Filler particles: Phonetic details, cross-linguistic comparisons, and the recall effect

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

vorgelegt von

Beeke Muhlack
geboren in Geesthacht

Saarbrücken, im November 2023

Dekanin der Fakultät P: Prof. Dr. Stefanie Haberzettl
Erstberichterstatter: Prof. Dr. Bernd Möbius
Zweitberichterstatterin: Prof. Dr. Angelika Braun
Tag der letzten Prüfungsleistung: 15. April 2024

‘A sound often heard in the search for a hard-to-find word,’
said Harry. ‘Er ... that’d be ... er ... hang on - “er”!
“Er”’s a sound!’

J.K. Rowling,
Harry Potter and the Goblet of Fire

Acknowledgements

At last, my PhD journey comes to an end. Blood, sweat and tears went into this work and I am relieved that this chapter is coming to a close. Luckily, I was not alone during this challenging time and I want to take this space to thank some people, without whom it wouldn't have been possible.

First, I want to thank my supervisors, Bernd Möbius and Jürgen Trouvain, for their guidance, helpful feedback and open doors during this project. Thank you also to Angelika Braun for agreeing to be my external reviewer.

I also want to thank Max Mangold for leaving a great deal of money behind and the Max Mangold trustees, William Barry, Markus Groß, Bernd Möbius, and Jürgen Trouvain, for their trust in me and my visions for this project.

Furthermore, I want to thank Michael Jessen from the BKA Wiesbaden for a fruitful collaboration and support on the project.

A huge thank you goes to my fellow PhD students, Raphael Werner and Mikey Elmers, for accompanying me every step of the way, for their encouraging words, for their jokes at the right time, and for their friendship without which I probably would have given up somewhere on the way. Thank you both, for sharing all the goods parts and the more difficult parts on this journey!

Thank you to Marina Frank, for holding my hand from a distance while we both fought our way through the academic jungle. Thank you for your moral and professional support ("Kann ich dir eine Praat-Frage stellen?"), for listening to my problems, for building me up again when I was questioning this project entirely, and for having my back at all times. I would never have applied for a PhD-position if it wasn't for you. Thank you for making me believe in myself!

Thank you to Ivan Yuen and Heiner Drenhaus for the support regarding statistical questions, thank you Marjolein van Os for being there during the writing phase, cheering me on, and pushing me towards the finish line. And thank you to Omnia Ibrahim and Wei Xue for your friendship and support.

Thank you to the rest of the Phonetics department, current and former colleagues, and our students assistants who worked so diligently on the speech data from several corpora. Thank you, Danielle Kopf-Giammanco from the Writing Centre, for reading most of this thesis and pointing out typos and grammatical errors.

Thank you also to my family and friends in every corner of the world, who are too numerous to name here, but please know that every one of you helped during this journey.

Ein großer Dank geht außerdem an meine Eltern, die immer an mich glauben, egal was ich mir vornehme oder an welchen Ort ich ziehe. Ich weiß, dass ihr Kiel jederzeit dem Saarland vorgezogen hättet. Danke, dass ihr mich auf diesem Weg begleitet und unterstützt habt!

Publications

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

Muhlack, B. (2023). Filler particles in English and Spanish L1 and L2 speech. Proc. International Congress of Phonetic Sciences (ICPhS '23), 2423-2427.

Muhlack, B. and Ibrahim, O. (2023). Filler particles and pausing behaviour in Egyptian Arabic. 31. Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Zurich.

Muhlack, B., Trouvain, J., Jessen, M. (2023). Distributional and Acoustic Characteristics of Filler Particles in German with Consideration of Forensic-Phonetic Aspects. Languages 8(2), 100. https://doi.org/10.3390/languages8020100

Muhlack, B., Trouvain, J., Jessen, M. (2022a). Acoustic characteristics of filler particles in German. 30. Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Prague.

Muhlack, B., Trouvain, J., Jessen, M. (2022b). Acoustic characteristics of filler particles in German. 18th Phonetik & Phonologie (P&P '22), Bielefeld.

Muhlack, B., Elmers, M., Drenhaus, H., Trouvain, J., Van Os, M., Werner, R., Ryzhova, M., and Möbius, B. (2021a). Revisiting recall effects of filler particles in German and English. Interspeech, Brno. pp. 3979-3983, doi: 10.21437/Interspeech.2021-1056

Muhlack, B., Elmers, M., Drenhaus, H., Trouvain, J., Van Os, M., Werner, R., Ryzhova, M., and Möbius, B. (2021b). "Uh/Uhm": Conflicting Recall Effects of Filler Particles in German and English. Conference for Architectures and Mechanisms for Language Processing (AMLaP), Paris.

Muhlack, B. and Trouvain, J. (2021). Annotationsschema für Häsitationspartikeln im phonetischen Kontext. Phonetik und Phonologie im deutschsprachigen Raum (P&P), Frankfurt.

This research was funded by private funds from the Max Mangold estate and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project "Pause-internal phonetic particles in speech communication" (project number: 418659027; project IDs: MO 597/10-1 and TR 468/3-1).

As an associate member of SFB1102 "Information Density and Linguistic Encoding" and its integrated graduate program, I received valuable support and benefited greatly from the scientific training and exchange offered to the PhD students (project number: 232722074).

Abstract

This thesis focuses on one specific type of disfluency, namely filler particles (FPs). FPs occur in many languages in similar forms, which is why phonetic differences and language-specific preferences are useful to investigate for fields such as second-language learning, speech synthesis, forensic phonetics, and language analysis for the determination of origin (LADO).

This work presents a new annotation scheme that includes different types of disfluencies and breathing noises, as well as a detailed segmentation of FPs into vowel and nasal. The annotation further includes the voice quality that is being used for FPs. The empirical part of the thesis includes four studies: a detailed investigation of the phonetic characteristics of FPs in a large corpus of native German speech, an investigation of FPs in L1 and L2 English and Spanish, a pilot study on FPs and other disfluencies in Arabic, and an evaluation of the previously reported recall effect using two different designs.

In the first study, different phonetic features of FPs such as the frequency distribution, the duration, the pause context, the voice quality, and the vowel quality are investigated using a corpus of 100 native German speakers with a normal speech condition and a Lombard speech condition. Results show that glottal FPs and tongue clicks are frequent phenomena that should not be neglected in disfluency research. Considering the pause context of the FPs *uh* and *um*, there seems to be an influence on the duration of these particles affecting both types to the same degree. FPs surrounded by pauses are the longest, followed by inter-pause unit (IPU) final FPs, and IPU-initial FPs. FPs that occur within speech stretches are the shortest in duration. This finding is called a duration hierarchy. We also see that the preferred type in the IPU-initial position is *um* or *hm* rather than the vocalic FP *uh*, suggesting that this latter type is dispreferred for introducing new content. An analysis of the voice quality within FPs shows that non-modal voice is very common in FPs, as over 40% of FPs begin with glottal stops or creaky voice of varying length. A comparison between the normal and the Lombard speech condition shows that Lombard speech promotes the production of tongue clicks but discourages the use of the FPs *uh, um*, and *hm*. Furthermore, the Lombard condition influences the vowel quality of FPs, which is most likely due to the increased muscle tension that is needed for the increased vocal effort in the Lombard condition. We see that in the normal condition, the vowel quality of FPs spread over a large area of the central vowel space. In the Lombard condition, the area of spread is much smaller, and the higher first formant (F1) additionally lowers the vowel space area for FPs. We investigated speaker-specificity on the basis of 12 sample speakers. Results show a high between-speaker specificity as well as a high within-speaker specificity which may be due to the mismatch condition (normal vs. Lombard speech) and/or the high variability of this feature for each speaker.

The second empirical study investigates the frequency distribution and the vowel quality of FPs using a conversational corpus including English and Spanish speech

from native speakers as well as second-language learners. Results show that native speakers of English prefer the vocalic-nasal FP *um* in their native language (L1) and foreign language (L2), while Spanish speakers prefer the vocalic FP *uh*. A reason for this may be the preferred syllable structures of the languages as well as other phonotactic rules. The analysis of FP vowel qualities shows very distinct vowels for each language: Native English FPs include a low central vowel similar to English /ʌ/ but native Spanish FPs are produced with a high unrounded front vowel very similar to Spanish /e/. When considering the L2-speech of the same speakers, we see that some speakers are approximating the vowel quality of the target language while a few are achieving the vowel quality of FPs in a native-like manner.

The third study investigates the disfluency pattern of seven Egyptian Arabic speakers in two different tasks: A spontaneous report of their daily activities and a map task in which the participants give directions from a famous sight in the city to the university building. These tasks are considered to have different cognitive loads that are hypothesised to influence the frequency distribution of the disfluencies. Results show that the Arabic speakers of this corpus prefer the production of silent pauses and the vocalic FP *uh* over all other disfluencies. The disfluency patterns seem to be quite consistent for each speaker in the two tasks which is a promising finding for forensic phonetic casework. A higher rate for the vocalic FP could be observed in the first task compared to the second task which may be due to the less restricted topic and a higher number of options to talk about, rather than a difference in cognitive load. A wide spread can be observed when looking at the vowel quality of FPs in this speaker group. Furthermore, an overlap with the German FPs was found, suggesting a similar vowel quality for FPs in these languages. Results for the Arabic corpus, however, should be taken cautiously as the corpus is very small and includes male, as well as female speakers.

The last empirical study consists of three consecutive experiments. The recall effect of FPs was assessed, i.e., whether particles such as *uh* and *um* help in the recollection of information as has been suggested before. For this purpose, a study by Fraundorf & Watson (J. Mem. & Lang.) was partially replicated in German and in English (Experiments 1a and b), and a list experiment (Experiment 2) was designed that was aimed at finding the same beneficial memory effect of FPs. In Experiment 1a with German data, an improvement of memory was found for the fluent condition, but in Experiment 1b with English data, no significant effect of condition could be detected. When testing a list recall paradigm, again no effect of FPs on the recall of information was found. That is, neither of the experiments confirmed the existence of a recall effect for FPs which may be due to several reasons, e.g. the web-based experiment design, the phonetic characteristics of the FPs, or the low power of the experiments. In conclusion, the improved memory effect of FPs reported in the literature is more inconsistent than expected.

Ausführliche Zusammenfassung

Unflüssigkeiten treten in der natürlichen Spontansprache sehr häufig auf, unabhängig davon, welche Sprache gesprochen wird. Es wird angenommen, dass Füllpartikeln (FPs) wie *uh* und *um* im Englischen in jeder Sprache vorkommen und dass sie außerdem eine ähnliche Form haben, die aus einem Vokal und einem optionalen Nasal oder nur einem Nasalkonsonanten (Clark & Fox Tree, 2002) besteht.

Untersuchungen der phonetischen Unterschiede und sprachspezifische Präferenzen können für Bereiche wie den Spracherwerb, Zweitspracherwerb, Sprachsynthese, forensische Phonetik, die Sprachanalyse zur Herkunftsbestimmung (LADO) oder die Sprachtherapie von Nutzen sein. Detaillierte Analysen von Disfluenzen liegen für das Englische (Shriberg, 1994) und das Schwedische (Eklund, 2004) vor, und speziell von FPs für das Englische (Clark & Fox Tree, 2002) und das Deutsche (Belz, 2021).

Füllpartikeln werden häufig als *gefüllte Pausen* bezeichnet und stehen damit im Kontrast zu stillen oder ungefüllten Pausen, d.h. Pausen, in denen kein akustisches Geräusch der Sprecherin[1] hörbar ist. Dies ist jedoch nicht immer der Fall. Stille Pausen können auch Atemgeräusche, Zungenschnalzen (clicks) oder anderes nicht näher spezifiziertes artikulatorisches Material enthalten (Belz & Trouvain, 2019). Der Begriff *gefüllte Pause* suggeriert demnach, dass ein Abschnitt artikulatorischer Stille auftritt, der mit zusätzlichem artikulatorischen Material gefüllt ist. Dies ist jedoch, wie diese Arbeit bestätigen wird, nicht immer der Fall. Ein großer Teil der FPs in der Spontansprache tritt innerhalb eines ansonsten fließenden Sprachabschnitts auf und wird überhaupt nicht in der Nähe einer Pause produziert (Belz et al., 2017). Darüber hinaus scheinen verschiedene FP-Typen einen unterschiedlichen Pausenkontext zu bevorzugen, d.h. die FP *uh* tritt am häufigsten innerhalb der Rede auf und die FP *um* am häufigsten innerhalb einer Pause (Clark & Fox Tree, 2002).

Es wird berichtet, dass FPs viele Funktionen erfüllen, wie z.B. den Turn zu halten oder abzugeben, Unsicherheit auszudrücken und die Aufmerksamkeit der Hörerin zu sichern (Clark & Fox Tree, 2002; Shriberg, 1994; Maclay & Osgood, 1959; Goodwin, 1981). Sie werden jedoch hauptsächlich als Ausdruck laufender kognitiver Prozesse während der Sprachproduktion angesehen, d. h. sie werden aufgrund von Verarbeitungsproblemen produziert, wie z. B. der Suche nach einem fehlenden Wort oder einer anderen Information, der Strukturierung der kommenden Sätze oder der Organisation der eigenen Gedanken (Maclay & Osgood, 1959; Goldman-Eisler, 1957, 1958; Beattie & Butterworth, 1979). Im wissenschaftlichen Diskurs wird jedoch viel darüber diskutiert, ob FPs als Signale für den Hörer verwendet werden, z.B. um ihn über die laufenden Probleme zu informieren, oder als bloße Symptome, die von den kognitiven Problemen herrühren (O'Connell & Kowal, 2005; Clark & Fox Tree, 2002; Corley & Stewart, 2008). Da es keine schlüssigen Beweise für die Signalhypothese gibt, scheinen sich die meisten Forscherinnen einig zu sein, dass FPs als Symptome

[1] Aus Gründen der Lesbarkeit wird bei Personenbezeichnungen hier die weibliche Form gewählt. Die männliche Form ist jedoch immer mitgemeint.

kognitiver Probleme zu interpretieren sind (O'Connell & Kowal, 2005; Corley & Stewart, 2008). Nichtsdestoweniger haben mehrere Studien gezeigt, dass Hörerinnen in der Lage sind, Informationen aus FPs abzuleiten und sie zu ihrem Vorteil zu nutzen, z.B. disfluente Informationen als Hinweis auf eine diskursneue oder unbekannte Referentin (Arnold et al., 2003, 2004, 2007; Bosker et al., 2014; Brennan & Schober, 2001; Corley et al., 2007; Collard et al., 2008; Corley & Hartsuiker, 2011; Fox Tree, 1995).

Um die Daten nach unseren Bedürfnissen zu annotieren, haben wir ein Annotationsschema in Praat (Boersma & Weenink, 2022) für das Pool2010-Korpus (Jessen et al., 2005) entwickelt, das in Kapitel 3 ausführlich vorgestellt wird. Dieses Schema besteht aus fünf Tiers und einem optionalen Tier für die Annotation von Vokalen in lexikalischen Wörtern jeder Sprecherin, wie z. B. die Eckvokale, die häufig als Referenzpunkte für FP-Vokale dienen. Die fünf Disfluency-Tiers sind stillen Pausen, Atemgeräuschen und Lachen, FPs und Zungenklicks sowie Disfluency-Phänomenen wie Wiederholungen, Reparaturen, Längerungen und Abbrüchen gewidmet. Da der Schwerpunkt dieser Arbeit auf den phonetischen Merkmalen von FPs liegt, wurde ein Tier hinzugefügt, um jede FP in Vokal und Nasalkonsonant zu segmentieren, aber auch um Knarrstimme und glottale Pulse innerhalb der FPs zu annotieren. Das Ziel der Präsentation eines neuen Annotationsschemas für Unflüssigkeiten war nicht, seine breite Anwendung in der wissenschaftlichen Gemeinschaft vorzuschlagen, sondern die Details des Schemas zu präsentieren, damit Kolleginnen Aspekte auswählen können, die für ihre eigene Forschung nützlich sein könnten. In vielen Fällen ist die Annotation aller Phänomene in unserem Schema für die Forschungsfrage nicht notwendig und daher zu zeit- und kostenaufwendig, um diese aufzunehmen. Ein detaillierter Bericht über die Annotationsrichtlinien ist jedoch in jeder Studie, die Sprachdaten beinhaltet, unerlässlich.

In dieser Arbeit wurden die FPs im Deutschen näher untersucht, wobei nicht nur normale Sprache, sondern auch eine Lombard-Sprachbedingung berücksichtigt wurde, d.h. eine Produktionsmethode, bei der die Teilnehmer während einer Sprechaufgabe weißes Rauschen über Kopfhörer hören. Dies führt dazu, dass die Teilnehmer lauter und mit erhöhtem Stimmaufwand sprechen (Lombard, 1911; Jessen et al., 2005). Der einzigartige Beitrag dieser Arbeit ist die Stichprobengröße, die für die Auswertung von FPs im Deutschen berücksichtigt wird. Die Sprache von 100 männlichen deutschen Muttersprachlern wird auf die Häufigkeitsverteilung der FPs *uh, um* und *hm*, aber auch auf die weniger erforschten glottalen FPs und Zungenklicks untersucht. Außerdem werden der Pausenkontext, die Dauer, die Stimmqualität und die Vokalqualität der FPs *uh* und *um* untersucht. In einer zweiten Studie werden drei weitere Sprachen auf ihre Verwendung von FPs untersucht: Spanisch, Englisch und Arabisch. Hier werden die Häufigkeitsverteilungen und die Qualität der Vokale in den FPs untersucht. Arbeiten über spanische FPs sind eher selten, während die Disfluency-Forschung im Arabischen eine Forschungslücke darstellt, so dass diese Arbeit einen Beitrag zu einem wenig erforschten Bereich auf diesem Gebiet leisten soll. Das letzte Kapitel wendet sich einer spezifischen Funktion von FPs zu, nämlich dem vermeintlich vorteilhaften

Erinnerungseffekt, der besagt, dass Informationen, die auf eine FP folgen, besser erinnert werden als Informationen, denen keine FP vorausgegangen ist (Fraundorf & Watson, 2011a; Corley et al., 2007; Collard et al., 2008; Diachek & Brown-Schmidt, 2022). Einzigartig ist die Untersuchung der phonetischen Details der FPs in den Stimuli, die in den Studien, die den Erinnerungseffekt berichten, nicht berücksichtigt wurden. Im Folgenden werden kurze Zusammenfassungen der einzelnen Kapitel des empirischen Teils dieser Arbeit gegeben.

Kapitel 4: Analyse von Füllpartikeln im Deutschen In diesem Kapitel wird eine eingehende Analyse einiger phonetischer Merkmale von FPs im Deutschen anhand des Pool2010-Korpus (Jessen et al., 2005) beschrieben, das aus Sprachaufnahmen von 100 männlichen deutschen Muttersprachlern unter zwei Bedingungen (normale und Lombard-Sprache) zusammengesetzt ist. Es basiert auf den Veröffentlichungen Muhlack et al. (2022a), Muhlack et al. (2022b) und Muhlack et al. (2023). Untersucht werden die Häufigkeitsverteilung von FPs, glottalen FPs und Zungenklicks, sowie die Dauer, der Pausenkontext, die Stimmqualität und die Vokalqualität der typischen FPs *uh* und *um*. Die Ergebnisse zeigen, dass glottale FPs und Zungenklicks genauso häufig vorkommen wie typisch beschriebene FPs, was sie zu wichtigen Merkmalen macht, die weiter untersucht werden sollten. Die Dauer der FPs scheint mit dem Pausenkontext zusammenzuhängen, in dem die FPs vorkommen. Beide FP-Typen, *uh* und *um*, sind in gleichem Maße betroffen, d.h. FPs in Isolation sind am längsten und FPs innerhalb einer Intonationsphrase sind am kürzesten. Die FP am Ende der Äußerung sind wiederum länger als die am Anfang der Äußerung.

Vergleiche zwischen der normalen und der Lombard Sprechbedingung zeigen, dass letztere die Produktion von Zungenklicks fördert, aber die Produktion der typischen FPs *uh, um* und *hm* verringert. Ein weiterer Effekt der Lombard-Bedingung ist die Verringerung des FP-Vokalraums im Vergleich zur normalen Bedingung. Die Variation des FP-Vokals nimmt drastisch ab, während der erwartete F1-Anstieg ebenfalls beobachtet werden kann (Van Summers et al., 1988). Die Gründe hierfür sind spekulativ, aber die erhöhte Muskelspannung aufgrund der erhöhten vokalen Anstrengung könnte eine Rolle spielen (Wohlert & Hammen, 2000). Die Analyse der Sprecherspezifität von FPs, die für Bereiche wie die forensische Phonetik wichtig ist, zeigt eine hohe Variation sowohl innerhalb als auch zwischen den Sprechern. Dies deutet darauf hin, dass FPs möglicherweise nicht zuverlässig zu einem forensischen Sprachvergleich beitragen, bei dem es fraglich ist, ob zwei Hörproben derselben Sprecherin angehören. Weitere Analysen unter ähnlichen und unähnlichen Bedingungen (matched vs. mismatched Bedingung) sind notwendig.

Kapitel 5: Sprachübergreifende Analyse von Füllpartikeln Dieses Kapitel basiert auf den folgenden Publikationen: Muhlack (2023) und Muhlack & Ibrahim (2023). Kapitel 5 präsentiert Datenanalysen zur Häufigkeitsverteilung und Vokalqualität der FPs in drei verschiedenen Sprachen unter Verwendung der Korpora von Cooke

et al. (2013) und Ibrahim et al. (2020). Ziel ist es, sprachspezifische Muster der FPs näher zu beleuchten sowie sprecherspezifische Tendenzen hinsichtlich der Fähigkeit zur Anpassung der FP-Vokalqualität in einer L2 und der Unflüssigkeitsmuster einer kleinen Gruppe von Arabischen Muttersprachlerinnen darzustellen. Die Ergebnisse zur Häufigkeitsverteilung von L1-Englisch und L1-Spanisch zeigen, dass die beiden Sprachen einen unterschiedlichen Typ von FP bevorzugen. Im Englischen ist die vokalisch-nasale FP *um* häufiger, während die vokalische FP *uh* im Spanischen häufiger vorkommt. Dieser Unterschied könnte mit dem häufigsten Silbentyp der jeweiligen Sprache zusammenhängen, d. h. Sprachen, die geschlossene Silben bevorzugen (z. B. Englisch), verwenden häufiger die FP *um*, während Sprachen, die häufiger offene Silben produzieren (z. B. Spanisch), die FP *uh* bevorzugen (Crystal & House, 1990; Gabriel, 2022). Vergleiche von L1- und L2-Sprache derselben Sprecherinnen zeigen, dass der bevorzugte FP-Typ in der L2 der Sprecherinnen gleich bleibt, was für die Hypothese spricht, dass selbst fortgeschrittene Sprecherinnen einige Aspekte ihres Disfluency-Profils in ihre L2 übertragen (Cenoz, 2000). Hinsichtlich der Vokalqualität der FPs wurde festgestellt, dass FPs in L1-Englisch einen Zentralvokal verwenden, während in L1-Spanisch der ungerundete halb-geschlossene Vokal /e/ verwendet wird, was im Vergleich zu anderen Sprachen einzigartig erscheint. Die FP-Vokalqualität der gleichen Sprecherinnen in ihrer L2 zeigt eine große Variation mit der Tendenz, sich in Richtung der L1-FP der Fremdsprache zu bewegen. Dies deutet darauf hin, dass die Sprecherinnen in der Lage sind, einen akustischen Unterschied zwischen ihren L1-FP und den L2-FP wahrzunehmen, denen sie sich anzunähern versuchen (Flege, 1995).

In den Analysen der arabischen Muttersprachlerinnen wurden nicht nur FPs, sondern auch stille Pausen, Wiederholungen, Verlängerungen und lexikalische FPs berücksichtigt, da die bisherige Literatur suggeriert, dass ein Disfluency-Profil sprecherspezifisch sein könnte (McDougall & Duckworth, 2018; Braun & Rosin, 2015; Braun et al., 2023). Vergleiche von zwei Aufgaben mit sieben Sprecherinnen zeigen in der Tat, dass die Disfluency-Profile für jede Sprecherin ziemlich konstant bleiben. Darüber hinaus bevorzugen die hier untersuchten arabischen Sprecherinnen die stille Pause gegenüber allen anderen Unflüssigkeiten und die vokalische FP gegenüber der vokalisch-nasalen FP. Die Bevorzugung von *uh* gegenüber *um* spricht für den Zusammenhang zwischen dem häufigsten Silbentyp (offene Silben im Arabischen) und dem bevorzugten FP-Typ (Hamdi et al., 2005). Der Vergleich der Vokalqualität von arabischen FPs und arabischen lexikalischen Vokalen zeigt, dass die FP die Form eines zentralen Vokals hat, aber auch, dass die lexikalischen Vokale eine große akustische Variation aufweisen. Dies ist möglicherweise auf die Tatsache zurückzuführen, dass im Arabischen die meisten Vokale in unbetonten Positionen auf einen zentralen Vokal reduziert werden können (Embarki, 2013).

Vergleiche der Vokalqualität von FPs in verschiedenen Sprachen (nur L1) zeigen, dass FPs im Deutschen, Englischen und Arabischen die Form eines zentralen Vokals haben. Deutsche und arabische FPs überschneiden sich in ihrer Vokalqualität, englische FPs zeigen eine niedrigere Vokalqualität als die beiden anderen Sprachen. Die

Vokalqualität der spanischen FPs ist ein ungerundeter halb-geschlossener Vokal, der sich nicht mit der FP-Vokalqualität der anderen drei Sprachen überschneidet. Gründe dafür könnten das Fehlen eines zentralen Vokals im Vokalinventar sowie die Häufigkeit des Phonems /e/ in der Sprache sein.

Kapitel 6: Evaluation des Recall-Effekts von Füllpartikeln Dieses Kapitel basiert auf den Publikationen Muhlack et al. (2021a) und Muhlack et al. (2021b). Es werden drei Experimente vorgestellt, die sich auf die Untersuchung des vorteilhaften Recall-Effekts von FPs konzentrieren, über den in der Literatur berichtet wurde (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2011a; Diachek & Brown-Schmidt, 2022). Diesem Effekt zufolge werden Informationen, die auf eine FP folgen, besser erinnert als Informationen, denen keine FP vorausgegangen ist. Da in der genannten Literatur die phonetischen Details der FPs nicht berücksichtigt werden und wir vermuten, dass z.B. die Dauer einen Einfluss auf diesen Effekt haben könnte, wurde eine Replikationsstudie von Fraundorf & Watson (2011a) im Deutschen und Englischen durchgeführt. Ein weitere experimentelle Methode wurde entwickelt, um den Erinnerungseffekt von FPs bei Listen zu ermitteln. In den drei Experimenten haben wir FPs in unsere Stimuli aufgenommen, die der Dauer von beobachteten FPs näher kamen.

Experiment 1a auf Deutsch ergab bessere Ergebnisse für die flüssige Bedingung und eine Tendenz, dass die Bedingung mit stillen Pausen besser erinnert wird als die Bedingung mit FPs. Dies steht in starkem Kontrast zur Originalstudie. In Experiment 1b auf Englisch wurden die Effekte von FPs natürlicher Länge und übermäßig langen FPs verglichen, um zu sehen, ob die Dauer einen Einfluss auf die Erinnerung von Informationen hat. Es wurde keine Auswirkung der Bedingung festgestellt, d. h. es kann kein positiver Erinnerungseffekt berichtet werden. Auch im letzten Experiment der Reihe, bei dem eine neue Listen-Methode verwendet wurde, konnte der berichtete Erinnerungseffekt der FPs nicht nachgewiesen werden. Stattdessen zeigen die Ergebnisse, dass die Reihenfolge der Elemente in der Liste und die allgemeine Fähigkeit der Teilnehmerinnen, sich Zahlen zu merken, Einfluss auf die Erinnerung von Listenelementen haben. Es wird geschlussfolgert, dass die phonetischen Details von FPs wichtige Faktoren sind, die in zukünftigen Experimenten, die diesen Effekt untersuchen, berücksichtigt werden sollten, und dass mögliche Experimente, die statistische Power berücksichtigen sollten, da die Effektgröße sehr klein ist (Diachek & Brown-Schmidt, 2022).

Contents

Part I

Preliminaries

Chapter 1

Introduction

Disfluencies occur in natural spontaneous speech very frequently regardless of the language which is spoken. Filler particles (FPs), such as *uh* and *um* in English, are hypothesised to occur in every language, and they furthermore take a similar form, consisting of a vowel and an optional nasal or a nasal consonant only (Clark & Fox Tree, 2002). Detailed analyses of disfluencies are available in English (Shriberg, 1994) and Swedish (Eklund, 2004) and analyses specifically regarding FPs are available in English (Clark & Fox Tree, 2002) and German (Belz, 2021).

Filler particles are frequently termed *filled pauses* and are thus set in contrast with silent or unfilled pauses, i.e., pauses in which no acoustic noise emitted from the speaker is audible. However, this is not always the case. Silent pauses may include respiration noises, tongue clicks, lip smacks, or other unspecified articulatory material (Belz & Trouvain, 2019). The term *filled pause* then suggests that a stretch of articulatory silence occurs which is filled with some (articulatory) material. But this, as this work will confirm, is not always the case. A large portion of FPs in spontaneous speech occurs within an otherwise fluent stretch of speech and are not produced within the vicinity of a pause at all (Belz et al., 2017). Furthermore, different FP types seem to prefer a different pause context, i.e., the FP *uh* occurs most frequently within speech and the FP *um* most frequently within a pause (Clark & Fox Tree, 2002).

FPs are reported to serve many functions such as holding the floor, ceding the floor, expressing uncertainty, and securing attention (Clark & Fox Tree, 2002; Shriberg, 1994; Maclay & Osgood, 1959; Goodwin, 1981). However, they are mainly seen as being reflective of ongoing cognitive processes during speech production, i.e., they are produced because of processing problems such as searching for a missing word or other information, structuring the coming sentences, or organising one's thoughts (Maclay & Osgood, 1959; Goldman-Eisler, 1957, 1958; Beattie & Butterworth, 1979). Much debated in the disfluency community is, however, whether FPs are used as signals for the listener, e.g., to let them know about the ongoing troubles, or as mere

symptoms stemming from the cognitive troubles (O'Connell & Kowal, 2005; Clark & Fox Tree, 2002; Corley & Stewart, 2008). As there is inconclusive evidence for the signal hypothesis, most researchers seem to be in agreement that FPs have to be interpreted as symptoms of cognitive troubles (O'Connell & Kowal, 2005; Corley & Stewart, 2008). Nevertheless, several studies have shown that listeners are able to derive information from FPs and use them for their benefit, e.g., disfluent information hints to a discourse-new or unfamiliar referent (Arnold et al., 2003, 2004, 2007; Bosker et al., 2014; Brennan & Schober, 2001; Corley et al., 2007; Collard et al., 2008; Corley & Hartsuiker, 2011; Fox Tree, 1995).

This thesis set out to investigate FPs in German in further detail, considering not only normal speech but also a Lombard speech condition, i.e., an elicitation method in which participants hear white noise over headphones during a speech task. This causes the participants to speak louder with increased vocal effort (Lombard, 1911; Jessen et al., 2005). The unique contribution of this thesis is the sample size that is considered for evaluating FPs in German. The speech of 100 native German male speakers is analysed for the frequency distribution of the FPs *uh, um* and *hm* and the less researched glottal FP and tongue clicks. Furthermore, the pause context, duration, voice, and vowel quality of the FPs *uh* and *um* are investigated. In a second study, three more languages are investigated for their use of FPs: Spanish, English, and Arabic. Here, the frequency distribution and the FP vowel quality are examined. Work on Spanish FPs is quite rare while disfluency research in Arabic forms a research gap, so this thesis aims at contributing to an under-researched area in the field. The last chapter turns to one specific function of FPs which is the supposed beneficial recall effect, meaning that information following an FP is better remembered than information that was not preceded by an FP (Fraundorf & Watson, 2011a; Corley et al., 2007; Collard et al., 2008; Diachek & Brown-Schmidt, 2022). Unique is the examination of the phonetic details of the FPs in the stimuli that have not been regarded in the studies that report the recall effect.

Chapter 2

Background

This chapter provides an overview of different disfluency phenomena and their functions (Chapter 2.1) to contextualise the phenomenon under investigation, filler particles (FP), in its research area. Chapter 2.2 describes some hypotheses regarding FPs in detail. A short overview of the debate on whether FPs are signals or symptoms is given in Section 2.2.1. Section 2.2.2 discusses the relationship of the vowel in an FP with the vowel inventory of a language and the central vowel schwa. The matter of the theoretical consideration of the vowel quality of FPs will be investigated through the exploration of different data sets in different languages in Part II of this thesis. One hypothesised function of FPs will be explored experimentally in Chapter 6, which adds to the discussion of whether FPs serve as signals or symptoms. Fields of application will be shortly discussed in Section 2.2.3.

2.1 Theoretical classification of disfluencies

The term *fluency* plays a role in different linguistic subfields, however, it is not always used with the same meaning. While fluency in second language learning typically describes a learner's ability to speak a foreign language, when dealing with pathological speech, fluency may simply mean the ability of a patient to utter (coherent) speech (e.g., fluent vs. non-fluent aphasia) (Lickley, 2015). The aim for both second language learning and speech therapy is usually to achieve the"typical fluency" of a healthy adult native speaker. So what is fluency when dealing with non-pathological native speech, i.e., typical speech? Segalowitz (2010) distinguishes between cognitive fluency, utterance fluency, and perceptual fluency. These three terms can be matched to the following areas of speech: cognitive fluency refers to the speaker's ability to efficiently plan and structure speech, utterance fluency refers to the acoustic output of the speaker, i.e., the utterance itself, and perceptual fluency refers to the perceived fluency of a listener. These three forms of fluency all refer to a "smooth" (Lickley,

2015) functioning of a process, i.e., the absence of complications that may bring the process to a halt. While these three aspects of fluency are linked to one another, it is not necessarily the case that a problem in one of these areas originates from or leads to a problem in another area. For example, a disfluency on the utterance level may not always be perceived as such (Betz et al., 2017). This thesis will mainly be concerned with utterance fluency, investigating the frequency and quality of disfluencies on the surface level of the speech signal. A later chapter (Chapter 6) will touch on perceived fluency, not in the form of fluency ratings, but by testing the recall of information in connection with disfluencies to investigate whether disfluencies may have benefits for the listener.

Fluency in typical spontaneous speech is a concept that is rarely achieved for longer stretches of speech. Eklund (2004) found that only 50% of eight-word utterances are produced fluently and with increasing utterance length; the probability of disfluencies rises considerably. Completely fluent speech output is seen more as an unrealistic "ideal" delivery than something that is actually encountered in spontaneous speech (Lickley, 2015). What we encounter most in the realm of spontaneous speech is disfluent speech: speech that is scattered with different disfluency phenomena, which occur when the process of speech encounters problems and the smooth delivery of speech fails.

2.1.1 Form

In general, disfluency phenomena are understood to be linguistic material that do not add to the propositional message itself, hence, this speech material could be deleted without altering the message of the speaker. Clark & Fox Tree (2002) divide the speech signal into a primary track and a collateral track. The primary track includes the linguistic message itself, everything that is necessary to convey the meaning of the utterance to the listener. The collateral track includes additional signals that do not include meaning themselves but refer to "the performance itself - to timing, delays, rephrasings, mistakes, repairs, intentions to speak, and the like" (Clark & Fox Tree, 2002).

The following example, adapted from Clark & Fox Tree (2002), illustrates the distinction between the primary track and the collateral track. In Example 1, a "hypothesised intended" (Shriberg, 1994) linguistic message is given to convey meaning from a speaker to a listener. If one constituent was left out, the meaning could change or some information could be lost. The actual delivery of the utterance is given in Example 2. There are lexical and non-lexical FPs included, as well as repetitions, false starts, and silent pauses (represented by full stops). These signals on the collateral track show that the speaker has difficulty structuring the sentence (= performance), but they do not add semantic information.

(1) Mallet said something about he felt it would be a good thing if Oscar went,

(2) well, . I mean this . uh Mallet said Mallet was uh said something about uh you know he felt it would be a good thing if u:h . if Oscar went,

Examples 1 and 2 from Clark & Fox Tree (2002)

Shriberg (1994) lists the following phenomena as disfluencies: filled pauses, repetitions, false starts, and repairs. She considers "cases in which a contiguous stretch of linguistic material must be deleted to arrive at the sequence the speaker 'intended,' likely the one that would be uttered upon a request for repetition" (Shriberg, 1994). Other scholars add unfilled (or silent) pauses as a counterpart to filled pauses (Maclay & Osgood, 1959; Goldman-Eisler, 1961; Crible et al., 2022) or add other phenomena to the disfluency category such as lengthenings, truncations, explicit editing terms (*sorry, I mean*), or discourse markers (*well, like*) (Braun & Rosin, 2015; Crible et al., 2022). Levelt (1983) also classifies the FP *uh* as an editing term when it occurs in repairs after an error was made and before it is repaired. Other fields of linguistic research may add other categories or alter the inventory based on the research question or the data. For example, when investigating disfluencies in pathological or disordered speech (i.e., dysfluencies), blocks may be added as a category when working with the speech of people who stutter or repetitions may be divided into sound and syllable repetitions (Staróbole Juste & Furquim De Andrade, 2011).

In the following, I will illustrate the different disfluencies by giving examples from the Pool2010 corpus (Jessen et al., 2005) and the Diapix-FL corpus (Cooke et al., 2013). Repetitions are the reiteration of phones, syllables, words, or even bigger clauses as seen in Example 3 (Maclay & Osgood, 1959). The utterance of the example sentence may not be perceived as disfluent as there is no pause or lengthening that would slow down the speech stream. Instead, the speech material including the FP *uh* and the repetition are delivered rather fluently, which may result in perceptual fluency, even though the fluency on the utterance level fails.

(3) hier uh seh ich <u>ein ein</u> Bild von drei Gegenständen
here uh I see <u>a a</u> picture of three items
(v05_fs1, at 129 s)

False starts, which are often synonymous with truncations, refer to the utterance of syllables, words, or clauses that are abandoned at some point during the utterance (Example 4). These false starts are often followed by fresh starts, as the speaker may have encountered problems during the planning and structuring of a sentence and decided that a different formulation of the sentence would be better suited.

