–70. DOI: [10.1111/j.2044-8317.1964.tb00244.x](https://doi.org/10.1111/j.2044-8317.1964.tb00244.x).
- Honorof, D. N., J. Weihing, and C. A. Fowler (2011). “Articulatory events are imitated under rapid shadowing.” In: *Journal of Phonetics* 39.1, pp. 18–38. DOI: [10.1016/j.wocn.2010.10.007](https://doi.org/10.1016/j.wocn.2010.10.007).
- Horn, J. L. (1965). “A rationale and test for the number of factors in factor analysis.” In: *Psychometrika* 30.2, pp. 179–185. DOI: [10.1007/bf02289447](https://doi.org/10.1007/bf02289447).
- Hotelling, H. (1933). “Analysis of a complex of statistical variables into principal components.” In: *Journal of Educational Psychology* 24.7, pp. 498–452. DOI: [10.1037/h0070888](https://doi.org/10.1037/h0070888).
- Hu, X. and S. M. Reiterer (2009). “Personality and pronunciation talent in second language acquisition.” In: *Language Talent and Brain Activity*. Ed. by G. Dogil and S. Reiterer. Berlin: Mouton de Gruyter, pp. 97–130. DOI: [10.1515/9783110215496.97](https://doi.org/10.1515/9783110215496.97).
- International Test Commission (2018). “ITC guidelines for translating and adapting tests (second edition).” In: *International Journal of Testing* 18.2, pp. 101–134. DOI: [10.1080/15305058.2017.1398166](https://doi.org/10.1080/15305058.2017.1398166).
- Jagdfeld, N. and S. Baumann (2011). “Order effects on the perception of relative prominence.” In: *International Congress of Phonetic Sciences (ICPhS)*. Hong Kong, pp. 958–961. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Jagdfeld/Jagdfeld.pdf>.
- Jilka, M. (2009a). “Assessment of phonetic ability.” In: *Language Talent and Brain Activity*. Ed. by G. Dogil and S. Reiterer. Berlin: Mouton de Gruyter, pp. 17–66. DOI: [10.1515/9783110215496.17](https://doi.org/10.1515/9783110215496.17).
- (2009b). “Talent and proficiency in language.” In: *Language Talent and Brain Activity*. Ed. by G. Dogil and S. Reiterer. Berlin: Mouton de Gruyter, pp. 1–16. DOI: [10.1515/9783110215496.1](https://doi.org/10.1515/9783110215496.1).
- John, O. P., E. M. Donahue, and R. L. Kentle (1991). *The Big Five Inventory – versions 4a and 54*.
- John, O. P., L. P. Naumann, and C. J. Soto (2008). “Paradigm shift to the integrative big five trait taxonomy.” In: *Handbook of Personality: Theory and Research* 3.2, pp. 114–158.
- Kaiser, H. F. (1958). “The varimax criterion for analytic rotation in factor analysis.” In: *Psychometrika* 23.3, pp. 187–200. DOI: [10.1007/BF02289233](https://doi.org/10.1007/BF02289233).
- Kelley, J. F. (1984). “An iterative design methodology for user-friendly natural language office information applications.” In: *ACM Transactions on Information Systems (TOIS)* 2.1, pp. 26–41. DOI: [10.1145/357417.357420](https://doi.org/10.1145/357417.357420).
- Kiesewalter, C. (2019). *Zur subjektiven Dialektalität regiolektaler Aussprachemerkmale des Deutschen*. Franz Steiner Verlag.
- Kim, D. and M. Clayards (2019). “Individual differences in the link between perception and production and the mechanisms of phonetic imitation.” In: *Language, Cognition and Neuroscience* 34.6, pp. 769–786. DOI: [10.1080/23273798.2019.1582787](https://doi.org/10.1080/23273798.2019.1582787).
- Kim, M., W. S. Horton, and A. R. Bradlow (2011). “Phonetic convergence in spontaneous conversations as a function of interlocutor language distance.” In: *Laboratory Phonology* 2.1, pp. 125–156. DOI: [10.1515/labphon.2011.004](https://doi.org/10.1515/labphon.2011.004).
- King, S., A. W. Black, P. Taylor, R. Caley, and R. Clark (1999). *Edinburgh Speech Tools Library*. Version 1.2. URL: http://www.cstr.ed.ac.uk/projects/speech_tools/.
- Kisler, T., U. Reichel, and F. Schiel (2017). “Multilingual processing of speech via web services.” In: *Computer Speech & Language* 45, pp. 326–347. DOI: [10.1016/j.cs1.2017.01.005](https://doi.org/10.1016/j.cs1.2017.01.005).
- Kleiner, S. (2011). *Atlas zur Aussprache des deutschen Gebrauchsstandards (AADG)*. Unter Mitarbeit von Ralf Knöbl. URL: <http://prowiki.ids-mannheim.de/bin/view/AADG/>.

- Krauss, R. M. and J. S. Pardo (2004). “Is alignment always the result of automatic priming?” In: *Behavioral and Brain Sciences* 27.2, pp. 203–204. DOI: [10.1017/S0140525X0436005X](https://doi.org/10.1017/S0140525X0436005X).
- Kunzendorf, R. G., E. Hartmann, R. Cohen, and J. Cutler (1997). “Bizarreness of the dreams and daydreams reported by individuals with thin and thick boundaries.” In: *Dreaming* 7.4, pp. 265–271.
- Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen (2017). “lmerTest package: tests in linear mixed effects models.” In: *Journal of Statistical Software* 82.13, pp. 1–26. DOI: [10.18637/jss.v082.i13](https://doi.org/10.18637/jss.v082.i13).
- Landis, B. (1970). “Ego boundaries.” In: *Psychological Issues* 6.4, pp. 1–172.
- Lawley, D. N. and A. E. Maxwell (1962). “Factor analysis as a statistical method.” In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 12.3, pp. 209–229. DOI: [10.2307/2986915](https://doi.org/10.2307/2986915).
- Le Maguer, S., I. Steiner, F. Tombini, P. Deb, M. Basu, and I. Kröger (2018). “Agile MaryTTS architecture for the Blizzard Challenge 2018.” In: *Blizzard Challenge*. Hyderabad. URL: http://festvox.org/blizzard/bc2018/MARY_BlizzardChallenge2018.pdf.
- Lee, C.-C., M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. S. Narayanan (2010a). “Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples.” In: *Interspeech*. Makuhari, pp. 793–796. URL: https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_0793.pdf.
- Lee, J., J. Jang, and L. Plonsky (2014). “The effectiveness of second language pronunciation instruction: A meta-analysis.” In: *Applied Linguistics* 36.3, pp. 345–366. DOI: [10.1093/applin/amu040](https://doi.org/10.1093/applin/amu040).
- Lee, M. K., S. Kiesler, and J. Forlizzi (2010b). “Receptionist or information kiosk: how do people talk with a robot?” In: *ACM Conference on Computer Supported Cooperative Work*, pp. 31–40. DOI: [10.1145/1718918.1718927](https://doi.org/10.1145/1718918.1718927).
- Levin, R., J. Galin, and B. Zywiak (1991). “Nightmares, boundaries, and creativity.” In: *Dreaming* 1.1, pp. 63–74. DOI: [10.1037/h0094318](https://doi.org/10.1037/h0094318).
- Levitan, R., Š. Beňuš, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg (2016). “Implementing acoustic-prosodic entrainment in a conversational avatar.” In: *Interspeech*. San Francisco, CA, pp. 1166–1170. DOI: [10.21437/Interspeech.2016-985](https://doi.org/10.21437/Interspeech.2016-985).
- Levitan, R., A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova (2012). “Acoustic-prosodic entrainment and social behavior.” In: *NAACL Conference on Human Language Technologies*, pp. 11–19. URL: <https://www.aclweb.org/anthology/N12-1002.pdf>.
- Levitan, R. and J. Hirschberg (2011). “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions.” In: *Interspeech*. Florence, pp. 3081–3084. URL: https://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_3081.pdf.
- Lewandowski, E. M. and L. C. Nygaard (2018). “Vocal alignment to native and non-native speakers of English.” In: *The Journal of the Acoustical Society of America* 144.2, pp. 620–633. DOI: [10.1121/1.5038567](https://doi.org/10.1121/1.5038567).
- Lewandowski, N. (2012). “Talent in nonnative phonetic convergence.” PhD thesis. Universität Stuttgart. DOI: [10.18419/opus-2858](https://doi.org/10.18419/opus-2858).
- Lewandowski, N. and M. Jilka (2019). “Phonetic convergence, language talent, personality & attention.” In: *Frontiers in Communication* 4.18. DOI: [10.3389/fcomm.2019.00018](https://doi.org/10.3389/fcomm.2019.00018).
- Lewin, K. (1935). *A dynamic theory of personality: Selected papers*. New York: McGraw Hill.
- Litman, D. and S. Silliman (2004). “ITSPOKE: An intelligent tutoring spoken dialogue system.” In: *Demonstration papers at HLT-NAACL 2004*, pp. 5–8. DOI: [10.3115/1614025.1614027](https://doi.org/10.3115/1614025.1614027).