(4) <u>das sind mehr so so</u> - wenn man Kisten transportieren möchte
<u>these are rather like like</u> - when one wants so transport crates
(v12_fs1, at 97 s)

Repairs usually include some speech errors on the word, syllable, or phoneme level, which are then corrected closely after the error occurs, as seen in Example 5. Sometimes the repair even includes an explicit editing term such as "sorry" or "I mean".

(5) aus diesem Aufpu- Auspuff entweichen
from this ex- exhaust leak
(v06_rs1, at 23 s)

Lengthenings (Example 6) are prolongations of speech sounds, which can occur both in the onset or rhyme of a syllable, thus affecting vowels as well as consonants (Braun & Rosin, 2015; Shriberg, 1999). Lengthenings are very frequent and probably the least noticeable disfluencies as they are often overlooked even by professional annotators (Betz et al., 2017).

(6) uh das is: schon ein uraltes Mittel, um:
uh that is: a very old medium to:
(v05_fs1, at 185 s)

Silent pauses (or unfilled pauses) are stretches of speech in which no verbal material is produced by the speaker. However, this does not necessarily imply that no acoustic material is present.

In a conversation, one speaker may be producing a silent pause (by not articulating) and another speaker, i.e., the interlocutor, may be speaking instead, and thus the silent pause of one speaker is the turn of another speaker, as seen in Figure 2.1 (Trouvain & Werner, 2022). In the marked section (5-8 seconds), speaker a) is speaking while speaker b) is not uttering speech, producing a silent pause. Another pause type, specific to conversations, is the gap between two turns (see Figure 2.2), which may indeed be silent or filled by non-verbal material such as breath noises (Trouvain & Werner, 2022). During the gap marked in Figure 2.2 (5.2-9.7 seconds), acoustic material with a very low amplitude is visible on speaker a)'s channel. While this is an example of a silent pause, as both speakers are not uttering speech, the pause is not acoustically silent. This acoustic material may be background noise or some other non-verbal material produced by speaker a) such as a breath noise.

When disregarding pauses that are due to turn taking, one is left with within-turn pauses. However, not all of these pauses can be counted as a disfluency since pauses are also used as markers for syntactic-prosodic breaks (Trouvain & Werner, 2022) or as rhetorical devices (O'Connell & Kowal, 2008). As both the Pool2010 corpus (Jessen et al., 2005) and the Diapix-FL corpus (Cooke et al., 2013) are highly communicative, the participants are asked to solve tasks that involve a communication partner in both corpora – silent pauses of one speaker are often filled with feedback utterances of the other speaker. This can be seen in Figure 2.3. While speaker b) uses the prosodic break for breathing, speaker a) produces a feedback utterance that overlaps with the last syllable of speaker b)'s utterance. In Figure 2.4, on the other hand, the within-turn pause is disfluent as becomes apparent when considering the content. Speaker b) is asking a question but cannot find the appropriate word to describe what she means. While this pause would be categorised as a silent or unfilled pause, it is neither because the pause is filled with an audible breath noise. This pause type usually means that no verbal material is being uttered. Tongue clicks, lip smacks,

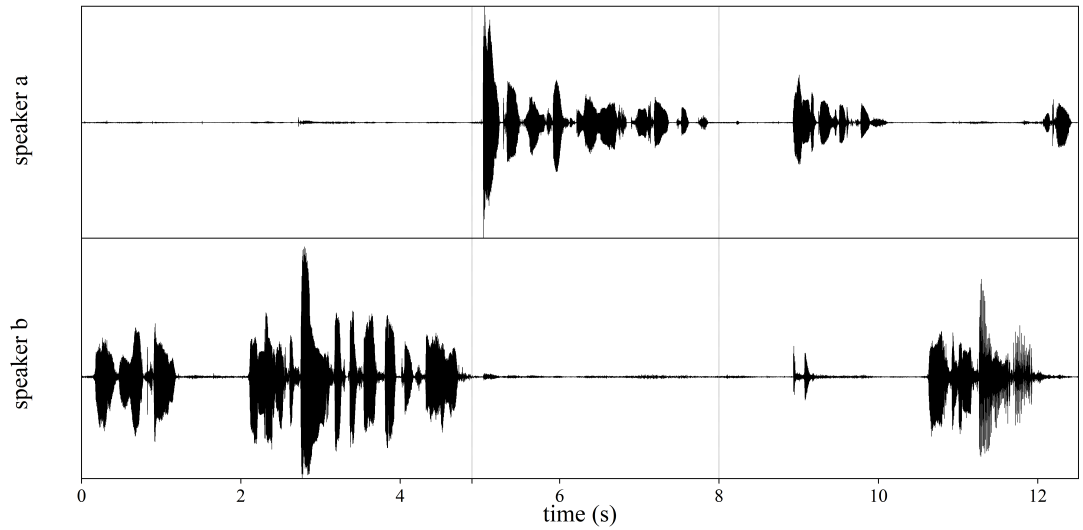

Figure 2.1: Silent pauses in turn-taking
Extract from the Diapix-FL corpus (Cooke et al., 2013) from dialogue En_spkg_En_1 (at 14s). Two speakers are shown illustrating turn-taking and the production of pauses.

and breath noises are usually discarded and pauses are labelled as "silent" or "unfilled" even though these phenomena occurred within the pause.

FPs are traditionally grouped into lexical and non-lexical FPs. Lexical FPs (*well, you know*) take the form of lexical words that could be used as "softening connectives" (Crystal & Davy, 1976) where they "maintain the continuity of discourse" (Crystal & Davy, 1976), but they may also express the attitude of the speaker or alter the intensity of the sentence (Crystal & Davy, 1976). However, they may also stem from uncertainty or difficulty in planning or processing in which case they would serve the function of hesitation similarly to other disfluencies (Crystal & Davy, 1976). Compare the use of *ja* in Example 7 with the affirmative meaning of the word. There is no question preceding the utterance of the speaker that would trigger an affirmative answer, but rather, *ja* is used to initiate a new turn and to gain more time to plan the next sentence.

(7) ja: hier geht's um ein uh
yes: this deals with a uh
(v05_rs1, at 58 s)

Non-lexical FPs can take many forms (Ward, 2006; Belz, 2021, 2023), but the most typical ones across languages include a vowel only, often a central vowel (*uh*), or a combination of a vowel and (bilabial) nasal consonant (*um*), as in Example 8 (Clark & Fox Tree, 2002). These FPs are at the centre of this work and will be described in more detail in Chapter 2.2.

(8) das Fahrzeug, das hier zu beschreiben ist, um

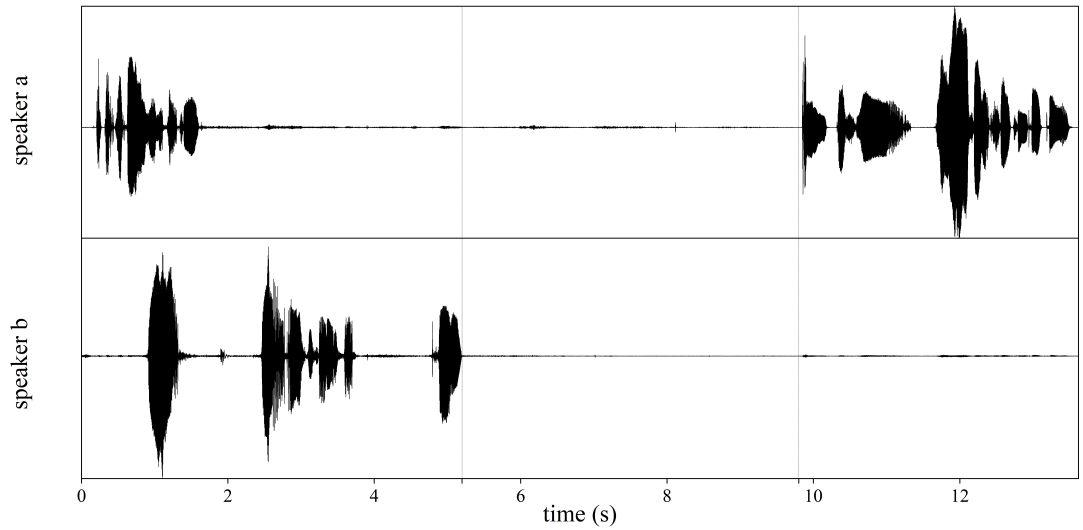

Figure 2.2: Gap in turn-taking with acoustic material at a low amplitude *Extract from the Diapix-FL corpus (Cooke et al., 2013) from dialogue En_spkg_En_1 (at 668s). Two speakers are shown illustrating the production of a gap between the turns of the two speakers.*

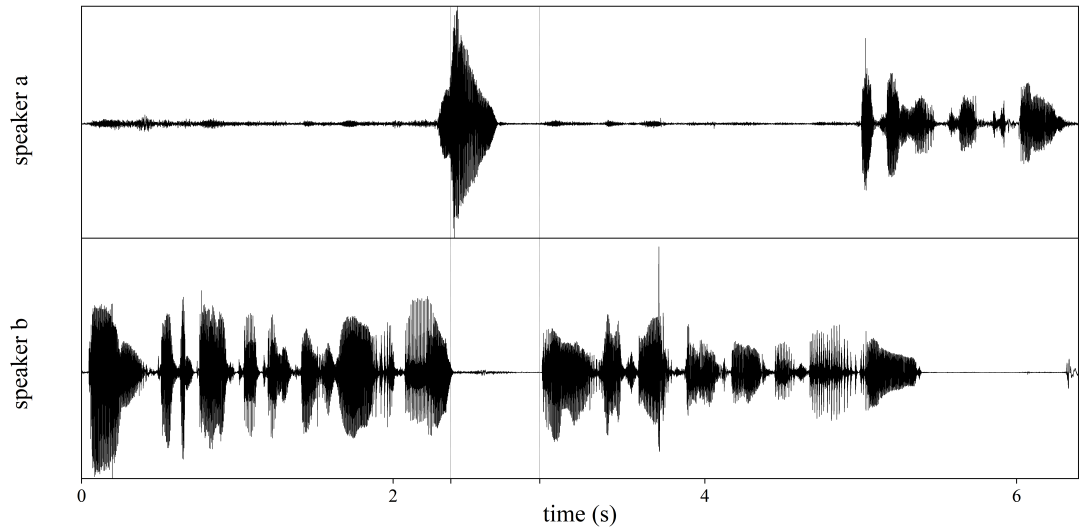

Figure 2.3: Silent pause of speaker b) filled with feedback utterance of speaker a) *Extract from the Diapix-FL corpus (Cooke et al., 2013) from dialogue En_spkg_En_1 (at 1120s). Two speakers are shown illustrating the production of a fluent within-turn pause.Speaker b) makes the pause, while speaker a) is the active listener, producing a feedback utterance. ("There is a boy stood in front of her as well (.) in like red short and a green t-shirt and")*

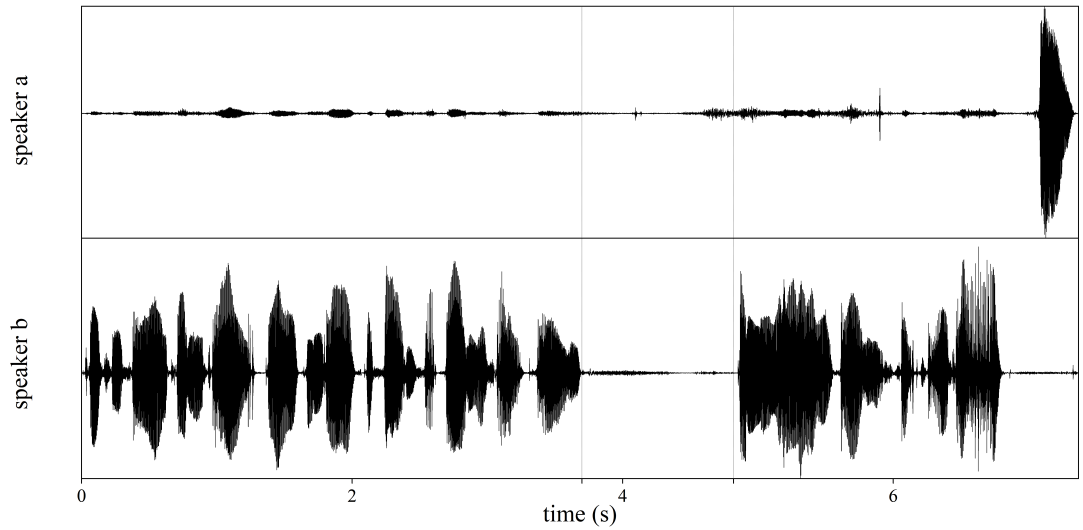

Figure 2.4: Disfluent pause by speaker b)
Extract from the Diapix-FL corpus (Cooke et al., 2013) from dialogue En_spkg_En_1 (at 716s). Two speakers are shown illustrating the production of a disfluent within-turn pause by speaker b), while speaker a) is in the listening role of the conversation. ("Did you say like on the door the open door of the farm shop that there is like a (.) on the yellow thing stuck to the door")

> the vehicle that is to be described here, um
> (v05_rs1, at 58 s)

2.1.2 Function

Many hypernyms are associated with the phenomena that are called disfluencies. Apart from *disfluency*, there are also the terms *hesitation*, *discourse marker*, or the less frequent, *fluenceme* (Götz, 2013). All these terms regard the phenomena from a specific angle. The term *disfluency* describes the phenomena on a surface level, describing the resultant failure in flow when these phenomena occur in an utterance. The terms *hesitation*, *discourse marker*, and *fluenceme*, however, attach a specific function to the phenomena in question. *Hesitation* suggests that the disfluencies, an FP or a lengthened syllable, are used to hesitate, or alternatively, that they originate from problems during the planning and structuring of the sentence. *Discourse marker*, on the other hand, refers to the structuring function of the phenomena within a larger context, e.g., a discourse, and further suggests that they are consciously used as a device to arrange the discourse and connect their parts with one another (Eklund, 2004; Swerts, 1998; Shriberg, 1994). The term *fluenceme* (Götz, 2013) proposes that these phenomena, contrary to what the term *disfluency* suggests, are used to keep up the flow of an utterance rather than inhibiting it. According to Götz (2013), "a fluenceme is a [...] feature of speech that contributes to the production and perception

of fluency."

Much research has been done to determine the function of disfluencies. One of the earlier works on disfluencies comes from Maclay & Osgood (1959) who investigated false starts, repetitions, and filled and unfilled pauses (the latter included lengthenings) in spontaneous spoken English according to the parts of speech with which they frequently occur. They found false starts usually occur with lexical items which may be due to the interference of phonologically or semantically similar words. Repetitions are usually reiterations of function words which is probably due to the same reason as just mentioned for false starts. The correct lexical item has not yet been selected so the speaker can buy time by repeating the current word or inserting an unfilled pause until the selection problem has been solved and the utterance can be resumed (Maclay & Osgood, 1959). Filled pauses occur at points of "highest uncertainty", i.e., at positions where the choice for the lexical word is most difficult or at phrase boundaries where structural planning is still ongoing and constructional and content decisions have to be made (Maclay & Osgood, 1959). Maclay & Osgood (1959) conclude that the disfluencies serve as auxiliary events that help in the sentence planning process. Thus, they function as "semantically empty" placeholders or time-buying devices until the processing troubles are overcome and the message can be resumed. This view is also supported by Eklund (2004) regarding FPs.

Goldman-Eisler' s work (1957; 1958) supports Maclay & Osgood's (1959) findings that disfluencies occur at places of highest uncertainty or with other words, lowest predictability. This is also supported by studies that prove a higher probability of disfluencies (FPs, unfilled pauses) before infrequent words, which inherently, have lower predictability (Beattie & Butterworth, 1979). Information theory (Shannon, 1948) infers a high information density (i.e., high surprisal) when predictability is low, and the finding that words with higher surprisal, i.e, lower predictability, follow the production of disfluencies has also been verified in the more recent literature (Dammalapati et al., 2019, 2021; Zámečník, 2019). Studies show that words preceding the disfluencies, in this case FPs, have a rather low surprisal value, which may be attributed to the fact that the speaker is trying to lighten the cognitive load when they anticipate upcoming processing problems (Dammalapati et al., 2019, 2021). Similarly, Zámečník (2019) found that the difference of surprisal values of the words preceding and following a disfluency is a better predictor of disfluencies compared to the surprisal of the following word alone. This attributes not only a time-buying function to disfluencies but also the reduction of the cognitive load.

Mahl (1956) studied disfluencies in patient's therapy sessions as an indicator of anxiety and he came to the conclusion that these phenomena are reliable indicators of which topics are anxiety-inducing for the patient. This may be closely connected with the placeholder function/time-buying function of disfluencies when processing problems occur. When a topic is particularly difficult to talk about for a patient, due to anxiety or trauma, the speaker may require more time to decide what to disclose to the therapist and choose a formulation that suits the topic best. Other functions

of FPs are also frequently reported in the disfluency literature including the floor-holding, but also floor-ceding function, and securing the listener's attention at the beginning of a turn (Maclay & Osgood, 1959; Clark & Fox Tree, 2002; Eklund, 2004; Sadanobu & Takubo, 1993).

While the previous studies reported the distribution of disfluencies on the surface level, several studies conducted behavioural experiments to draw conclusions regarding the impacts of disfluencies on the cognitive processing of speech. Some selected studies will be reported below.

Fox Tree (1995) compared the speech comprehension of sentences including repetitions or false starts by using a word monitoring task. In this task, subjects press a button as soon as they hear a previously given word in an aurally presented utterance. These utterances either included false starts or repetitions in the disfluent condition or no disfluencies in the fluent condition. Fox Tree (1995) found that while false starts seem to slow down reaction times, repetitions aided in the task and sped up reaction times. The same beneficial effect was found for the vocalic FP *uh* but not for the vocalic-nasal FP *um* (Fox Tree, 2001). This is also supported by Corley & Hartsuiker (2011) who found an improved word recognition effect after disfluent instructions. Considering the additional processing time and beneficial effect, the type of disfluency (filled pause, unfilled pause, or pure tone) did not matter.

Corley et al. (2007) conducted an EEG (Electroencephalography) experiment in which they presented subjects with spoken sentences that ended with either predictable or unpredictable words and were, furthermore, sometimes preceded by a disfluency (FP) and sometimes not. The resultant N400 effect, which measures final word integration, was higher for unpredictable words than for predictable words, however, this effect was reduced for unpredictable words that were preceded by a disfluency. A similar effect was found when the last words were acoustically manipulated to differ from their preceding context, and the authors concluded that the FPs oriented the listener's attention to the upcoming material, which is why the effect was reduced (= better word integration) compared to the fluent versions (Collard et al., 2008).

A number of behavioural studies by Arnold et al. show that listeners can use disfluencies to make predictions about upcoming information. In an eye-tracking study (Arnold et al., 2003, 2004), they show that disfluent instructions ("Now put theee, uh, candle below the salt shaker") lead to a higher fixation rate on discourse-new referents, i.e., items that have not been mentioned before. A following experiment (Arnold et al., 2007) suggested that disfluencies do not only make fixations on discourse-new items more probable, but it showed the same effect for unfamiliar items. Participants fixated on the unfamiliar items ("squiggly shapes") more often upon hearing disfluent instructions than on familiar items (e.g., ice cream cones). This effect was reduced in follow-up experiments when participants were told that the speaker in the experiment had trouble remembering familiar objects. Similarly, Bosker et al. (2014) found that listeners expect the speaker to refer to a low-frequency item when encountering dis-

fluencies, however, this prediction does not apply when the speaker has a non-native accent. It seems that listeners are able to use disfluencies as predictors of upcoming information and that they are able to adapt these predictions based on the current situation (Arnold et al., 2007; Bosker et al., 2014). This is also supported by Brennan & Schober (2001), who found that disfluencies help listeners delete misleading information quicker compared to the same repair that did not include an FP.

Another listener benefit has been reported, namely the influence of disfluencies (here foremost FPs) on the recognition memory of words or larger discourse elements (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2011a). Corley et al. (2007) and Collard et al. (2008) have conducted similar follow-up experiments after participants took part in an EEG experiment in which they were aurally presented with fluent and disfluent sentences. Results show that words preceded by FPs were better recalled in the follow-up recognition test than those words that occurred in fluent sentences. A different approach was taken by Fraundorf & Watson (2011a) who used short stories instead of single sentences to test participants' ability to recall crucial plot points of the story rather than recognising words. The stories included six disfluencies in the disfluent condition and none in the fluent condition. Their results show that subjects who heard the disfluent stories were better able to recall the plot of the story than subjects who heard the fluent stories. Furthermore, whether the disfluencies occurred in typical or atypical locations within the sentence did not have an effect on the recall, and listeners still benefited from their inclusion when FPs occurred in atypical locations. The authors argue that these results support the attention-orienting function of FPs rather than the predictive function or the simple addition of extra processing time (Fraundorf & Watson, 2011a).

As becomes clear from the research above, there is no unique function that one type of disfluency serves and thus no one-to-one matching of form and function. Rather, many disfluencies can serve several functions and the same function can be achieved by several disfluencies. This poses a problem for disfluency research as it may be the case that the characteristics of one type of disfluency change according to the function in which they are used. For example, the duration of FPs could depend on whether they are used as hesitation or a discourse marker. However, as assigning a function could be influenced by the characteristics of the disfluencies, the undertaking is rather circular and thus difficult to accomplish, even without taking the multi-functionality of disfluencies into account (Belz, 2021). For this reason, the detailed consideration of function lies outside of the scope of this thesis, and the investigation of the form of one type of disfluency, i.e., FPs, will be the focus of this work. One specific hypothesised function of FPs, namely the benefits of recall on the listener part, will be investigated in Chapter 6.

2.2 Filler particles

The terminological variations around disfluencies have been described in the previous section, however, there is also no terminological consensus when it comes to the phenomenon that we call a filler particle (FP). Most frequently, the term *filled pause* is encountered which may lead to confusion when only the filler particle is investigated and not the surrounding silent phases. The term *filled pause* refers to a pause that is filled with some acoustic material, but the particles that are frequently referred to are not surrounded by silent phases, so a pause in the typical meaning of the word often does not occur (O'Connell & Kowal, 2005; Belz, 2023). In the schematic representation in Figure 2.5 this problem is visualised. We see one filled pause, which is actually the gap between two speech phases. Within this gap, several pause-internal phonetic phenomena occur: one breath noise, two filler particles, a tongue click, and three short silent phases between these phenomena. Figure 2.6 shows FPs in spontaneous speech, showing that there are four pause contexts in which FPs can occur: within a fluent stretch of speech (a), after a stretch of speech and before a silence (b), between two silences in isolation (c), or after a silence and before a stretch of speech (d). Another term often used is the term *filler* (Clark & Fox Tree, 2002; Corley & Stewart, 2008; Barr & Seyfeddinipur, 2010; Braun & Rosin, 2015; Lomotey, 2021). I refrain from using this term as it is often used synonymously with the term *distractor* in psycho-linguistic studies, in which it refers to material that is included in an experiment to distract the participant from the relevant material and to conceal the purpose of the study. The term *filler particle* (FP) is preferred and will be used throughout this work.

FPs can take many forms but most of them include a vowel part and an optional nasal (Ward, 2006; Belz, 2023). These segments are usually longer than the phonemes that occur in lexical material, and they can occur in both the speech and pause context (Hughes et al., 2016; Pätzold & Simpson, 1995; Gósy & Silber-Varod, 2021; Belz, 2021). Orthographic representations depend on the language (and accent) under investigation, but the following forms have been encountered:

- *uh, u(h)m* for English; sometimes also *er, erm* (esp. for British English) (Shriberg & Lickley, 1993; Kjellmer, 2003; McDougall & Duckworth, 2017)
- *äh, ähm* for German (Klug & König, 2012; Jessen, 2012; Belz, 2021)
- *euh, eu(h)m* for French (Lo, 2020; Fuchs & Rochet-Capellan, 2021)
- *eh, e(h)m* for Spanish (Garciá-Amaya & Lang, 2020; Erker & Vidal-Covas, 2022)
- *ö(h), ö(h)m* for Hungarian (Gósy et al., 2017)

While the orthographical variation suggests acoustical differences across languages, the vowel quality of FPs in many languages is reported to be a central vowel, similar to schwa (Maclay & Osgood, 1959; Künzel, 1987; Shriberg, 1994; McDougall &

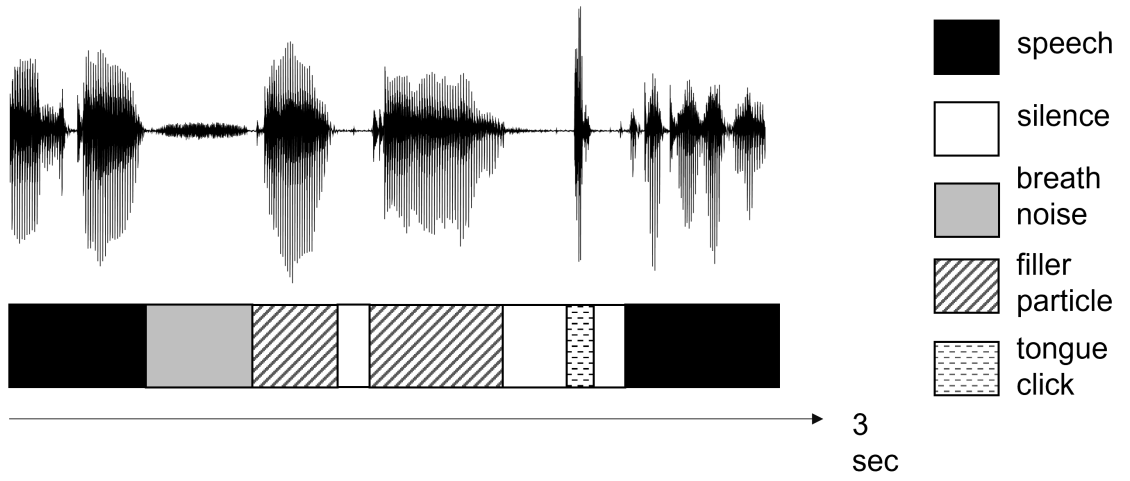

Figure 2.5: Schematic representation of pause-internal phenomena
Extract of spontaneous speech taken from the Kiel Corpus (IPDS, 2006). Bottom: schematic representation of a gap between two articulation phases which can be seen as a complex 'filled pause' including two filler particles, three silent phases, one breath noise, and a tongue click.

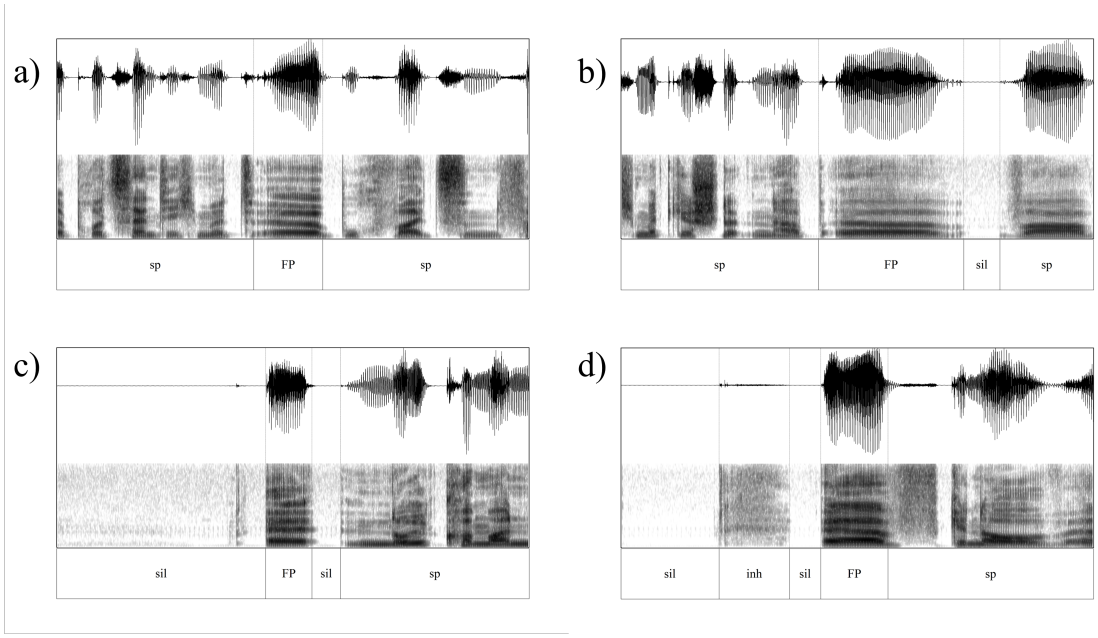

Figure 2.6: Pause contexts of FPs
2-sec sections (spectrogram: 0-8 kHz) from the Lindenstrasse corpus (speaker: l06 ch1) from IPDS (2006) showing filler particles (FP) a) with speech as left and right context (time stamp: at around 49 sec of the file), b) with speech (sp) as left and silence (sil) as right context (at 24 sec), c) with silence as left and right context (at 196 sec), and d) with silence and preceding inhalation noise (inh) as left and speech as right context (at 277 sec).

Duckworth, 2017; Pätzold & Simpson, 1995), or some other vowel quality that differs across languages:

- [ɛ] or [ɐ] for German (Belz et al., 2017; Belz, 2021; Pätzold & Simpson, 1995)
- [ʌ] for English (Shriberg, 1994)
- [ø] or [œ] for French (Candea et al., 2008; Vasilescu & Adda-Decker, 2007)
- [e] for Spanish (Erker & Bruso, 2017)
- [ɛ] for Cantonese (Cao & Mok, 2023)
- [a] (or some variant like [ɐ ʌ]) for Mandarin (Cao & Mok, 2023)

This work focuses on the FPs that consist of either a vowel, a nasal, or a combination of the two segments, which will be referred to by their orthographic forms *uh, um, hm*.

The purely nasal variant *hm* occurs in both English and German but to a lesser extent than the typical FPs *uh* and *um* (de Leeuw, 2007). The phonetic form of this FP does not necessarily match the orthographic form as two components are not usually observed in this FP but only one, namely a nasal consonant. The orthographic form *hm* (and also *mh*) is often also used for other phenomena like feedback utterances or discourse particles. Functions like a reaction signal ("What did you say?"), turn holding ("Let me think."), completion signal ("Done!"), and expressing appraisal ("This tastes/smells good!") have been described (Pistor, 2017). These discourse particles mainly vary in their intonational contour and duration. Note that the turn holding signal may not be identical to a hesitation, as the former is used intentionally while FPs are usually produced more subconsciously (Kjellmer, 2003). In line with Schmidt (2001), *hm* is considered a consonant with a closed mouth and with only intonation as the carrier of phonetic information.

Additionally, glottal FPs will be taken into account when looking at the German Pool2010 corpus (Jessen et al., 2005), as there is sufficient data to also investigate these less frequent FPs (Chapter 4). This type stands out due to its specific voice quality, that is not modal voice, but creaky voice up to the extent that it only consists of a series of glottal pulses with no parts of modal voice at all (see Figure 3.7). Belz (2017) describes this phenomenon as "glottal pulses and creak phonation without coarticulated vowels that seem to be used in a similar way to other FPs". As observed in the Pool2010 corpus (Jessen et al., 2005), glottal FPs seem to be produced with both an open and closed mouth. They may then be seen as variants of *uh* and *hm*, entirely produced with a creaky voice quality. In a study with 7 female participants, Belz (2017) found that approximately 20 % of the FPs were glottal FPs and that they did not differ in duration from the vocalic FP *uh*. In a different study with 12 female and 12 male participants, Belz (2021) found that about 5 % of the FPs were

glottal FPs and that the females produced them more often ($\sim$ 7 %) compared to males ($\sim$ 3 %).

For the same reason as for glottal FPs, tongue clicks will also be considered in the analysis of the Pool2010 corpus (Jessen et al., 2005). Tongue clicks are produced by creating a small pocket of air with the tongue against the alveolar ridge. By moving the tongue downwards, the air pocket is enlarged "and the pressure drop in the trapped air generates a short but quite strong inflow of air as the closure is released" (Clark et al., 2007), which results in the production of a click. In some languages, clicks are included as phonemes in the sound inventory, however, clicks also occur in other languages as part of the non-linguistic message. As Belz (2023) points out, those discourse clicks "are not prototypically used interchangeably with filler particles, but presumably serve different functions in dialogue". Though Belz (2023) regards clicks as 'candidates of filler particles', they are ignored in the rest of his study. In contrast to Belz (2023), clicks are taken into account in the analysis of the Pool2010 corpus (Jessen et al., 2005) (Chapter 4) to get an idea about the relative frequency of this phenomenon. However, clicks are only considered as *potential* FPs but no specific function is attributed to single instances of clicks. They can serve functions such as claiming the turn, displaying a stance, such as disapproval (e.g., tutting), or signalling difficulty during word search (Ogden, 2013). These types of clicks, especially when displaying stance, are used as intentional messages from the sender and thus clearly serve as signals (Bühler et al., 2017; Ogden, 2013). Closely related to clicks are percussives. In contrast to clicks, these are not produced deliberately but are a "by-product" of articulation; they often occur in preparation for speech when the articulators are separating (Ogden, 2013). As function is excluded for the following analyses, no distinction is made between clicks and percussives but both are grouped under tongue clicks.

The phenomena under investigation here occur frequently in spontaneous speech (Bortfeld et al., 2001; Fox Tree, 1995; Ogden, 2013). FPs are often grouped under the umbrella terms *disfluency*, *hesitation*, or even *discourse marker*. All these terms suggest a specific function of the FPs, namely the indication of production problems of some kind for the former two terms and a structuring function for the latter. Determining the function of FPs can be difficult as they may not only serve one function or the functions seem to overlap (Belz, 2021). FPs are often said to serve functions like signalling the search for a word, serving as an editing phase when repairing a speech error ("show flights from Boston - uh - Denver" (Shriberg, 1994)), holding the floor or ceding the floor, expressing uncertainty, or securing attention (Clark & Fox Tree, 2002; Shriberg, 1994; Maclay & Osgood, 1959; Goodwin, 1981). As the function of an FP is difficult to determine and heavily depends on the conversational situation or speech task, we are not taking the different possible functions into account. However, due to the task of the corpus, the FPs under consideration here are likely used to repair a speech error, search for the right word, or conceptualise the next sentence rather than for turn-taking.

2.2.1 Filler particles as linguistic signals or symptoms

A question that is prevalent when working with disfluencies is whether the phenomena are symptoms that originate from planning and processing troubles on the cognitive side or whether the phenomena function as signals that are used by the speaker to convey a certain meaning to the listener.

The concepts of signal and symptom go back to Bühler's Organon model (1965) (Braun et al., 2023). The concrete speech signal can serve three semantic functions: *symbol*, *symptom*, and *signal*. The linguistic sign is used as a symbol when it refers to objects and concepts in the outside world, it is used as a symptom when it expresses the inner state of the speaker (e.g, emotions), and it is used as a signal when it influences the listener's external or internal behaviour (Bühler, 1965). There is academic consensus that FPs are not linguistic symbols as they do not refer to objects or concepts of the outside world (like, e.g. the word "tree" does), in fact, the idea that they could be symbols never occurs in discussions revolving around the symptom or signal hypothesis (Clark & Fox Tree, 2002; O'Connell & Kowal, 2005; de Leeuw, 2007; Corley & Stewart, 2008; Belz, 2021; Braun et al., 2023). Whether FPs can be categorised as signals or symptoms, however, is under constant debate within the disfluency community.

Clark & Fox Tree (2002) argue that filler particles are used by the speaker intentionally and must therefore be linguistic signals. The speaker monitors their speech plans, and once they discover upcoming processing troubles likely to cause a delay, FPs can be used to comment on the delay and signal to the listener that the speech will continue in due course. The authors argue that FPs are common English interjections and speakers plan and produce them just as other English words and that the vocalic and the vocalic-nasal FPs signal delays of different severity (Clark & Fox Tree, 2002). While the vocalic FP *uh* signals a minor delay, the vocalic-nasal *um* signals a major delay, which Clark & Fox Tree (2002) support by the fact that the major delay signal *um* is also accompanied more often by pauses (which are longer than for minor delays). As a reason against the symptom hypothesis, they mention that speakers seem to have control over them, as trained speakers (e.g., US presidents giving their inaugural address) speak without any FPs (Kowal et al., 1997). Similarly, lecturers speak more fluently in formal speech (lectures) than in less formal face-to-face interviews (Schachter et al., 1991). However, Clark & Fox Tree's (2002) claim that FPs are signals rather than symptoms has received disapproval from the disfluency community. O'Connell & Kowal (2005) could not confirm in their data that *uh* and *um* are used in different pause contexts, but rather that in the majority of cases, the FPs are produced with no pauses surrounding them at all, which the authors take as evidence that FPs cannot signal delay. Furthermore, the duration of the pauses that occurred after the FPs did not show a large enough difference between the types (*uh* vs. *um*) for the FPs to be reliable predictors of a minor and a major delay, respectively (O'Connell & Kowal, 2005). So, if speakers did use the type of

FP to predict the gravity of a delay, they would more often be incorrect than correct (O'Connell & Kowal, 2005). Moreover, it is argued that the FPs do not behave like English interjections as they do not occur in the same positions and are also not used to convey emotions or initiate cited speech, as common interjections like *oh* do (O'Connell & Kowal, 2005). Corley & Stewart (2008) argue that there is inconclusive evidence for the assumption that FPs are, in fact, words, but it cannot be denied that they convey some information to the listener. However, the information from the FP is not that there is an upcoming delay but rather that the delay is already in progress. It seems that listeners have learned to interpret the FPs to their benefit when making predictions about the upcoming material as was shown in the previous section (Arnold et al., 2003, 2004, 2007; Bosker et al., 2014; Brennan & Schober, 2001; Corley et al., 2007; Collard et al., 2008; Corley & Hartsuiker, 2011; Fox Tree, 1995). FPs are then not necessarily signals intentionally used by speakers like words, but they are symptoms reflecting cognitive processes from which listeners have learned to derive information about the upcoming material.

2.2.2 Schwa as a predictor of vowel quality in filler particles

The concept of the central vowel schwa is often discussed in connection with the vowel quality of FPs, as many languages employ a central vowel quality in their FPs. Schwa, however, is also said to differ across languages, perhaps due to language-specific articulatory settings or the vowel inventory of the language itself. The question of how the concept of the central vowel schwa relates to the vowel quality in FPs is reviewed in the following by consulting the literature and giving examples from different languages.

The vowel inventory differs across languages, however, most languages show a symmetric vowel inventory containing five to seven vowels (Schwartz et al., 1997; Maddieson, 2005). Schwartz et al. (1997) claim that the central vowel schwa is the only vowel that does not influence the (non-)existence of other vowels, which they call the "transparency" rule. All other vowels in a vowel inventory influence each other, e.g., the lack of the high front vowel /i/ entails the lack of its rounded counterpart /y/ (Schwartz et al., 1997). It seems that the central vowel schwa occupies a special place in vowel systems, but also in phonetic research itself (Oostendorp, 2014).

The central vowel, schwa, is transcribed with the IPA symbol [ə]. The name "schwa" is a borrowed Hebrew term used to describe two phenomena regarding vowels: the silent and the mobile schwa (Laufer, 2019). The silent schwa "indicates a zero vowel sound" (Laufer, 2019) at the end of a syllable, and the mobile schwa is, as phoneticians would describe it, a diacritic that is placed below vowel symbols (at the beginning of syllables) to indicate the shortness of the vowel. The vowel quality, however, is still determined by the specific letter, not by the diacritic. The term schwa was first used in English by Peter Giles in 1895 but the symbol <ə> had been in use long before that (Gósy, 2004). The German linguist Johann Andreas Schmeller first used the symbol

in 1821 to refer to the reduced vowel as phoneticians know it today (Gósy, 2004; Laufer, 2019). In modern phonetics, the symbol <ə> and the term "schwa" are used to refer to the most central vowel in the vowel chart of the IPA. It is produced with a central tongue position, i.e., neither front nor back, and with a central tongue height, i.e., neither open-mid nor close-mid. The vowel is often reported to be unrounded (Silverman, 2011; Gósy, 2004) but as the vowel stands alone in the IPA vowel chart, it also has no rounded counterpart as is the case for many other vowels. So in this dimension, too, the schwa can be considered a central or neutral vowel.

In German, the schwa holds a special place in the vowel inventory as it is debatable, whether this sound is a true phoneme in German (Hirschfeld & Wallraff, 2002). The fact that minimal pairs can be found (*Bässe - besser, Rolle - Rollo*) speaks for the central vowel being a phoneme rather than merely an allophone of another phoneme. However, the phone [ə] only occurs in unstressed syllables, which is often mentioned when arguing against schwa's phoneme status along with its acoustic variability in context (Hirschfeld & Wallraff, 2002). Furthermore, some sounds that may appear as [ə] on the surface level may actually stem from vowel reduction from other vowel phonemes. Meinhold & Stock (1980) argue for a dual status of schwa in German: [ə] can be the realisation of the phoneme /ə/ but also the realisation of the phonemes /ɛː/, /eː/, /iː/ and /ɪ/ when these vowels undergo vowel reduction.

The schwa has a similar status in English as it occurs only in unstressed syllables or weak word forms (Roach, 2009). Monosyllabic words like *a* or *the* can be produced as strong or weak forms depending on whether they receive sentence stress or emphasis. While the strong forms of these sample words are produced with a full vowel as /eɪ/ and /ðiː/, respectively, in the weak forms the central vowel is used instead (i.e., /ə/, /ðə/).

Another question in relation to schwa is whether the vowel has a target or if it is unspecified (Cohen Priva & Strand, 2023; Bates, 1995; Gick, 2002). While Gick (2002) found that schwa is different from the articulatory rest position in American English, which he takes as evidence for an acoustic target, Bates (1995) argues that the high context dependency of schwa is evidence against schwa's target. The large acoustic variance of schwa-realisations also is evidence against schwa having a defined acoustic target (Koopmans-van Beinum, 1994). More recently, Cohen Priva & Strand (2023) replicated Bates's (1995) findings that the schwa (in American English) is more context-dependent than other vowels, however, a prolonged schwa does not occupy the same acoustic space that the "full" vowels are reduced to when their duration gets shorter. The authors speculate that another factor, e.g., dispersion theoretic pressures, may play a role here (Cohen Priva & Strand, 2023).

Barry (1995) suggests that schwa is produced with "an equilibrium in the tonus of the pro- and antagonistic muscles" and thus results in a "physiological 'relaxation target'" (Barry, 1995). As the central vowel in German and English always occurs in unstressed positions, as outlined above, its duration is considerably shorter than that of other lexical vowels which could also result in a target-undershoot (Lindblom,

1963; Bates, 1995). It may be the case, that schwa has a target, but it is never reached because of its short duration.

It was established above that the existence of schwa in a language seems to be independent of the other vowel phonemes in the language (Schwartz et al., 1997), however, the vowel quality of schwa seems to differ slightly, even if it remains a central vowel (Gósy, 2004). It may be the case that the vowel inventory influences the exact position of schwa so that the central vowel is sufficiently different from all other vowels, or equally distanced from all other vowels. The existence of other central vowels, /ɜ/ and /ɜˠ/ for British and American English and [ɐ] for German, may then influence the position of schwa. Another possibility is that the language-specific articulatory setting may influence the exact position of schwa (Gick et al., 2004).

To conclude, the central vowel schwa seems to be the vowel that is produced with the least muscular effort, as its production may require muscle relaxation rather than tension (Barry, 1995). Because of this, it serves as the perfect candidate for a filler particle, as cognitive effort is aimed to be reduced and processing time is gained with minimal articulatory effort (Dammalapati et al., 2019, 2021; Bates, 1995). The question that arises is then whether the central vowel is always used in filler particles across languages if the particular language possesses schwa in its vowel inventory as a phoneme or even only as an allophone. Moreover, which vowel quality is used in the FPs in languages that do not show a central vowel in their inventory? Is it still a central vowel, or another vowel of the vowel inventory? Possible candidates may be the vowel with the least articulatory effort or the most frequent vowel. Languages that do not include schwa in their vowel inventory, such as Spanish, Italian, and Japanese, may shed more light on the phenomenon.

Schwa is not used as a phoneme in Standard European Spanish, and it does not occur as an allophone of the full vowels as it does in other languages, due to vowel reduction in unstressed syllables (Gabriel, 2022; Roach, 2009; Hirschfeld & Wallraff, 2002). The vowel quality used in native Spanish FPs is often reported to be close to the front vowel /eː/ rather than a realisation of a central vowel (Erker & Bruso, 2017; Erker & Vidal-Covas, 2022; Garciá-Amaya & Lang, 2020). This may be due to the lack of the central vowel schwa in this variety. The varieties of Spanish spoken by Sephardic Jews in Bulgaria known as Judeo-Spanish adapted loan words from the surrounding Bulgarian language, and with it, the central vowel schwa (Andreeva et al., 2019; Gabriel, 2022). In Bulgarian this vowel does not only occur in unstressed syllables but also as a full vowel in stressed positions with the same vowel quality (Simeonova, 1989). An investigation into the vowel quality in the Judeo-Spanish FPs in comparison to native Spanish and Bulgarian FPs may shed more light on the relationship between the central vowel and the FP.

Another language lacking the central vowel in its vowel inventory is Italian. However, some southern Italian dialects produce schwa as the result of vowel reduction (Russo & Barry, 2008; Giannini, 2003). Giannini (2003) showed that speakers of

two southern Italian dialects (Bari, Naples) produce a central vowel in their FPs but speakers of two central Italian dialects (Pisa, Rome) produce a non-central vowel in their FPs that ranges between the front open-mid and close-mid lexical vowels. Even though the speaker sample is rather small (n = 8), this study delivers some initial evidence that the hypothesis - the existence of schwa predicts the vowel quality of FPs - may be true.

A third language whose vowel inventory does not include schwa is Japanese (Sugiura, 2015). The language uses five vowel qualities in its inventory but does not make use of central vowels (Sugiura, 2015). The most frequently reported FPs in Japanese are *ano* and *eeto* (Sadanobu & Takubo, 1993; Rose & Watanabe, 2019; Watanabe et al., 2021; Li et al., 2022). A typical FP consisting of a central vowel only or an additional nasal does not seem to exist in Japanese, judging from the existing literature on Japanese disfluencies. There are, however, other vocalic and nasal forms used as FPs: the vowel [ɛ] and the nasal [ŋ] (Rose, 2017). Here, again, the non-existence of schwa as a phoneme of Japanese seems to influence the form and vowel quality of FPs.

Assuming that languages that do include a central vowel in their vowel inventory use this vowel in their FPs still begs the question of which vowel "schwa-less" languages use. The most frequent vowel would be a suitable candidate, and in fact, Spanish speakers use the most frequent vowel to hesitate: /e/ (Guirao & García Jurado, 1990). However, this hypothesis is contradicted by Tamaoka & Makioka's findings (2004) that the unrounded open-mid front vowel /ɛ/ is actually the least frequent among the Japanese vowels. Another open question, that cannot be answered in this thesis is whether there are languages that possess a schwa-phoneme in their vowel inventory but use a different vowel in their FPs or the opposite, "schwa-less" languages that still use a central vowel in their FPs regardless.

2.2.3 Fields of application

The relevance of disfluency research becomes apparent when considering different fields of application, including the development of natural-sounding voice assistants, second language learning, rhetorical speaker coaching, forensic phonetic casework, language analysis for the determination of origin (LADO), and possibly many other fields (e.g., speech therapy, language acquisition).

FPs are an integral part of spontaneous speech as they not only occur very frequently but are also considered to escape the speaker's conscious control in unscripted speech (Bortfeld et al., 2001; Fox Tree, 1995; Butterworth, 1975; Künzel, 1987; Jessen, 2012). FPs occur as a symptom when a speaker is structuring and planning their next utterance and listeners seem to be able to interpret FPs as such, so the listener knows not to interrupt the speaker (Maclay & Osgood, 1959). The inclusion of FPs into the speech of voice assistants may result in speech that is more natural sounding and could then lead to a system that uses verbal material to signal to the listener

that the request is being processed. FPs in voice assistants can also be used to signal uncertainty as Elmers et al. (2023) showed for synthesised speech, similar to the effect in natural speech (Brennan & Williams, 1995). Similarly, the system could use FPs and other disfluencies to politely decline requests that lead to no result, because FPs are more frequently used in polite/formal speech than informal speech in human-human interaction, e.g., as hedging elements (Grawunder et al., 2014). Betz (2020) has attempted to include disfluencies in a conversational voice assistant showing promising results for lengthenings and pauses. The quality of FPs in the system needs further improvement and the introduction of variation to make the system more natural-sounding. Adell et al. (2010) implemented the insertion of FPs into an existing text-to-speech (TTS) model by predicting their segmental duration and pitch contour from the training data and accounting for variation by including an additional regression model. Székely et al. (2019) tested several synthesis options: 1) explicitly stating the FP type in the prompt, 2) indicating the place only and 3) not indicating neither place nor type. The best output is achieved by indicating the place of the FP but not its type, letting it be chosen by the system automatically. The authors furthermore highlight that all three types result in good TTS output; the option to be chosen depends on the application of the system (Székely et al., 2019). Another benefit of annotating disfluencies in the training data of a TTS model is that the production of fluent speech will improve as a result (Székely et al., 2019). Training a synthesis model on a large dataset of a single speaker can result in synthesised speech that is almost indistinguishable from natural speech (Kirchhübel & Brown, 2022). Using such a synthesised model and including the annotation of disfluencies in the training data could pose useful in experimental research as well, by creating speech stimuli using TTS synthesis systems that may offer more control over the speech output than over natural speech stimuli (Elmers et al., 2023).

As outlined before, fluency is an important factor in second language learning where it is regularly used to assess a non-native speaker's second language proficiency (Council of Europe, 2011). As it is not the case that native speakers are completely fluent, the aim of any learner is to approximate the level of fluency that native speakers show. In order to sound native-like, achieving the disfluency characteristics of the target language may be a relevant step on the way towards language proficiency. Teaching the specific aspects of disfluencies in the second language classroom to students, e.g., the specific FP types and vowel qualities used in the target language, may assist the students in acquiring a more native-like hesitation pattern and thus reduce the non-native accent (Trebon, 2022). Rose (2020) created an online feedback tool for second language learners that aims to help improve their L2 fluency by giving them immediate feedback on several disfluency measures. While this tool works well for silent pauses and speaking tempo, the underlying method for detecting FPs still needs improvement; there must also be more research that learners benefit from this way of feedback (Rose, 2020). There is various research on disfluencies in second language learning, but studies usually focus on the patterns of the L2-learners and

the perception of their fluency to rate their proficiency (De Jong et al., 2015; Brand & Götz, 2011), or they compare the disfluency patterns of L1 and L2 speakers (Temple, 2000; Riazantseva, 2001; Belz et al., 2017). To the best of my knowledge, there has been little to no work done to investigate the success of teaching language-specific disfluency patterns to learners to reduce their L2-accentedness.

Outside of the disfluency research community, FPs are usually considered elements that should be avoided during speaking, especially in formal situations such as public talks or job interviews (Fox Tree, 2002; Niebuhr & Fischer, 2019). In rhetorical training and textbooks on the topic, speakers may be advised to avoid FPs altogether or to reduce them to a minimum (Martin, 2019; Ditko & Stauch, 2023). Understanding the characteristics of FPs better and investigating the advantages and disadvantages of FPs for the listener may help to improve speaker coaching strategies. For example, if listeners actually benefit from FPs to a certain extent, it may not be necessary to teach speakers to avoid FPs but rather to use them more strategically in the right places, or in a form that benefits the listener rather than distracts from the content of the talk. As mentioned above, FPs in spontaneous speech are usually considered to escape the speaker's conscious control. If this is truly the case, training speakers to avoid FPs may not actually be effective but teaching speakers to effectively prepare and rehearse their talk will result in a talk with fewer FPs automatically (Ditko & Stauch, 2023).

In forensic phonetic casework and the field of LADO, language experts typically face speech data produced by unknown speakers whose language background may be equally unknown. Investigating the speaker- and language-specificity across a large dataset and a wide variety of languages may be useful for both fields. If FPs are revealed to be a robust feature, i.e., a feature that shows high between-speaker variation and low within-speaker variation, it may help in forensic phonetic casework to strengthen the conclusion (Künzel, 1987; Rose, 2002; Jessen, 2012) (see Chapter 4.1.2 for more details). In the field of LADO, analysts are asked to confirm the language background of an asylum seeker based on their speech samples, as asylum seekers frequently arrive in a country without identity papers (Fraser, 2012). If FPs show high language-specific characteristics that also occur in an L2, the analysis of FPs may shed light on a speaker's origin, i.e., native language and possibly the country of origin.

2.2.4 Interim conclusion

We have established that FPs are part of disfluencies with a time-buying function but also function as discourse markers, turn-taking and attention-grabbing devices, and possibly can be used for other purposes. In Chapter 2.2.1, the debate of whether FPs function as signals or symptoms was summarised shortly, coming to the tentative conclusion that they are most likely symptoms of a speaker's ongoing processing trouble that can, nevertheless, convey information to the listener. The recall function

of FPs in connection with their form will be explored in Chapter 6.

The question of how FPs' vowel quality and the central vowel schwa are connected was discussed by reviewing the literature on the central vowel in different languages, its dependence (or independence) of the vowel inventory, and studies on the FP's vowel quality in different languages. The matter will be addressed further in Part II of this thesis with experimental studies on data sets in different languages, namely German, English, Spanish, and Arabic. The most typical FPs are those consisting of either a vowel, a nasal, or a combination of the two parts. Less researched, but not necessarily less frequent, are the glottal FPs (Belz, 2021). They consist of a sequence of glottal pulses without any discernible vowel quality. FPs are also frequently produced with creaky voice portions at the beginning or end of the particle. Here, the middle part of the FP is still produced with modal voice so a vowel quality can be identified, which distinguishes this typical FP from the glottal FP. A detailed investigation into the FP form with considerations of creaky voice portions and glottal FPs will be undertaken in Chapter 4 on the Pool2010 corpus (Jessen et al., 2005), which consists of 100 native male German speakers. This corpus is large enough to explore the distribution of different types of FPs in a forensically relevant sample population. These results can then be consulted as a reference in forensic phonetic casework.

A cross-language analysis of FPs will be explored in Chapter 5, in which data sets from English, Spanish, and Arabic native speakers will shed more light on the question of how FPs can differ between languages. A short look at FPs in a second language will be given in the Chapter 5.2.

Chapter 3

Annotation scheme

Annotation schemes for spontaneous speech corpora vary considerably based on their research question. When disfluencies are not part of this research question, they are often ignored or conflated in one group that may even include other phenomena like feedback utterances (see Diapix-FL annotations; Cooke et al., 2013). Another factor that introduces variation between annotation schemes is the data itself. Dialogue data needs to account for both speakers in its annotation scheme, a factor that does not need consideration when annotating monologue data. We developed our scheme for monologue data, but it may be adapted to dialogue data by making a few changes (see Chapter 3.3). The annotation scheme presented here was developed in the project *Pause-internal phonetic particles in speech communication*, which is investigating not only FPs, but also other phenomena occurring in pauses such as inhalations. This is why FPs are not only considered in the annotation scheme but also phenomena such as tongue clicks, breathing noises, and other disfluency phenomena such as repairs, truncations, repetitions and lengthenings. As the focus of this work is on FPs, respiratory noises, and the other disfluency phenomena mentioned above are considered in the annotation scheme for future research but are not part of the analyses in this thesis. The annotations of inhalations have been used in Werner et al. (2023) to describe human respiration noises.

In an ideal research area of disfluencies, there would be one standardised annotation scheme for all phenomena relevant to this area. However, as the field is quite diverse in its research questions no such ideal exists and, consequently, no such standardised annotation scheme is available. The aim of creating this annotation scheme is not to propose it as the field's standard but to present it to other researchers in the field with its benefits and drawbacks to guide their own development of an annotation scheme for their research questions. Another limitation that speaks against a standardised annotation scheme is the time-consuming manner of the task itself. Manual annotation of spontaneous speech data demands a lot of the researcher's time and, more often than not, student assistants are trained to take on the task. As their

time (and cost) needs to be managed efficiently, it is simply not feasible to annotate phenomena that are not immediately relevant to the research question. In the following, we take a look at different annotation schemes (3.1) before our own scheme is described in detail (3.2). The chapter continues to describe changes that can be made to account for different data types such as dialogues (Chapter 3.3).

3.1 Review of different annotation schemes

Sharing speech corpora is a popular and efficient trade in the research community as the data collection of speech takes a large amount of time and the researcher is often only interested in one specific phenomenon of speech. Giving the data to other researchers to conduct further analyses on different research questions reduces the time needed for data collection, as well as the cost of the project. Therefore, a detailed description of the annotations done for the speech data is needed, so the second researcher (and one's future self) understands which phenomena have been annotated and which ones may be missing. Furthermore, disfluencies are more prone to erroneous annotations than fluent speech (Zayats et al., 2019) which is why annotators in this area might need more training, detailed instructions, and close supervision until they gain some experience and become attuned to the disfluency phenomena. In the following, we will compare annotation schemes of several spontaneous speech corpora concerning the annotation of disfluencies (for a comparison of annotations see also Trouvain & Truong, 2012; Trouvain & Belz, 2019; Trouvain & Werner, 2020, 2022).

The **Diapix-FL corpus** (Cooke et al., 2013) consists of annotated dialogue data from a spot-the-difference task (Diapix task) in the speakers' L1 and L2 (each English or Spanish). The corpus was compiled to investigate cross-linguistic differences between the languages. The bi-directionality of the corpus allows for the investigation of language factors controlled for native and second language use (Garcia Lecumberri et al., 2017). Each speaker was annotated using one TextGrid per recording. The TextGrid only consists of one single tier on which speech is transcribed orthographically in one interval and pauses are annotated using an interval, which is labelled with a plus sign. Pauses that are caused by the interlocutor talking are labelled with a minus sign (see Figure 3.1). Only selected disfluency phenomena are labelled: lengthenings using a colon after the lengthened segment (*a:nd*) and non-lexical FPs using the per cent sign for all types as can be seen in Figure 3.1. However, it is important to note that feedback utterances were also labelled using the per cent sign even though they differ from FPs in their pitch contour quite distinctly (Pistor, 2017). Segment elisions and word truncations are annotated using an asterisk (*an** for *and*). Analyses on the FPs in the Diapix-FL corpus (Cooke et al., 2013) have been conducted (see Chapter 5.2) using the existing annotations. Additional annotations had to be made to determine the type of FP and to segment the FPs into vowel and nasal consonants. Thanks to the existing annotations, our additional annotation work was accelerated immensely.

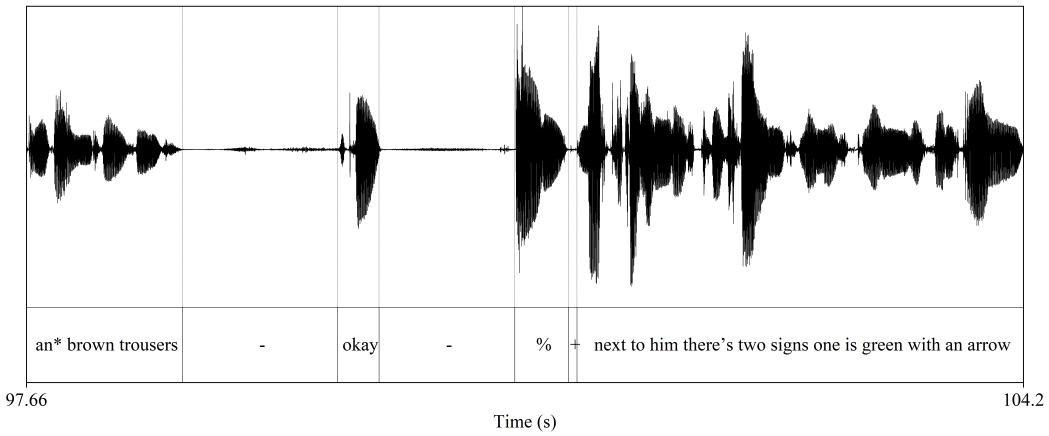

Figure 3.1: Annotation scheme: Diapix-FL
Example from the annotation scheme for the Diapix-FL corpus (Cooke et al., 2013), file En_spkg_En_1a. Tier name: speech

The **Buckeye corpus** (Pitt et al., 2007) is a corpus of American English speakers including 40 speakers from Columbus, Ohio and was compiled to investigate pronunciation variation in conversational speech. Speech was elicited in an interview set-up, and orthographic annotations of the interviewee's speech were done manually. Not many disfluencies were annotated in this corpus, but the label <SIL> can be found for silences that "occur during running speech, and those which occur between stretches of running speech" (Kiesling et al., 2006). FPs are orthographically transcribed using the labels *uh* and *ah* for vocalic FPs, *mm* and *hm* for nasal FPs, and *um* for vocalic-nasal FPs, according to which "transcription the actual sound is closest to" (Kiesling et al., 2006). The use of the different types for the vocalic and the nasal FPs seems subject to individual variation between annotators as no clear description is provided as to when to use which type. Truncations, laughter, laughed speech, and non-verbal vocalisations (sighs, throat clearing) receive their own category each (Kiesling et al., 2006). Similar annotation guidelines can be found for the Switchboard corpus (Godfrey & Holliman, 1997; Hamaker et al., 1998).

The **GECO corpus** (Schweitzer & Lewandowski, 2013) includes 46 dialogues in two conditions (unimodal and multimodal) from 18 native German female speakers. The corpus has been compiled to investigate phonetic convergence in German spontaneous conversations in correlation with the speakers' personality traits. The accompanying TextGrid files include annotations on the word, syllable, and phone level (see Figure 3.2). Silent pauses are marked in separate intervals, FPs are written out orthographically as *äh*, *ähm*, and *mhm* or *hm*. Lengthenings of the FP are marked by reduplicating the letter for the vowel (*ääääh*) so there is some variation in the data set regarding the different types of FPs and their notation. The notation of nasal FPs is also not a one form to one function match, as feedback utterances may also be transcribed as *mh* or *mhm*. Additional annotations for eight dialogues are

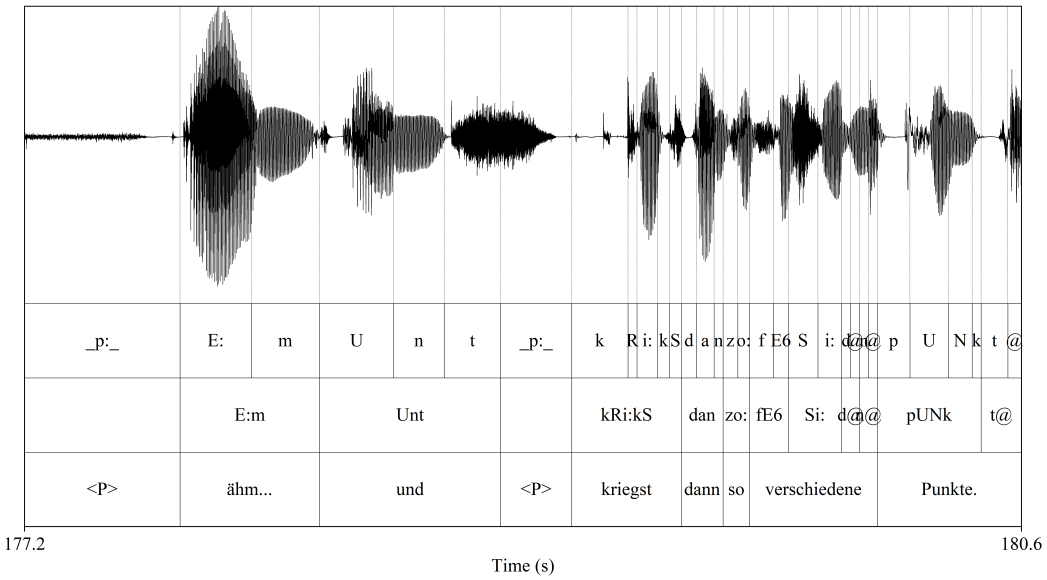

Figure 3.2: Annotation scheme: GECO
Example from the annotation scheme for the GECO corpus (Schweitzer & Lewandowski, 2013), file multi_A-C_left. Tier names from top to bottom: phones, syls, words

provided by Belz (2019b) with the **GECO-FP** annotation files. These annotations were created with the purpose of investigating FPs and their context. A detailed description of the annotation guidelines can be found in Belz (2019a). The first tier is the word-level annotation from the original annotation files; five more tiers are added for the GECO-FP annotations. The second tier marks the FP and its sequential context, and the third tier gives a detailed segmentation of the phenomena on the second tier as seen in Figure 3.3. The fourth tier includes a label for the FP's phonation type, e.g., modal voice, glottalised, or laryngealised phonation. The fifth tier annotates lexical vowels for each speaker as a way to compare the FP's vowel quality, and the sixth tier includes information about the dialogue structure of the current turn in which the FP is produced, i.e., whether it is an initialising turn like a question or a statement or a responsive turn like an answer.

Crible et al. (2022) present an annotation scheme that tries to account for disfluencies in typical as well as atypical speech, such as speech by people who stutter. They propose a Praat TextGrid (Boersma & Weenink, 2022) with seven tiers (see Figure 3.4). On the first tier, an orthographical transcription is provided in which pauses can already be marked by using a specific symbol (here: -). The second tier includes the speaker label in an interval that stretches along the entire turn of the speaker. This is a practical way to allocate turns to individual speakers in multi-speaker conversations. The third tier is a summary of the following two tiers in which different disfluency phenomena are annotated. On the fourth tier, "simple" disfluency phenomena are annotated, i.e., those that only affect one word or one syllable:

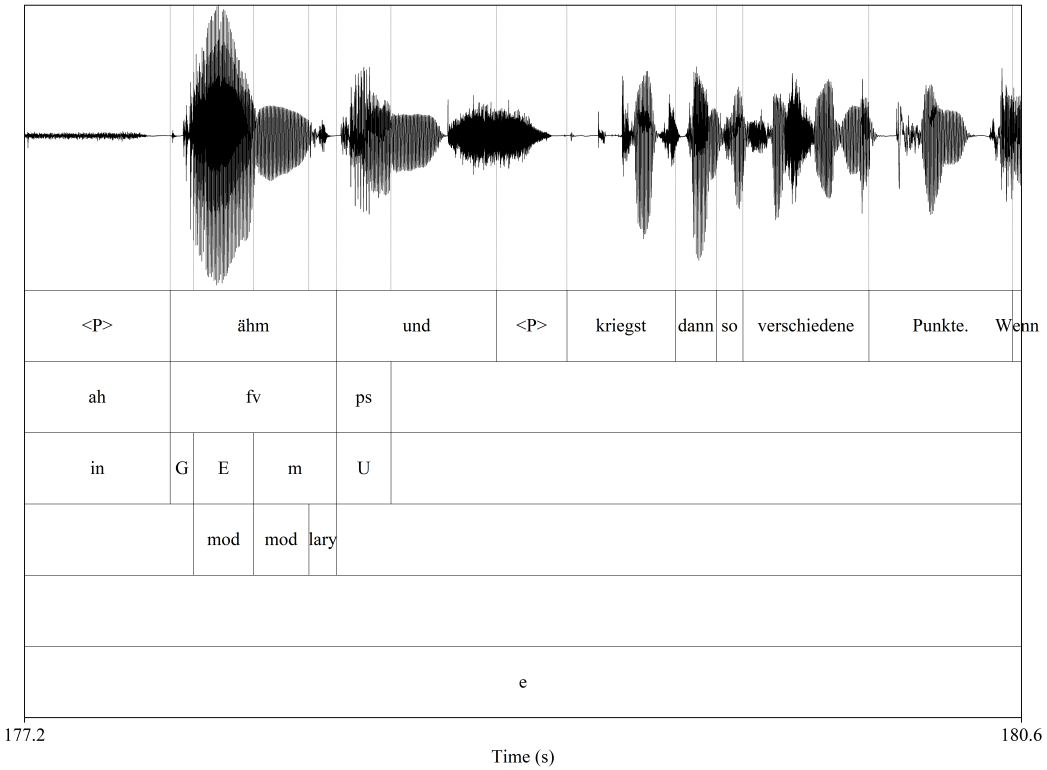

Figure 3.3: Annotation scheme: GECO-FP

Example from the annotation scheme for the GECO-FP corpus (Belz, 2019b), file multi_A-C_left. Tier names from top to bottom: wordscor, hes, segm, coart, vowel, func

"[L]ocal, one-part disfluencies" (Crible et al., 2022). Local disfluencies are the following: blocks[1], discourse markers (*well, you know*), silent pauses, filled pauses (= filler particles), and self-interruptions. Compound disfluencies, meaning those disfluencies that consist of at least two words, are annotated on the fifth tier. These are repetitions and modifications. A repetition is defined here as "the reiteration of one or several elements of various sizes without modification" (Crible et al., 2022), i.e., the material is not altered when repeated. A modification is classified as "a substitution, addition or deletion of linguistic material in the context of a repair sequence" (Crible et al., 2022). As these disfluencies stretch over multiple elements, it is possible that simple disfluencies such as FPs occur within them, which justifies the categorisation into simple and compound disfluencies quite well, and the annotation on two tiers is more practical. The third tier, as mentioned before, summarises tiers 4 and 5 by concatenating all the labels from one turn into one sequence. This allows for an analysis of disfluency patterns, although the timing aspect is not considered, i.e., the researcher cannot tell from the sequence tier how closely together the disfluencies were produced, only that they occur in the same turn. As lengthenings can affect compound as well as simple disfluencies, they are annotated in another tier that is only dedicated to this phenomenon (tier 6). Lengthenings are annotated in cases where a phone or a syllable is perceived as abnormally long, explicitly not annotated are what the authors call phonological lengthenings such as phrase final syllable lengthenings (Crible et al., 2022). The reason for this may be that not all lengthenings function as disfluencies, they are also used for accentuation (Betz, 2020). The annotation of lengthenings seems particularly difficult as they are, in many cases, not perceived as disfluent phenomena and therefore often missed by human annotators (Betz, 2020). Annotations of lengthenings are thus subject to the annotator's perception and judgement of a lengthening's fluency, which should be addressed in the annotator's training.

The last tier, tier 7, is dedicated to so-called paraverbal events, examples put forward by the authors are laughter, coughs, and tongue clicks, and can also include non-lexical vocalisations ("mm, creaky voice"; Crible et al., 2022), respiratory noises, and explicit editing terms. This last category seems to include quite a wide range of different phenomena, which may cause problems during annotations or else the specific annotations will not be used in any analyses as the category is too broad.

By reviewing some selected corpus annotations, the difficulties of annotating disfluencies and working with existing corpus annotations were established. Corpora are usually compiled and annotated with a specific research application in mind. If the research question does not involve disfluencies, it is likely that disfluencies were either not annotated at all, only some phenomena were annotated (e.g., FPs), or they shared one category with several other phenomena such as feedback utterances, laughter, or non-verbal vocalisation (e.g., coughs, sighs, lip smacks). Another difficulty is the variety of labels used for the same phenomena across corpora, but also within corpora

[1]"disruptive silent pause characterised by visible or audible tension before or within a word" (Crible et al., 2022) characteristic for speech by people who stutter

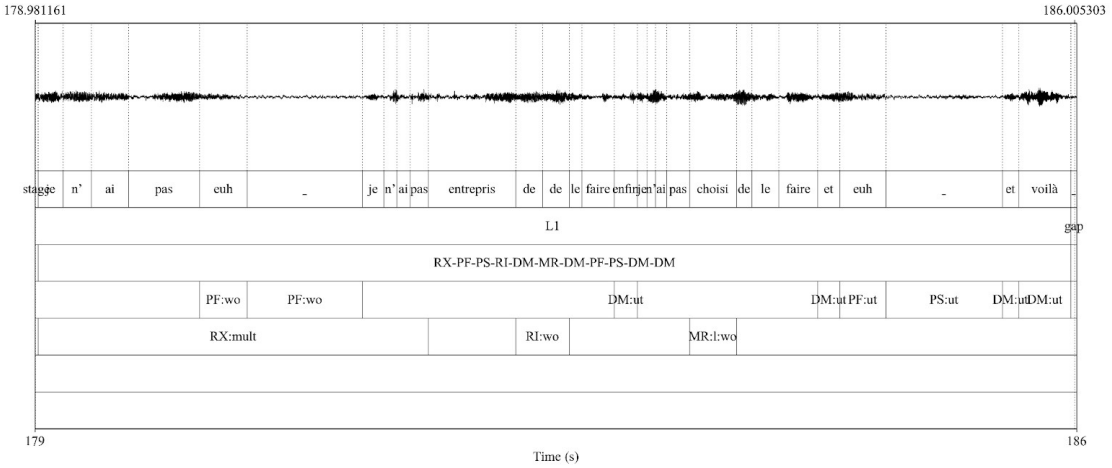

Figure 3.4: Annotation scheme: Crible et al. (2022)
Example annotation taken from (Crible et al., 2022) for illustration purposes. Tier names from top to bottom: word, speaker, seq, disf, rm, leng, para.

due to annotator preferences or different perceptions. All the more important is a thorough documentation of the annotations and an in-depth annotator training and feedback system.

3.2 Presenting the annotation scheme in use

This annotation scheme was developed specifically for a subcorpus of the Pool2010 corpus (Jessen et al., 2005) using TextGrids of the speech analysis software Praat (Boersma & Weenink, 2022). The specific task of the subcorpus, a guessing game similar to the game "Taboo", makes it necessary to include some pause phenomena that may not be needed in other speech data with different tasks. Possible adaptations of the annotation scheme for different data, e.g., dialogues, are outlined below in section 3.3.

The annotation scheme consists of five interval tiers, each focusing on a different phenomenon and one additional optional interval tier (*comments*) dedicated to the highlighting of questionable speech samples, etc. for the annotators (student assistants) to receive feedback from the researcher. A seventh tier was added to annotate 10 tokens of the German corner vowels /aː/, /iː/ and /uː/ in each file. An image of the applied scheme for the Pool2010 corpus (Jessen et al., 2005) can be found in Figure 3.5, and a comprehensive list of labels used in each tier can be found in Table 3.1 along with a short description of the labels.

The first tier of the TextGrid labels all pauses and leaves intervals containing speech empty with no label. We will call these speech stretches inter-pausal units (IPU). As syntactic information was not annotated, information is lacking whether complete grammatical phrases are produced and if every pause acts as a phrase boundary.

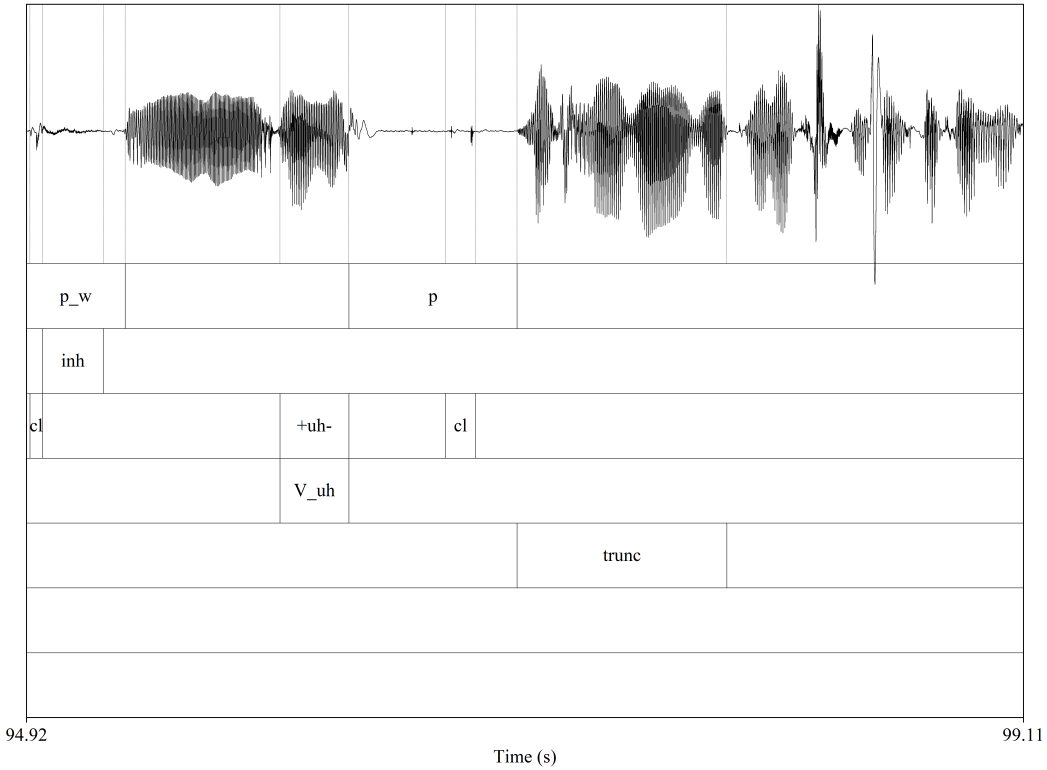

Figure 3.5: Annotation scheme: Pool2010
Example from the annotation scheme for the Pool2010 corpus (Jessen et al., 2005), file v12_fs1. Tier names from top to bottom: pau, resp, fp, V-N, disfl, com, CV.