- Lubold, N. and H. Pon-Barry (2014). “Acoustic-prosodic entrainment and rapport in collaborative learning dialogues.” In: *ACM workshop on Multimodal Learning Analytics*, pp. 5–12. DOI: [10.1145/2666633.2666635](https://doi.org/10.1145/2666633.2666635).
- Lubold, N., E. Walker, and H. Pon-Barry (2016). “Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion.” In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 255–262. DOI: [10.1109/HRI.2016.7451760](https://doi.org/10.1109/HRI.2016.7451760).
- MacLeod, B. (2021). “Problems in the Difference-in-Distance measure of phonetic imitation.” In: *Journal of Phonetics* 87, p. 101058. DOI: [10.1016/j.wocn.2021.101058](https://doi.org/10.1016/j.wocn.2021.101058).
- Manson, J. H., G. A. Bryant, M. M. Gervais, and M. A. Kline (2013). “Convergence of speech rate in conversation predicts cooperation.” In: *Evolution and Human Behavior* 34.6, pp. 419–426. DOI: [10.1016/j.evolhumbehav.2013.08.001](https://doi.org/10.1016/j.evolhumbehav.2013.08.001).
- McCrae, R. R. (1994). “Openness to experience: expanding the boundaries of Factor V.” In: *European Journal of Personality* 8.4, pp. 251–272. DOI: [10.1002/per.2410080404](https://doi.org/10.1002/per.2410080404).
- McCrae, R. R. and P. T. Costa (1985). “The NEO personality inventory manual.” In: *Psychological Assessment Resources*.
- Michalsky, J. and H. Schoormann (2017). “Pitch convergence as an effect of perceived attractiveness and likability.” In: *Interspeech*. Stockholm, pp. 2253–2256. DOI: [10.21437/Interspeech.2017-1520](https://doi.org/10.21437/Interspeech.2017-1520).
- Miller, R. M., K. Sanchez, and L. D. Rosenblum (2013). “Is speech alignment to talkers or tasks?” In: *Attention, Perception, & Psychophysics* 75.8, pp. 1817–1826. DOI: [10.3758/s13414-013-0517-y](https://doi.org/10.3758/s13414-013-0517-y).
- Mitterer, H. and M. Ernestus (2008). “The link between speech perception and production is phonological and abstract: evidence from the shadowing task.” In: *Cognition* 109.1, pp. 168–173. DOI: [10.1016/j.cognition.2008.08.002](https://doi.org/10.1016/j.cognition.2008.08.002).
- Mitterer, H. and J. Müsseler (2013). “Regional accent variation in the shadowing task: evidence for a loose perception-action coupling in speech.” In: *Attention, Perception & Psychophysics* 75.3, pp. 557–575. DOI: [10.3758/s13414-012-0407-8](https://doi.org/10.3758/s13414-012-0407-8).
- Möbius, B. (1993). *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Niemeyer.
- Möhler, G. (1998). “Describing intonation with a parametric model.” In: *International Conference on Spoken Language Processing*. Sydney, pp. 2851–2854. URL: https://www.isca-speech.org/archive/archive_papers/icslp_1998/i98_0205.pdf.
- Möhler, G. and A. Conkie (1998). “Parametric modeling of intonation using vector quantization.” In: *ESCA/COCOSDA Workshop on Speech Synthesis (SSW)*. Blue Mountains, Australia, pp. 311–316. URL: https://www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_311.pdf.
- Morise, M., F. Yokomori, and K. Ozawa (2016). “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications.” In: *IEICE TRANSACTIONS on Information and Systems* 99.7, pp. 1877–1884.
- Namy, L. L., L. C. Nygaard, and D. Sauerteig (2002). “Gender differences in vocal accommodation: the role of perception.” In: *Journal of Language and Social Psychology* 21.4, pp. 422–432. DOI: [10.1177/026192702237958](https://doi.org/10.1177/026192702237958).
- Nass, C. and K. M. Lee (2001). “Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction.” In: *Journal of experimental psychology: applied* 7.3, p. 171. DOI: [10.1037/1076-898x.7.3.171](https://doi.org/10.1037/1076-898x.7.3.171).
- Nass, C. and Y. Moon (2000). “Machines and mindlessness: social responses to computers.” In: *Journal of Social Issues* 56.1, pp. 81–103. DOI: [10.1111/0022-4537.00153](https://doi.org/10.1111/0022-4537.00153).

- Nass, C., J. Steuer, and E. R. Tauber (1994). “Computers are social actors.” In: *SIGCHI Conference on Human Factors in Computing Systems*, pp. 72–78. DOI: [10.1145/191666.191703](https://doi.org/10.1145/191666.191703).
- Nguyen, N., S. Dufour, and A. Brunelière (2012). “Does imitation facilitate word recognition in a non-native regional accent?” In: *Frontiers in Psychology* 3.480. DOI: [10.3389/fpsyg.2012.00480](https://doi.org/10.3389/fpsyg.2012.00480).
- Nielsen, K. Y. (2011). “Specificity and abstractness of VOT imitation.” In: *Journal of Phonetics* 39.2, pp. 132–142. DOI: [10.1016/j.wocn.2010.12.007](https://doi.org/10.1016/j.wocn.2010.12.007).
- Olive, J., J. van Santen, B. Möbius, and C. Shih (1998). “Synthesis.” In: *Multilingual text-to-speech synthesis: the Bell Labs approach*. Ed. by R. Sproat. Dordrecht: Kluwer Academic. Chap. 7, pp. 191–228.
- Ostendorf, F. and A. Angleitner (2004). *Neo-Persönlichkeitsinventar nach Costa und McCrae: Neo-PI-R; Manual (Revidierte Fassung)*. Göttingen: Hogrefe.
- Oviatt, S., C. Darves, and R. Coulston (2004). “Toward adaptive conversational interfaces: modeling speech convergence with animated personas.” In: *ACM Transactions on Computer-Human Interaction* 11, pp. 300–328. DOI: [10.1145/1017494.1017498](https://doi.org/10.1145/1017494.1017498).
- Pardo, J. S. (2006). “On phonetic convergence during conversational interaction.” In: *Journal of the Acoustical Society of America* 119.4, pp. 2382–2393. DOI: [10.1121/1.2178720](https://doi.org/10.1121/1.2178720).
- Pardo, J. S., A. Urmanche, S. Wilman, and J. Wiener (2017). “Phonetic convergence across multiple measures and model talkers.” In: *Attention, Perception, & Psychophysics* 79.2, pp. 637–659. DOI: [10.3758/s13414-016-1226-0](https://doi.org/10.3758/s13414-016-1226-0).
- Pardo, J. S., A. Urmanche, S. Wilman, J. Wiener, N. Mason, K. Francis, and M. Ward (2018). “A comparison of phonetic convergence in conversational interaction and speech shadowing.” In: *Journal of Phonetics* 69, pp. 1–11. DOI: [10.1016/j.wocn.2018.04.001](https://doi.org/10.1016/j.wocn.2018.04.001).
- Pearson, K. (1901). “LIII. On lines and planes of closest fit to systems of points in space.” In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572. DOI: [10.1080/14786440109462720](https://doi.org/10.1080/14786440109462720).
- Pickering, M. J. and S. Garrod (2004). “Toward a mechanistic psychology of dialogue.” In: *Behavioral and Brain Sciences* 27.2, pp. 169–190. DOI: [10.1017/S0140525X04450055](https://doi.org/10.1017/S0140525X04450055).
- (2013). “An integrated theory of language production and comprehension.” In: *Behavioral and Brain Sciences* 36.4, pp. 329–347. DOI: [10.1017/s0140525x12001495](https://doi.org/10.1017/s0140525x12001495).
- (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press. ISBN: 978-1108473613.
- Plaisant, O., R. Courtois, C. Réveillère, G. A. Mendelsohn, and O. P. John (2010). “Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente avec le NEO-PI-R.” In: *Annales Médico-psychologiques* 168.2, pp. 97–106. DOI: [10.1016/j.amp.2009.09.003](https://doi.org/10.1016/j.amp.2009.09.003).
- Plaisant, O., S. Srivastava, G. A. Mendelsohn, Q. Debray, and O. P. John (2005). “Relations entre le Big Five Inventory français et le manuel diagnostique des troubles mentaux dans un échantillon clinique français.” In: *Annales Médico-psychologiques* 163.2, pp. 161–167. DOI: [10.1016/j.amp.2005.02.002](https://doi.org/10.1016/j.amp.2005.02.002).
- Raveh, E., I. Gessinger, S. Le Maguer, B. Möbius, and I. Steiner (2017a). “Investigating phonetic convergence in a shadowing experiment with synthetic stimuli.” In: *Conference on Electronic Speech Signal Processing (ESSV)*. Ed. by J. Trouvain, I. Steiner, and B. Möbius. Saarbrücken, pp. 254–261.
- Raveh, E., I. Siegert, I. Steiner, I. Gessinger, and B. Möbius (2019). “Three’s a crowd? Effects of a second human on vocal accommodation with a voice assistant.” In: *Interspeech*. Graz, pp. 4005–4009. DOI: [10.21437/Interspeech.2019-1825](https://doi.org/10.21437/Interspeech.2019-1825).