However, it is assumed that the waiting pauses and the task changes also mark a phrase boundary as either a response from the interviewer is expected or a new task begins.

The first pause boundary is set where the last phoneme of the previous speech stretch ends, and the last boundary is set where the first phoneme begins. This was ensured by consulting the oscillogram and spectrogram as visual aids. When the first phoneme of the next IPU is a plosive, either voiced or voiceless, the last boundary of the pause is moved 50 ms to the left in order to account for the closure of the plosive (see Belz & Trouvain, 2019). Place of articulation and voice influences the closure duration of the plosive (Kuzla & Ernestus, 2011), however, using different durations for different plosives is not practical during annotation. As the task-a guessing game-requires the speaker to describe a word that the interlocutor (i.e., the interviewer) must guess, overly long pauses occur when the speaker waits for a response from the interviewer, or, once a word is correctly guessed, when the speaker moves on to the next word. These pauses are specifically labelled as it is assumed that they would influence the pause duration considerably when being grouped together with simple pauses. The labels used are *waiting pause* (p_w) and *task change* (tc), respectively. Moreover, two other pause types are labelled: *starting pause* (p_start) and *ending pauses* (p_end). This is due to the recording conditions of this specific corpus as the speakers performed several tasks in a row, and the recording was not stopped in-between tasks. It is assumed that the speakers took a short break, listened to explanations for the next task, or changed to the Lombard setting between the two tasks. Here, only a subset of the corpus containing the semi-spontaneous speech task, which was taken from the original full recording, is used so that initial and final pauses of the present files are part of longer pauses resulting from the changes between tasks in the recording session. These pauses are labelled accordingly so they can be excluded during the analysis.

The second tier is dedicated to respiratory noises and laughter. When laughter (laugh), inhalations (inh), and exhalations (exh) are labelled on the second tier, a pause interval is labelled on the first tier, as these phenomena require the interruption of the speech stream. Breath noises are labelled from their audible beginning to their audible end, again taking the visual aid of oscillogram and spectrogram into consideration. Laughter is subsumed under respiratory noises as often a series of inhalations and (voiced) exhalations can be observed.

Filler particles (FPs) are annotated on the third tier, including any initial or final glottal pulses or creaky voice and are labelled according to their type as vocalic (*uh*), nasal (*hm*), or vocalic-nasal (*um*) FPs. The vocalic and the vocalic-nasal FPs receive additional labels according to their pause context. When speech occurs at the left or right edge, the FP-label receives a + on the left or right side, respectively; when a pause occurs at the left or right edge, the FP-label receives a - on the left or right side, e.g., +uh+ for a vocalic FP surrounded by speech or -um- for a vocalic-nasal FP in isolation (i.e., between two pauses) (see Figure 3.6 for examples). When FPs

Tier	Name	Labels	Description
1	pau		annotation of pauses:
		p	simple pause
		p_w	waiting pause
		tc	task change
		p_start	pause at the beginning of recording file
		p_end	pause at the end of recording file
2	resp		annotation of respiratory noises:
		inh	inhalation
		exh	exhalation
		laugh	laughter
3	fp		typical FP with context annotation:
		+uh+, +um+	uh/um within speech
		+uh-, +um-	uh/um preceded by speech and followed by silence
		-uh+, -um+	uh/um preceded by silence and followed by a pause
		-uh-, -um-	uh/um in isolation (i.e., flanked by silences on both sides)
		hm	nasal FP
		cl	tongue click
		gl	glottal FP
4	V-N	V_uh, V_um,	segmentation of FPs into vocalic and nasal parts
		N_um, N_hm,	according to FP type
		crv, gl	annotation of glottal plosives/ creaky voice within FP
5	disfl		annotation of other disfluencies:
		rep	repair
		trunc	truncation
		lngth	lengthenings
		repeat	repetition
6	com	*free text*	any questions/comments the annotators may have for the supervising researcher
7	CV	a, i, u	corner vowels /aː/, /iː/, /uː/; 10 tokens of each vowel for every file

Table 3.1: Detailed overview of the annotation scheme
Tiers of the annotation scheme, and the labels used for each phenomenon including a short description.

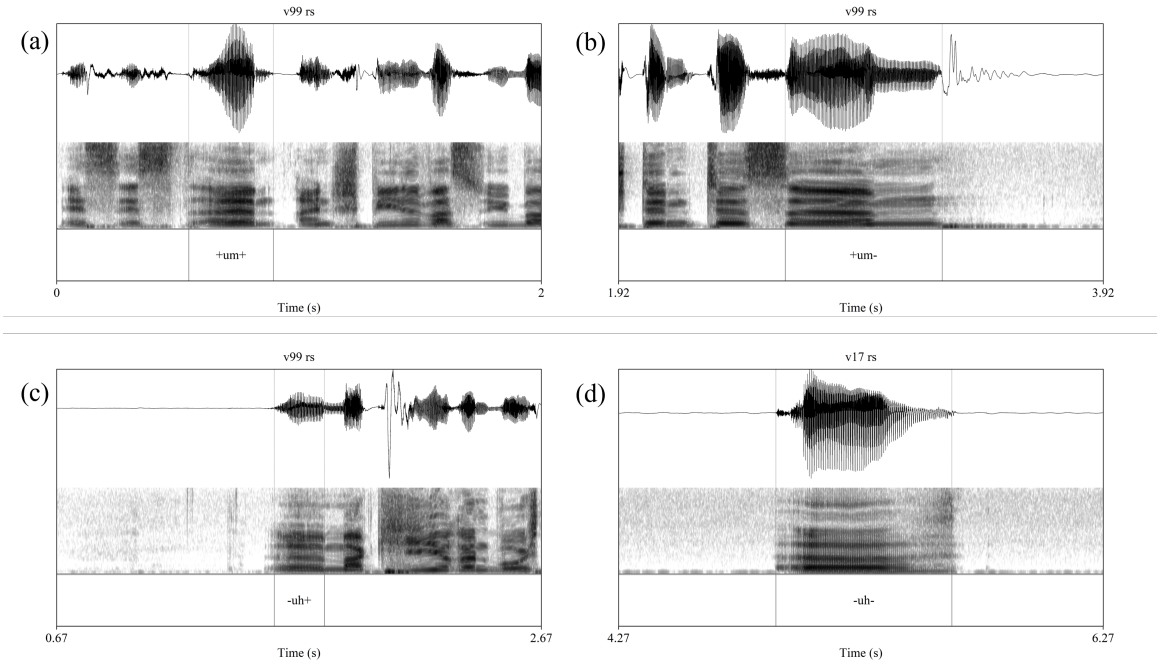

Figure 3.6: Examples of FPs in their pause context

2-sec sections (spectrogram: 0-8 kHz) from the Pool2010 corpus (speakers: v99, v17) showing FPs with a) speech as left and right context (+FP+), b) speech as left and silence as right context (+FP-), c) silence as left and speech as right context (-FP+), and d) silence as left and right context (-FP-).

are annotated on the third tier no pause interval on tier 1 overlaps with this tier, which means FPs are always labelled as speech material occurring between pauses, not within pauses. Glottal FPs (gl) are also annotated on this tier. They are defined as a sequence of glottal plosives or creaky voice in isolation, i.e., independent from other lexical material, without a clear vowel quality being discernible (Belz, 2017) (see Figure 3.7). Another phenomenon annotated on this tier is tongue clicks, which are regarded as FP candidates (Belz, 2023) in this work. The tokens of clicks are only annotated broadly so no duration measurements will be provided.

As FPs are the focus of the annotation scheme, they are segmented into their parts on the fourth tier. Vowels and nasals are marked using the audio and visual aids available, as well as creaky voice and glottal pulses during the FPs. Investigations of voice quality of FPs are sparse (Belz, 2021; Cataldo et al., 2019), which is why the annotation of this feature is included in this annotation scheme. A distinction between glottal pulses and creaky voice is made based on Belz (2021), who proposes to annotate glottal pulses at the beginning or end of FPs when three or fewer glottal pulses (gl) are visible in the spectrogram and to annotate creaky voice (crv) when more than three pulses occur (see Figure 3.8). The individual glottal plosives are usually further apart in glottal pulses than in creaky voice.

The fifth tier is included to annotate other disfluency phenomena for further re-

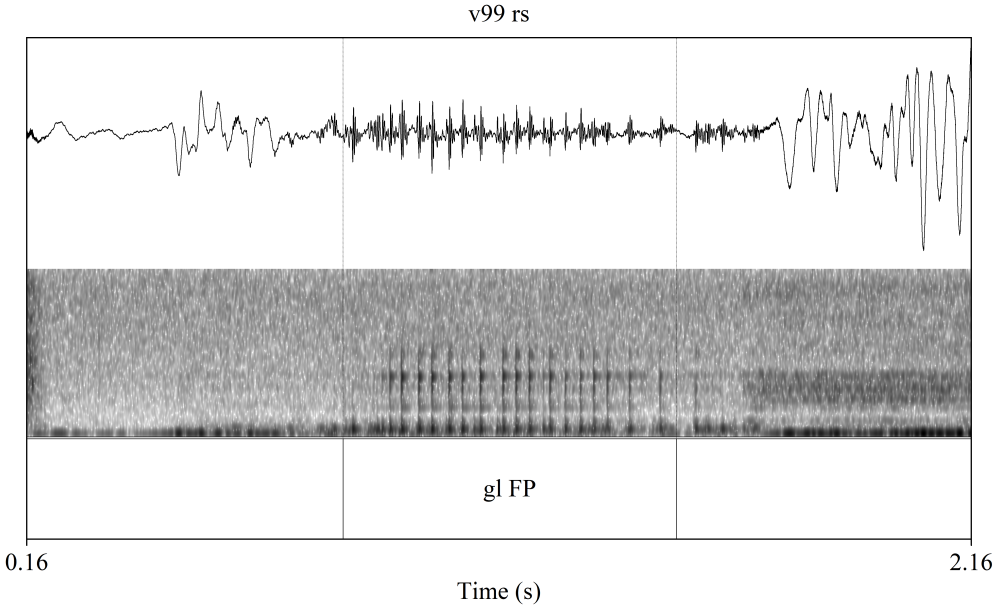

Figure 3.7: Example of glottal FP
2-sec section (spectrogram: 0-8 kHz) from the Pool2010 corpus (speaker: v99 in Lombard condition) showing a glottal filler particle (gl FP).

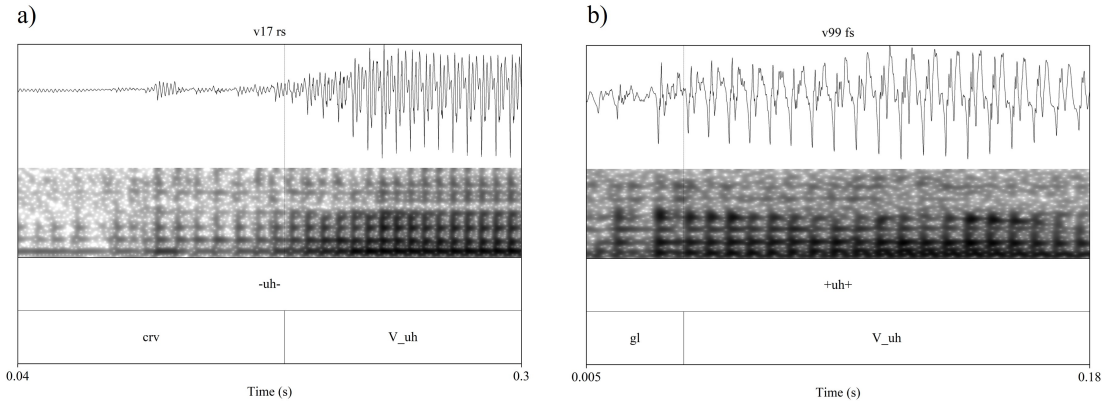

Figure 3.8: Comparison of creaky voice vs. glottal pulses within an FP
Sections (spectrogram: 0-8 kHz) from the Pool2010 corpus (speakers: v99, v17) showing filler particles a) with initial creaky voice (crv) and b) with two initial glottal pulses (gl). Note that in 2a only the first 260 ms of the FP are shown.

search projects which are not considered in this thesis. As these phenomena may include FPs, they receive their own tier, which is also done in the annotation scheme by Crible et al. (2022). The reason for annotating them was the investigation of disfluency clusters and co-occurrences of different disfluencies with specific FPs. Disfluencies that were considered in the annotation scheme are lengthenings, truncations, and repairs (see Chapter 2.1.1 for descriptions of the phenomena). Repetitions are annotated as a separate class, as this type of repair is very frequent.

The sixth tier is used as a form of communication between the annotators and the leading researcher and is not of particular interest here. The seventh tier was added to the first draft of the annotation scheme as it is common to compare the vowels in FPs with the language-specific corner vowels. A similar tier was added in the FP annotation scheme for the GECO corpus (Schweitzer & Lewandowski, 2013) by Belz (2019b). While on all other tiers, every token of the respective phenomenon should be annotated; only 10 tokens of each corner vowel /a:/, /i:/, and /u:/ are annotated on this last tier. Tokens are annotated when they occur in a stressed syllable based on the phonemic transcription of Standard High German. As the speakers from the Pool2010 corpus may display a regional Hessian accent (Jessen et al., 2005), annotators were instructed not to take a corner vowel into account when its vowel quality clearly deviates from Standard High German.

3.3 Adapting the scheme for different data

Adapting the annotation scheme for different data can be accomplished as follows: As other monologue data may not include the same task, the specific pause labels for waiting pauses and task changes may be superfluous, though task changes may also occur in different data sets. Furthermore, the labels for initial and final pauses may also be omitted though a preparation phase, and a final phase may also occur in different data.

The adaptation of the annotation scheme to dialogue data can be accomplished by adding a label for the sections in which the interlocutor is speaking. I propose to have one audio file per speaker and, if possible, different recording channels for each speaker as the annotation is then facilitated by different intensity levels for each speaker. The sections where the interlocutor (or interviewer) is speaking can be labelled on the pause tier by using a label such as “p_int”. FPs would still be labelled using plus and minus signs, and the pause types would then show the distribution of FPs in connection with the interlocutor’s speech.

Lexical FPs, if of interest, can be included on the third FP tier. To distinguish lexical and non-lexical FPs from one another, their labels could be marked with “_lex” and “_nonlex” respectively. Another option could be adding another tier for lexical FPs to clearly separate them from non-lexical FPs.

The previous chapter presented the theoretical background of disfluencies and FPs, gave an introduction to the research topic by presenting the signal vs. symptom hypothesis, and presented a new hypothesis regarding the vowel quality of FPs and the central vowel schwa. The current chapter introduced a new annotation scheme which was used to annotate the data of the Pool2010 corpus (Jessen et al., 2005) presented in Chapter 4.

The next part of this thesis presents four empirical studies examining the phonetic details of FPs (Chapter 4), comparing them between languages (Chapter 5), and investigating one specific function of FPs together with the phonetic details of the FPs (Chapter 6).

The next chapter evaluates the phonetic features of FPs, frequency distribution, pause context, duration, voice quality, and vowel quality in a large German corpus consisting of speech by 100 male speakers in two conditions. Chapter 5 presents the frequency distribution of FPs and their vowel quality in L1 and L2 English and Spanish, as well as the frequency distribution of some disfluencies and the vowel quality of the vocalic FP in a small corpus of Arabic speech. The vowel qualities of the FPs are compared across languages. The last chapter of the empirical part presents a series of three experiments which are all aimed at finding the beneficial recall effect of FPs which has been reported in previous literature (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2011a; Diachek & Brown-Schmidt, 2022).

Part II

Empirical Studies

Chapter 4

Analysis of filler particles in German

4.1 Introduction

The focus of this chapter is on filler particles (FPs) such as *uh* and *um* and tongue clicks, phenomena that are typical for, and often observed in spontaneous speech. Frequently, but not exclusively, FPs occur in the vicinity of speech pauses. The distribution of FPs connected to pauses and stretches of speech will be investigated, as well as the question of whether the different FP types behave differently. A large corpus of German spontaneous speech was analysed concerning the frequency of occurrence of various FP types, pause context, durations, voice quality, and vowel quality.

Typical FPs can occur within two stretches of speech production silence[1]. Still, they are not limited to this position and do frequently occur within a speech utterance with no silence on either side. Another phenomenon that serves a similar function as *uh* and *um* and occurs in similar positions as these typical FPs is the purely nasal FP type, *hm*. Glottal FPs and tongue clicks are less often described, but see, for instance, Smith & Clark (1993). We consider tongue clicks as *potential* FPs that can be used for self-repairs (Li, 2020) or when having trouble finding a word (Trouvain & Malisz, 2016; Ogden, 2020) as situations that are typical when using FPs.

Research on pauses, FPs, and disfluencies, in general, has become more frequent in the last decades, and most work has focused on the frequency distribution, duration, and vowel quality of FPs. However, most of these studies focus on English data, and research on German data mostly focuses on one or two aspects, such as the frequency distribution (Braun & Rosin, 2015; Bellinghausen et al., 2019), fundamental frequency and duration (Batliner et al., 1995), the vowel quality (Pätzold & Simpson, 1995), the vowel quality and fundamental frequency (Klug & König, 2012), or vowel quality, duration and frequency (Niebuhr & Fischer, 2019). In some cross-language comparisons, German was investigated alongside other European languages, such

[1] As opposed to "true" acoustic silence where background noise is absent.

as English, Dutch and French (de Leeuw, 2007; Lo, 2020; Gerstenberg et al., 2018; Muhlack, 2020a). Other studies have looked at FPs in first vs. second languages (L1 vs. L2), e.g., in L2 German by speakers with several different L1 backgrounds (Belz & Klapi, 2013; Belz et al., 2017; Reitbrecht, 2017; Muhlack, 2020a). A comprehensive phonetic analysis of FPs in German was recently provided by Belz (Belz, 2021; Belz & Reichel, 2015; Belz, 2017, 2018; Belz & Trouvain, 2019) who investigated several features of FPs (especially *uh* and *um*), also among them was the voice quality of FPs and the occurrence of glottal FPs. The study reported in this chapter follows (Belz, 2021) in describing the phonetic characteristics of FPs in German by examining different features such as the frequency distribution, their duration, the pause context, the voice quality, and vowel quality uttered by 100 male native German speakers[2]. The features of the pause context and voice quality are under-researched aspects of FPs, which is why they are included here. Benefits from this study are the number of speakers examined, the inclusion of a Lombard condition, and the number of features that are investigated, for both their general trends and on an individual speaker level. Since data from 100 speakers also poses problems for qualitative analyses, speaker-specificity will be examined by means of comparing individual patterns of 12 sample speakers. In the following, a brief (and not exhaustive) overview of the literature on FP research in different languages is given.

Frequency distribution

Several studies report disfluency rates per minute, however, they are often not directly comparable because different phenomena were considered. Braun & Rosin (2015) report a disfluency rate of 4.5-12.3 per minute (for 10 speakers) when looking at typical FPs (*uh*, *um*), the nasal FP *hm*, as well as initial and final vowel and consonant lengthenings in German. Belz (2021) reports an FP rate of 2.9 FPs per minute (range: 1.4-4 disfl./min) for the GECO-FP corpus (Belz, 2019b) and a rate of 4.3 FPs per minute (range: 1.9-11.3 disfl./min) for the BeDiaCo (Belz & Mooshammer, 2020), both of which are German dialogue corpora. For English, Clark & Fox Tree (2002) report an FP rate (*uh* and *um* only) of 17.3 per 1000 words ranging from 1.2 to 88.5 FPs per 1000 words for 65 speakers. This rate translates to 2.6 FPs per minute when assuming an average of 150 words [3] are produced per minute. By considering FPs, pauses, repetitions, and false starts, Maclay & Osgood (1959) found a mean disfluency rate of 10.97 per 100 words (16.5 disfl./min) ranging from 5 to 15 disfluencies per 100 words for 13 different speakers of English. Shriberg (1994) reports a disfluency rate of 0.01-0.08 disfluencies per word (1.5-12 disfl./min) for three corpora in English by not only including FPs, repetitions, false starts, and repairs. McDougall & Duckworth (2017) report an FP rate for English (including

[2]The majority of forensic phonetic casework deals with male voices which is why most research in this area also focuses on this speaker group.

[3]Maclay & Osgood (1959) found a mean rate of 152 words/minute

only *uh* and *um*) ranging approximately from 2 to 8 FPs per 100 syllables (appr. 5-20 disfl./min)[4]. The overview of these studies shows the large variation of disfluencies considered for each study and proves that there is no standard unit in which the rate is reported. Time units can be minutes, however, a rate per word, per 100 words (or syllables) or even 1,000 words has been encountered. It should be noted that a disfluency rate per word/s may not be the most useful unit as the word length may differ considerably (especially in German where compounding is very frequent), and some languages show differences in word length which makes a cross-language comparison more difficult (Trouvain, 2004). A rate per e.g., 100 syllables (or even phones) may be more useful, but the syllable structure and complexity may still pose problems for cross-language comparisons.

We consider tongue clicks potential FPs (Belz, 2023) as they frequently occur in pauses and their function as a hesitation device has also been reported (Trouvain & Malisz, 2016). A large variation between studies and individuals on the use of tongue clicks can be seen. A rate of 1.3 clicks per minute was reported for English dialogues (Ogden, 2013), however, a high variation between speakers is observable. Trouvain & Malisz (2016) found a click rate of 6-12 per minute for one native English speaker who was regarded as a heavy clicker. However, Zellers (2022) found a rate ranging between 1-5.4 clicks per minute for 12 Swedish speakers. It seems that speakers vary in their clicking behaviour, however, Gold et al. (2013) argue that speakers (of English) do not vary sufficiently in the click distribution and that audio material available in forensic cases may be too short for the click frequency to be a useful feature in a forensic phonetic analysis.

Duration of filler particles

When looking at the duration of FPs, usually only the FPs *uh*, *um*, and *hm* are considered. Belz (2021) reports the following values for these FPs in German data: a mean of 262 ms (sd = 121 ms) for *uh*, 396 ms (sd = 140 ms) for *um*, and 450 ms (sd = 183 ms) for *hm*. The same duration pattern (*uh* shortest, *hm* longest) is reported by de Leeuw (2007) for German (but not for English and Dutch): 317 ms (sd = 113 ms) for *uh*, 457 ms (sd = 161 ms) for *um*, and 470 ms (sd = 234 ms) for *hm*. It is often reported that the vowel in vocalic-nasal FP types is shorter than in purely vocalic FPs (Belz, 2021; Hughes et al., 2016) and that the former type is more often surrounded by silences (Clark & Fox Tree, 2002; Hughes et al., 2016). These silences also tend to be longer for *um* than for *uh* (Clark & Fox Tree, 2002).

[4]Converting this unit to a rate per minute is more difficult than for the rate per 100 words. Syllable durations highly depend on the syllable structure, the stress, and pause context (Crystal & House, 1990). An approximation was reached by taking the most frequent syllable structure (CVC) reported by Crystal & House (1990) and calculating the mean duration of the CVC type before and after pauses in stressed and unstressed positions (mean = 250 ms).

Vowel quality

The vowel quality of FPs is generally considered to be a central vowel in most languages but with language-specific tendencies. If not reported as central schwa [ə], the following vowels are used to describe the vowel quality in some languages: the front vowel [ɛ] or central vowel [ɐ] for German, the rounded front vowels [ø] or [œ] for French, and the close-mid front vowel [e] for Spanish (Belz, 2021; Künzel, 1987; Simpson, 2007; Candea et al., 2008). English, especially American English, typically uses a vowel similar to the open-mid vowel [ʌ] in their FPs (Shriberg, 1994); it is important to note that in phonetic transcription for English the symbol ʌ is usually used to describe an open-mid central vowel (Roach, 2009). German FP vowels are also reported to have a lower F1 than English and French FP vowels (Muhlack, 2020a; Lo, 2020). French FP-vowels are usually produced with more lip rounding than the German FPs (Lo, 2020).

Voice quality

A phenomenon that is less researched in the disfluency area is the voice quality with which FPs are produced. Belz (2017, 2021) introduced glottal FPs into the field and with it the proportion of creaky voice and glottal plosives within each FP. It was found that about two thirds of all FPs are realised with creaky phonation in Italian tourist guide speech (Cataldo et al., 2019). Shriberg (2001) remarks that FPs may be subject to a decrease in amplitude and a drop in pitch, which supports the production of creaky voice in the FP-final position. However, Belz (2021) finds that FP-initial creaky voice is more frequent than FP-final creaky voice and is more frequent with the vocalic FP *uh* than the vocalic-nasal FP *um*.

4.1.1 Hypotheses

With this study, we hope to shed light on the frequency distribution, the duration of FPs, their vowel quality, and voice quality in connection with Lombard speech, the pause context, and the speech tempo. We apply an exploratory analysis, as many aspects of our analysis are under-researched, e.g., disfluencies in the Lombard condition, the production of creak in FPs, and the influence of pause context on the acoustic measures of the FPs. The literature guides us in some aspects, as it is expected to find more (and longer) vocalic-nasal FPs in the data than vocalic FPs (Wieling et al., 2016; Hughes et al., 2016). As Gósy & Silber-Varod (2021) found an effect of pause context on the duration of vocalic FPs, it is expected to see similar trends in that FPs that are surrounded by pauses should be longer than FPs that occur within speech. It is not expected to find a difference in the vowel quality between *uh* and *um*, as well as any influence of FP-duration on the vowel quality (Hughes et al., 2016). In line with Belz (2021) and Cataldo et al. (2019), we anticipate observing a

high rate of creaky voice within the FPs but more so in the FP-initial position than in the FP-final position.

4.1.2 Importance for forensic phonetics

FPs are frequent in spontaneous speech (Bortfeld et al., 2001; Fox Tree, 1995), which makes them interesting for forensic phonetic casework. If they were considered words, as suggested by Clark & Fox Tree (2002), they would probably be grouped under high-frequency words. It is usually assumed that speakers do not produce them intentionally (Kjellmer, 2003; Braun & Rosin, 2015). The lack of control that speakers usually exert over FPs make them a practical feature for forensic phonetics, as they are considered to be unaffected by voice disguises (Butterworth, 1975; Künzel, 1987; Jessen, 2012). In forensic phonetic casework, a suspect's voice recording is often compared to a recording of a questioned speaker (forensic voice comparison). The task of the phonetic expert is to assess whether the linguistic/phonetic features observed are more likely under the hypothesis of speaker identity or non-identity (and how much more likely in one or the other direction). FPs and other disfluencies could be one of the features of speaker characteristics to include in the voice comparison analysis. Specific features to be considered may be the general frequency distribution of the FPs, their type and duration, the pause context, i.e., whether they occur within a pause or speech, the proportion of creaky voice produced during the FPs, and the vowel quality of the vowel produced in FP types like *uh* and *um*. In order for the features to be applied in forensic phonetic casework, the distribution of the features in the relevant population must be known to the expert to assess the typicality and similarity between the compared recordings (Rose, 2002). For example, suppose a particular feature instantiation (e.g., a certain value of *uh* per minute) is rare in the population, and similarity in the recordings is high. In that case, the strength of evidence in the direction of speaker identity is higher than if, given the same similarity, the feature instantiation is quite common. For all features that may prove useful in forensic casework, the general distribution and variation within the population (between-speaker variation) has to be known as well as the within-speaker consistency.

In this chapter, we do not provide full documentation of between-speaker variation for each of the many disfluency features investigated here, nor do we document within-speaker variation to the fullest level of detail. Plots showing how many speakers display which feature instantiation can be found in Appendix A. This chapter focuses on mean patterns in two conditions (normal and Lombard), which allow the reader to see typical patterns across speakers and decipher variable speaker-internally across the conditions. A more in-depth analysis of within- and between-speaker variation in a subset of the speakers investigated in this chapter is also provided (Section 4.5). The fact we chose a Lombard condition is forensically relevant as well because variations in vocal effort are common in forensic casework.

4.2 Materials

4.2.1 Corpus

The data used here is part of the Pool2010 corpus that was compiled in 2001 by the German Federal Criminal Police Office (Bundeskriminalamt, Wiesbaden) during a research project (Jessen et al., 2005). The recordings of 107 male native German speakers, most of them employees from this office, were collected during several tasks. Seven speakers were excluded due to technical issues or problems such as a speaker's voice disorder. While several speech tasks were recorded for the corpus, this analysis only uses the semi-spontaneous speech task. This speech task was similar to the "taboo" game, in which the speaker has to describe terms in their own words without using the two to three given taboo words[5]. Each speaker described seven words in the mean per condition (range: 2-11 words). The number of taboo words may have influenced the difficulty of the task, but as each speaker described several words with a differing number of taboo words, the influence of difficulty per speaker was considered balanced. A female interviewer served as the interlocutor for the guessing game, providing the answers when the speaker gave sufficient information. The speakers did not know that she was from the research team and knew the words previously. She extended the speaker's description to a certain degree by not providing the correct word immediately.

The speakers completed the task in two conditions, one normal speech condition and one Lombard speech condition, for which the speakers heard white noise (80 dB_{SPL}) over headphones, which led to louder speech by the participants. The order of the two conditions was changed from subject to subject to prevent potential serial order effects. The Lombard effect describes the phenomenon in which a noisy environment causes speakers to increase their level of vocal effort, which results in, not only, louder speech but also a raised fundamental frequency (Lombard, 1911). The presentation of white noise also leads to an impaired feedback loop during speech production. Usually, a speaker hears their own speech and monitors it for possible errors, e.g., slips of the tongue or the like. It is possible that this shortcoming has an impact on the production of disfluencies and also on FPs. It has been shown that fluency increases in people who stutter when the subjects are presented with noise over headphones (Adams & Hutchinson, 1974). Furthermore, the level of noise was negatively correlated with the number of disfluencies.

The microphone was attached to a helmet that also included the headphones for the Lombard condition. This allowed the distance from the microphone to remain constant and allowed for a quick and easy transition between the conditions. The recordings were made with a sampling rate of 16 kHz and a sampling depth of 16 bits. The resulting audio files have a mean duration of 3:52 minutes (sd = 1:03 minutes), with a range of 1:43 to 7:53 minutes. The part of the corpus used for this study

[5]In some cases also 1 or 4 taboo words.

consists of 12 hours and 56 minutes (both conditions together in total). Annotations have been made as described in Chapter 3. The corner vowels /a:/, /i:/, and /u:/ of each speaker were annotated in selected syllables carrying the lexical stress where they would appear in standard German. The aim was to collect 10 tokens for each vowel and condition per speaker, this aim, however, could not be reached for some speakers as the recordings did not provide enough tokens.

4.2.2 Speaking tempo

The speaking tempo was measured in each file with the help of the Praat script "Praat Script Syllable Nuclei v2" created by de Jong & Wempe (2009). As the data includes long pauses (p_w, tc) that influence the speech rate, these pause types are excluded using a Praat script, and the resulting new (shorter) files were used for the speaking tempo measurements. Simple pauses (p) were kept intact. The script detected intensity peaks surrounded by intensity dips, which were then considered syllable nuclei. Three measures of the script were adjustable: the silence threshold, the minimum dip between peaks, and the minimum pause duration. To find the optimal settings for the script, ten files of the corpus were manually annotated, marking the phonetic syllables and the intervals where speech occurs in line with the output TextGrids that the script produces. This allowed us to determine the syllable number per file and the total speaking time, and these values were then used to test the settings of the script. The settings that resulted in a minimum deviance from the manual values on the ten files were as follows: a silence threshold of -20 dB, a minimum dip of 2, and a minimum pause duration of 500 ms. The mean deviance from the manual values was an over-counting of syllables by 13.8 and a misdetection of speaking time by 7.2 seconds per file. The articulation rate and the speech rate were determined for every file using the optimal settings of the script. Two files resulted in particularly low values for these measures. These were inspected further, and loud bursts and laughs were deleted, which improved the performance of the script with the previously mentioned settings. It is possible that the script performed better on some files than on others, but we assume that the values are a good approximation of the true speaking tempo and a time-efficient alternative to manual annotations.

Figure 4.1 shows that for most speakers the articulation rate in the Lombard condition is actually faster than in the normal condition, contrary to previous findings (Tuomainen et al., 2021). Although the authors in Tuomainen et al. (2021) raise the question of whether the differences in articulation rate are perceptually salient (cf. Quené, 2007), deviations in the articulation rate reported in Jessen (2007) on the same corpus can be explained using the script. The authors of the script report that automatic values are generally lower than values that were manually obtained and need to be multiplied by 1.28 to predict the manual values (de Jong & Wempe, 2009). This is due to the failure of the script to detect some unstressed syllables and, thus, fewer syllables are counted, which leads to a lower articulation rate. Jessen (2007)

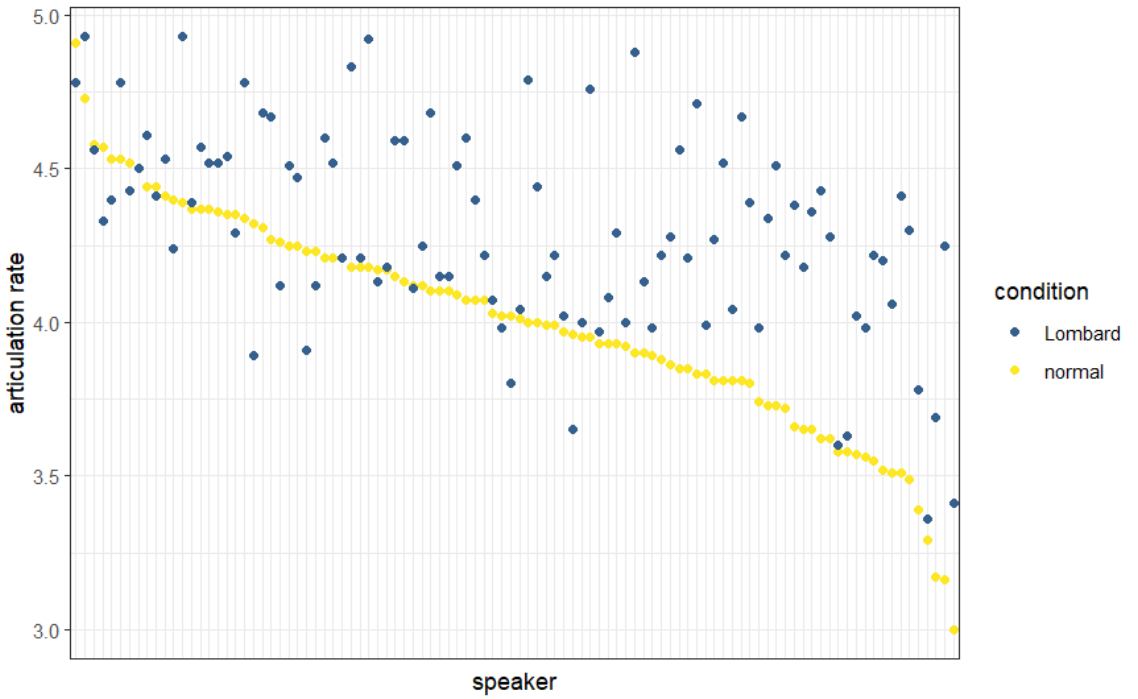

Figure 4.1: Speaking tempo in Pool2010 corpus
Articulation rate (syll/s) per speaker as a function of condition.

reports a mean articulation rate of 5.21 syll/s in the normal speech condition, while the automatically obtained mean values for the same data is 4.0 syll/s. When using the factor reported above to predict the manual mean values from the automatically obtained mean, it becomes clear that this conversion approximates the manual value quite well (4.0*1.28 = 5.12).

4.2.3 Statistical methods

Analyses and plots were done using R (R Core Team, 2022) (version 4.1.3) and the tidyverse package (Wickham et al., 2019). Linear mixed models were created using the lme4 package (Bates et al., 2015) with FP duration, F1, F2, and frequency as dependent variables and FP type, condition, pause context, FP duration (for F1 and F2), and the speaking tempo measures as independent variables. The speaker was included as a random effect, allowing the intercept (but not the slope) to vary between subjects. P-values were obtained using the lmerTest package (Kuznetsova et al., 2017). Models were built by including the condition and FP type with the interaction term and adding the other variables as control factors without the interaction term. We are aware that, when building linear mixed models, one should aim for maximal models. However, when including all interaction terms of numerous factors, the results become difficult to interpret and the model would not serve the research question. As speech and articulation rate are co-dependent variables, only one of these measures was included. The articulation rate was chosen, as this measure is independent of any

pause rate and is considered to be more independent from the disfluency rate than the speech rate measure. The difference between the speech conditions is also larger for articulation rate than for speech rate.

Furthermore, a Pillai score is calculated as a measure of overlap between vowels (Kelley & Tucker, 2020). It is determined by calculating a Multivariate Analysis of Variance (MANOVA) using the first and the second formant as response variables and the FP type as a predictor variable (using the manova-function from the stats package (R Core Team, 2022)).

4.3 General results

In the following, a general overview of the FPs in the Pool2010 corpus will be given using the normal and Lombard conditions. A comparison of normal speech to Lombard speech follows. Another section focuses on between-speaker variation and within-speaker consistency. Therefore, the Lombard speech condition is regarded as a second speech mode in order to compare two files of the same speaker.

Frequency distribution

In the entire corpus material, 6,734 FPs have been detected (12h 56 min), which results in a ratio of 8.67 FPs per minute. When looking at the rate of FPs per type (Table 4.1), it becomes clear that tongue clicks are most frequent, closely followed by the vocalic FPs. The vocalic-nasal FP type *um* is only half as frequent as the vocalic type *uh*. The number of *um* instances is just as high as the number of the glottal FPs and the nasal type *hm* taken together.

Considering only the FPs *uh* and *um*, as done in Figure 4.2 with their pause context, we can see that the FP *um* amounts to only around one-third of all typical FPs. The most frequent FP type in the pause context is *uh* surrounded by speech, i.e., within no silence.

Table 4.1: Frequency and duration of FPs

*Absolute numbers and rate per minute of FPs (*uh*,* um*,* hm*, glottal FPs, and tongue clicks), and mean (sd) durations (in ms) of the phenomena, together with the vowel duration of* uh *and* um *types. NA (not applicable) means that the duration was not measured (e.g., for clicks) or that the phenomena did not include a vowel. Creaky voice portions are included in the total duration and vowel duration.*

FP type	absolute	rate: FPs/min	duration mean (sd)	vowel duration mean (sd)
uh	2,250	2.9	382 (180)	382 (180)
um	1,054	1.4	559 (234)	281 (125)
hm	314	0.4	442 (224)	NA
glottal FP	757	1.0	244 (332)	NA
clicks	2,359	3.0	NA	NA

Duration

The duration of FPs is highly variable as seen by the high standard deviation values in Table 4.1 for each phenomenon. Glottal FPs have the shortest duration (but the highest standard deviation), followed by the vocalic type (*uh*), the nasal type (*hm*), and the vocalic-nasal type (*um*), with the latter being the longest. The vowels in *uh* are longer than in *um*.

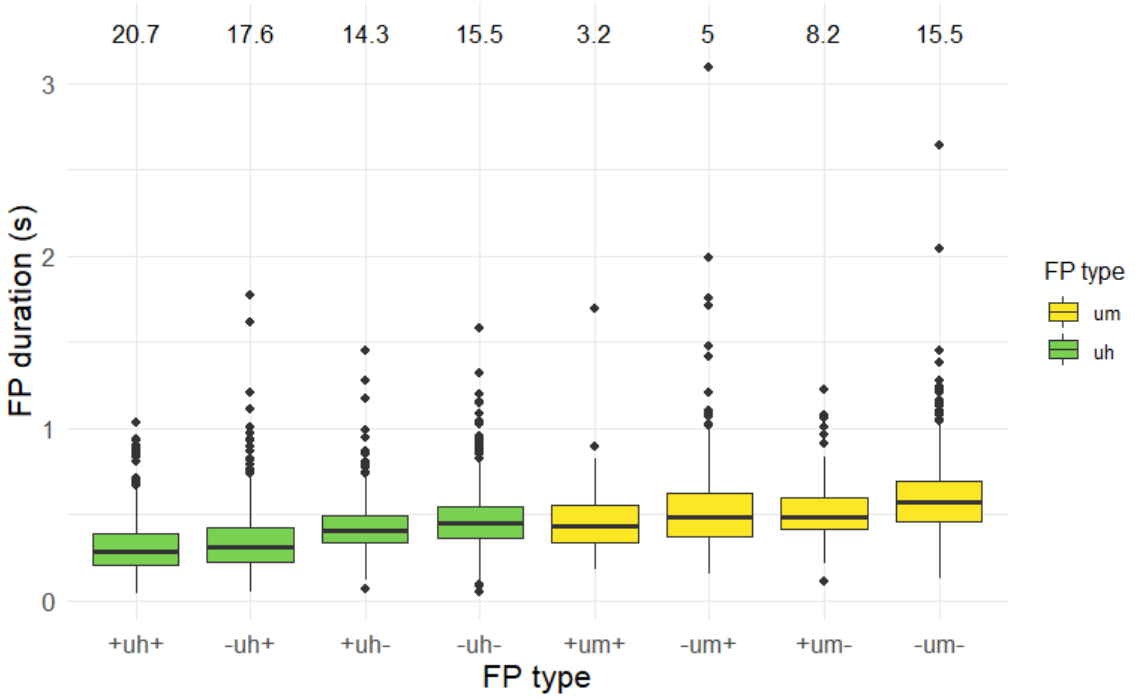

Figure 4.2: Duration hierarchy
Duration of the FPs uh *and* um *in their pause contexts in seconds (s). Context types are described using a + to denote speech and a - to denote a silent phase surrounding the FP. The values at the top refer to the percentage of each displayed FP type for all FPs* uh *and* um.

Considering only the FPs *uh* and *um* in their pause context (see Figure 4.2), it becomes apparent that not only are vocalic-nasal types longer than vocalic types, but that there seems to be a duration hierarchy depending on pause context. FPs in speech (+FP+) are shorter than FPs surrounded by pauses (-FP-). Moreover, IPU-final FPs (+FP-) are longer than IPU-initial FPs (-FP+). This pattern applies to the vocalic type *uh* as well as the vocalic-nasal type *um*.

Pause context

The values at the top of Figure 4.2 show the percentage of each FP type in the pause context. We see that the most frequent type is the vocalic FP within speech (+uh+), followed by the vocalic FP at the beginning of IPUs (-uh+). Both FP types, *uh* and *um*, occur in isolation in 15.5% of all cases. For the vocalic-nasal FP, this is by far the most frequent type. For the vocalic FP, the most frequent type is the within-speech type. Figure 4.3 shows how many FPs were preceded and followed by a pause as well as the type of pause. Simple pauses are more frequent than the other pause types, and pauses, in general, are more frequent surrounding the nasal and the vocalic-nasal FP types than surrounding the vocalic type. Additionally, the pause type "tc" is more frequent preceding an FP than following an FP, which means a task is more

Table 4.2: Pause durations (in ms)

Durations of different pause types pooled over both conditions (normal, Lombard).

pause type	pre FP	post FP
simple pause (p)	1,177 (1,222)	1,083 (1,227)
waiting pause (p_w)	3,182 (2,302)	2,285 (1,563)
task change (tc)	3,962 (2,706)	3,560 (2,283)

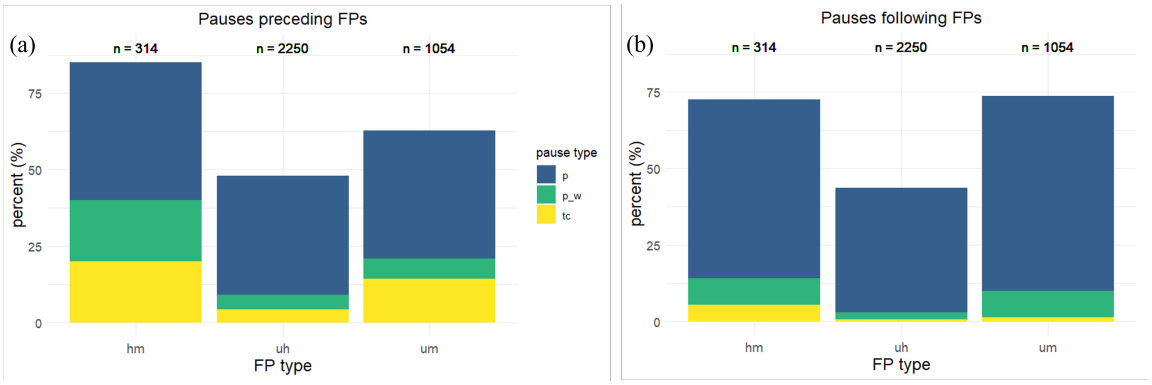

Figure 4.3: FPs surrounded by different pause types

Percentage of different FP types that are preceded (a)/followed (b) by a pause. The colours show different types of pauses (simple pause (p), waiting pause (p_w), task change (tc)). The values at the top show the values representing 100% for each FP type.

often started with an FP than closed with an FP. This observation is not surprising, as turn-initial FPs are reported more frequently than turn-final ones, also in other corpora (Swerts, 1998; O'Connell & Kowal, 2005).

The duration of the pauses surrounding an FP varies to a high degree as indicated again by the high standard deviation values (Table 4.2). A linear model including pause type (simple, waiting, task change) and pause position (pre FP, post FP) as predictors for pause duration indicated that the three different pause types (simple, waiting, task change) are significantly different from one another (Table 4.3). Furthermore, while simple pauses (p) and task changes (tc) do not differ in their duration depending on the pause position, waiting pauses (p_w) do differ significantly in their duration. These pauses are longer before an FP than after an FP.

Voice quality

A large percentage of FPs are produced with an initial creaky voice. Nearly 46 % of *uhs* include initial creaky voiced portions or glottal pulses, while 41 % of *ums* include creaky voice sections. Not even 7 % of *uhs* and as little as 1.14 % of *ums* include final creaky voiced portions or glottal pulses. 189 tokens of *uhs* and *ums* (5.7 %)

Table 4.3: LM of pause duration

Model output of the linear model for the pause duration as dependent variable and pause position (pre/post), pause type (p, p_w, tc) as independent variables.

	Estimate	Std. Error	t-value	Pr(<|t|)
(Intercept)	1.083	0.036	30.32	<0.001 ***
typepre	0.094	0.053	1.77	0.08 .
pausetypep_w	1.202	0.121	9.96	<0.001 ***
pausetypetc	2.477	0.227	10.92	<0.001 ***
typepre:pausetypep_w	0.803	0.16	5.02	<0.001 ***
typepre:pausetypetc	0.308	0.246	1.26	0.21

are produced with 100 % creaky voice, these are included under the initial creaky voice category.[6] Figure 4.4 shows the ratio of FPs that are produced with creaky voice portions or glottal pulses in the beginning (a) or at the end (b) of the FP as a function of pause context. It becomes apparent that creaky voice is frequent in the beginning of FPs and is relatively rare at the end of FPs. The vocalic-nasal type *um* shows especially low numbers of creaky voice portions in the final position. It seems that the nasal consonant leads to an FP that is less likely to be produced with creaky voice. Duration measurements of creaky voice portions and glottal pulses seem to be stable across FP types (*uh* and *um*). The difference between the mean duration values of initial and final creaky voice portions/glottal pulses is (just) not significant, as determined by a t-test (initial 122 ms vs. final 145 ms; t = -1.95, p = 0.05). Standard deviation values are 144 ms against 141 ms, which is quite long considering the mean values. This again shows the high variation within the feature of non-modal voice quality.

Vowel quality

Vowel quality was measured at the temporal midpoint of the vowel for the FPs *uh* and *um* using the Burg method provided by Praat[7](Boersma & Weenink, 2022). FPs are considered to have rather stable formants (*uh* more so than *um*), which is why a midpoint measurement was chosen (Hughes et al., 2016). For the analysis of vowel quality, and to reduce any measurement errors, the dataset (n = 3,304) was reduced to only those observations within three standard deviations from the mean of the first and second formant (n = 2,996; 308 observations excluded). Measurements for corner vowels were done in the same way. The aim was to annotate 10 tokens for each corner vowel in each file of the speakers. This aim was not always achieved, as the close rounded back vowel is under-represented in the data set of lexical vowels (token

[6]Note that these are not the same as glottal FPs, as a vowel may still be discernible.

[7]Maximum formant: 5000 Hz, maximum number of formants: 5, window length: 0.025 s, dynamic range: 50 Hz

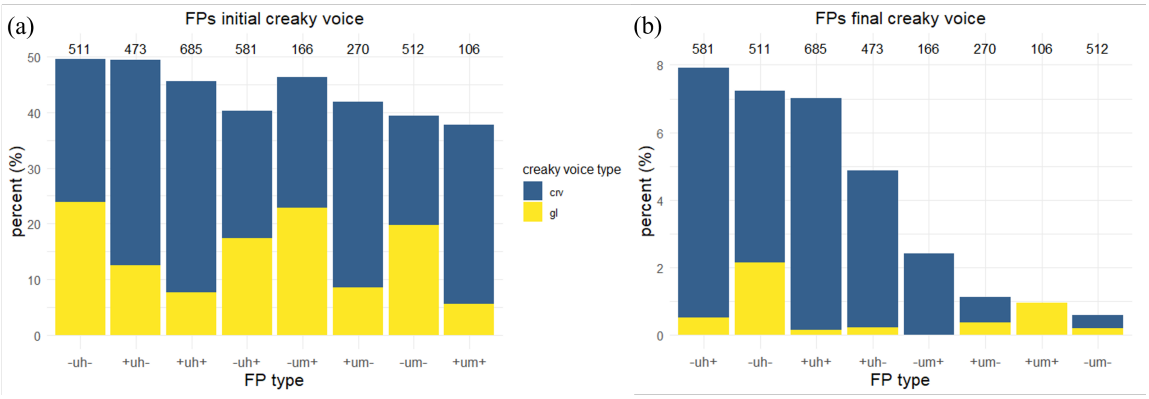

Figure 4.4: Voice quality in FPs

Creaky voice portion at the beginning (a) and end (b) of FPs. The values at the top show the values representing 100% for each FP type. Note that the scales of graphs a) and b) are different as there are considerably fewer FPs including creaky voice or glottal pulses at the end of the FP.

numbers after reduction: /iː/ = 1,214, /aː/ = 1,460, /uː/ = 514). In the graphical representations of vowel quality, the first formant values are plotted on the y-axis and the second formant values on the x-axis; both axes are inverted to better match the vowel quadrilateral, as is common practice in phonetics research. In this chapter, ellipses are drawn to represent in which area 95 % of the data are distributed, but each data point is represented by a coloured point in the plots as well.

As seen in Figure 4.5, the vowels of the FPs *uh* and *um* show a high degree of overlap which is also supported by the Pillai-score of 0.03. The Pillai score is a measure of overlap ranging from 1 (no overlap) to 0 (complete overlap) (Kelley & Tucker, 2020). It is determined by calculating a multivariate analysis of variance (MANOVA) using the first and the second formant as response variables and the FP type as a predictor variable. A Pillai score was also calculated for each corner vowel to measure the overlap of each lexical vowel with the FP vowel. The values support the relationship between vowels visible in Figure 4.5. The high front vowel has little overlap with the FPs (Pillai = 0.8), and the German central /aː/ has more overlap (Pillai = 0.5), along with the high rounded back vowel (Pillai = 0.3). It is important to note that the number of tokens for /uː/ is considerably lower than for the other two corner vowels.

Discussion

To conclude the general results, it can be seen that tongue clicks are frequent in this corpus, as is the FP type *uh*. The overall disfluency rate of 8.67 FPs per minute is driven considerably by the high rate of clicks in the corpus. When excluding clicks (as is often done when looking at FPs), the rate of 5.7 FPs per minute is closer to the rates reported in previous literature (Belz, 2021). Contrary to our assumption

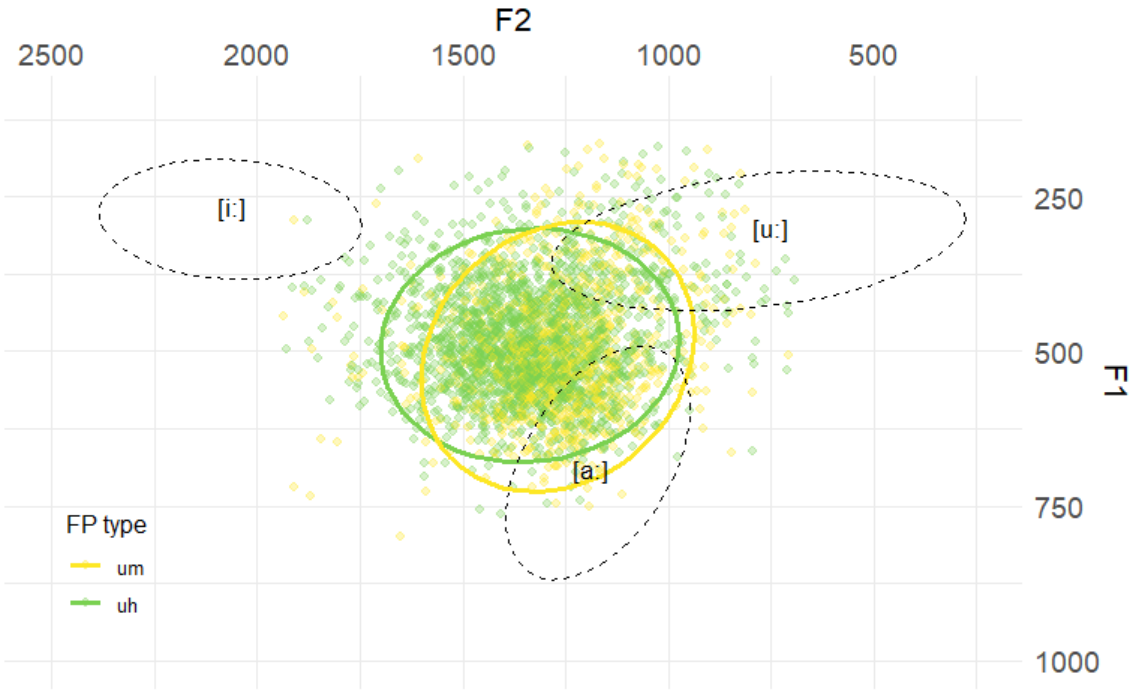

Figure 4.5: Vowel quality of FPs

Vowel quality of the FPs uh *and* um *in an F1-F2 chart in comparison to the corner vowels* /aː/, /iː/ *and* /uː/ *by the same speakers. Ellipses include 95 % of all data points. Formant values in Hz.*

(see 4.1.1), the vocalic-nasal type *um* is less frequent than the vocalic type. This is surprising as Wieling et al. (2016) and de Leeuw (2007) report a change towards *um* being more frequent than *uh* in Germanic languages. This may be due to the fact that our corpus represents the FP use in the year of recording, which was 2001. Furthermore, our participants are only male speakers, while language change is more often led by females (Labov, 1990; Wieling et al., 2016). The least frequent FP type is the nasal type, even glottal FPs are more frequent. Glottal FPs should be considered in analyses of disfluencies, as they are nearly as frequent as the FP type *um* in this corpus.

As expected, vocalic-nasal FPs are produced with the longest duration (Hughes et al., 2016), followed by nasal FPs and vocalic FPs. Glottal FPs have the shortest duration. Furthermore, for the typical FPs *uh* and *um*, a duration hierarchy can be seen where FPs surrounded by pauses are the longest (-FP-), followed by IPU-final FPs (+FP-), and IPU-initial FPs (-FP+). Within-speech FPs (+FP+) show the shortest duration. Similar trends were also found by Gósy & Silber-Varod (2021) for Hungarian vocalic FPs. For *uh*, the within-speech context is the most frequent, while for *um* the within-pause context is the most frequent.

The most frequent pause context for the vocalic FP is the within-speech context. For the vocalic-nasal FP, however, it is the type in isolation, i.e. within pauses on both sides. In general, simple pauses are more frequent than waiting pauses or task changes. After task change pauses, the FP types *um* and *hm* are much more frequent than the vocalic *uh*. This pause type (task change) occurs infrequently after FPs, which suggests that FPs are used when the speaker starts their speech. Pauses occur more often with the nasal or the vocalic-nasal type than with the vocalic FP type. In terms of pause duration, we find a significant difference in the duration of the waiting pause type, meaning that these pauses before an FP are longer than after an FP. This suggests that when an IPU ends with an FP before a waiting pause, the speaker is quicker in picking up their thoughts than after a waiting pause. In this case, the speaker may use the FP to buy time for formulating their next thought.

Creaky voice and glottal pulses are considerably more frequent in the particle-initial position than in the particle-final position which is also reported in Belz (2021). In the particle-final position, only a small portion of *um* shows creaky voice, while the percentage for *uh* is higher. A possible explanation for this striking difference is that since the FPs *uh* and *um* begin with a vowel, this vowel-initial position corresponds to the context in which the glottal stop can occur in German (i.e., words beginning with vowels). According to Kohler (1994), the glottal stop in German is frequently realised as creaky voice.

In terms of vowel quality, the vowels of *uh* and *um* show a high degree of overlap with each other. Furthermore, they spread over a large portion of the central vowel space. The similarity between the vowels of the two FP types was also found by Hughes et al. (2016) in British English.

4.4 Normal vs. Lombard speech condition

Lombard speech is produced with a higher vocal effort, which typically results in a rise of fundamental frequency (Jessen et al., 2005). Hyperarticulation, especially of the jaw and the tongue, has been reported (Šimko et al., 2016). This increased jaw opening accounts for the increase in first formant values that are typically reported in Lombard speech (Van Summers et al., 1988). Other effects of Lombard speech include a slower speaking tempo, which includes a lower articulation rate (Tuomainen et al., 2021), and possibly a higher pause rate. Our data reveals a lower speech rate in Lombard speech as determined by a t-test (t = -2.13, $p < 0.03$; 2.84 syll/s vs. 2.68 syll/s), but a higher articulation rate in Lombard speech (t = 6.07, $p < 0.001$; 4.01 syll/s vs. 4.3 syll/s) (see Table 4.1). Effect sizes reveal a large effect for the difference in articulation rate (d = 0.86) and a small effect size for speech rate (d = -0.3), which is why articulation rate is the preferred factor for the following linear mixed models.

Frequency distribution

The rate of FPs in normal (8.67 FPs/min) and Lombard speech (8.69 FPs/min) appear stable. However, when looking at the different FP types, it becomes apparent that they are not as stable as the overall rate suggests. This is determined by using a linear mixed model with the rate of each FP type per speaker as a dependent variable and an FP type, condition and articulation rate as independent variables, and speaker as a random intercept: $lmer(freq\_rate \sim fp\_type * condition + articulationrate + (1 \mid speaker), data = data)$. The effect plot in Figure 4.6 shows that the rate for clicks and glottal FPs increases in the Lombard condition, though the model (see Table 4.4) shows the difference in conditions is not significant in the glottal FPs. The rate of the FPs *uh*, *um*, and *hm* decreases in the Lombard condition, this difference is statistically significant. Articulation rate was included in the model, but the factor did not reach statistical significance, i.e., it had no influence on the frequency of FPs. Possible reasons for these results are outlined in the discussion of this section.

Duration

To determine the influence of condition on the duration of FPs, the following model was fit: $lmer(freq\_dur \sim condition * fp\_type + articulationrate + prepause + postpause + (1 \mid speaker), data = data)$. The model output can be seen in Table 4.5. It shows that FPs in the Lombard condition are on average 48 ms longer than in the normal condition, but also that the vocalic-nasal FP *um* is 162 ms longer than the other FPs. The pause context also has an effect on the duration in that FPs are shorter when no pause precedes or follows the FP. The pause after an FP has a larger effect on duration (-109 ms) than the pause before an FP (-37 ms). The articulation rate was included as a control variable but did not influence the duration of FPs.

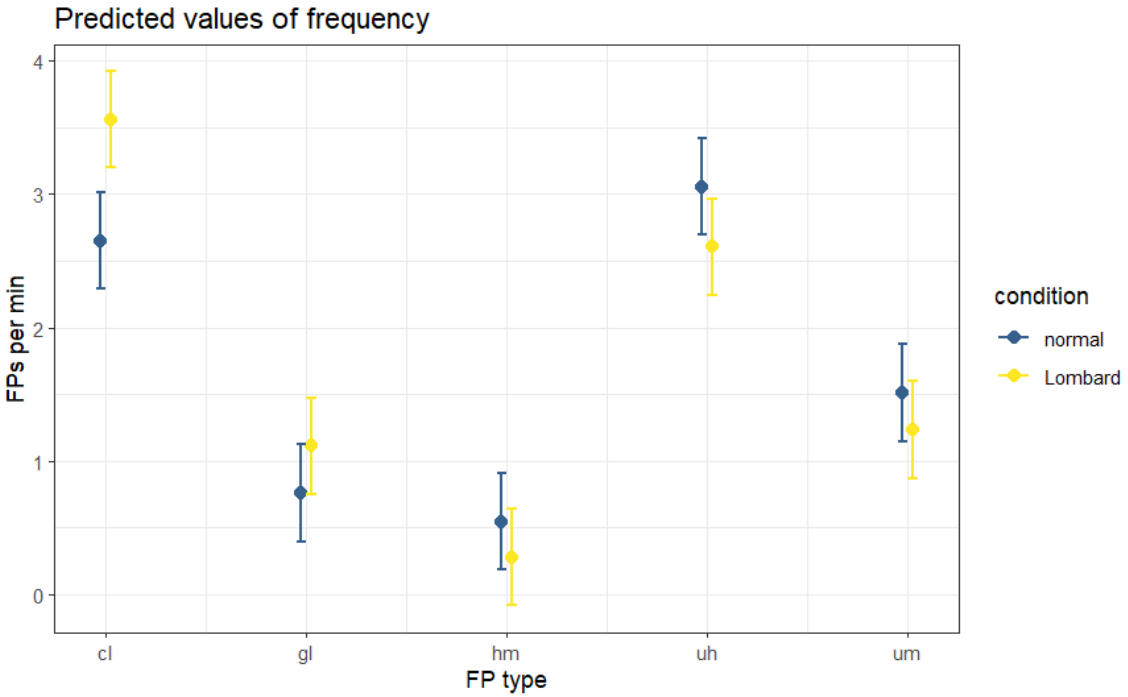

Figure 4.6: Predicted values of frequency
Effect plot of frequency (FPs/min) as a function of FP type and condition.

Table 4.4: LMM of FP frequency

Model output of the linear mixed model for frequency with the rate per minute of FPs as a dependent variable and FP type, condition and articulation rate as independent variables.

	Estimate	Std. Error	df	t-value	Pr(<|t|)
(Intercept)	3.31	0.87	280.07	3.81	<0.001 ***
fp_typegl	-1.89	0.24	890.83	-7.81	<0.001 ***
fp_typehm	-2.1	0.24	890.83	-8.7	<0.001 ***
fp_typeuh	0.41	0.24	890.83	1.67	0.09
fp_typeum	-1.14	0.24	890.83	-4.71	<0.001 ***
conditionLombard	0.91	0.25	947.46	3.65	<0.001 ***
articulationrate	-0.16	0.21	266.15	-0.75	0.45
fp_typegl:conditionLombard	-0.56	0.34	890.83	-1.63	0.1
fp_typehm:conditionLombard	-1.18	0.34	890.83	-3.45	<0.001 ***
fp_typeuh:conditionLombard	-1.36	0.34	890.83	-3.98	<0.001 ***
fp_typeum:conditionLombard	-1.19	0.34	890.83	-3.47	<0.001 ***

Table 4.5: LMM of FP duration

Model output of the linear mixed model for duration with the duration of FPs (in seconds) as a dependent variable and FP type, condition and articulation rate as independent variables.

	Estimate	Std. Error	df	t-value	Pr(<|t|)
(Intercept)	0.502	0.065	724.9	7.7	<0.001 ***
conditionLombard	0.048	0.009	3,027	5.55	<0.001 ***
fp_typeum	0.162	0.01	3,289	16.11	<0.001 ***
articulationrate	-0.017	0.016	772	-1.04	0.3
prepause+	-0.037	0.007	3,296	-5.63	<0.001 ***
postpause+	-0.109	0.007	3,280	-16.26	<0.001 ***
conditionLombard:fp_typeum	-0.023	0.013	3,261	-1.73	0.08 .

Furthermore, the aforementioned duration hierarchy is still applicable, where pause context affects the duration of *uh* and *um* to a similar degree.

Pause context

A similar distribution of pause types preceding and following FPs is observed in Lombard compared to normal speech. The types *hm* and *um* are more often followed and preceded by a pause than the vocalic type. The Lombard condition does not seem to affect the rate of FPs that are surrounded by pauses, nor the type of pauses.

Pause duration, however, is affected by the Lombard condition as determined by a linear mixed model with pause duration as a dependent variable and pause position (before/after FP) and condition and pause type as predictor variables (Table 4.6). A significant effect of pause type shows that waiting pauses are longer than simple pauses by 1.2 seconds on average, and task change pauses are longer than simple pauses by 2.2 seconds on average. Furthermore, the Lombard condition increases the pause duration by 200 ms, however, there is an interaction between the pause type, condition, and pause position, which means that waiting pauses before FPs are yet again longer in the Lombard condition than in the normal condition (see Table 4.7).

Voice quality

The percentages of FPs that include creaky voice portions or glottal pulses in normal compared to Lombard speech seem to only differ for the FP type *uh*. In normal speech, 49 % of all vocalic FPs include creaky voice or glottal pulses while only 42 % are affected in Lombard speech. The difference for the FP type *um* is not as large, however, a slightly higher portion is also affected in Lombard speech compared to normal speech (normal = 40 %, Lombard = 42 %). There seems to be no pattern in the pause context (see Figure 4.7). The 189 FPs that include 100 % creaky voice are

Table 4.6: LMM of pause duration

Model output of the linear model for the pause duration as a dependent variable and pause position (pre/post), pause type (p, p_w, tc) and condition (normal/Lombard) as independent variables.

	Estimate	Std. Error	t-value	Pr(<|t|)
(Intercept)	0.985	0.048	20.53	<0.001 ***
typepre	0.045	0.072	0.62	0.54
pausetypep_w	1.204	0.165	7.31	<0.001 ***
pausetypetc	2.206	0.284	7.78	<0.001 ***
conditionLombard	0.211	0.07	2.99	<0.01 **
typepre:pausetypep_w	0.408	0.213	1.92	0.06 .
typepre:pausetypetc	0.152	0.31	0.49	0.62
typepre:conditionLombard	0.09	0.105	0.85	0.39
pausetypep_w:conditionLombard	-0.012	0.238	-0.05	0.96
pausetypetc:conditionLombard	0.765	0.46	1.66	0.1 .
typepre:pausetypep_w:conditionLombard	1.048	0.317	3.31	<0.001 ***
typepre:pausetypetc:conditionLombard	0.253	0.496	0.51	0.61

Table 4.7: Comparisons of pause durations in two conditions

Pause durations (mean and standard deviation) in normal vs. Lombard speech, divided by their position (-FP: preceding an FP, and FP-: following an FP) and the type of pause (simple pause = p; waiting pause = p_w; task change pause = tc).

Pause Position	Pause type	normal mean (sd) in ms	Lombard mean (sd) in ms	Difference in ms
-FP	p	1,030 (919)	1,330 (1,457)	300
	p_w	2,642 (1,473)	3,978 (2,984)	1,336
	tc	3,389 (1,903)	4,706 (3,346)	1,317
FP-	p	985 (1,123)	1,196 (1,329)	211
	p_w	2,189 (1,313)	2,388 (1,795)	199
	tc	3,192 (1,807)	4,167 (2,863)	975

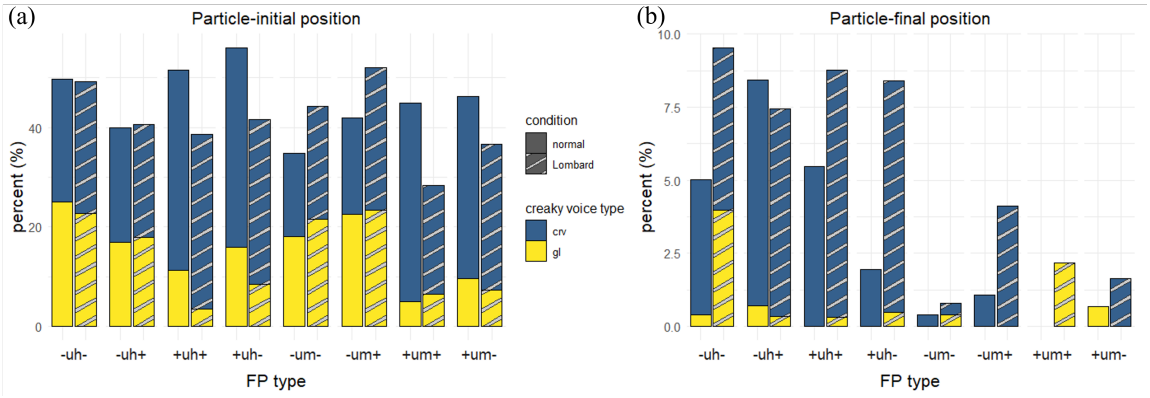

Figure 4.7: Voice quality in normal and Lombard speech
Creaky voice portions and glottal pulses (colours) in the FPs uh *and* um *divided by their pause context (+/-) and the speech condition (patterns).*

included in Figure 4.7a as initial creaky voice portions (132 tokens in normal speech, 57 in Lombard). However, creaky voiced portions (or glottal pulses) in the FP-final position increase, while in the FP-initial position this is only true for IPU-initial *um* (*-um+*) and *um* in isolation (*-um-*).

Vowel quality

In Figure 4.8, it can be seen that the Lombard condition has relatively little influence on the vowel space taken up by the three corner vowels. The close front vowel is distributed similarly in both conditions, the vowel space for the open central vowel decreases slightly, and the ellipse for the close back vowel is somewhat more round in the Lombard condition than in the normal speech condition. However, the vowel space taken up by the vowels in the FPs *uh* and *um* has drastically decreased. Fitting a linear model (Table 4.8) also suggests that the Lombard condition has a significant effect on the vowel height of FPs: $lmer(f1 \sim condition * fp\_type + articulationrate + prepause + postpause + (1 \mid speaker), data = data)$. The F1 is in the mean 97 Hz higher in the Lombard condition than in the normal condition, i.e., the vowels are produced with a lower tongue position or with a more open jaw. The articulation rate also significantly affects the F1 in that with every 1-unit increase (e.g., from 3 syll/s to 4 syll/s), the F1 increases by 30 Hz. The effect of the occurrence of a pause before the FP also reaches significance, though the difference of an increase by 10 Hz is considered to be negligible (Whalen et al., 2022). A raising of the F1 values due to the greater jaw aperture can also be observed in the literature (Šimko et al., 2016). A correlation between signal amplitude (resulting from speech production alone, not environmental or technical factors) and the F1 value has been observed by Ibrahim et al. (2022) which suggests that the increased F1 is not independent of the increased intensity. There is no significant difference in F2 values between the conditions that were determined by the following model (Table 4.9): $lmer(f2 \sim condition * fp\_type +$

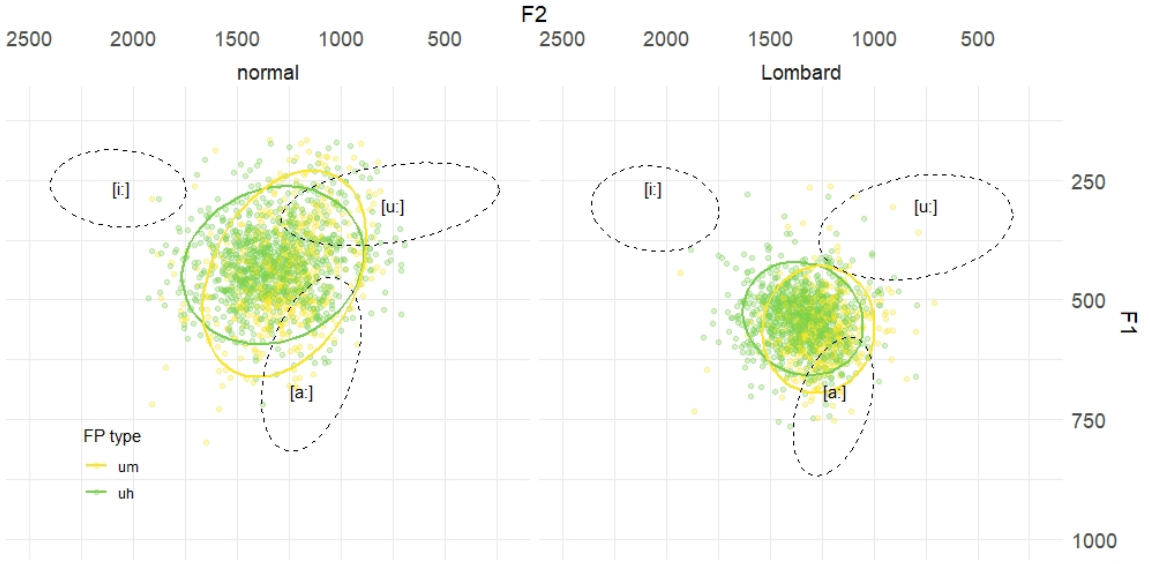

Figure 4.8: Vowel quality of FPs in normal and Lombard speech
Vowel quality of the FPs uh *and* um *in normal vs. Lombard speech in comparison to the corner vowels* /aː/, /iː/, *and* /uː/ *by the same speakers. Ellipses include 95 % of all data points. Formant values in Hz.*

$articulationrate + prepause + postpause + (1 \mid speaker), data = data$). A main effect of duration revealed that a longer FP duration results in a lower F2, i.e., when the FP duration increases by one unit (= 1 second) the F2 decreases by 107 Hz or using more realistic numbers for this field, when the FP duration increases by 100 ms the F2 decreases by approx. 11 Hz. The significant interaction between the Lombard condition and FP type suggests that the condition affects the vocalic-nasal FP but not the vocalic FP. The difference of 34 Hz is rather small and may also have a negligible effect (Whalen et al., 2022).

Discussion

To conclude this section, the results are briefly summarised and some unexpected results are explained in detail. The rate of the typical FPs (*uh*, *um*) and *hm* decreases in Lombard speech while the rate of the glottal FPs and tongue clicks increases. The duration of FPs also increases in the Lombard condition by 48 ms; the articulation rate is not a significant factor for this difference. Pause context affects the duration of the FPs; they are longer when surrounded by pauses and shorter when not. The FP-following pause has a greater effect on the duration of the FP than the FP-preceding pause, which may be explained by the common phenomenon of final lengthening (Lindblom, 1968). Pause durations of both types, FP-preceding and following, also increase but more so for FP-preceding pauses. Furthermore, pause types are 200 ms longer on average in the Lombard condition than in the normal condition. The waiting pause is affected to a different degree: in the FP-preceding position, the

Table 4.8: LMM of F1

Model output of the linear mixed model for the first formant with the F1 (in Hz) as a dependent variable and FP type, condition and articulation rate as independent variables.

	Estimate	Std. Error	df	t-value	Pr(<|t|)
(Intercept)	312	34	1115.03	9.27	<0.001 ***
conditionLombard	97	4	2978.47	22.32	<0.001 ***
fp_typeum	0.6	5	3061.22	0.12	0.9
articulationrate	30	8	1181.06	3.61	<0.001 ***
fp_dur	10	9	3054.38	1.18	0.24
prepause+	-10	3	3046.97	-3.08	<0.01 **
postpause+	-0.5	3	3031.99	-0.14	0.89
conditionLombard:fp_typeum	12	7	3010.67	1.85	0.06 .

Table 4.9: LMM of F2

Model output of the linear mixed model for the second formant with the F2 (in Hz) as a dependent variable and FP type, condition and articulation rate as independent variables.