- Raveh, E., I. Steiner, and B. Möbius (2017b). “A computational model for phonetically responsive spoken dialogue systems.” In: *Interspeech*. Stockholm, pp. 884–888. DOI: [10.21437/interspeech.2017-1042](https://doi.org/10.21437/interspeech.2017-1042).
- Rawlings, D. (2001). “An exploratory factor analysis of Hartmann’s Boundary Questionnaire and an empirically-derived short version.” In: *Imagination, Cognition and Personality* 21.2, pp. 131–144. DOI: [10.2190/3xm9-1ga6-mj76-x658](https://doi.org/10.2190/3xm9-1ga6-mj76-x658).
- RCore Team (2018). *R: a language and environment for statistical computing*. Vienna, Austria. URL: <https://www.r-project.org>.
- Reeves, B. and C. Nass (1996). *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge, UK: Cambridge University Press. ISBN: 157586052X.
- Reichel, U. D., Š. Beňuš, and K. Mády (2018). “Entrainment profiles: Comparison by gender, role, and feature set.” In: *Speech Communication* 100, pp. 46–57. DOI: [10.1016/j.specom.2018.04.009](https://doi.org/10.1016/j.specom.2018.04.009).
- Reiterer, S. M. (2019). “Neuro-psycho-cognitive markers for pronunciation/speech imitation as language aptitude.” In: *Language Aptitude*. Ed. by Z. Wen, P. Skehan, A. Biedroń, S. Li, and R. L. Sparks. Routledge, pp. 277–298. DOI: [10.4324/9781315122021-14](https://doi.org/10.4324/9781315122021-14).
- RStudio Team (2016). *RStudio: Integrated development environment for R*. RStudio, Inc. Boston, MA. URL: <https://www.rstudio.com/>.
- Sanker, C. (2015). “Comparison of phonetic convergence in multiple measures.” In: *Cornell Working Papers in Phonetics and Phonology*, pp. 60–75.
- Schredl, M. (2004). *Traumerinnerung: Modelle und empirische Untersuchungen*. Marburg: Tectum Verlag.
- Schredl, M., A. Bocklage, J. Engelhardt, and T. Mingeback (2009). “Psychological boundaries, dream recall, and nightmare frequency: a new Boundary Personality Questionnaire (BPQ).” In: *International Journal of Dream Research* 2.1, pp. 12–19. DOI: [10.11588/ijodr.2009.1.162](https://doi.org/10.11588/ijodr.2009.1.162).
- Schredl, M. and H. Engelhardt (2001). “Dreaming and psychopathology: dream recall and dream content of psychiatric inpatients.” In: *Sleep and Hypnosis* 3.1. no DOI found, pp. 44–54.
- Schredl, M. and D. Erlacher (2004). “Lucid dreaming frequency and personality.” In: *Personality and Individual Differences* 37.7, pp. 1463–1473. DOI: [10.1016/j.paid.2004.02.003](https://doi.org/10.1016/j.paid.2004.02.003).
- Schweitzer, A., G. Möhler, G. Dogil, and B. Möbius (in press). “The PaIntE model of intonation.” In: *Prosodic Theory and Practice*. Ed. by J. A. Barnes and S. Shattuck-Hufnagel. In press. MIT Press.
- Schweitzer, A. and N. Lewandowski (2014). “Social factors in convergence of F1 and F2 in spontaneous speech.” In: *International Seminar on Speech Production*. Cologne.
- Schweitzer, A., N. Lewandowski, and G. Dogil (2014). “Advancing corpus-based analyses of spontaneous speech: switch to GECO!” In: *LabPhon*. Tokyo.
- Schweitzer, A., N. Lewandowski, D. Duran, and G. Dogil (2015). “Attention, please! Expanding the GECO database.” In: *International Congress of Phonetic Sciences (ICPhS)*. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0620.pdf>.
- Schweitzer, K., M. Walsh, and A. Schweitzer (2017). “To see or not to see: interlocutor visibility and likeability influence convergence in intonation.” In: *Interspeech*. Stockholm, pp. 919–923. DOI: [10.21437/Interspeech.2017-1248](https://doi.org/10.21437/Interspeech.2017-1248).
- Shen, J., R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu (2018). “Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions.” In: *2018 IEEE Inter-*

- national Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. DOI: [10.1109/ICASSP.2018.8461368](https://doi.org/10.1109/ICASSP.2018.8461368).
- Shepard, C. A., H. Giles, and B. A. Le Poire (2001). “Communication accommodation theory.” In: *The New Handbook of Language and Social Psychology*. Ed. by W. P. Robinson and H. Giles. Wiley, pp. 33–56.
- Shockley, K., L. Sabadini, and C. A. Fowler (2004). “Imitation in shadowing words.” In: *Perception & Psychophysics* 66.3, pp. 422–429. DOI: [10.3758/BF03194890](https://doi.org/10.3758/BF03194890).
- Simon, J. R. (1990). “The effects of an irrelevant directional CUE on human information processing.” In: *Advances in Psychology* 65, pp. 31–86. DOI: [10.1016/S0166-4115\(08\)61218-2](https://doi.org/10.1016/S0166-4115(08)61218-2).
- Smith, B. L., B. L. Brown, W. J. Strong, and A. C. Rencher (1975). “Effects of speech rate on personality perception.” In: *Language and Speech* 18.2, pp. 145–152. DOI: [10.1177/002383097501800203](https://doi.org/10.1177/002383097501800203).
- Staum Casasanto, L., K. Jasmin, and D. Casasanto (2010). “Virtually accommodating: Speech rate accommodation to a virtual interlocutor.” In: *32nd Annual Meeting of the Cognitive Science Society (CogSci)*. Cognitive Science Society, pp. 127–132.
- Steiner, I. and S. Le Maguer (2018). “Creating new language and voice components for the updated MaryTTS text-to-speech synthesis platform.” In: *Language Resources and Evaluation Conference (LREC)*. Miyazaki, pp. 3171–3175. URL: <http://www.lrec-conf.org/proceedings/lrec2018/summaries/1045.html>.
- Stoet, G. (2010). “PsyToolkit: A software package for programming psychological experiments using Linux.” In: *Behavior Research Methods* 42.4, pp. 1096–1104. DOI: [10.3758/brm.42.4.1096](https://doi.org/10.3758/brm.42.4.1096).
- (2017). “PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments.” In: *Teaching of Psychology* 44.1, pp. 24–31. DOI: [10.1177/0098628316677643](https://doi.org/10.1177/0098628316677643).
- Strauch, I. and B. Meier (1999). “20 Jahre Traumforschung. Bericht Nr. 46 aus der Abteilung Klinische Psychologie (pp. 66–73).” Report available from the Department of Psychology, University of Zurich.
- Strobel, A., A. Beauducel, S. Debener, and B. Brocke (2001). “Eine deutschsprachige Version des BIS/BAS-Fragebogens von Carver und White.” In: *Zeitschrift für Differentielle und diagnostische Psychologie*. DOI: [10.1024//0170-1789.22.3.216](https://doi.org/10.1024//0170-1789.22.3.216).
- Suzuki, N. and Y. Katagiri (2007). “Prosodic alignment in human-computer interaction.” In: *Connection Science* 19.2, pp. 131–141. DOI: [10.1080/09540090701369125](https://doi.org/10.1080/09540090701369125).
- Talkin, D. (1995). “A robust algorithm for pitch tracking (RAPT).” In: *Speech Coding and Synthesis*, pp. 497–518.
- Taylor, P. (2009). “Text-to-speech synthesis.” In: Cambridge University Press. Chap. 14, pp. 422–445. DOI: [10.1017/CB09780511816338](https://doi.org/10.1017/CB09780511816338).
- Thomason, J., H. V. Nguyen, and D. Litman (2013). “Prosodic entrainment and tutoring dialogue success.” In: *International Conference on Artificial Intelligence in Education*. Springer, pp. 750–753. DOI: [10.1007/978-3-642-39112-5_104](https://doi.org/10.1007/978-3-642-39112-5_104).
- Tokuda, K., H. Zen, and A. W. Black (2002). “An HMM-based speech synthesis system applied to English.” In: *IEEE Workshop on Speech Synthesis*. Santa Monica, CA, pp. 227–230. DOI: [10.1109/WSS.2002.1224415](https://doi.org/10.1109/WSS.2002.1224415).
- Trouvain, J., S. Schmidt, M. Schröder, M. Schmitz, and W. J. Barry (2006). “Modelling personality features by changing prosody in synthetic speech.” In: *International Conference on Speech Prosody*. Dresden. URL: https://www.isca-speech.org/archive/sp2006/papers/sp06_088.pdf.

- Tsang, S., C. F. Royse, and A. S. Terkawi (2017). “Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine.” In: *Saudi Journal of Anaesthesia* 11.Suppl 1, pp. 80–89. DOI: [10.4103/sja.SJA_203_17](https://doi.org/10.4103/sja.SJA_203_17).
- van den Oord, A., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu (2016). “WaveNet: a generative model for raw audio.” In: *arXiv preprint arXiv:1609.03499*.
- Wade, T., G. Dogil, H. Schütze, M. Walsh, and B. Möbius (2010). “Syllable frequency effects in a context-sensitive segment production model.” In: *Journal of Phonetics* 38.2, pp. 905–945. DOI: [10.1016/j.wocn.2009.10.004](https://doi.org/10.1016/j.wocn.2009.10.004).
- Walker, A. and K. Campbell-Kibler (2015). “Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task.” In: *Frontiers in Psychology* 6.546. DOI: [10.3389/fpsyg.2015.00546](https://doi.org/10.3389/fpsyg.2015.00546).
- Wang, Y., R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous (2017). “Tacotron: towards End-to-End Speech Synthesis.” In: *Interspeech*. Stockholm, pp. 4006–4010. DOI: [10.21437/Interspeech.2017-1452](https://doi.org/10.21437/Interspeech.2017-1452).
- Ward, A. and D. Litman (2007). “Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora.” In: *Workshop on Speech and Language Technology in Education (SLaTE)*. Farmington, PA, pp. 57–60. URL: https://www.isca-speech.org/archive_open/archive_papers/slate_2007/sle7_057.pdf.
- Weiner, I. B. and W. E. Craighead (2010). *The Corsini encyclopedia of psychology*. Vol. 2. John Wiley & Sons.
- Weise, A. and R. Levitan (2018). “Looking for structure in lexical and acoustic-prosodic entrainment behaviors.” In: *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 2*. New Orleans, LA, pp. 297–302. DOI: [10.18653/v1/n18-2048](https://doi.org/10.18653/v1/n18-2048).
- Weise, A., S. I. Levitan, J. Hirschberg, and R. Levitan (2019). “Individual differences in acoustic-prosodic entrainment in spoken dialogue.” In: *Speech Communication* 115, pp. 78–87.
- Więckowska, A. (2011). “Ego boundaries as determinants of success in foreign language learning – a state-of-the-art perspective.” In: *Acta Universitatis Wratislaviensis. Anglica Wratislaviensis XLIX*. University of Wrocław, pp. 199–208.
- Williams, J. N. (2009). “Implicit learning in second language acquisition.” In: *The new handbook of second language acquisition*. Ed. by W. C. Ritchie and T. K. Bhatia. Emerald. Chap. 14, pp. 319–353. ISBN: 978-1848552401.
- Wochner, D., J. Schlegel, N. Dehé, and B. Braun (2015). “The prosodic marking of rhetorical questions in German.” In: *Interspeech*. Dresden, pp. 987–991. URL: https://www.isca-speech.org/archive/interspeech_2015/papers/i15_0987.pdf.
- Yu, A. C. L., C. Abrego-Collier, and M. Sonderegger (2013). “Phonetic imitation from an individual-difference perspective: subjective attitude, personality and autistic traits.” In: *PloS one* 8.9, e74746. DOI: [10.1371/journal.pone.0074746](https://doi.org/10.1371/journal.pone.0074746).
- Zen, H. and T. Toda (2005). “An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005.” In: *European Conference on Speech Communication and Technology (Eurospeech)*. Lisbon. URL: <http://www.festvox.org/blizzard/bc2005/IS052192.PDF>.
- Zen, H., K. Tokuda, and A. W. Black (2009). “Statistical parametric speech synthesis.” In: *Speech Communication* 51, pp. 1039–1064.

A SHADOWING: TEXT MATERIAL

Overview of the text material used in the shadowing experiment consisting of target and filler sentences. The underlined graphemes in the target sentences correspond to the three variations of segmental pronunciation: [ɛ:]/[e:], [ɪç]/[ɪk], and [ɲ]/[ən]. The words in bold type were used for the amplitude envelope analysis.

I Target sentences

► [ɛ:] vs. [e:]

- 1) Die **Bestätigung** ist für Tanja. (*The confirmation is for Tanja.*)
- 2) Der **Schädling** sieht aber komisch aus. (*The pest looks funny though.*)
- 3) Ich mag die **Qualität** deiner Tasche. (*I like the quality of your bag.*)
- 4) Wie viel **Verspätung** hat der Zug? (*How much is the train delayed?*)
- 5) War das **Gerät** sehr teuer? (*Was the device very expensive?*)

► [ɪç] vs. [ɪk]

- 6) Es ist ganz schön **staubig** im Keller. (*It is pretty dusty in the basement.*)
- 7) Der **König** hält eine Rede. (*The king makes a speech.*)
- 8) Ich bin **süchtig** nach Schokolade. (*I am addicted to chocolate.*)
- 9) Kommt **Essig** in den Salat? (*Does vinegar go in the salad?*)
- 10) Kommt **Ludwig** heute Abend mit? (*Is Ludwig coming tonight?*)

► [ɲ] vs. [ən]

- 11) Wir **reden** ohne Unterbrechung. (*We talk without interruption.*)
- 12) Wir **besuchen** euch bald wieder. (*We will visit you again soon.*)
- 13) Sie **begleiten** dich zur Taufe. (*They accompany you to the baptism.*)
- 14) Sind die **Küchen** immer so groß? (*Are kitchens always this big?*)
- 15) Sind die **Affen** denn zutraulich? (*Are the monkeys trusting?*)

II Filler sentences

- 16) Ich hätte gern zwei kleine **Brüder**. (*I would like to have two little brothers.*)
- 17) Das **Heft** war gestern noch da. (*Yesterday, the booklet was still here.*)
- 18) Die **Glühbirne** ist leider kaputt. (*Unfortunately, the light bulb is broken.*)
- 19) Sucht sich Karin eine neue **Arbeit**? (*Is Karin looking for a new job?*)
- 20) Wird die **Wohnung** noch renoviert? (*Will the apartment still be renovated?*)
- 21) **Sara** hat eine andere Meinung. (*Sara has a different opinion.*)

- 22) Ich **täusche** mich so gut wie nie. (*I am almost never wrong.*)
- 23) Keiner glaubt diese **Geschichte**. (*No one believes this story.*)
- 24) Habt ihr das rote **Auto** erkannt? (*Did you recognize the red car?*)
- 25) Kommt Fabian auch zu dem **Fest**? (*Will Fabian also come to the party?*)
- 26) Die **Katze** weckt mich immer auf. (*The cat always wakes me up.*)
- 27) Der **Kaffee** war ja schon kalt. (*The coffee was already cold.*)
- 28) Das wird ein schönes **Geschenk**. (*This will be a nice gift.*)
- 29) Wer fliegt heute in den **Urlaub**? (*Who is going on vacation today?*)
- 30) **Warum** regt er sich denn so auf? (*Why is he getting so upset?*)

B

WOZ: EXPLAINING UTTERANCES

Utterances Mirabella uses to explain tasks 3 and 4 — played by the wizard before the respective task — and utterances she says in closing of the experiment (see *III Goodbye screen*). During the explanations, the participants are asked to provide feedback, which ensures understanding and increases interactivity. Note: tasks 1 and 2 are introduced in writing only.

I Task 3: Q&A

- Jetzt kommt die dritte Aufgabe. (*Now it's time for the third task.*)
- Bist du bereit? (*Are you ready?*)
- [*participant feedback*]
- Super! (*Great!*)
- Die Tiere haben sich in den Häusern versteckt. (*The animals hid in the houses.*)
- Wir wollen wissen, wo sie sich versteckt haben. (*We want to know where they hid.*)
- Erst frage ich dich nach einem Tier und du antwortest.
(*First I ask you about an animal and you answer.*)
- Dann fragst du mich nach einem Tier und ich antworte.
(*Then you ask me about an animal and I answer.*)
- Wir spielen zwei Runden. (*We will play two rounds.*)
- Verwende jedes Tier einmal pro Runde. (*Use each animal once per round.*)
- Die Reihenfolge ist egal. (*The order does not matter.*)
- Ich gebe mal ein Beispiel. (*Let me give you an example.*)
- “Wo hat sich der Hund versteckt?” (*Where did the dog hide?*)
- “Der Hund hat sich in Haus Nummer drei versteckt.”
(*The dog hid in house number three.*)
- Willst du das Beispiel noch einmal hören? (*Do you want to hear the example again?*)
- [*participant feedback*]
- Ok. (*Ok.*)
- Dann fangen wir jetzt an. (*Then let's get started.*)
- Ich stelle die erste Frage. (*I'll ask the first question.*)
- [*Q&A round 1*]
- Das war die erste Runde. (*That was the first round.*)
- Jetzt kommt die zweite Runde. (*The second round is coming up.*)
- Ich fange wieder an. (*I'll start again.*)
- [*Q&A round 2*]
- Das war Aufgabe 3. (*That was task 3.*)

II Task 4: map task

- Jetzt kommt die vierte Aufgabe. (*Now it's time for the fourth task.*)
- Alles klar? (*All right?*)
- [**participant feedback**]
- Wunderbar! (*Wonderful!*)
- Schau dir diese Karte an. (*Take a look at this map.*)
- Du startest im Haus und gehst den roten Weg entlang.
(*You start in the house and follow the red path.*)
- Beschreibe den Weg, den du gehst. (*Describe the path you are taking.*)
- Verwende dazu die passenden Präpositionen.
(*To do so, use the appropriate prepositions.*)
- Die Präpositionen findest du auf der rechten Seite des Bildschirms.
(*You can find the prepositions on the right side of the screen.*)
- Hast du sie gefunden? (*Did you find them?*)
- [**participant feedback**]
- Ok. (*Ok.*)
- Ich gebe mal ein Beispiel. (*Let me give you an example.*)
- “Ich gehe aus dem Haus heraus.” (*I am going out of the house.*)
- Beschreibe danach das Bild mit dem angegebenen Adjektiv.
(*Then describe the picture with the given adjective.*)
- “Das Haus ist leer.” (*The house empty.*)
- Willst du das Beispiel noch einmal hören? (*Do you want to hear the example again?*)
- [**participant feedback**]
- Ok. (*Ok.*)
- Wenn ein Bild oder ein Adjektiv versteckt ist, helfe ich dir.
(*If a picture or adjective is hidden, I'll help you.*)
- Frag mich einfach! (*Just ask me!*)
- Zum Beispiel so: (*For example, like this:*)
- “Mirabella, was ist hinter der blauen Box?” (*Mirabella, what is behind the blue box?*)
- Ok? (*Ok?*)
- [**participant feedback**]
- Ok. (*Ok.*)
- Wir gehen von Bild zu Bild. (*We'll go from image to image.*)
- Das aktuelle Bild ist immer gelb markiert.
(*The current image is always marked in yellow.*)
- Wir spielen vier Runden. (*We'll play four rounds.*)
- Ich gebe noch einmal das Beispiel. (*Let me give the example again.*)
- “Ich gehe aus dem Haus heraus.” (*I am going out of the house.*)

- “Das Haus ist leer.” (*The house empty.*)
- Willst du das Beispiel noch einmal hören? (*Do you want to hear the example again?*)
- [*participant feedback*]
- Ok. (*Ok.*)
- Dann fangen wir jetzt an. (*Then let’s get started.*)
- Wiederhole das Beispiel, um das erste Bild zu beschreiben.
(*Repeat the example to describe the first image.*)
- [*map task round 1*]
- Das war die erste Runde. (*That was the first round.*)
- Jetzt kommt die zweite Runde. (*The second round is coming up.*)
- [*map task round 2*]
- Das war die zweite Runde. (*That was the second round.*)
- Jetzt kommt die dritte Runde. (*The third round is coming up.*)
- [*map task round 3*]
- Das war die dritte Runde. (*That was the third round.*)
- Jetzt kommt die vierte Runde. (*The fourth round is coming up.*)
- [*map task round 4*]
- Das war Aufgabe 4. (*That was task 4.*)

III Goodbye screen

- Wir haben alle Aufgaben gelöst. (*We have solved all tasks.*)
- Es hat mir viel Spaß gemacht, mit dir zu arbeiten.
(*I had a lot of fun working with you.*)
- Vielen Dank, dass du teilgenommen hast! (*Thank you so much for participating!*)
- Du kannst jetzt die Kopfhörer absetzen und die Kabine verlassen.
(*You may now take off the headphones and leave the booth.*)
- Bis bald! (*See you soon!*)

C WOZ: GUIDING UTTERANCES

Utterances that are available to the wizard during the experiment in order to react spontaneously to the behavior of the participants.

Using the available utterances, the experimenter was able to manage the interaction well and respond to all events. The only statement which was missing in retrospect was “Mein Name ist Mirabella.” (*My name is Mirabella.*) in case a participant would forget this information. However, this only happened once and was clarified by the experimenter via the talkback microphone.