	Estimate	Std. Error	df	t-value	Pr(<|t|)
(Intercept)	1307	74	1,383.6	17.63	<0.001 ***
conditionLombard	15	9	3,020.62	1.61	0.11
fp_typeum	10	11	3,056.78	0.93	0.35
articulationrate	10	18	1,506.81	0.55	0.59
fp_dur	-107	19	3,047.17	-5.63	<0.001 ***
prepause+	-4	7	3,039.32	-0.54	0.59
postpause+	29	7	3,025.15	3.99	<0.001 ***
conditionLombard:fp_typeum	-34	14	3,007.09	-2.47	<0.05 *

Lombard condition increases its duration by over a second while in the FP-following position this difference is considerably smaller (200 ms).

When looking at voice quality, the most noticeable effect was the increase of creaky voice portions and glottal pulses in the particle final position. In terms of vowel quality in Lombard speech, an increase in F1 mean values was detected for the vowels in *uh* and *um* and a reduction of vowel space used for these vowels. While the increase in F1 values can be explained by the greater jaw aperture (Šimko et al., 2016), the effect on vowel space is more surprising. The vowel target for the FP seems to be reached in the Lombard condition more easily than in the normal condition. This was unexpected, as it is still unclear whether FPs have vowel targets at all (Gick et al., 2004). One could argue that this reduction of vowel space is an act of articulatory precision consistent with Lombard speech as a form of clear speech. However, Lombard speech does not show all characteristics of clear speech (for the concept of clear speech, see Smiljanić & Bradlow (2009)). In particular, there is no general effect of increased formant dispersion in Lombard, as shown here with the corner vowels as well as other studies in which no robust Lombard effect on F2 was found (Gully et al., 2019; Šimko et al., 2016; Garnier et al., 2006; Van Summers et al., 1988; Hay et al., 2017). It could be that the reduction of vowel space in FPs is a secondary effect of the Lombard condition. For example, the increased muscle tension needed for an increased vocal effort may play a role (Wohlert & Hammen, 2000).

One of the results in need of an explanation is that the FP rate for *uh*, *um*, and *hm* is lower in Lombard than in normal speech. One possible explanation could be as follows. Although it is undisputed that FPs can be used by the listener, it is not yet clear whether the speaker actively produces FPs as a signal to the listener, or whether FPs are the result of planning processes etc. on the part of the speaker, which can be detected by the listener, but are not actively intended by the speaker (see Chapter 2.2.1 for discussion). If at least some FPs of the set *uh*, *um*, and *hm* are caused by active signalling from the speaker, it could mean that communicative interaction in the Lombard situation of the corpus was somewhat inhibited, resulting in the reduction of the active signalling of FPs.

This explanation would not cover the behaviour of glottal FPs and clicks, which are more frequent in Lombard than normal speech. These two types of FPs may be difficult to produce deliberately and are, generally, more difficult to perceive (unless they are very long or loud) than the other FP types. In terms of production, creaky voice and clicks are partially the by-product of aerodynamic principles; it is probably difficult to learn the gestural coordination patterns that are necessary to actively and deliberately produce them (both evidenced by the fact that they are infrequently used phonemically in languages). If glottal FPs and clicks are more difficult to control in production, the active signalling function might not work well enough, and if they are more difficult to perceive, speakers would realise that their signalling intention might not reach the listener. Therefore, active signalling of glottal FPs and tongue clicks are unlikely to occur, and the above explanation would no longer apply. This still does not

explain why glottal FPs and clicks are more frequent in Lombard than normal speech and not merely equal in number. For glottal FPs, the effect was non-significant but clicks are significantly more frequent in Lombard than normal speech. One possible explanation for this could be that, due to increased jaw lowering (Schulman, 1989) and presumably some associated tongue lowering in loud speech, the negative air pressure is increased and thereby the production of clicks is enhanced.

A similarly difficult situation occurs when trying to explain the patterns of creaky voice (and glottal pulses) in *uh* and *um*, shown in Figure 4.7a. Keating et al. (2015) in their typology of different kinds of creaky voice, make (among other criteria) a distinction between types that are associated with glottal constriction and those without (cf. Jessen (2012): 52ff., for the forensic relevance of the distinction between constricted and non-constricted types of creaky voice). For predictions about the effect of Lombard speech, it matters whether creaky voice in FPs is of the constricted or the non-constricted type. For the non-constricted type, the prediction is that creaky voice is reduced in the Lombard condition; Lombard is associated with increased P_{sub}, which would have an inhibiting effect on non-constricted creak, which is associated with lowered P_{sub}. For the constricted type the prediction is that creaky voice is increased in the Lombard condition. A clear glottal adduction gesture is useful in Lombard in order to withstand the increase in P_{sub} coming from Lombard (i.e., with increased vocal effort). The particle-initial position is a context in which the constricted type of creaky voice is to be expected as a correlate of the glottal stop in German (Kohler, 1994). This would predict more creaky voice in Lombard than normal, which cannot be seen as a general trend. However, when looking separately at FPs preceded by a pause (-FP) and by speech (+FP), creaky voice gains ground in Lombard in the post-pausal (-FP) position. This is a position in which the glottal stop is more likely to occur than without a preceding pause (Kohler, 1994; Krech, 1968). This means the second explanation seems to have some effect, although it appears to interact with other factors. The patterns in Figure 4.7b (particle-final position) are particularly difficult to understand as it is not clear what kind of creaky voice occurs, but the particle-final position could involve non-constricted creak, especially when occurring before a pause. Nonetheless, there is generally more creak in Lombard, and there is no trend across *uh* and *um* that this effect is smaller before a pause than without a following pause.

4.5 Speaker-specificity

As outlined before, in phonetic casework, it is important to be aware of the distribution of a certain feature within a relevant population as well as within-speaker differences. In this section, we will show differences between speakers, for 12 sample speakers of the corpus, that were selected due to their heavy use of one of the five phenomena (*uh*, *um*, *hm*, glottal FPs, clicks). For each phenomenon, we chose two speakers with the highest rate per minute, and we additionally selected two more

speakers that also occurred among the heavy users in more than one category. The question was if the frequent use of one phenomenon had an effect on the use of the other FPs. Within-speaker differences were addressed by comparing the two conditions, normal vs. Lombard, by examining per-speaker standard deviations across FP-tokens, or both.

Frequency distribution

Figure 4.9 shows the frequency distribution of the five phenomena under investigation in the normal and Lombard speech conditions in 12 sample speakers. The general tendency discussed in section 4.4 is that the frequency of the typical FPs (*uh*, *um*) and *hm* decreases while the frequency of tongue clicks increases. This general trend, however, cannot be seen in all individual speakers. For some, the rate of tongue clicks increases visibly (e.g., v12, v26, v37), and for speaker v75, both clicks and glottal FPs, but here the rate of *uh* and *um* increase in Lombard speech. The rate of the FPs *uh*, *um*, *hm* also increases for some speakers (v12, v45, v69) in Lombard speech, contrary to the general trend. Individual patterns are consistent for some speakers (e.g., v05, v45, v62, v69) for which the frequently used FPs in normal speech also occur in Lombard speech, even though frequencies still change somewhat. For other speakers, one FP type does not occur in Lombard speech at all, like *um* for v26, *hm* for v47, or glottal FPs for v30.

Duration

FP durations for all tokens in normal and Lombard condition combined vary considerably between the speakers (Table 4.10): Mean glottal FP durations range from 93 ms to 629 ms while standard deviations per speaker range from as little as 24 ms to also over 600 ms. A speaker's small standard deviation means that the speaker is quite consistent in producing FPs with similar durations while a high standard deviation means that the speaker varies considerably. Mean duration for the nasal FP range from 160 ms to 721 ms (sd: 82 - 350 ms) for the 12 sample speakers. Durations for the vocalic FP are less variable, they range from 190 ms to 450 ms, and within-speaker variation is also smaller (sd: 70 ms to 280 ms). The general trend that *um* is longer, can also be seen for the speakers selected here (mean: 330-650 ms, sd: 46-315 ms). Glottal FPs seem to vary the most in duration between speakers, while the vocalic FPs *uh* vary the least.

To illustrate the speakers' individual patterns, Figure 4.10 shows the pooled durations of the typical FPs *uh* and *um* in the normal and the Lombard condition. It becomes apparent when looking at the sample set that the durations of FPs in both conditions seem to not depend on each other. It is not the case that FPs are always longer in one condition than in the other nor that the variance is always higher in one of the two conditions. While speakers v12 and v47 show similar standard deviations in both conditions and a similar increase in duration in the Lombard condition, other

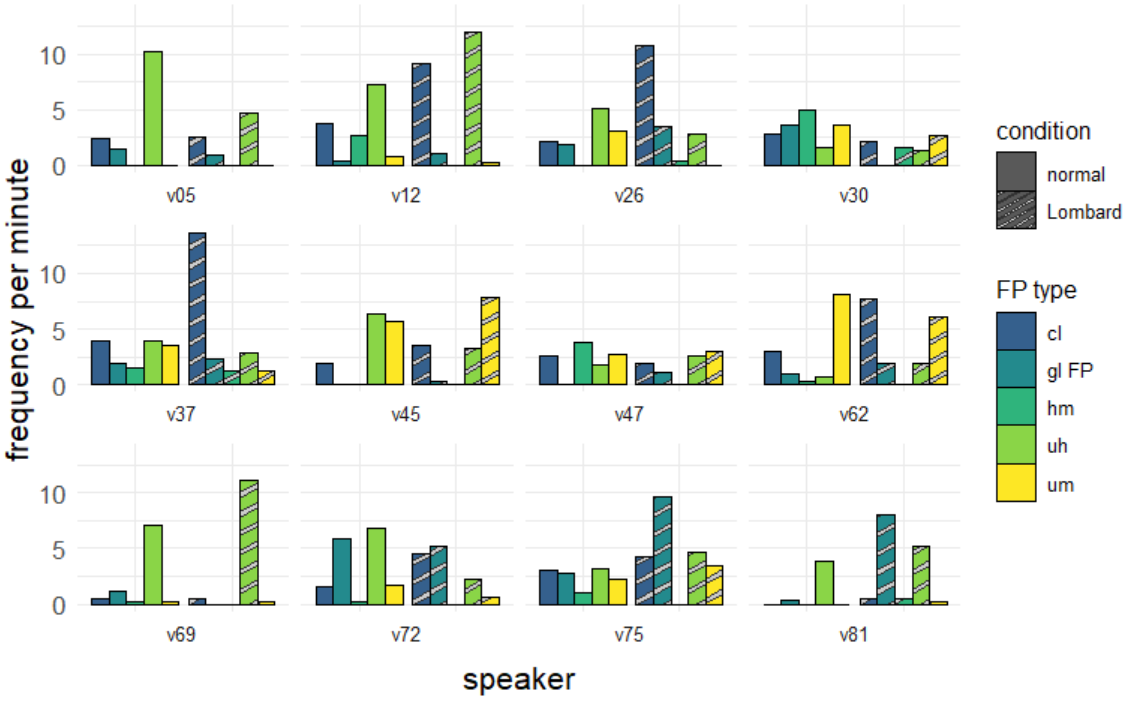

Figure 4.9: FP frequency of sample speakers

Frequency distribution of FPs for 12 sample speakers, comparing their production of FPs in the normal (left-hand side for each speaker) vs. the Lombard condition (diagonal stripes; right-hand side for each speaker).

Table 4.10: Range of FP durations of sample speakers

Minimum (min) and maximum (max) mean duration values and standard deviation (in ms) of the FPs uh, um, hm, *and the glottal FP in the sample set of 12 speakers.*

	min. duration	max. duration
uh	188 (69)	448 (278)
um	334 (46)	648 (315)
hm	161 (82)	721 (349)
gl FP	93 (24)	629 (615)

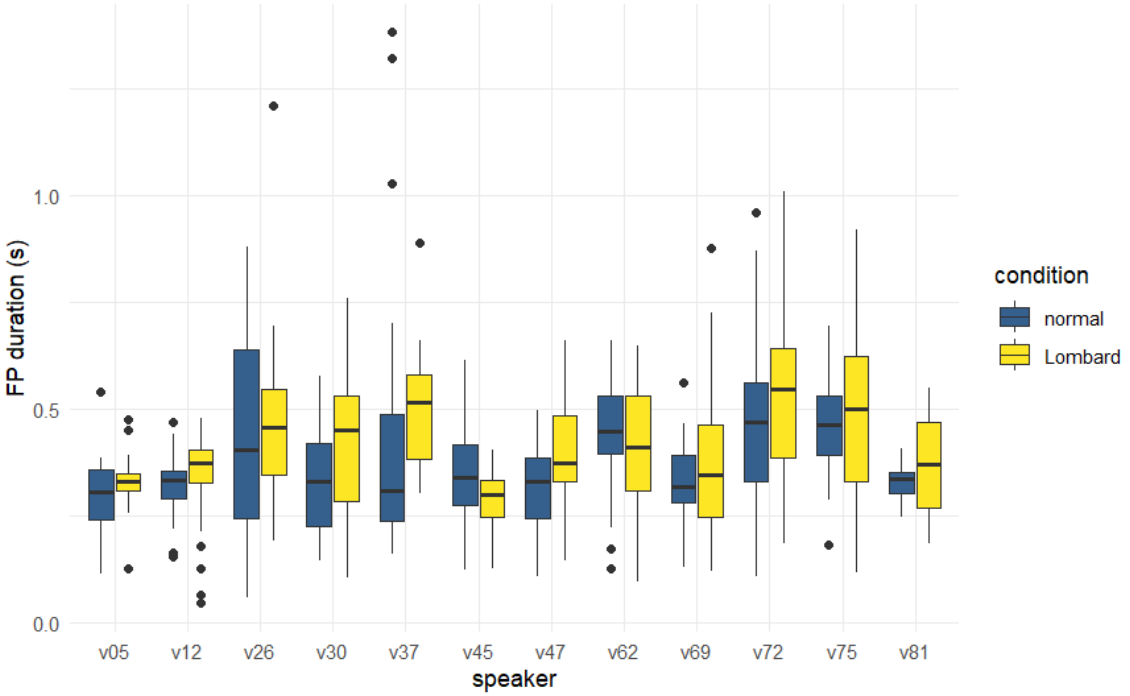

Figure 4.10: FP durations of 12 sample speakers
FP duration (in seconds) of the typical FPs uh *and* um *(pooled) per speaker as a function of condition (normal condition vs. Lombard condition).*

speakers do not show this pattern (v26, v81). Between-speaker differences in FP duration seem to be slightly larger than within-speaker differences that result from the normal-Lombard difference. For example, v12 has a relatively low mean FP duration in both the normal and Lombard conditions, whereas v72 has relatively high values in both conditions. Overall, however, the difference between the within-speaker and between-speaker variation is not large.

Pause context

The mean durations of pauses per speaker, again for all tokens in the normal and Lombard conditions combined, seem to be even more variable than the mean durations of the FPs themselves (see Figure 4.11 for simple pauses). FP-preceding pauses, regardless of their type (simple pauses, waiting pauses, and task changes), range from almost 800 ms to 2,831 ms with the standard deviation being equally variable (sd: 562-2,977 ms). The high standard deviation value seems to be caused by speaker v37 (see Figure 4.11), as the other participants show smaller variations from the mean. FP-following pauses are generally shorter as they range from 474 ms to 1,561 ms (sd: 309-1895 ms). The addition of a condition to this variable does not add any insightful information other than more variance, as also illustrated for FP duration above, which is why we refrain from adding the factor to Figure 4.11.

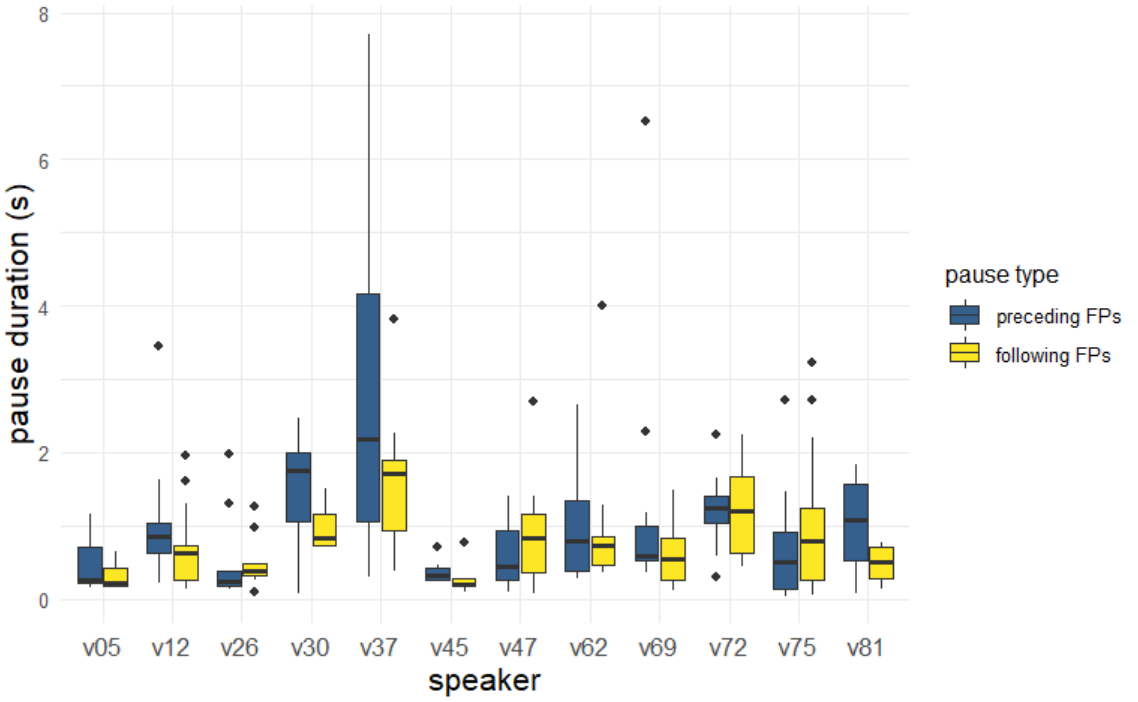

Figure 4.11: Pause durations of 12 sample speakers
Duration of pause surrounding FPs per speaker. Only simple pauses (p) are considered, waiting pauses and task changes are excluded.

Voice quality

We will only look at particle-initial creaky voice or glottal pulses in this section, as they are very infrequent in the final position, as shown above (section 4.4: Voice quality). The production of creaky voice and glottal pulses in the set of sample speakers is highly variable. Some speakers (v26, v30, v37, v45, v47, v81) produce less than 10 instances in each condition with a portion of this voice quality, independent of their total number of FPs produced (see Figure 4.12). A speaker to be highlighted here is v45 who produces 86 FPs (*uh* and *um*) in total, but only three instances of which are produced with initial creaky voice or glottal pulses. All of them are produced in the Lombard condition. Equally low is the number for v37, this participant, however, produces only 32 *uh* and *um* in total. The speakers who produce more than 10 tokens using creaky voice or glottal pulses in their FPs tend to vary more between conditions. For example, speaker v05 produces around 70 % of all their FPs with a non-modal voice in the beginning of FPs. The majority of this is produced with glottal pulses, instead of creaky voice, and more so in the normal condition than in the Lombard condition. Speaker v12, in contrast, produces the non-modal voice portions more frequently in the Lombard condition. Two speakers (v69, v72) produce more than 20 FP tokens with creaky voice or glottal pulses in the normal condition and none or very few in the Lombard condition. Only four out of 12 speakers (v37, v47, v62, v81) did not show differences between the two conditions. Thus, between- and within-speaker

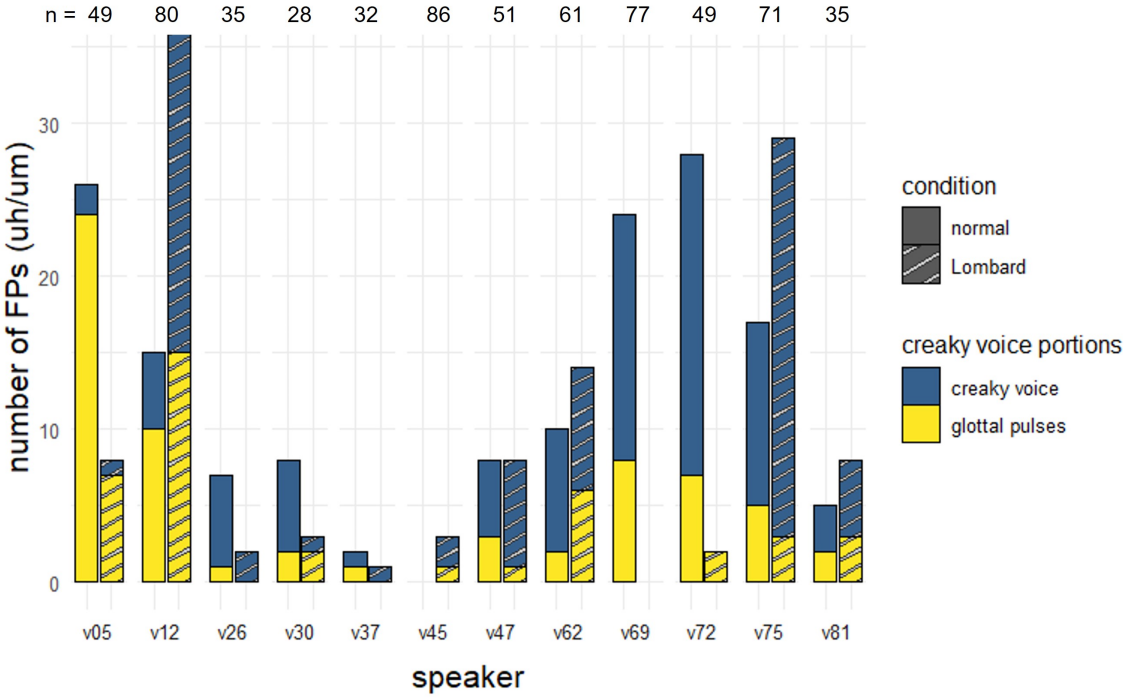

Figure 4.12: Voice quality of FPs produced by 12 sample speakers
Number of particle-initial creaky voice portions and glottal pulses in the 12 sample speakers. The values at the top denote the speakers' total number of uh *and* um*, the FPs where creaky voice and glottal pauses were annotated.*

differences seem to be quite high for this feature.

Vowel quality

The general trend for the Lombard condition was a decreased vowel space and a lowering of the mean F1 value. For our sample speakers, we can see a decrease in the vowel space for nine out of 12 speakers (Figure 4.13). Out of the remaining three, two (v69, v81) show a rather similar-sized vowel space for the vowels in the FPs while only one (v75) shows a clear increase of vowel space in the Lombard condition. All of the sample speakers show a lowering of the vowel space in the Lombard condition compared to the normal condition which is equivalent to an F1 increase. A change in F2 seems to be absent. Speakers vary considerably in how much of the vowel space is used for the FP vowels. While some speakers (v05, v12, v75, v81) produce their FP vowels in a quite limited vowel space, other speakers' vowels (v45, v47, v72) expand over a large portion of the central vowel space.

Discussion

The glimpse into speaker-specificity on the basis of 12 sample speakers shows high variation in the FP frequency distribution, FP duration, pause duration, voice quality,

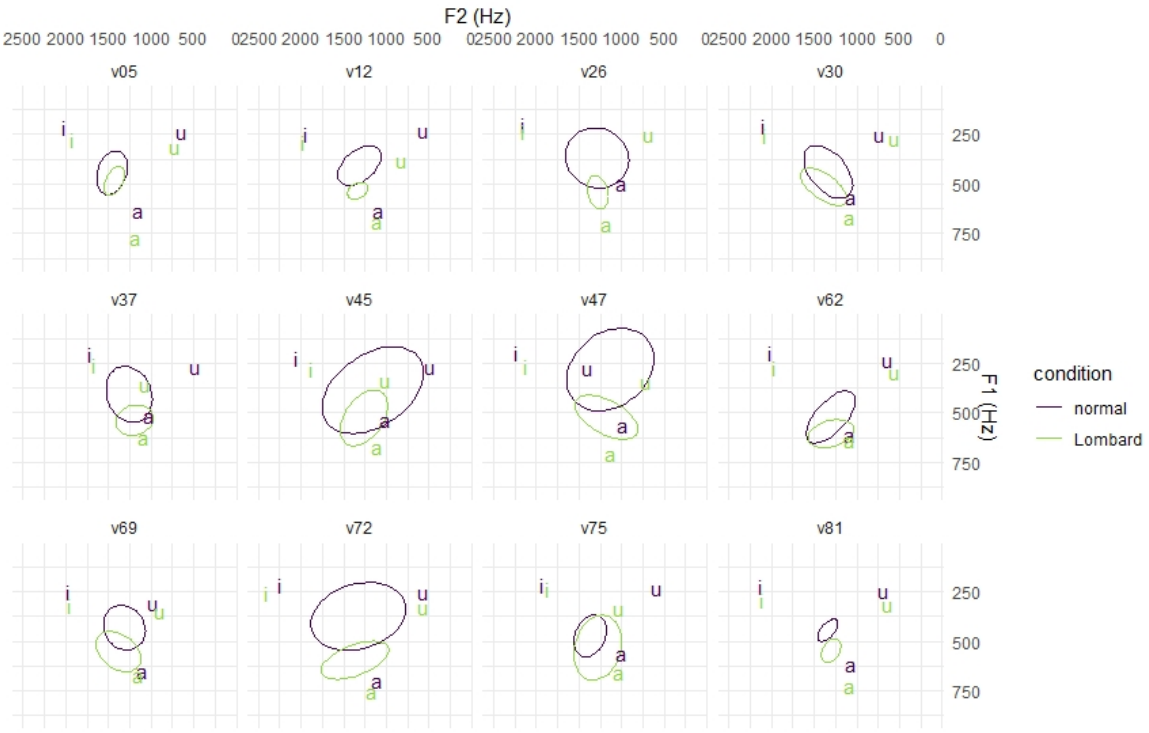

Figure 4.13: Vowel quality of FPs produced by sample speakers
Vowel quality of the FPs uh *and* um *in normal vs. Lombard speech in comparison to the corner vowels* /aː/, /iː/, *and* /uː/ *for 12 sample speakers. Ellipses include 95 % of all data points.*

and vowel quality. Some of this variation is between speakers, which has positive implications for forensic voice comparison, but there is also substantial, though slightly less, variation within speakers. Some part of this within-speaker variation follows the statistical patterns addressed in the previous section, but some of it does not and the degree to which the patterns are congruent or non-congruent with the general trend differs. There are mixed findings in the literature in terms of speaker-specificity using a disfluency profile. McDougall & Duckworth (2018) find a consistent pattern in two tasks (interview and telephone conversation) while Harrington et al. (2021) show deviating patterns for the same speakers in a third task (voicemail message).

A specific characteristic of the study lies in the fact that two conditions (normal and Lombard) were investigated, and these conditions a quite different from one another. These differences are considered to have a relatively strong impact on several speaker characteristics and are therefore classified as "mismatched conditions" (Alexander et al., 2005). This is in contrast to matching conditions (e.g., same or similar speech task, perhaps two weeks apart). Intra-speaker variation is expected to be stronger in mismatched conditions than in matching conditions. It would be possible to reduce some of the intra-speaker variation by applying a normalisation procedure that takes into account the statistically dominant patterns of the normal-Lombard distinction. For example, the duration of the FPs could be increased in the normal condition or decreased in the Lombard condition before the voice comparisons are made. This would increase the comparison scores (reduce the difference) when speakers who follow the dominant trend closely are compared but can decrease them for some other speakers, but altogether such a normalisation procedure would probably increase speaker discrimination. Had the same methods been used on a data set using matching conditions instead, intra-speaker variation for the FP features would be expected to be lower. The results of both the present and aforementioned studies (McDougall & Duckworth, 2018; Harrington et al., 2021)suggest the difference between matching and mismatched conditions is somewhat of a continuum, at least as far as FP patterns are concerned.

4.6 Interim conclusion

As outlined above, we investigated filler particle (FP) use in a spontaneous speech corpus of 100 male native German speakers. For the analyses, we looked into the use of the five phenomena *uh*, *um*, *hm*, glottal FPs, and tongue clicks and the features of the frequency distribution of FPs, FP duration, pause duration, voice quality, and vowel quality.

All in all, it was shown that the phenomena which are often disregarded in disfluency research, glottal FPs and tongue clicks, are equally frequent as the FPs *uh*, *um*, and *hm*. In the Pool2010 corpus (Jessen et al., 2005), the vocalic type *uh* and tongue clicks occur as the most frequent FP phenomena, and glottal FPs are nearly as com-

mon as the FP type *um*, though it should be noted that percussives are grouped under tongue clicks here and may contribute to the high rate of clicks in the corpus. Nasal FPs *hm* are quite rare. Considering the duration of the FP types *uh* and *um*, the pause context seems to influence their duration in that FPs that are produced within pauses are the longest, followed by IPU-final FPs, and then IPU-initial FPs. FPs occurring within speech stretches are the shortest. We call this a duration hierarchy. Previous findings that noted that the FP *uh* is less often flanked by a pause than the FP *um* can also be confirmed (Belz, 2021; Clark & Fox Tree, 2002). The vocalic FP occurs most often within speech, while the vocalic-nasal FP occurs most often within a pause context. Furthermore, *uh* is used very rarely after waiting pauses and task changes, meaning the introduction of new content entails the FPs *um* and *hm* rather than *uh*. We see that non-modal voice is very common at the beginning of the FPs *uh* and *um*, as over 40 % of each FP type include initial glottal stops or creaky voice of varying length. The vowel quality of the typical FPs (*uh*/*um*) spread over a large area of the central vowel space, and they show a very high degree of overlap.

The main findings of the comparison between the normal speech condition and the Lombard condition are the following: Lombard speech promotes the production of tongue clicks, while the frequency of the FPs *uh*, *um*, and *hm* decreases. Durational measures of FPs and surrounding pauses increase but more so for the longer pause type, *waiting pause*. An interesting effect on vowel quality can be observed in Lombard speech. The vowel space in the Lombard condition taken up by the vowels of the FPs is drastically reduced compared to the normal condition, along with the typical F1 increase. Reasons for this are not yet known, but the increased muscle tension needed for the increased vocal effort and higher intensity may play a role (Wohlert & Hammen, 2000). Articulation rate was included as a control variable in the linear mixed models but is only a significant factor in predicting the first formant values; the difference of 30 Hz per increase in the articulation rate may, however, be rather small considering the large increase in speaking tempo. For example, a change of articulation rate in 2 units (from 3 syll/sec to 5 syll/sec) only increases the F1 by 60 Hz according to the linear mixed model.

Speaker-specificity was investigated based on 12 sample speakers. We saw a high degree of between-speaker variation and a substantial amount of within-speaker variation in terms of the features investigated and the different FPs. Some of the within-speaker variation is caused by the conditions (expressed as standard deviations per condition), and some are found across the two conditions, normal and Lombard. Details have been shown in Chapter 4.5 in the section on speaker-specificity. The two conditions were of the mismatched type, which means that strong intra-speaker variation can be expected. It would be possible to reduce some of this variation by implementing a normalisation procedure in which the statistical differences are taken into consideration. Speaker discrimination performance would probably still be lower compared to matching conditions even then. The patterns of intra- and inter-speaker variation may look different to some extent if, instead of the selection process de-

scribed at the beginning of the section on speaker-specificity, we had selected the 12 speakers randomly. Further research into speaker-specific disfluency patterns in German using several recordings from the same speaker in various conditions could give a more complete picture of the nature and size of within-speaker effects. Another research goal, executable with the available data set, would be to use all 100 speakers, instead of 12, for a full investigation of the speaker discrimination potential of the FP features. An advantage of using the likelihood ratio framework for such an investigation would be that the implications for forensic phonetics can be fully worked out. Such a study was out of scope for the present thesis, but even with the current results two implications for forensic phonetics shall be pointed out. Firstly, the average values shown for the different FPs and their features in the two conditions, along with the standard deviations, indicate the typicality patterns addressed in section 4.1.2, i.e., which feature values are frequent in the relevant population of speakers (here male, adult, German-speaking) and which are non-typical. Secondly, the articulation rate was measured and included in the statistical models, but it had almost no effect on the features. This indicates a strong independence between the disfluencies studied here and the articulation rate. Since both disfluency and the articulation rate are frequently used speaker characteristics in voice comparison casework, this independence is beneficial when combining results from different characteristics (Gold & French, 2011; Jessen, 2018).

The present chapter reported an in-detail analysis of FPs in German based on a large corpus of 100 male native speakers. The next chapter will present the results of two features, the frequency distribution and the vowel quality, in three different languages: English, Spanish, and Arabic. The analysis of the English-Spanish comparison is accompanied by the additional factor of second language speech, while the analyses of the Arabic corpus take additional disfluencies into account, such as repetitions, lengthenings, and lexical FPs. A comparison of the FP vowel quality across the investigated languages concludes Chapter 5.

Chapter 5

Analysis of filler particles across languages

5.1 Introduction

Disfluencies are an integral part of any language. A language without filler particles (FPs) is yet to be found. FPs may be a universal feature that occurs in every language but which, nevertheless, carries language-specific characteristics such as the specific vowel quality. Many languages use non-lexical FPs alongside lexical FPs, and these non-lexical FPs often take a similar form, namely a (central) vowel, a nasal consonant (often bilabial), or a combination of the two (Clark & Fox Tree, 2002). Differences across languages primarily occur in the preferred FP type and the vowel quality that is used in the FPs. As outlined in Chapter 2, this vowel quality may be determined by the language's vowel inventory and, more specifically, by a central vowel such as schwa in the language. If FPs across languages can be shown to carry language-specific information, this may be relevant for fields in which the language background of a speaker is unknown and must be deciphered, such as forensic phonetic casework or language analysis for the determination of origin (LADO) of asylum seekers. In these fields, language experts may be required to make informed judgments about the plausibility of a suggested language background or provide this information themselves along with references on which they based their judgments. FPs and a general disfluency pattern could be taken as one feature to include in the analysis to add further insights towards the speaker's language background. We have seen in Chapter 4, that the FPs in German show a high within- and between-speaker variation, which may be detrimental when used to tell two German speakers apart in a forensic case. In comparing FPs across languages, we hope to add to the discussion of whether the feature can be used to differentiate between speakers from different language backgrounds.

Another aspect investigated in this chapter is the FP in a second language context.

If it can be proven that languages make use of different FPs, be it in the type that they prefer or in their vowel quality, the question of whether second language learners are able to adapt to the FP used in their foreign language arises. According to Flege's Speech Learning Model (SLM) (Flege, 1995), speakers form categories for each native phoneme during language acquisition and are able to add categories once they start learning a second language. The premise to form a new category for a phoneme of the foreign language is, that speakers are able to perceive a large enough difference to the closest L1 sound. If this is not the case, speakers are likely to use the L1 phoneme as no new category is formed, but the foreign and native sounds are represented by the same category (Flege, 1995). It is hypothesised that speakers also form a category for their FP vowel quality that is loosely connected with the vowel category of their native language. This category may allow much more variation in terms of quality, as no semantic meaning is connected with the FP and differences in vowel quality cannot lead to misunderstandings. As the ability to form phoneme categories remains intact according to the SLM (Flege, 1995), speakers may use this ability to establish a new FP category for their foreign language, but only when they are able to perceive a difference between their native and the foreign FP vowel quality. This shall be investigated by looking at one sample language pair: English and Spanish.

This chapter aims to exemplify language differences in acoustic characteristics of FPs by investigating FPs in three different languages from three different language families: Spanish as a Romance language, English as a Germanic exemplar, and Arabic as a sample from the Semitic language branch. Spanish was chosen to represent a language that does not use a central vowel in its vowel inventory and is thus interesting to investigate. The availability of the Diapix-FL corpus (Cooke et al., 2013), which offers the chance to investigate the second language aspect, made English an optimal candidate alongside Spanish. Egyptian Arabic was chosen, as this language is highly underrepresented in many research areas, especially in disfluency research. It forms a research gap which will not be closed with this thesis, but it is a first step towards a better understanding of disfluency use in this language. Comparisons will be drawn to the previously presented Pool2010 corpus (Jessen et al., 2005) on German presented in Chapter 4.

Previous literature suggests that the use of FPs may be language-specific regarding the type that is used and the vowel quality (de Leeuw, 2007; Wieling et al., 2016; Clark & Fox Tree, 2002; Belz et al., 2017). De Leeuw (2007) shows that L1 speakers of English and German prefer the vocalic-nasal FP type *um* while L1-Dutch speakers prefer the vocalic FP *uh*. Wieling et al. (2016), however, show that social factors also influence the use of FPs to a similar degree in several languages. They investigated the use of the FP types *uh* and *um* in several Germanic language corpora. They found consistent patterns that young, female, and higher-educated participants prefer the vocalic-nasal FP type *um* over the vocalic type *uh*. Whether this is also the case in other language families, e.g., Romance languages, is still an open question. An influence of the preferred syllable type (open vs. closed syllables) and other

phonotactic constraints on the favoured FP type is plausible but has never been investigated systematically.

Another aspect introduced in this chapter is the use of FPs by second language (L2) learners. The data presented before, while unique for its quantity, only consisted of German speech by native speakers; the Diapix-FL corpus (Cooke et al., 2013) used in the following contains L1 and L2 speech by native English and native Spanish speakers. A research question that requires further investigation is whether L2 learners of a language can produce FPs that are similar to those of native speakers of the target language. That is, are learners of e.g., Spanish able to produce FPs like a native Spanish speaker given that there is a difference between the target language and their native language?

Work on L2 learning has shown that the use of disfluencies, and thus also FPs, is typically higher in an L2 (Brand & Götz, 2011; Gilquin, 2008; Temple, 2000; Wiese, 1984) but decreases for pauses with rising L2 proficiency (Riazantseva, 2001). An L2 effect for pauses, but not for FPs, was also found by Belz et al. (2017) for advanced English learners of German. The preference of the vocalic FP in L2 English (intermediate and advanced learners) by native Spanish speakers is reported by Cenoz (2000). This seems to suggest that learners transfer their native FPs to their L2. The results in Muhlack (2020a) show that advanced learners of an L2 (English and German) are able to adapt the vowel quality of the target language in their FPs, but intermediate learners are less likely to adapt. The latter group seems to transfer their L1 FPs to the L2.