1. Ja. (*Yes.*)
2. Nein. (*No.*)
3. Doch. (*Yes, it is.*)
4. Ok? (*Ok?*)
5. Ok. (*Ok.*)
6. Mhm. [\rightarrow backchannel]
7. Fast! (*Almost!*)
8. Genau! (*Exactly!*)
9. Super! (*Great!*)
10. Wunderbar! (*Wonderful!*)
11. Sehr gut! (*Very good!*)
12. Weiter so! (*Keep it up!*)
13. Fast geschafft! (*Almost done!*)
14. Geschafft! (*Done!*)
15. Einen Moment. (*Just a moment.*)
16. Lass mich überlegen... (*Let me think...*)
17. Das weiß ich leider nicht. (*Unfortunately, I do not know that.*)
18. Versuch's nochmal! (*Try again!*)
19. Zurück zur Aufgabe! (*Back to the task!*)
20. Ich bin dran. (*It is my turn!*)
21. Du bist dran! (*It is your turn!*)
22. Gib die Antwort! (*Provide the answer!*)

23. Stell' eine Frage! (*Ask a question!*)
24. Sehr gute Frage! (*Great question!*)
25. Wir müssen leider trotzdem weitermachen. (*Unfortunately, we have to continue anyway.*)
26. Verwende die angegebenen Wörter! (*Use the given words!*)

► only for the Q&A

27. Verwende die gleichen Sätze wie im Beispiel. (*Use the same sentences as in the example.*)
28. Dieses Tier war schon dran. (*We already talked about this animal.*)
29. Nimm ein anderes Tier. (*Pick another animal.*)

► only for the map task

30. Was machst du jetzt? (*What are you doing next?*)
31. Beschreibe das gelb markierte Bild! (*Describe the image marked in yellow!*)
32. Frag nach der roten Box! (*Ask for the red box!*)
33. Frag nach der schwarzen Box! (*Ask for the black box!*)
34. Frag nach der lilanen Box! (*Ask for the der purple box!*)
35. Frag nach der blauen Box! (*Ask for the blue box!*)
36. Frag nach der weißen Box! (*Ask for the white box!*)
37. Frag nach der gelben Box! (*Ask for the yellow box!*)
38. Frag nach der grünen Box! (*Ask for the green box!*)
39. Frag nach der grauen Box! (*Ask for the gray box!*)

► only for the goodbye screen

40. Vielen Dank! (*Thank you very much!*)
41. Wir sind jetzt fertig. (*We are done.*)
42. Das kannst du die Versuchsleiterin fragen.
(*That is something you can ask the experimenter [female].*)
43. Das kannst du den Versuchsleiter fragen.
(*That is something you can ask the experimenter [male].*)
44. Du kannst die Kopfhörer absetzen. (*You may put down the headphones.*)
45. Du kannst die Kabine verlassen. (*You may leave the booth.*)

D

WOZ: TARGET AND FILLER WORDS

Overview of the 71 target and filler words presented in task 1. The ten animals (see *II Filler words*) are used in the Q&A game. With the exception of *Affe* and *Hase*, all words are used in the map task. The target words contain the allophonic contrasts [ɛ:]/[e:] and [ɪç]/[ɪk]. Corresponding graphemes are set in bold.

I Target words

► [ɛ:] vs. [e:]

- 1 Sä**g**e (*saw*)
- 2 Mä**d**chen (*girl*)
- 3 Kä**f**er (*beetle*)
- 4 B**ä**r (*bear*)
- 5 Universit**ä**t (*university*)
- 6 Kä**s**e (*cheese*)
- 7 J**ä**ger (*hunter*)
- 8 Gl**ä**ser (*glass, pl.*)
- 9 vers**p**ätet (*delayed*)
- 10 **ä**hnlich (*similar*)
- 11 gef**ä**hrlich (*dangerous*)
- 12 gew**ä**hlt (*elected*)

► [ɪç] vs. [ɪk]

- 1 Kön**ig** (*king*)
- 2 Hon**ig** (*honey*)
- 3 mut**ig** (*brave*)
- 4 schatt**ig** (*shady*)
- 5 schmutz**ig** (*dirty*)
- 6 vorsicht**ig** (*cautious*)
- 7 hung**rig** (*hungry*)
- 8 lust**ig** (*funny*)
- 9 traur**ig** (*sad*)
- 10 neugier**ig** (*curious*)
- 11 bill**ig** (*cheap*)
- 12 biss**ig** (*likely to bite*)

II Filler words

- Pferd (*horse*)
- Fisch (*fish*)
- Kuh (*cow*)
- Maus (*mouse*)
- Hund (*dog*)
- Katze (*cat*)
- Löwe (*lion*)
- Vogel (*bird*)
- Hase (*rabbit*)
- Affe (*monkey*)
- Haus (*house*)
- Baum (*tree*)
- Autos (*car, pl.*)
- Kuchen (*cake*)
- Bahnhof (*train station*)
- Bus (*bus*)
- Apfelsaft (*apple juice*)
- Blumen (*flower, pl.*)
- Zwillinge (*twin, pl.*)
- See (*lake*)
- Flughafen (*airport*)
- Computer (*computer*)

- Wald (*forest*)
- Politiker (*politician*)
- Museum (*museum*)
- leer (*empty*)
- schwer (*heavy*)
- schlau (*smart*)
- laut (*loud*)
- müde (*tired*)
- rund (*round*)
- neu (*new*)
- kalt (*cold*)
- berühmt (*famous*)
- wild (*wild*)
- schön (*beautiful*)
- groß (*big*)
- teuer (*expensive*)
- alt (*old*)
- gesund (*healthy*)
- nass (*wet*)
- modern (*modern*)
- klein (*small*)
- dunkel (*dark*)
- süß (*sweet*)
- sauber (*clean*)
- interessant (*interesting*)

E WOZ: QUESTIONS FROM FRAGMENTS AND ANSWERS

Questions to be formulated by the participants in task 2 with the provided fragments (•) and corresponding answers given by Mirabella (○).

- Wann hat Italien den Euro eingeführt?
(*When did Italy introduce the Euro?*)
 - Italien hat den Euro 1999 eingeführt.
(*Italy introduced the Euro in 1999.*)
- Was ist die Hauptstadt von Lettland?
(*What is the capital of Latvia?*)
 - Die Hauptstadt von Lettland is Riga.
(*The capital of Latvia is Riga.*)
- Wo sind die Brüder Grimm geboren?
(*Where were the Brothers Grimm born?*)
 - Die Brüder Grimm sind in Hanau geboren.
(*The Grimm brothers were born in Hanau.*)
- Wer war die erste Frau im Weltall?
(*Who was the first woman in space?*)
 - Walentina Tereschkowa war die erste Frau im Weltall.
(*Valentina Tereshkova was the first woman in space.*)
- Wie viele Tage hat der August?
(*How many days are in August?*)
 - Der August hat 31 Tage.
(*August has 31 days.*)

F

WOZ: MAP TASK PREPOSITIONS

The prepositions used in task 4 govern either the accusative case [ACC] or the dative case [DAT].

- um [ACC] herum (*around*)
- aus [DAT] heraus (*out of*)
- in [ACC] hinein (*into*)
- an [DAT] vorbei (*past*)
- durch [ACC] hindurch (*through*)

G WOZ: MAPS

Task 4 maps with and without boxes hiding target objects and adjectives. All drawings by Christine Mangold.

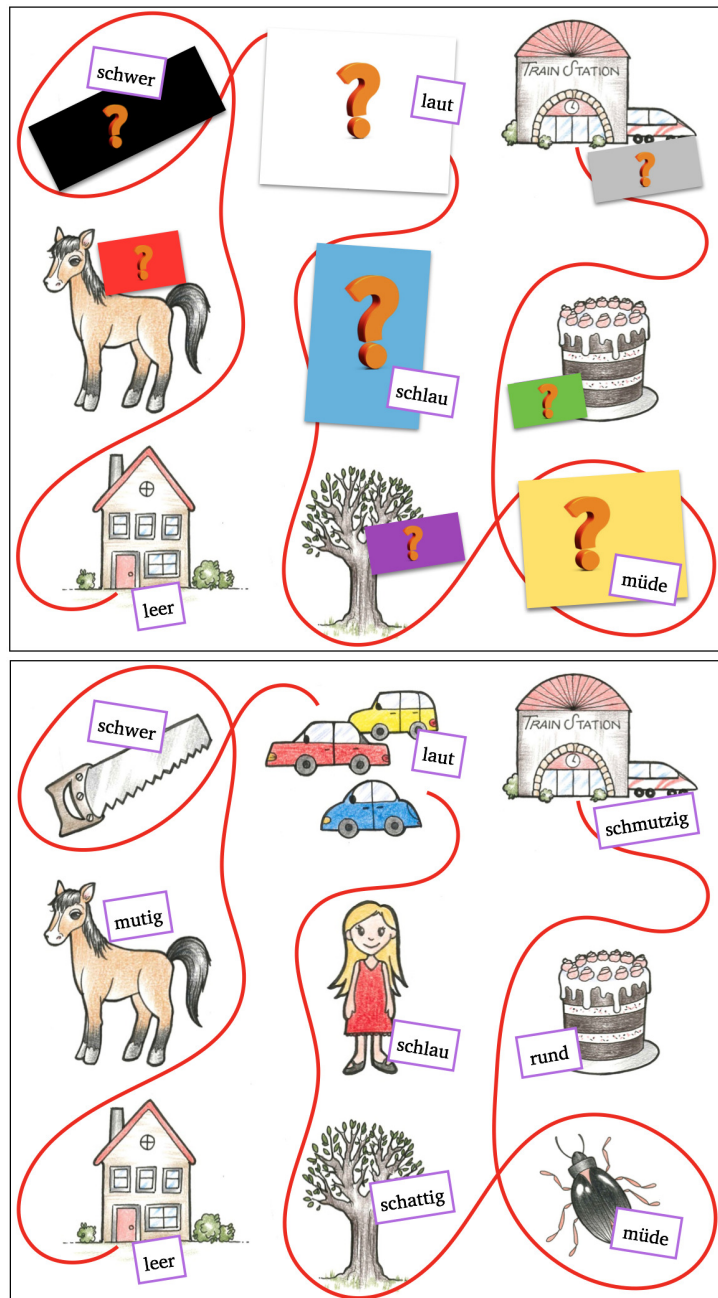


Figure 32: Map 1 — with and without boxes.

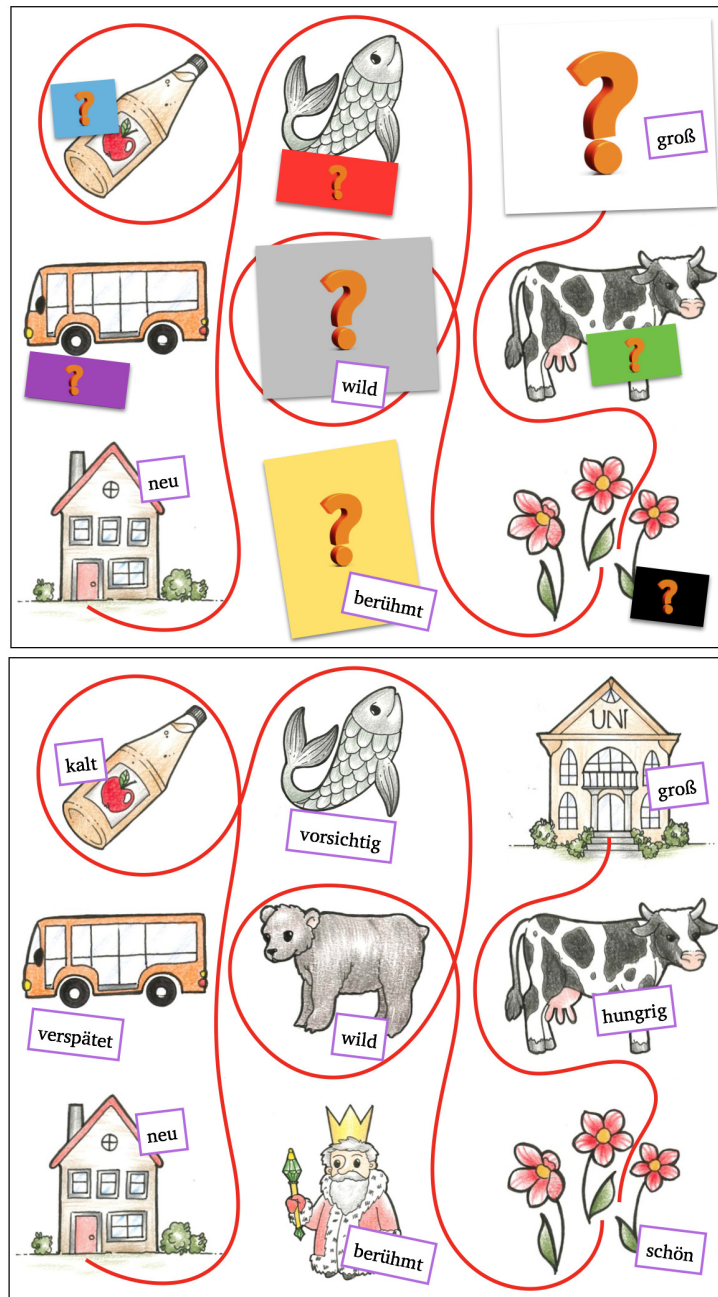


Figure 33: Map 2 — with and without boxes.

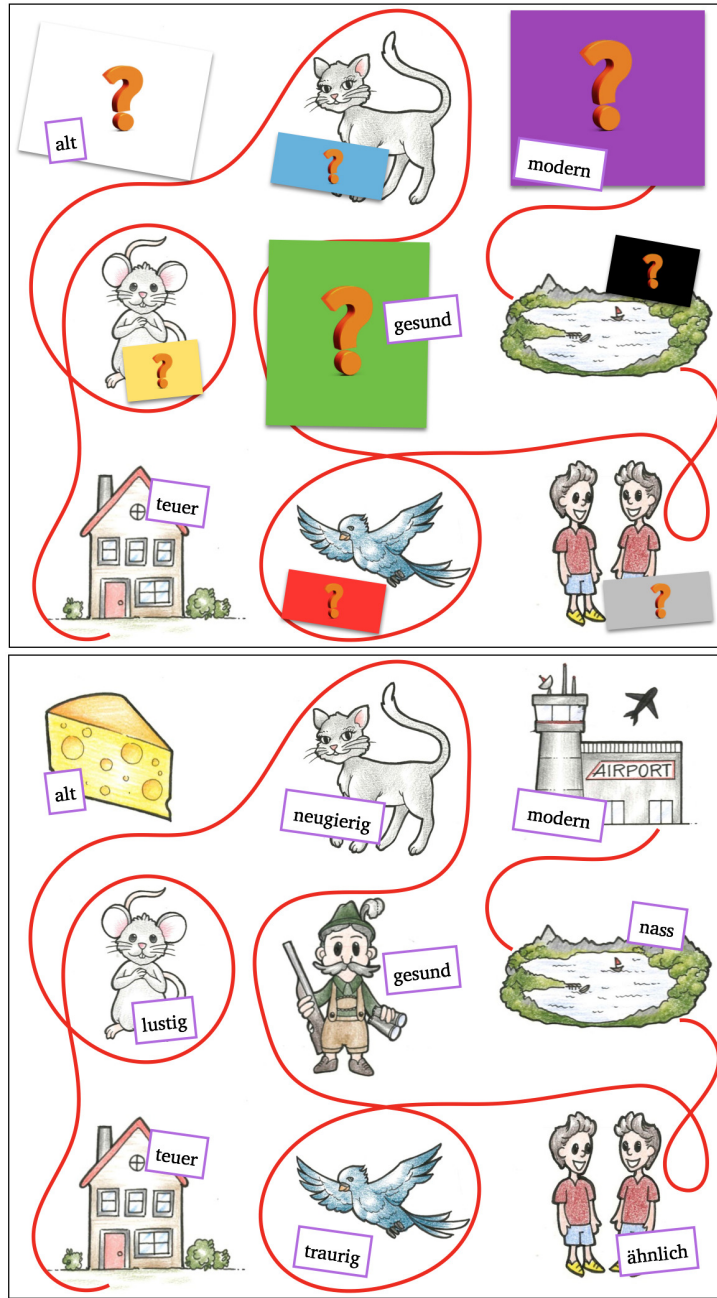


Figure 34: Map 3 — with and without boxes.

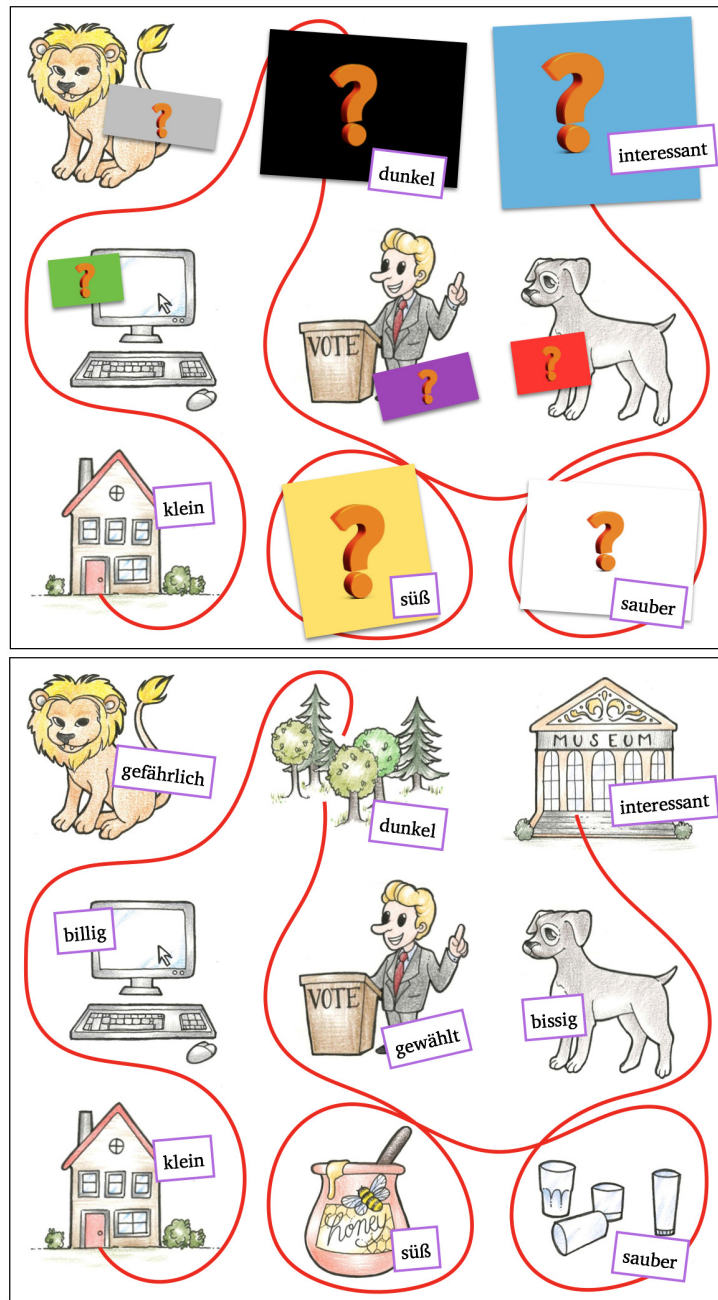


Figure 35: Map 4 — with and without boxes.