FPs in L1 Dutch and L2 English by the same speakers are investigated by de Boer & Heeren (2020). They take the minimal F1 difference in vowel qualities of FPs as evidence that advanced L2 learners can adapt their FPs to the target language. However, in their control experiment, they compare L1 English with L1 Dutch and find a difference of below 50 Hz in the mean values of first formant measurements. This cannot be taken as sufficient evidence that the two languages differ in the vowel quality of FPs, as differences below 50 Hz can be the result of erroneous measurements (Whalen et al., 2022). So it is crucial that a comparison between L1 and L2 speech is preceded or at least accompanied by a comparison of L1 speech in both languages and that minimal acoustic differences should not be over-interpreted.

A comparison between FPs in L1-Hungarian speech and L2-English speech by the same speakers revealed that the speakers produce more FPs in their L2 and that the FPs are longer than in their L1 (Gósy et al., 2017). Surprisingly, the language proficiency had an effect on the frequency of FPs in the expected direction (fewer FPs with increasing proficiency contrary to (Belz et al., 2017)) but not on the duration of FPs. Only small effects could be found regarding the vowel quality adaptation from L1 Hungarian to L2 English, which could be taken as evidence of transfer processes. Unfortunately, this study is lacking an English control group.

Furthermore, Lo (2020) found that German-French bilinguals produce distinct vowel qualities in the two languages but the weaker language shows a shift towards

the vowel quality of the dominant language. For Afrikaans-Spanish bilinguals from Patagonia, Argentina, Garciá-Amaya & Lang (2020) found that the speakers do not share the vowel quality of Afrikaans and Spanish monolinguals, but they still produce separate vowel qualities for their FPs in both languages.

Considerations of vowel normalisation

As the Pool2010 corpus (Jessen et al., 2005) was compiled for a forensic purpose, it contains speech by male participants only. The subset of the Diapix-FL corpus used here only contains the data from female participants, while the Arabic data presented in chapter 5.3 contains male as well as female participants. Different speaker genders include different spectral properties which are brought about by the differences in physiology. A male speaker's vocal tract and larynx are usually longer than those of a female, and the vocal folds themselves are also longer and thicker in males than in females (Titze, 1989). These differences result in a lower fundamental frequency on average in male speakers than in female speakers as well as differences in resonance frequencies. Due to the higher formants, female acoustic vowel spaces are lower in the vowel chart than the male vowel spaces (Coleman, 1971; Titze, 1989). These differences can be eliminated by a number of normalisation methods. As the two larger corpora in this thesis, the Pool2010 corpus (Jessen et al., 2005) and the Diapix-FL corpus (Cooke et al., 2013), only represent one gender each and normalisation would not be needed for within-corpus comparisons, introducing normalisation for the small-scale Arabic corpus was refrained from.

For a comparison between the corpora, the Nordström method is used (Nordström, 1977) as reported in Adank et al. (2004). It is a vowel- and formant-extrinsic method which means that information from multiple vowels is used to normalise a single vowel token, and that information from other formants is used to put the other formants in relation. In the case of the Nordström method, the average of the third formant is calculated across vowels (all tokens with an F1 $>$ 600 Hz) for male and female participants separately. The ratio is then calculated by dividing the male F3 average by the female F3 average to get a value (k) that indicates the ratio of how much smaller a female vocal tract is than a male's. All vowel formants by female speakers are then multiplied with this value k to align the formants with those produced by male speakers. The full formula is given in 5.1 and 5.2.

$$F_{\mathrm{i}} = kF_i^{female} \tag{5.1}$$

$$k = \frac{L^{male}}{L^{female}} = \frac{\mu_{F^3}^{male}}{\mu_{F^3}^{female}} \tag{5.2}$$

Adank et al. (2004) compared the method with other normalisation methods regarding their ability to preserve phonemic and sociolinguistic contrasts while eliminating gender differences due to physiological differences between men and women.

They conclude that vowel-extrinsic and formant-intrinsic methods perform the task of vowel normalisation best, such as the methods by Lobanov (1971) and Gerstman (1968). The authors do not recommend the Nordström method in their conclusion as it did not perform very well at normalising fundamental frequency values of male and female talkers. When f0 was not considered, and only the vowel formants F1-F3 were used, the method performed equally well as the method reported in Gerstman (1968) and nearly as good as the method by Lobanov (1971).

The Lobanov and the Gerstman methods were not used because vowel tokens of the vowel inventory or at least of the corner vowels are necessary to include in the calculation of the normalisations. Gerstman (1968) normalisation procedure uses the minimum and maximum values of each formant per speaker, while Lobanov (1971) uses the average of each formant across all available vowels. As the data sets used here include data in different languages that possess a different vowel inventory, the risk in using these methods was that differences in vowel inventory would influence the normalisation procedure. To avoid this, the Nordström method was selected, and only FP vowels were used in the normalisation procedure. Another benefit is that the resultant unit of the scale remains as Hertz and the values can easily be interpreted. Nevertheless, using only FP vowels could influence the normalisation procedure, though the results show the expected effect.

5.2 Spanish vs. English

The aim of this chapter is to compare the production of filler particles (FPs), i.e., the frequency distribution and vowel quality, in English and Spanish as a native (L1) and second language (L2). Work on FPs and their phonetic characteristics has increased over the last few decades. However, languages other than English (e.g., Spanish (Erker & Bruso, 2017)) are still under-researched in this area. Relevant studies for Spanish FPs show the following: the preferred type of FP in Spanish seems to be the vocalic FP with approx. 76% of all FPs being realised as *uh*, 13% as the nasal FP *hm*, and only 11% with a vocalic-nasal FP *um* (Garciá-Amaya & Lang, 2020). The investigation of FP usage and quality of 24 L1 Spanish speakers living in the US shows that the vowel quality of FPs in their L1 changes from [e], to a lower, more central variant ([ə] or /a/) with an increase of English usage (proficiency and length of stay) (Erker & Bruso, 2017).

Similarly, Garciá-Amaya & Lang (2020) showed that Spanish monolinguals use an FP vowel with a lower F1 and a higher F2 than Spanish-Afrikaans bilinguals. It seems that Spanish speakers produce a fronted and more close vowel in their FPs compared to other languages (Candea et al., 2008). In comparison with vowels in lexical material, Vasilescu et al. (2007) show that the FP vowel is closest to the realisations of /e/ but claim it is produced with a higher F1 and a lower F2 than the lexical vowels. Unfortunately, in Candea et al. (2008) and Vasilescu et al. (2007),

the degree of overlap between the realisations of /e/ and the vowel in FPs was not reported.

From the numerous studies on FPs in English, for instance, de Leeuw (2007) reports the vocalic-nasal FP *um* as the preferred type for British English speakers (81% *um*, 18% *uh*). The vowel quality of English FPs is often reported to vary between the central vowels [ə] and [ʌ][1] and the back vowel [ɑ] (Shriberg, 1994). The vowels in the FPs usually spread over a large part of the central vowel space (Hughes et al., 2016). Results in Muhlack (2020b) suggest that the vowel used in English FPs is closest (in terms of smallest Euclidean distance) to the open-mid central vowel [ʌ]; however, the English speaker group is quite heterogeneous in terms of their English accents (which may introduce further variation).

Based on the reviewed literature, it is assumed that native English speakers prefer the vocalic-nasal FP *um* while Spanish native speakers prefer the vocalic FP *uh*. It is expected that language learners also employ their preferred FPs in their L2 (Cenoz, 2000). Distinct vowel qualities for each language are expected to be observed, and a shift towards the L2 vowel quality is likely in the second language context.

Material

For this study, a subset of the Diapix-FL corpus (Cooke et al., 2013) is used, which consists of 20 female speakers (10 native English, 10 native Spanish/Basque) completing a dialogue task in both their L1 and L2 (Spanish and English, respectively). Native speakers of English were recorded at the University of Edinburgh, and native Spanish/Basque speakers were recorded at the University of the Basque Country in Vitoria. The speakers were proficient in their L2 (B2-C1 on CEFR[2]) as both groups studied their L2 in their second year at university. The native Spanish participants all passed a B2 proficiency test in English as part of their study programme to reach their second year, and the native English speakers self-reported their proficiency level in Spanish as their L2. Four male speakers (2 English, 2 Spanish) from the corpus were excluded in order to keep the speaker group as homogeneous as possible.

The speakers were grouped in same-L1 pairs and given a spot-the-difference task in their L1 and L2. Each speaker was given a picture; all the pictures showed different versions of the same scene. The participants had to cooperate to find 12 differences without being able to see each other's pictures. Dialogue partners solved three spot-the-difference tasks per language. Each speaker was recorded on a separate channel. For more details on the recording method, see Garcia Lecumberri et al. (2017) and Wester et al. (2014).[3]

[1]In phonetic transcription for English the symbol ʌ is commonly used to describe an open-mid central vowel (Roach, 2009) as opposed to a back vowel as described by the IPA.

[2]The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2011) provides guidelines for the categorisation of L2-fluency and proficiency.

[3]The DIAPIX-FL corpus is freely available at http:// datashare.is.ed.ac.uk/handle/10283/346

Orthographic annotations are included in the corpus along with annotations of silent pauses (including breath noises), elongations, and FPs. All FPs are marked with one symbol; for the re-annotation, which was done by one annotator, distinct labels for vocalic (*uh*), nasal (*hm*), and vocalic-nasal (*um*) FPs were used. Additionally, ten tokens (annotated in stressed positions) of each corner vowel (/i u a/ for Spanish and /i u ɑ/ for English) for all L1 speech were annotated along with ten tokens of one additional vowel that is frequently reported to occur in the FPs of the respective language (/e/ for Spanish, /ʌ/ for English).

Results

A total of 2,737 FPs were found in the subset of the corpus: 245 nasal FPs (*hm*), 1,118 vocalic FPs (*uh*), and 1,374 vocalic-nasal FPs (*um*).

Frequency distribution

FPs are more frequent in each L2 than in the respective L1 speech. L1 English shows an overall FP rate of 1.9 per minute, while L1 Spanish shows a lower rate with 1.0 FP per minute. Native English speakers increase their FP rate only slightly in their L2 (2.2 FPs/min for L2 Spanish), whereas native Spanish speakers show a rate of 2.5 FPs per minute in their L2, which is considerably higher compared to their low L1 rate. The frequency distribution of the four FP types in both speaker groups (English and Spanish native speakers) and both conditions (L1, L2) is shown in Figure 5.1. In their respective native languages (left panel), L1 Spanish speakers use more vocalic FPs than the other FP types, while L1 English speakers use the vocalic-nasal FP type *um* more frequently (4.8 vs. 0.7 *um*/min). The nasal FP *hm* is used rarely but more often by Spanish speakers. The same pattern is also visible in L2 speech (right panel), with numbers increasing overall.

During the classification of the FPs, it was observed that speakers, when producing a vocalic-nasal FP, do not always use the bilabial nasal /m/ but also an alveolar /n/ in some instances. An auditory inspection by one annotator (the author) revealed that 3.4% (i.e., 38/1,108) of all vocalic-nasal FPs used by the native English speakers (in both L1 and L2) were produced with a nasal that was not bilabial, and 18.1% (i.e., 48/266) of those used by the Spanish native speakers were produced with a nasal that was not bilabial. The majority of non-bilabial nasals were alveolar, while a few were also categorised as labiodental, velar, or a sequence of alveolar and bilabial nasals. To shed more light on the phenomenon, the annotation of several annotators should be taken into account in future studies.

Vowel quality

The first and second formant (F1, F2) of all purely vocalic FPs and the corner vowels of L1 English and L1 Spanish were measured in Praat (Boersma & Weenink, 2022)

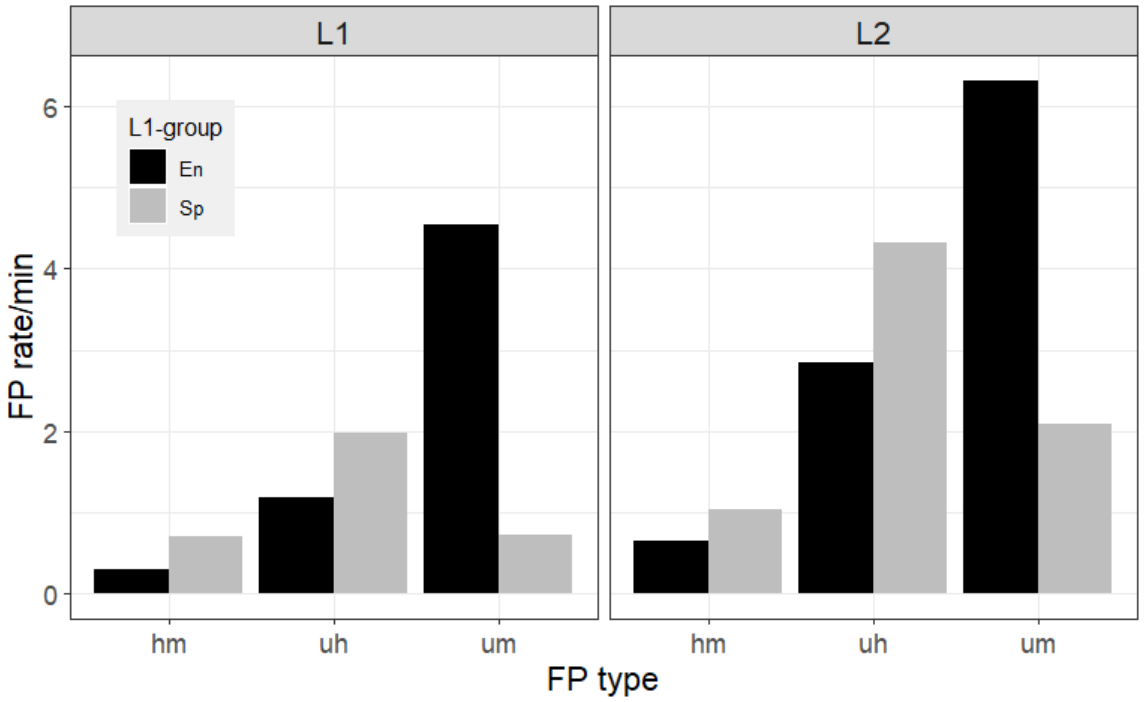

Figure 5.1: FP frequency in the Diapix-FL corpus
FP rate per minute in L1 and L2 speech by both speaker groups. Native English speakers in black; native Spanish speakers in grey.

at the midpoint of the vowel, using the Burg method.[4] To visualise the vowel quality produced in the FP *uh*, two-dimensional (F1, F2) kernel density distributions of the lexical vowels and the FP vowel were plotted in R (R Core Team, 2022) using ggplot2 (Wickham, 2016). These plots show in which area the majority of the data is assembled (in the inner-most ring of the ellipses), and in which area the data distribution is thinner (the outer rings of the ellipses). In the graphical representations of vowel quality, the first formant values are plotted on the y-axis and the second formant values on the x-axis; both axes are inverted to better match the vowel quadrilateral, as is common practice in phonetics research.

Figure 5.2 shows that the vowel in native English FPs partly overlaps with the realisations of the lexical vowels (also L1 English) /ɑ/ and /ʌ/. Based on this observation, the overlap in the two-dimensional vowel spaces between these two lexical vowels and the FP vowel was calculated using the Pillai-score (Kelley & Tucker, 2020), which ranges from 1 (no overlap) to 0 (complete overlap). The vowel quality of the FP *uh* overlaps more with the open-mid vowel /ʌ/ (Pillai: 0.27) than the open vowel /ɑ/ (Pillai: 0.46). The /u/-fronting has been described before for the same data set (Garcia Lecumberri et al., 2017).

Figure 5.3 shows that the vowel in native Spanish FPs overlaps with the two front

[4]Maximum formant: 5.5 kHz, max. number of formants: 5, window length: 0.025 s, dynamic range: 50 Hz

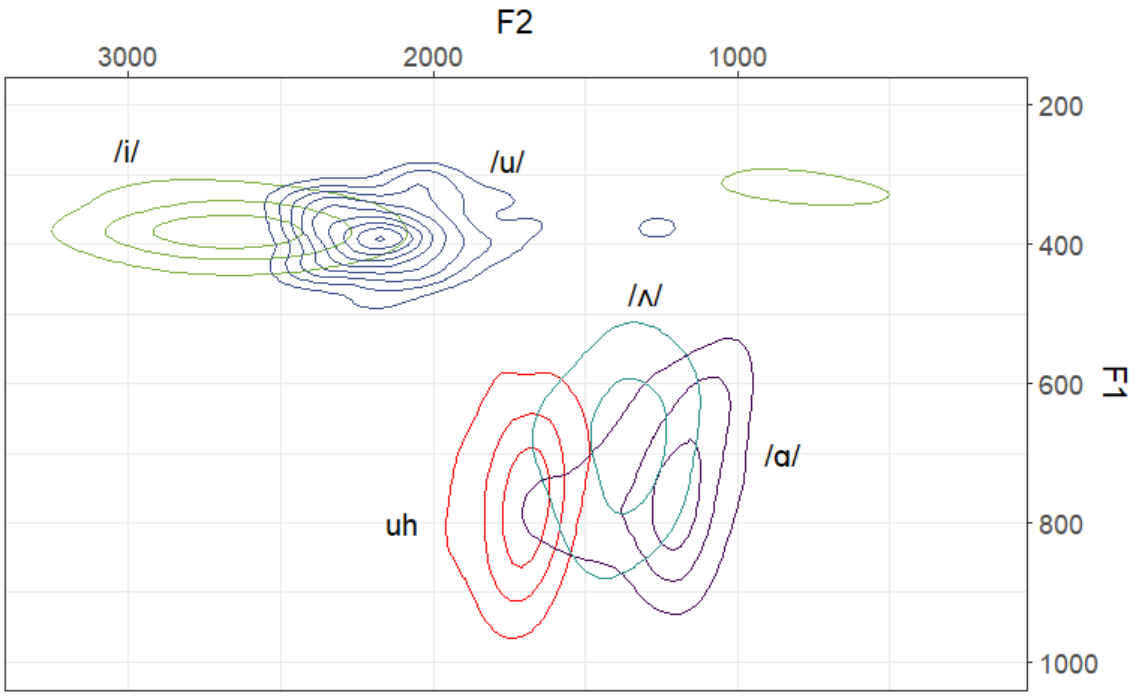

Figure 5.2: Vowel quality of FPs and corner vowels in L1 English. *Formant values (Hz) of realisations of* uh *(red) in English L1 compared to those of the lexical vowels* /ɑ i u ʌ/.

vowels /i/ and /e/ of native Spanish speech to a considerable degree. The Pillai scores show more overlap with the close-mid vowel /e/ (Pillai: 0.22) than the close vowel /i/ (Pillai: 0.34).

In order to answer the question of whether speakers transfer the vowel quality of their L1 FP to their L2, Figure 5.4 compares the FP *uh* by both speaker groups in their L1 and L2 speech. The left panel shows that FPs in English and Spanish L1 are produced with distinct vowel qualities (Pillai: 0.67). The L2 FPs (right panel) show vowel qualities that spread over a larger area of the central vowel space for both speaker groups. However, they also show a bimodal distribution for both groups. Some FPs in L2 Spanish used by English speakers are produced with a lower F1 and a higher F2, while the majority of FPs are produced with the same vowel quality as observed in their L1. Conversely, for L2 English by Spanish speakers, some of the FPs are produced with a native-like vowel quality (similar to the close-mid /e/); however, a large portion is produced with a higher F1 and a lower F2, approximating the native English vowel quality.

The question that remains is whether all learners behave similarly in how they adapt to the vowel quality of the target language. Individual speakers likely adapt the vowel quality of the target language while others keep their native vowel quality. Another possibility is that speakers do not switch from one vowel quality to the other entirely, but they employ a mixture of native-like and foreign FPs. To shed more light

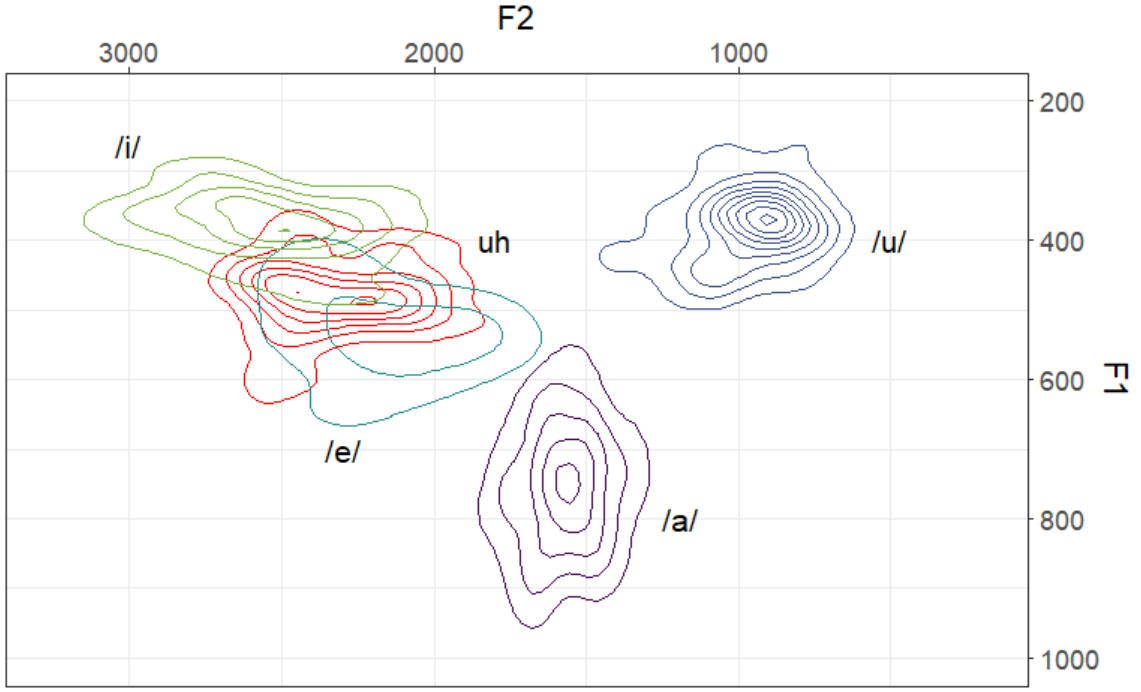

Figure 5.3: Vowel quality of FPs and corner vowels in L1 Spanish.
Formant values (Hz) of realisations of uh *(red) in Spanish L1 compared to those of the lexical vowels* /a i e u/.

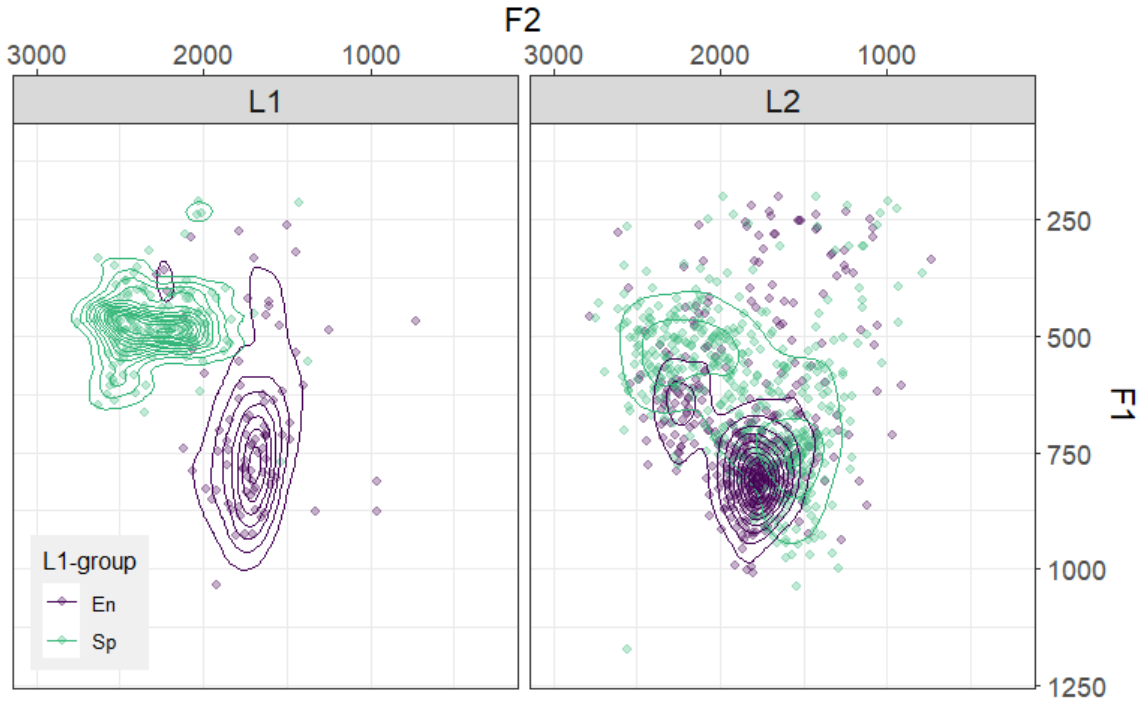

Figure 5.4: Vowel quality of FPs in L1 and L2 English and Spanish.
Vowel quality of uh *in L1 and L2 speech by native English speakers (purple) and native Spanish speakers (green). Formant values in Hz.*

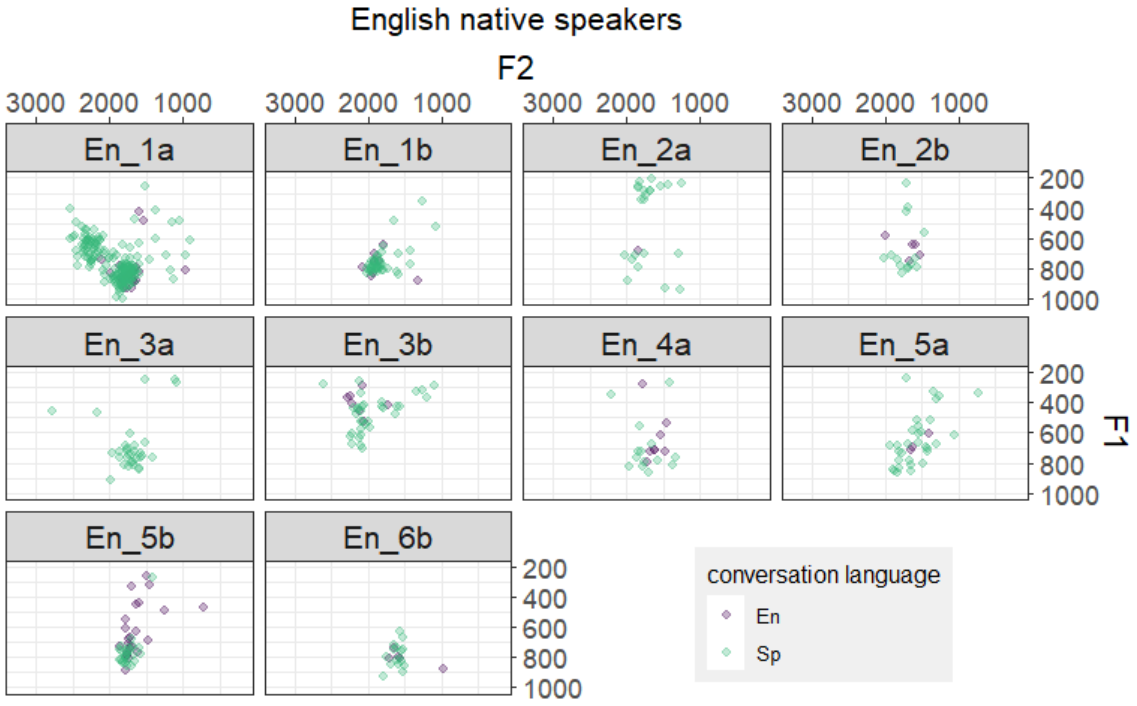

Figure 5.5: Vowel quality of FPs in L1 English and L2 Spanish per speaker. *Vowel quality of* uh *in L1 English (purple) and L2 Spanish (green). Each panel shows the FPs of one individual speaker. Formant values in Hz.*

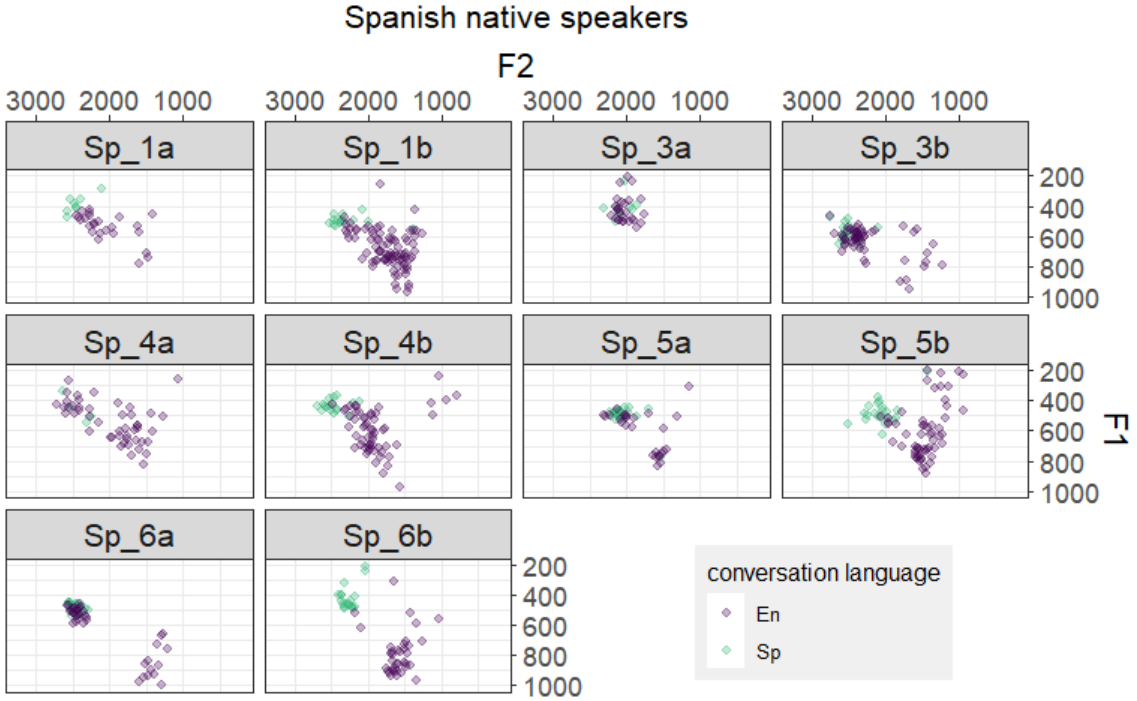

Figure 5.6: Vowel quality of FPs in L1 Spanish and L2 English per speaker. *Vowel quality of* uh *in L1 Spanish (green) and L2 English (purple). Each panel shows the FPs of one individual speaker. Formant values in Hz.*

on this phenomenon, individual speakers are shown in Figures 5.5 and 5.6 divided by their native language. As the native English speakers in Figure 5.5 produced more vocalic-nasal FPs, especially in their native speech, and the figure only plots the vocalic FPs' vowel qualities, English reference tokens are sparse. Nevertheless, the individual performance can be seen for the L2 Spanish tokens. Some speakers (1b, 3a, 4a, 5b, 6b) are quite consistently using an open central vowel in their Spanish FPs, which can be found in the pooled L1 English FPs (see Figures 5.2). Other speakers (2a, 2b, 5a) show more variation. One speaker to be highlighted is speaker 1a. She produces more tokens in her Spanish L2 speech than the other speakers, and a bimodal distribution becomes visible when plotting the FP vowel qualities (Figure 5.5). While she also shows a lot of variation, she may switch between Spanish-like vowels and English-like vowels quite a lot. The native Spanish speakers (Figure 5.6) show similar patterns. Most speakers (1a, 1b, 3b, 4a, 4b, 5b) show quite a lot of variation, but a pull from their high native front vowel towards lower central vowels is visible. Speaker 3a, however, does not show any adaptation of English-like vowel qualities, but she uses her native Spanish FP in her L2 without variation. Speakers 5a and 6a show a bimodal distribution but still use their native FP vowel in their English speech in some instances, unlike speaker 6b. This speaker seems to have mastered a clear-cut use of their FPs: native vowel qualities in their Spanish L1 and English-like FP vowels in their English L2.

Discussion

The results of the current study confirm that L1 English prefers the vocalic-nasal FP (*um*) and L1 Spanish prefers the vocalic FP (*uh*) (de Leeuw, 2007; Böttcher & Zellers, 2023; Garciá-Amaya & Lang, 2020). Nasal FPs are not very frequent in either speaker group but for L1 Spanish *hm* is as infrequent as *um*. The frequency of FPs in L2 speech increases as reported before (Brand & Götz, 2011; Gilquin, 2008; Temple, 2000; Wiese, 1984), and moreover, the preferred type of FP remains consistent, with native Spanish speakers preferring the vocalic FP *uh* and the native English speakers preferring the vocalic-nasal type *um* in both their L1 and L2 (Cenoz, 2000). A reason for the preference of the specific FP type may be the syllable structures in both languages. In English, closed syllables seem to be more common than open syllables as observed from the data reported in Crystal & House (1990), while in Spanish closed syllables do occur, but most often word-internally (Gabriel, 2022). Another reason for the preference of *uh* over *um* may be the restriction for Spanish word-final closed syllables, which only allows the consonants /l, n, r, s, d/ in the coda[5] (Gabriel, 2022). This restriction would also explain the occurrence of alveolar nasals in the vocalic-nasal FPs of the native Spanish speakers, yet not those produced by the native English speakers.

The results of the comparison of vowel qualities between lexical vowels and the

[5]Other consonants may occur in loanwords.

FP *uh* in English show that the realisations of the open-mid central vowel /ʌ/ best represent the vowel of the FP, although vowel spaces do not entirely overlap. The FP vowel in English is produced more centrally than the vowel /ʌ/. The Spanish FP vowel, on the other hand, is produced as a close-mid front vowel, similar to the realisations of the Spanish lexical vowel /e/. In Spanish, [e] and [ɛ] are allophones of the phoneme /e/ (Gabriel, 2022), so the annotated corner vowels in Spanish include tokens of both allophones. The FP vowel overlaps with (the upper) half of the vowel space covered by the vowel phoneme /e/, suggesting that the vowel quality of *uh* is represented by [e] rather than [ɛ]. This is also supported by previous literature (Candea et al., 2008; Garciá-Amaya & Lang, 2020). However, the claim by Vasilescu et al. (2007) that the Spanish FP is produced with a lower F2 is not supported by the data presented here. Although a higher F1 can be observed, this may be due to the inclusion of the allophone [ɛ] in the set of corner vowels. Note that native Spanish speakers were bilinguals in Spanish and Basque, which may have influenced the vowel quality of the FPs. However, as the vowel systems of Spanish and Basque are the same (Hualde & Ortiz de Urbina, 2003; Gabriel, 2022), this influence is assumed to be rather small.

The comparison between L1 and L2 speech shows that while the FP vowels in L1 English and Spanish are very distinct, this cannot be observed in the speakers' L2 speech. The distributions of the FP vowels of Spanish and English L1 speakers show two peaks when the speakers produce FP vowels in their L2, which suggests that some hesitations approximate the realisations of the Spanish /e/ and some approximate the realisations of the English /ʌ/. It seems that advanced learners are able to adapt the FP's vowel quality, but this is not the case for all the data. The results suggest that L1 Spanish speakers may be better at approximating the vowel quality of English FPs and that L1 English speakers are more likely to keep their native FP in their L2. This may be due to the fact that the English vowel inventory does not have the phoneme /e/, only the more open vowels /ɛ/ and /æ/ (Roach, 2009).

To shed light on individual speaker performances, FP vowels were plotted per speaker. It is shown that speakers show individual patterns: most vary considerably while moving towards L2 vowel qualities, fewer manage a clear distinction between L1 and L2, and one speaker even transferred their L1 FP to their L2 without variation. Factors that may influence the switch from L1 to L2 FP vowel quality may be L2 proficiency, exposure to the target language by native speakers, the speakers' stance towards the foreign language, or even the thickness of their mental boundaries. This last concept was investigated by Gessinger (2022) in the context of phonetic accommodation. The hypothesis is that thinner mental boundaries would allow a higher degree of phonetic accommodation towards an interlocutor. It could be the case that thinner mental boundaries, which are determined by the factor's openness and neuroticism, are correlated with the accommodation of some phonetic features. Gessinger (2022) could not prove that this was the case when investigating the question intonation and the phonemic accommodation of words ending in [ɪk]/[ɪç]. She proposes taking a

more holistic approach which could be looking at the influence of mental boundary strength on foreign accentedness and thus also at the transfer and adaptation of FP vowels.

5.3 Arabic

The research area of disfluencies and filler particles has gained interest across many languages in the last decades. However, languages that are not part of the Indo-European language family are considerably under-researched. Disfluency research in healthy adults in Arabic is, to the best of our knowledge, a research gap. This study delivers the first puzzle piece in filling this research gap, starting with a closer look into filler particles and pausing behaviour in the Egyptian dialect. The Arabic language is the fourth most spoken language in the world, with an estimated number of 400 million speakers distributed over 23 countries (Bateson, 2003); disfluency research in this language is, however, still a research gap. Previous studies for British English and German suggest that disfluency patterns may be speaker-specific and thus, could aid as a feature in forensic phonetic casework (McDougall & Duckworth, 2018; Braun & Rosin, 2015; Braun et al., 2023; Braun & Elsässer, 2023). It is important to investigate language- and speaker-specific disfluency patterns to attribute possible similarities or differences in a forensic case to the same speaker or two different speakers with the same language background.

Material

Spoken data of Arabic was acquired from the Arabic Speech Rhythm Corpus (Ibrahim et al., 2020), which includes read as well as spontaneous speech. To investigate disfluencies, only the two spontaneous speech tasks were used. In the first task, participants were asked to give an account of their daily schedule and talk about it for 1-2 minutes (= *Daily life*). The second task was a description of directions from a famous sight in the city to a university building, which was of similar length as the first task (= *Map task*). Participants received a map as a visual aid for this task which may have resulted in a higher cognitive load during this second task compared to the first task. The speech total amounts to approximately 19 minutes. Both spontaneous tasks including Praat TextGrids (Boersma & Weenink, 2022) are available for seven speakers (4 male, 3 female). Two more speakers from the corpus were excluded, one due to the lack of accompanying TextGrids and one due to the lack of one task. Provided TextGrids include the annotation on sentence, syllable, and phone level including the phonetic transcription made by a native Egyptian speaker who is also a trained phonetician. All speakers exhibit the same Egyptian dialect, namely Egyptian Colloquial Arabic from the city of Alexandria. Participants were all students in the Linguistics & Phonetics department at Alexandria University, ranging from 20 to 21 years of age. The recording session took place in a soundproof room

using a large membrane condenser microphone at a distance of approximately 40 cm from the speakers' mouth; the microphone was directly connected to a desktop PC. Task order was counter-balanced for the participants, both tasks were recorded on the same day for each participant. Recordings were made using a 44.1 kHz sampling frequency and 16-bit quantisation.