H WOZ: INDIVIDUAL DIFFERENCE IN DISTANCE DISTRIBUTIONS

Individual distributions of difference in Euclidean distance in the F1–F2 space (in Hz) between participant realizations of ⟨-ä-⟩ and the respective realizations by Mirabella in the baseline compared to the map task. Positive values indicate convergence, negative values divergence. Participants have a baseline preference for either [ɛ:] ■ or [e:] ■ and are ordered by increasing median of their individual distribution. For the participants marked in **green** or **gray**, the difference in distance (DID) distributions differ significantly from zero according to a Wilcoxon signed rank test. For the participants marked in **green** or **blue**, a kernel density estimation (KDE) test showed that their productions during baseline phase and map task differed significantly from each other. Thus, a **green** highlight means that both tests yielded a significant result. The figures also show p-values that were adjusted ([corr.]) using the Benjamini–Hochberg procedure, based on the number of speakers per group. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, *n.s.* $p \geq 0.05$.

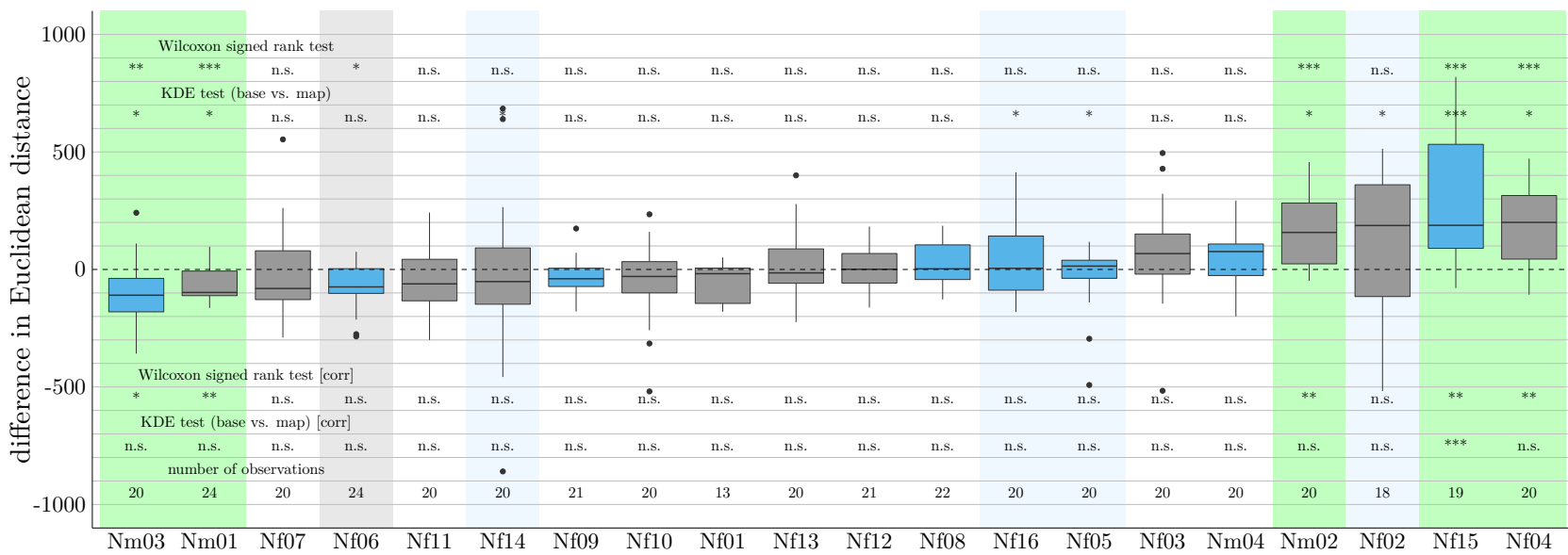


Figure 36: L1 natural group.

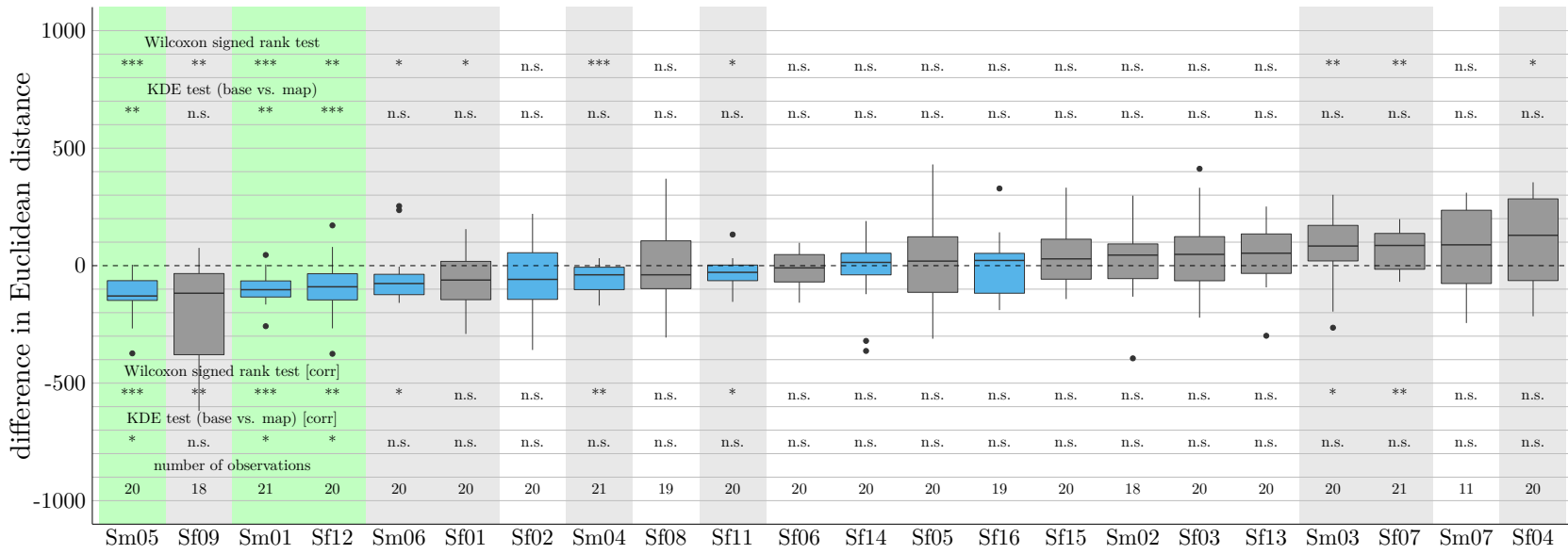


Figure 37: L1 synthetic group.

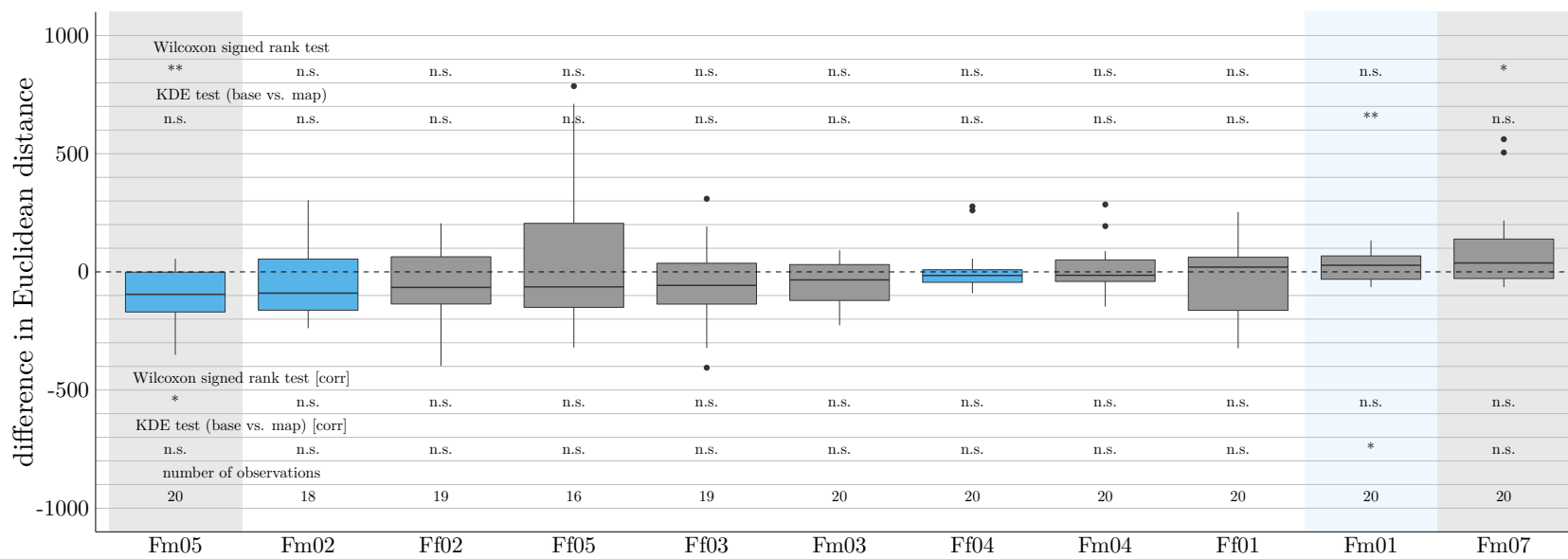


Figure 38: L2 natural group.

I THE GERMAN ADAPTATION OF THE BQ-SH: BQ-SH-G

Listed below are the 46 items of the German adaptation of the BQ-Sh (BQ-Sh-G) ordered by its six subscales “Ungewöhnliche Erfahrungen” (I, *Unusual Experiences*), “Bedürfnis nach Ordnung” (II, *Need for Order*), “Vertrauen” (III, *Trust*), “Wahrgenommene Kompetenz” (IV, *Perceived Competence*), “Kindlichkeit” (V, *Childlikeness*), and “Sensitivität” (VI, *Sensitivity*).

The number specified next to each item indicates the order of occurrence in the BQ-Sh-G. The list also contains the corresponding English items of the BQ-Sh (Rawlings, 2001) with a numbering that refers back to the original Boundary Questionnaire (BQ; Hartmann et al., 1987; Hartmann, 1989; Hartmann, 1991).

BQ-Sh item 108 was split into BQ-Sh-G items 11 and 34; BQ-Sh item 33 was split into BQ-Sh-G items 10 and 27. BQ-Sh items that have no counterpart in the BQ-Sh-G were excluded (13 and 90).

Items are to be rated on a five-point scale from 0 (“starke Ablehnung” *strong rejection*; “trifft überhaupt nicht auf mich zu” *not at all true of me*) to 4 (“starke Zustimmung” *strong agreement*; “sehr zutreffend für mich” *very true of me*). An “R” next to an item number indicates that the item must be reverse-scored (i.e., 0 = 4; 1 = 3; 3 = 1; 4 = 0).

To calculate the total BQ-Sh-G score *SumBound*:

1. reverse the score of items 1, 2, 3, 4, 5, 7, 9, 11, 12, 13, 15, 17, 23, 25, 26, 30, 32, 33, 34, 40, 41, 45, and 46
2. sum all scores of the UE, BnO^r, WK^r, Ki, and Se subscales

note: The Ve subscale is excluded from *SumBound*.

The BnO^r and WK^r scales are marked with a superscript “r” for *reversed* since the concepts they measure are negatively correlated with *SumBound* and the entire scales are therefore inversely included in its calculation. As a result, higher scores on the BnO^r scale denote a *lower* need for order, and higher scores on the WK^r scale denote *lower* perceived competence.

BQ-Sh-G		BQ-Sh	
Ungewöhnliche Erfahrungen (UE)		Unusual Experiences (UE)	
16	Jedes Mal, wenn mir etwas Furchteinflößendes passiert, habe ich Alpträume, Fantasien oder Flashbacks, die mit dem Erlebten zu tun haben.	49	Every time something frightening happens to me, I have nightmares or fantasies or flashbacks involving the frightening event.
19	Es ist mir schon passiert, dass ich nicht wusste, ob ich mir etwas nur einbilde oder ob es tatsächlich passiert.	126	I have had the experience of not knowing whether I was imagining something or it was actually happening.
21	Die Dinge um mich herum scheinen ihre Größe und Form zu ändern.	73	Things around me seem to change their size and shape.
24	Manchmal scheint mein Körper seine Größe und Form zu ändern.	83	My body sometimes seems to change its size and shape.
28	In meinen (Tag-)Träumen kommt es manchmal vor, dass sich eine Person in eine andere verwandelt.	82	In my daydreams, people kind of merge into one another or one person turns into another.
35	Ich habe Tages-Alpträume.	112	I have daymares.
36	Ich habe Träume, die ineinander übergehen.	113	I wake from one dream into another.
38	Meine Träume wirken so lebendig, dass ich sie selbst später kaum mehr von der Realität im Wachzustand unterscheiden kann.	119	My dreams are so vivid that even later I can't tell them from waking reality.
39	Ich habe oft die Erfahrung gemacht, dass verschiedene Sinne sich verbinden. Zum Beispiel hatte ich schon das Gefühl, dass ich eine Farbe riechen oder ein Geräusch sehen oder einen Geruch hören kann.	120	I have often had the experience of different senses coming together. For example, I have felt that I could smell a color, or see a sound, or hear an odor.

- 43 Es ist vorgekommen, dass jemand nach mir ruft oder meinen Namen sagt und ich mir nicht sicher war, ob es wirklich passiert ist oder ich mir das nur eingebildet habe.
- 44 In meinen (Tag-)Träumen verschwimmen irgendwie die Grenzen zwischen verschiedenen Menschen.



- 131 I have had the experience of someone calling me or speaking my name and not being sure whether it was really happening or I was imagining it.
- 92 In my dreams, people sometimes merge into each other or become other people.
- 13 I have dreams, daydreams, nightmares in which my body or someone else's body is being stabbed, injured, or torn apart.

Bedürfnis nach Ordnung (BnO^r)

Need for Order (NfO^r)

- 2R** In einer Organisation sollte jeder einen festen Platz und eine bestimmte Rolle haben.
- 7R** Es ist sehr wichtig, ordentlich und gepflegt gekleidet zu sein.
- 13R** Ich mag Geschichten, die einen klaren Anfang, Mittelteil und Schluss haben.
- 15R** Es gibt einen Platz für alles und alles sollte an seinem Platz sein.
- 22 Ich kann mir vorstellen, mit einer Person aus einer anderen ethnischen Gruppe zusammenzuleben, oder diese zu heiraten.
- 23R** Ich mag klare, eindeutige Grenzen.
- 25R** Ein guter, stabiler Rahmen ist sehr wichtig für ein Bild oder ein Gemälde.
- 26R** Ich denke, dass Kinder strenge Regeln brauchen.

- 10R** In an organization, everyone should have a definite place and a specific role.
- 23R** Being dressed neatly and cleanly is very important.
- 44R** I like stories that have a definite beginning, middle, and end.
- 48R** There is a place for everything and everything should be in its place.
- 79R** I cannot imagine living with or marrying a person of another race.
- 137R** I like clear, precise borders.
- 87R** Good solid frames are very important for a picture or a painting.
- 88R** I think children need strict discipline.

- 30R** Ich mag am liebsten Filme und Fernsehsendungen, in denen es Gute und Böse gibt und man immer weiß, wer zu welcher Kategorie gehört.
- 40R** Ein Mann ist ein Mann und eine Frau ist eine Frau; es ist sehr wichtig, diesen Unterschied beizubehalten.
- 46R** Ich mag Häuser, in denen die Räume klar definierte Wände haben und jeder Raum eine bestimmte Funktion hat.



Vertrauen (Ve)

- 1R** Ich bin vorsichtig damit, was ich Menschen erzähle, bis ich sie wirklich gut kennengelernt habe.
- 3R** Ich erwarte, dass andere Menschen eine gewisse Distanz wahren.
- 29 Es ist leicht für mich, anderen Menschen zu vertrauen.
- 31 Ich bin ein sehr offener Mensch.
- 33R** Ich bin immer zumindest ein bisschen auf der Hut.
- 37 Manchmal treffe ich jemanden und vertraue ihm oder ihr so vollkommen, dass ich beim ersten Treffen so ziemlich alles über mich mitteilen kann.

Wahrgenommene Kompetenz (WK^r)

- 4R** Ich denke, ich wäre ein guter Psychotherapeut.

- 97R** The movies and TV shows I like the best are the ones where there are good guys and bad guys and you always know who they are.
- 124R** A man is a man and a woman is a woman; it is very important to maintain that distinction.
- 140R** I like houses where rooms have definite walls and each room has a definite function.
- 90R** East is East and West is West, and never the twain shall meet. (Kipling)

Trust (Tr)

- 5R** I am careful about what I say to people until I get to know them really well.
- 17R** I expect other people to keep a certain distance.
- 95 I trust people easily.
- 103 I am a very open person.
- 107R** I am always at least a bit on my guard.
- 116 Sometimes I meet someone and trust him or her so completely that I can share just about everything about myself at the first meeting.

Perceived Competence (PC^r)

- 18R** I think I would be a good psychotherapist.

- 5R** Ich halte meinen Schreibtisch und meinen Arbeitsplatz sauber und ordentlich.
- 9R** Ich komme pünktlich zu Terminen.
- 11R** Ich bin vernünftig.
- 12R** Ich kann gut den Überblick über meine finanziellen Einnahmen und Ausgaben behalten.
- 17R** Ich habe ein gutes Gedächtnis, was meine Vergangenheit betrifft. Ich könnte problemlos erzählen, was in welchem Jahr passiert ist.
- 32R** Es gibt klare Trennlinien zwischen normalen Menschen, Menschen mit Problemen und Menschen, die als psychotisch oder verrückt angesehen werden.
- 34R** Ich bin bodenständig.
- 41R** Ich weiß genau, welche Teile der Stadt, in der ich lebe, sicher und welche unsicher sind.
- 45R** Ich habe ein gut ausgeprägtes Zeitgefühl.

Kindlichkeit (Ki)

- 6 Ein guter Lehrer muss Kinder dabei unterstützen, etwas Besonderes zu bleiben.
- 10 Kinder und Erwachsene haben viele Gemeinsamkeiten.
- 14 Ich denke, ein Künstler muss zu einem gewissen Maße Kind bleiben.
- 20 Gute Eltern müssen auch ein bisschen Kind geblieben sein.
- 27 Kinder und Erwachsene sollten sich gegenseitig erlauben, ohne strikte Rollen beisammen zu sein.

- 19R** I keep my desk and worktable neat and well organized.
- 31R** I get to appointments right on time.
- 108aR** I am a no-nonsense, [...] (*see 108b*)
- 43R** I am good at keeping accounts and keeping track of my money.
- 52R** I have a clear memory of my past. I could tell you pretty well what happened year by year.
- 105 There are no sharp dividing lines between normal people, people with problems, and people who are considered psychotic or crazy.
- 108bR** [...] down-to-earth kind of person. (*see 108a*)
- 125R** I know exactly what parts of town are safe and what parts are unsafe.
- 139R** I have a clear and distinct sense of time.

Childlikeness (Ch)

- 21 A good teacher needs to help a child remain special.
- 33a** Children and adults have a lot in common. [...] (*see 33b*)
- 45 I think an artist must in part remain a child.
- 68 A good parent has to be a bit of a child too.
- 33b** [...] They should give themselves a chance to be together without any strict roles. (*see 33a*)

42 Ich denke, dass ein guter Lehrer teilweise Kind
bleiben muss.

56 I think a good teacher must remain in part a child.

Sensitivität (Se)

8 Ich fühle mich schnell verletzt.

18 Ich bin eine sehr sensible Person.

Sensitivity (Se)

30 I am easily hurt.

54 I am a very sensitive person.

ERKLÄRUNG

Hiermit erkläre ich, dass ich die Dissertation mit dem Titel „Phonetic Accommodation of Human Interlocutors in the Context of Human-Computer Interaction“ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet.

Die Fragestellung der Dissertation wurde in Abstimmung mit meinem Betreuer, Herrn Prof. Dr. Bernd Möbius, entwickelt und dann selbstständig bearbeitet.

Ich versichere außerdem, dass ich die Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und, dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Saarbrücken, Januar 2022

Iona Gessinger