A native speaker of Egyptian Arabic provided further annotations for the speech data which included the following disfluency phenomena: pauses, lengthenings, repetitions, lexical FPs, filled pauses (which were defined as filler particles flanked by pauses on both sides), and nasal, vocalic, vocalic-nasal, and glottal FPs. Because the results in Chapter 4 showed a high variation between speakers and within speakers in the use of FPs alone, the small corpus here offered the opportunity to take a closer look into more types of disfluencies. The phonetic transcriptions on a segment level were used to take formant measurements of the Arabic long vowels /iː aː uː/ (Embarki, 2013) at the midpoint of the segments, using the Burg method[6] provided by Praat (Boersma & Weenink, 2022). Formant measurements were taken for the vocalic FPs using the same method. FP frequencies and measurements were extracted using a custom Praat script which provided durations and formant measurements of the FPs as well as durational measurements for the annotated pauses.

Results

This section will show the results of the data analysis using descriptive statistics and inferential statistics where appropriate. All analyses are done in R (R Core Team, 2022) using the stats and the tidyverse package (R Core Team, 2022; Wickham et al., 2019).

Figure 5.7 shows the frequency of the different disfluency phenomena in the two tasks as number per minute. The phenomena *filled pause, glottal FP* (glottal), *nasal FP* (nasal), *lexical FP* (lexical), *vocalic-nasal FP* (uh), *lengthening* (length), and *repetition* only make up a small portion of the entire disfluency palette for each speaker. The *vocalic FP* (uh) and, especially, the *silent pauses* (p) predominate the disfluency pattern of these speakers. When comparing the disfluency patterns of each speaker per task, the visual inspection suggests that the speakers remain quite consistent across tasks in their frequency of each disfluency type and overall count. Only speaker Ar12 deviates from the pattern shown in the first task (*Daily life task*) in that they resort to only using two types of disfluencies (the vocalic FP and pauses) in the second task *(Map task*) instead of also including a vocalic-nasal FP, a nasal FP, a glottal FP, or a filled pause, as was the case in the first task. Speaker Ar04 shows a higher rate of disfluencies in the *Daily life task*, but the use of different disfluency types remains rather stable. The speaker decreases the rate of the vocalic-nasal FP in the second task but makes use of syllable lengthenings instead. A factor that is not

[6]Maximum formant: 5000 Hz for males, 5500 Hz for females; maximum number of formants: 5; window length: 0.025 s; dynamic range: 50 Hz

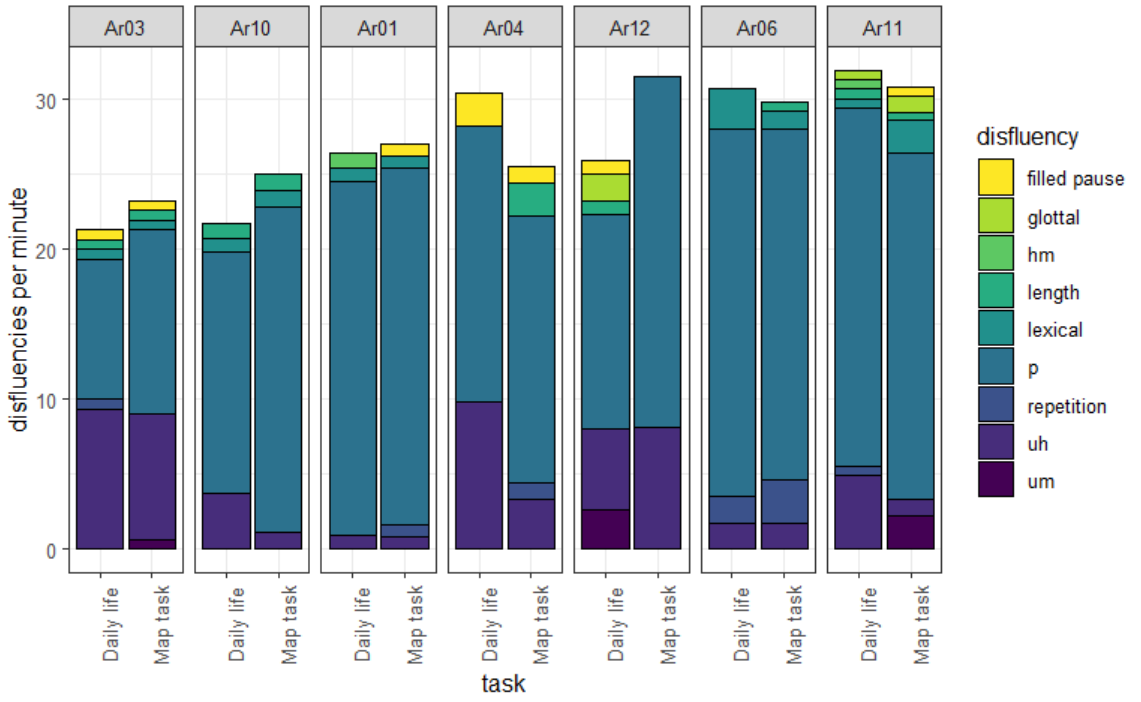

Figure 5.7: Frequency of disfluencies (per min) of L1 Arabic speakers. *Number of disfluencies per minute for each individual speaker in the* Daily life *and the* Map *task. Colours denote different types of disfluencies.*

taken into account in Figure 5.7 is the speaking tempo of the speakers per task. It may be that the tasks' topics influenced the speaking tempo, e.g., the alleged higher cognitive load of the *Map task* resulted in a slower speaking rate. To investigate this assumption, the disfluency frequencies were normalised to a rate per 100 syllables instead of minutes. The results can be seen in Figure 5.8. For most speakers, this conversion did not have a large impact apart from speaker Ar11 as this speaker's tempo changed significantly between the two tasks.

To investigate the question of where in the acoustic vowel space the Arabic FPs are located, the plot in Figure 5.9 shows the FPs of the individual speakers (as coloured dots) compared to the Arabic long vowels /iː aː uː/ of lexical material (ellipses) (Embarki, 2013). The vowel quality of the FPs seems to assemble in the central vowel space with some speakers clustering closer together (e.g., speaker Ar03) than others (e.g., speaker Ar12). Another phenomenon visible in this plot is the wide spread of each Arabic lexical long vowel. This may be because in Arabic all vowels show a high degree of allophonic variation depending on phonetic context and prosody (Embarki, 2013).

Comparing the Arabic FP vowels to other languages may shed more light on the language-specificity of the vowel quality in FPs, which may be useful in situations where the speaker and their language background are unknown. In Figure 5.10, the Arabic FP vowel quality is compared to the ones presented before in Chapters 4 and

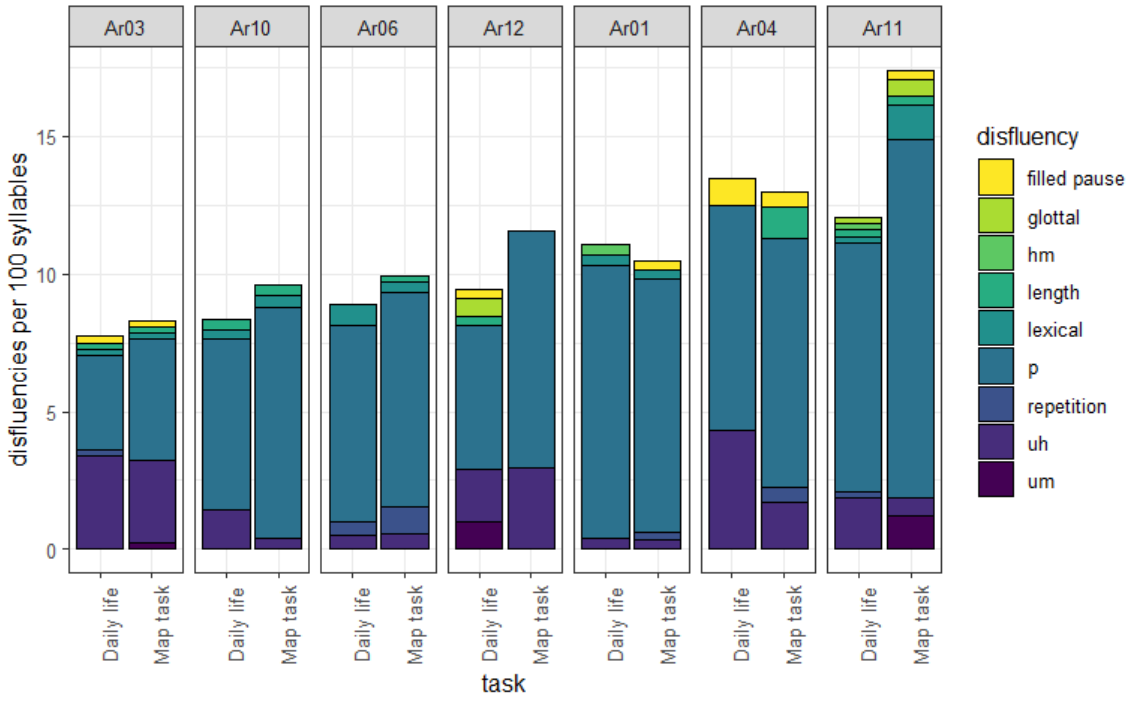

Figure 5.8: Frequency of disfluencies (per 100 syll) of L1 Arabic speakers. *Number of disfluencies per 100 syllables for each individual speaker in the* Daily life *and the* Map *task. Colours denote different types of disfluencies.*

5.2. Figure 5.11 shows the same data, but female formant values were normalised using the Nordström method as explained above. The figures show the Arabic FP vowel compared to Spanish, English, and German FP vowel qualities. It becomes apparent that in the two-dimensional vowel space, the Arabic FPs overlap most with the German FP vowels and the least with the FPs produced by native Spanish speakers. To measure the overlap of the data, Pillai scores were calculated on the normalised data. Pillai scores were measures of overlap that range between 0-1 (0 denotes complete overlap of two areas and 1 denotes complete separation) (Kelley & Tucker, 2020). The scores were acquired by calculating a Multivariate Analysis of Variance (MANOVA) in R (R Core Team, 2022). The Pillai scores showed what can also be seen in Figure 5.10. The least amount of overlap of the Arabic FPs occurred with the Spanish FP (Pillai: 0.65); considerable overlap occurred with the English FPs (Pillai: 0.12) and a very high degree of overlap occurred with the German FPs (Pillai: 0.05).

The comparison of disfluency durations between the tasks has to be taken cautiously. Figure 5.12 shows that the disfluency phenomena in the *Daily life task* are longer than in the *Map task*. However, all disfluency phenomena, except for the vocalic FP *uh*, are quite rare in the data (see Table 5.1). For the vocalic FP, we see the same trend: the particle is longer in the *Daily life task*, but this difference is not significant according to a two-sample t-test ($t = 1.32$, $p = 0.19$). The difference between the means is merely 7 ms and the variance is quite high as visible in Figure 5.12.

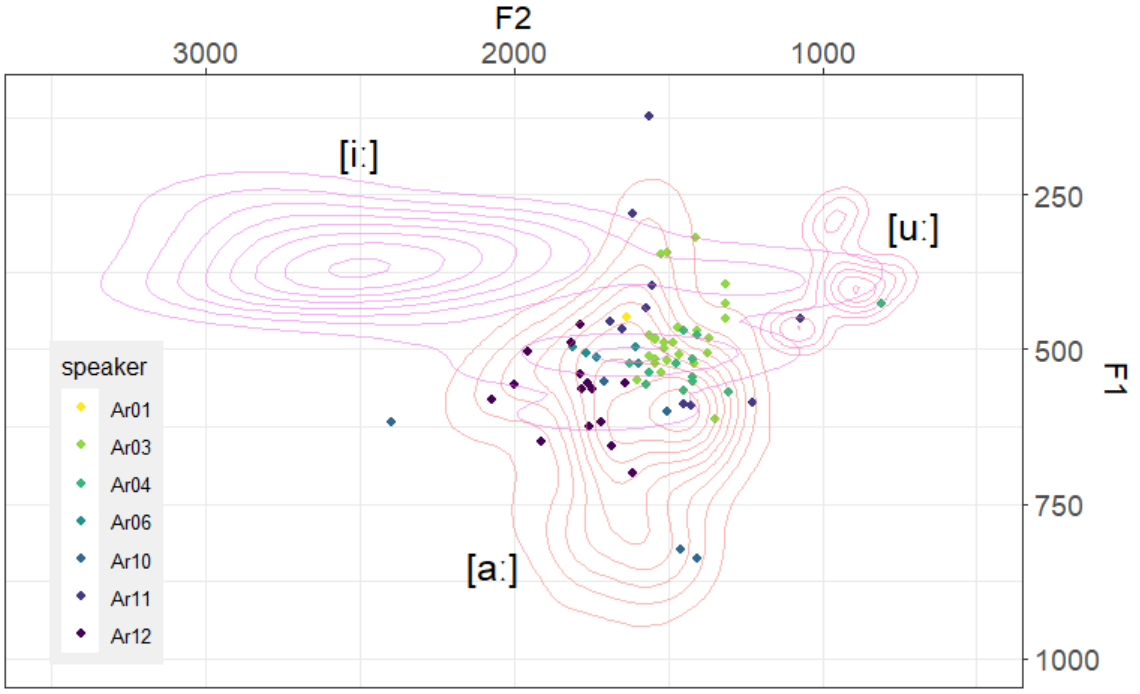

Figure 5.9: Vowel quality of FPs and corner vowels produced by Arabic speakers. *Vowel qualities of the vocalic FP* uh *of each individual speaker (coloured dots) compared to Arabic long vowels (ellipses). Formant values in Hz.*

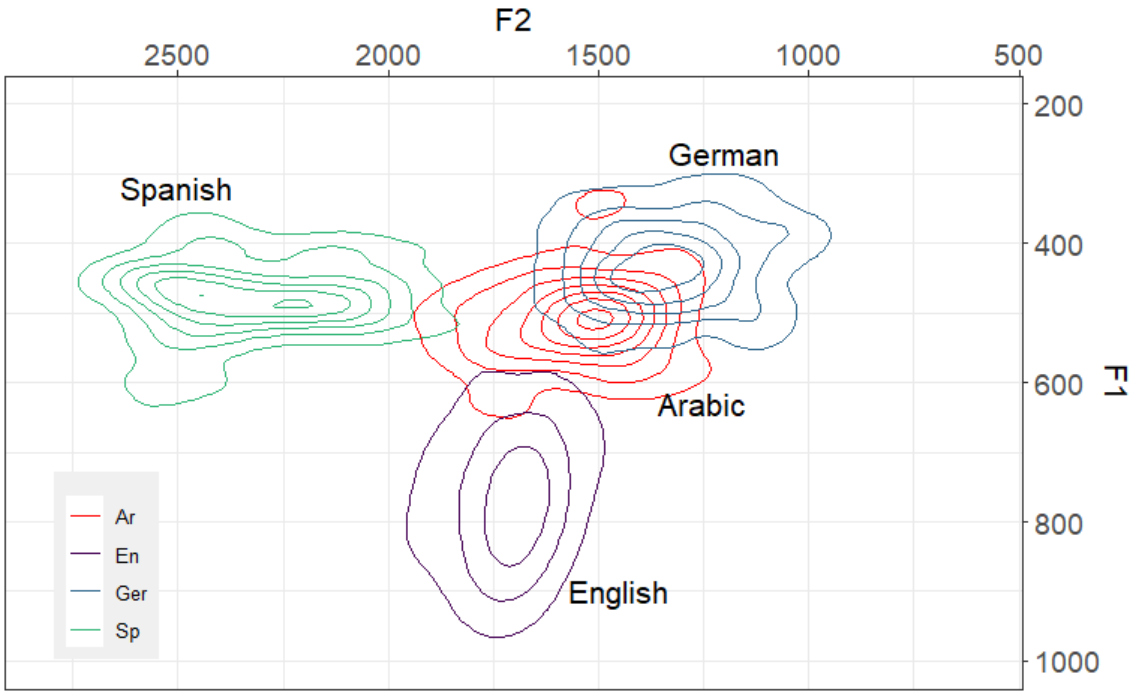

Figure 5.10: Vowel quality of FPs across languages.
Vowel qualities of the vocalic FP uh *of the Arabic speakers compared to the Spanish, English, and German data. Formant values in Hz.*

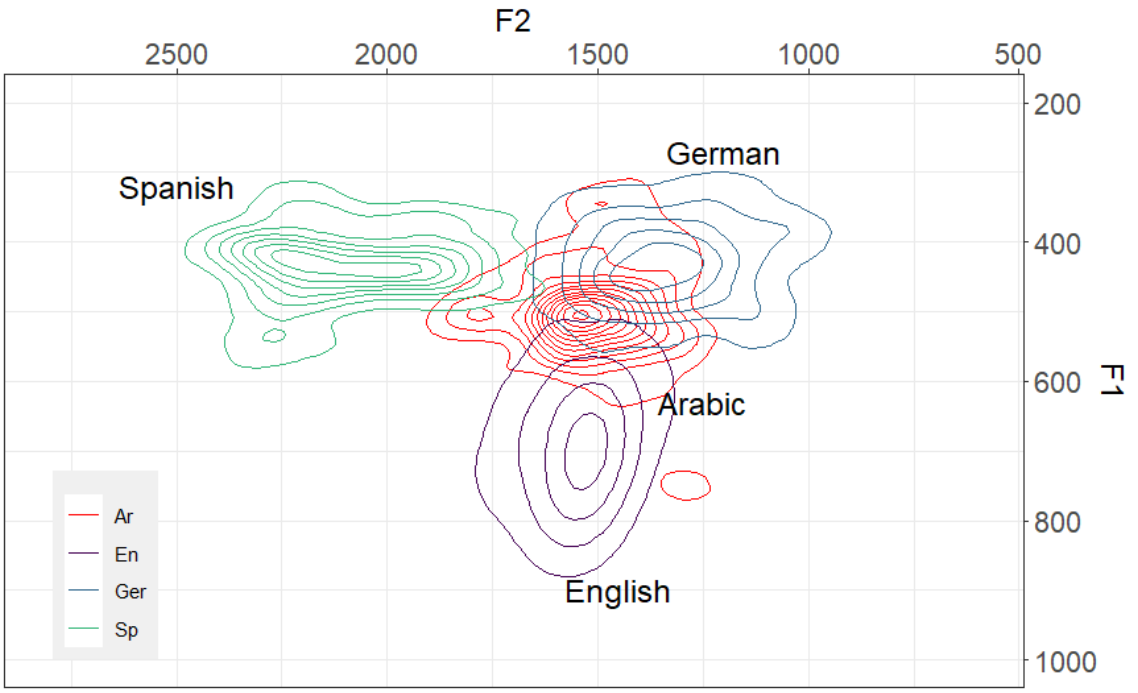

Figure 5.11: Normalised vowel quality of FPs across languages.
Vowel qualities of the vocalic FP uh *of the Arabic speakers compared to the Spanish, English, and German data. Female data points were normalised using the Nordström method (Nordström, 1977). Formant values in Hz.*

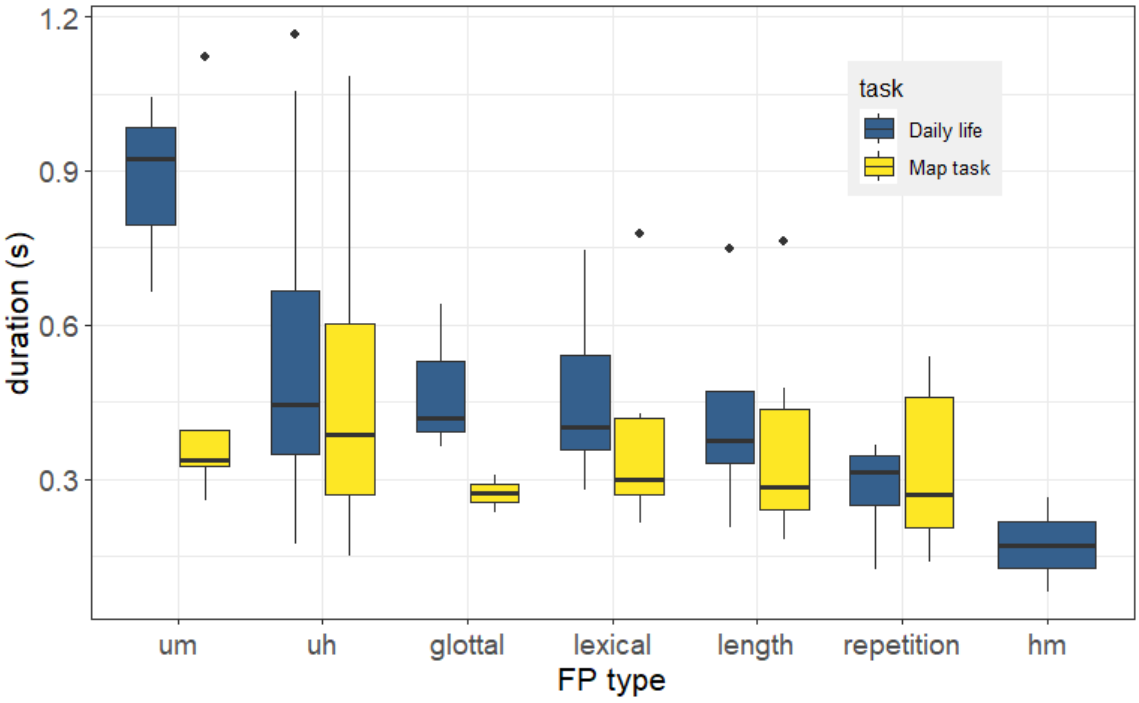

Figure 5.12: Disfluency durations in the Arabic corpus
Durations (in seconds) of the different disfluency types uh, um, *glottal FPs, lexical FPs, lengthenings, repetitions, and the nasal FP* hm *as a function of task.*

Table 5.1: Frequency count of Arabic disfluencies

Absolute numbers of the disfluency phenomena uh, um, *glottal FPs, lexical FPs, lengthenings, repetitions, and the nasal FP* hm *in the Arabic data set divided by task.*

Disfluency	um	uh	glottal	lexical	length	repetition	hm
Daily life	3	44	3	7	4	4	2
Map task	5	31	2	9	6	7	0
total	**8**	**75**	**5**	**16**	**10**	**11**	**2**

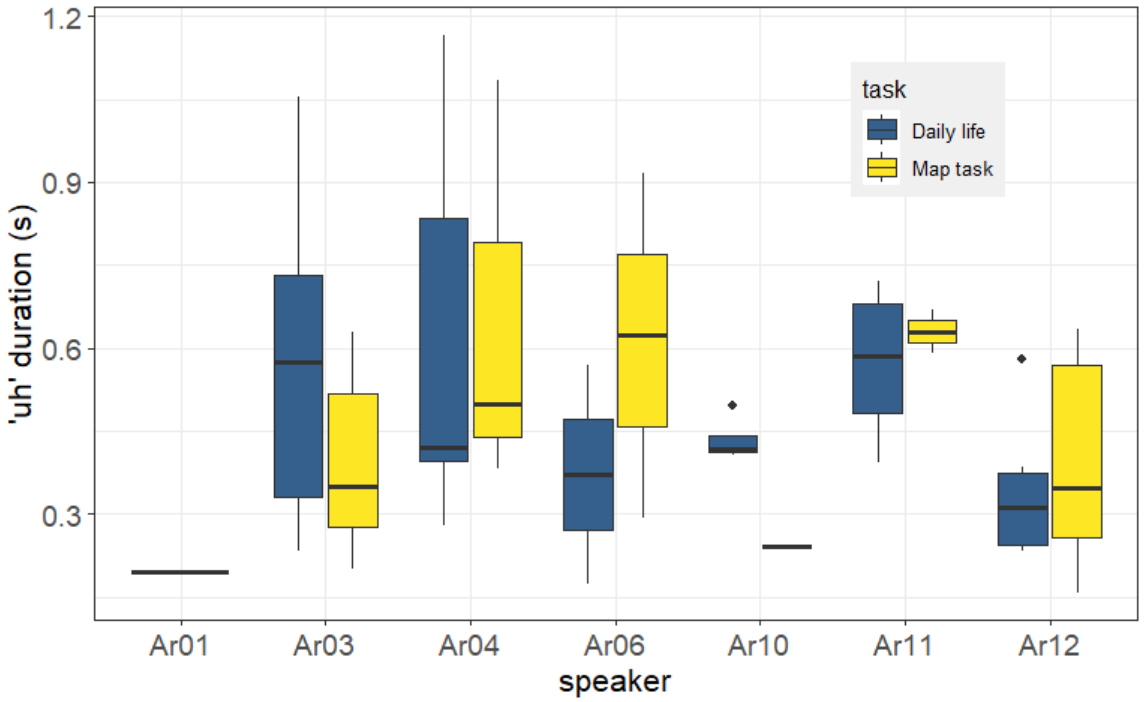

Figure 5.13: Durations of vocalic FPs in the Arabic corpus
Durations (in seconds) of the vocalic FP uh per speaker as a function of task (Daily life/Map task).

As the vocalic FP is the most frequent disfluency phenomenon in this data set, we use this particle to look at the individual differences in duration. The overall frequency of *uh* per minute is higher in the *Daily life task* than in the *Map task* (5.19 vs. 3.36 FPs/min). In Figure 5.13, durations of the vocalic FP are plotted per speaker and task. Table 5.2 includes additional information about the number of tokens used for each boxplot. When taking this information into account, we see that the duration's variance increases with a higher number of tokens, which means that each individual has a high within-speaker variability regarding the duration of vocalic FPs. The observed trend, that FPs are longer in the *Daily life task*, can be seen in the speaker with the most FPs, Ar03. A larger data set is needed to verify any observations made here, as the current data set is too small for reliable results.

Table 5.2: Frequency count of vocalic FP per speaker

Absolute numbers of the FP uh in the Arabic data set per speaker, divided by task.

speaker	Ar01	Ar03	A04	Ar06	Ar10	Ar11	Ar12
Daily life	1	14	9	2	4	8	6
Map task	0	13	3	3	1	2	9
total	**1**	**27**	**12**	**5**	**5**	**10**	**15**

Discussion

The previous analysis indicates that native Arabic speakers use silent pauses more frequently than any disfluency investigated here. The vocalic FP *uh* is the most frequent disfluency aside from the silent pause. The disfluency patterns per speaker are quite consistently in line with previous literature on British English and German (McDougall & Duckworth, 2018; Braun & Rosin, 2015; Braun et al., 2023). For the vowel quality of the Arabic FP, a wide spread across the central vowel space could be observed as before for German (Chapter 4) and the variation of lexical vowels is quite high too. For lexical vowels, this can be explained by the high allophonic variation that is allowed in the Arabic vowel system. These results on the vowel quality of Arabic FPs have to be considered carefully, as the data set is very small, with only seven speakers. Furthermore, Figure 5.9 shows the data of male and female speakers together without a normalisation method applied. When comparing the Arabic FP vowel quality to the data of the Pool2010 corpus (Jessen et al., 2005) and the Diapix-FL corpus (Cooke et al., 2013), the highest degree of overlap was observed with the German data. For Figure 5.11, the Nordström normalisation method was used to eliminate formant differences due to physiological differences between males and females. Future studies would benefit from a large speaker pool representing male and female speakers in sufficient numbers and all languages investigated here, especially Arabic, as disfluencies in this language are less researched than in Germanic and Romance languages.

The analyses on the durational measures of Arabic disfluencies can only show tendencies, which must be verified by investigating larger data sets. For the frequent vocalic FP, a large variance in duration was observed without a significant difference between the tasks. This could mean that the difference in cognitive load does not affect the duration of FPs but only their frequency of occurrence. Another possibility could be that the tasks did not differ in cognitive load at all and thus, no difference in duration was to be expected. Another tendency observed here was the higher rate of the vocalic FP *uh* in the *Daily life task* than in the *Map task*. This was initially surprising as we expected the higher cognitive load of the *Map task* to increase the frequency of disfluencies. However, FPs are also produced at points of higher uncertainty and with a higher number of options (Goldman-Eisler, 1957, 1958; Maclay & Osgood, 1959; Dammalapati et al., 2019, 2021; Zámečník, 2019). The higher rate

in the less scripted *Daily life task* could be explained by this.

5.4 Interim conclusion

To conclude, this chapter has shown that different languages prefer different FP types and vowel qualities. FPs in native English and Spanish speech are quite distinct in (i) the type that is preferred (*um* vs. *uh*) and (ii) the vowel quality that is used, approximating the realisations of the lexical vowels /ʌ/ in English and /e/ in Spanish. Furthermore, it is shown that learners of an L2 are able to produce FPs with a native-like vowel quality, even though a full adaptation was not observed. The degree of the target language FP adaptation is speaker-specific, most speakers produce FP vowels that move in the direction of those of the target language, and only a few achieve the target consistently. It may be the case that speakers form a new category for their FP vowel, but this category is not very specific in terms of the allowed vowel quality as suggested by the variety of the L2 FPs, also within speakers. With increasing proficiency and language use, the L2 FP category may decrease in size so the foreign FP category can be better approximated.

First findings from an Arabic data set suggest that speakers of this variety prefer the vocalic FP *uh* as a particle for hesitation, but generally also produce silent pauses frequently. The analyses of the seven speakers furthermore suggest that the patterns of how speakers use disfluencies may be speaker-specific. Speakers were quite consistent in using their preferred disfluency pattern in both tasks. This has to be confirmed by a large-scale study on Arabic, but preliminary results are promising. Comparisons of vowel qualities across languages show a large amount of overlap between the FP vowel in German and Arabic. English FP vowels overlap partially with Arabic FP vowels but less with German FP vowels. The Spanish vowel quality seems to be unique in this comparison as overlap with the Arabic, English, and German data is minimal.

The results discussed here offer insights into the realisations of FPs in three languages, suggesting that different languages employ specific FP paradigms. The tendency that native-like qualities of FPs (their vowel quality and preferred type) transfer to the speaker's L2 even for advanced learners may be highly relevant in fields where the speaker's background is unknown (forensic phonetic casework, LADO). The results also support the view that disfluencies should be discussed in the L2 classroom to raise the learners' awareness of the foreign FP realisations. Whether the discussion of FPs in the L2 context helps the learners better approximate the L2 FPs is still an open question in L2 research.

This chapter presented cross-linguistic comparisons of FPs in English, Spanish, and Arabic data, including the German data presented in the previous chapter for comparisons of the FP vowel quality. For the next chapter, we turn to the perception

of FPs and, specifically, to the beneficial recall effect of FPs that is frequently reported in the literature (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2011a; Diachek & Brown-Schmidt, 2022). As these studies do not focus on the phonetic details of the FPs in the stimuli, the next chapter is aimed at shedding more light on the recall effect in combination with the duration of FPs. It is hypothesised that longer FPs result in a more pronounced recall effect, than shorter FPs, as the former may be more salient in the stimuli and the latter may escape the listeners' attention.

Chapter 6

Re-evaluation of the recall effect of filler particles

6.1 Introduction

Spoken material immediately preceded by a filler particle (FP), such as *uh* or *um*, can lead to an improvement in verbal recall as briefly outlined in Chapter 2.1.2. This chapter reports on a series of experiments designed to evaluate the improved memory effect of FPs. The aim was to develop an easy-to-use test scheme for detecting this beneficial effect that can be applied across languages and speakers. We tested the scheme in English and German, but it should be applicable to other languages as well.

We regard recall as the ability to retrieve previously encountered information. Information recall plays an important role in all aspects of our lives, starting from an early age. Throughout the education of a person, it is common to conduct exams that focus on the recall of previously taught information. Naturally, learners and researchers alike attempt to find factors that aid in remembering information, thereby simplifying the learning process. For example, taking notes while listening to a lecture is associated with better recall (Fisher & Harris, 1973). This example is a measure that the active listener can take in order to improve their recall. However, the speaking style of the speaker may also influence the recall of the listener. It has been reported that FPs – such as *uh* and *um* in English – improve the recall of the information that follows the hesitation (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2011a). Notably, this research only uses English as the experimental language. The studies presented in this chapter are aimed at replicating the positive recall effect of FPs reported by Fraundorf & Watson (2011a), i.e., information preceded by an FP improves recall compared to fluent information. The aim is using a similar experimental design in German and English as well as finding a suitable

experimental design for testing this recall effect across languages.

6.1.1 Related work

FPs are often associated with undesirable delays in speaking that speakers try to avoid, especially when talking publicly (Fox Tree, 2001, 2002; Niebuhr & Fischer, 2019). Therefore, it might seem counter-intuitive that the symptoms of processing difficulties (de Leeuw, 2007) may have benefits for the listener. In an English word-monitoring task, Fox Tree (2001) found that the FP *uh* increased the speed of word detection, indicating an improvement in online word processing. While the same effect could not be found for the particle *um*, the results suggest that it does not inhibit processing either. In a behavioural experiment, Corley & Hartsuiker (2003) found that participants reacted faster in choosing the correct picture when the instructions included the FP *um* (1,078 ms in duration). Brennan & Schober (2001) also found this effect in a similar study, but they suggest that this benefit may be due to the additional processing time an FP provides since conditions with a long silence instead of an FP led to similar reaction times. The studies presented in Corley & Hartsuiker (2011) support this suggestion.

The possible benefits of FPs regarding word processing may also lead to long-term effects on memory. Corley et al. (2007) and Collard et al. (2008) found a recall benefit for words that were preceded by FPs in a surprise memory test. Their main experiments were EEG experiments in which participants listened to sentences including predictable and unpredictable final words that were sometimes preceded by an FP. Collard et al. (2008) also included a condition of acoustically manipulated final words. The results showed that word integration was more difficult for unpredictable (and acoustically manipulated) final words, which is to be expected. This effect was reduced when the unpredictable (and acoustically manipulated) words were preceded by an FP. Both Corley et al. (2007) and Collard et al. (2008) conducted a subsequent memory test where the subjects had to indicate whether the target words and some distractors occurred in the previous EEG experiment. Recognition of the target words, which were preceded by an FP in the main experiment, was higher than for target words without an FP. Further evidence for the recall effect of FPs come from Diachek & Brown-Schmidt (2022). They conducted four studies testing the effect of different disfluency types (FPs, silent phases, and repetitions) on the recall of single words in a similar set-up as Corley et al. (2007) and Collard et al. (2008). The results of these studies show that all disfluency types boost the recall of the target words similarly when these phenomena occur directly before the item in the initial experiment, but this is only the case for sentence-final words. The recall of items that occurred earlier in the sentence was not affected by preceding disfluencies. A beneficial effect of FPs was also found at the level of discourse by Fraundorf & Watson (2011a) who employed an experimental design using short stories instead of single sentences.

While all previously mentioned papers used English as the experimental language, Chen et al. (2022b) could not replicate this effect for Mandarin Chinese, in a study similar to Fraundorf & Watson (2011a), for human nor synthesised speech when using a humanoid social robot as “speaker". Another study examined the recall effect for Dutch (Bosker et al., 2014). Following an eye-tracking experiment, they conducted a post-experiment memory test similar to Corley et al. (2007) and Collard et al. (2008). Participants were presented with a combination of pictures with high- and low-frequency referents and fluent or disfluent instructions. Following native-accented instructions to click on one of the referents, participants' gaze turned towards the low-frequency items more often after hearing a disfluency. It seems that participants attributed the hesitation to problems with lexical retrieval, which is more likely with low-frequency words. The effect did not occur when the instructions were given in a non-native accent, suggesting that participants, in this case, did not use the hesitation to predict the referent, possibly because a non-native speaker struggles more to find words than a native speaker. Importantly for the current studies, the post-experimental memory test did not show any beneficial effect of FPs. One reason for this might be the difference in design in this study compared to previous studies. Seeing the pictures in the eye-tracking experiment might have affected the memory results and thus the effect from the FP.

Experiments 1a and 1b are partial replication studies of Fraundorf & Watson (2011a). Experiment 1a was conducted in German while Experiment 1b was conducted in English with the original stimuli used by Fraundorf & Watson (2011a). Experiment 2 introduces a new paradigm to test the recall effect of FPs with fluent and disfluent lists. This experiment was also accompanied by a digitspan test to assess the subjects' memory capacity and a pretest to collect the stimuli for the main experiment. Due to the Covid-19 pandemic, all experiments were conducted online.[1]

6.1.2 Aims of the replication experiments

Experiments 1a and 1b of this chapter are based on the first experiment by Fraundorf & Watson (2011a). This section describes their experiment in further detail as it forms the basis for Experiments 1a and 1b. Their aim was to assess whether the recall effect of FPs could be observed when using a longer discourse rather than short sentences as stimuli. Furthermore, they tried to assess the reason behind the recall effect on the basis of three hypotheses. Is the recall effect due to (1) added processing time, do FPs (2) help in predicting upcoming information, or merely (3) direct attention to the upcoming speech stream? The processing time hypothesis attributes the recall effect of FPs to the time that is added by the FPs which can be beneficial for processing information. The predictive processing hypothesis states that the recall effect may be because FPs suggest certain kinds of information following them, like new information

[1]Our own stimuli, data sets and the analysis code can be found in the OSF repository under the following link: https://osf.io/wqjv5/

and less accessible referents. The last hypothesis, the attentional-orienting hypothesis, states that FPs orient attention to the following speech stream which is why the information is better recalled. This hypothesis differs from the previous, as it supposes that the listener does not necessarily make predictions about the upcoming speech (cf. Fraundorf & Watson (2011a) for a detailed explanation of the three hypotheses). To test these hypotheses, Fraundorf & Watson (2011b) developed an experimental design that included three short stories in three different conditions.

They created an experimental design using three short passages from *The Adventures of Alice in Wonderland* (Carroll, 1916) that were retold in one's own words, rather than read verbatim. The retelling was done by a female native English-speaking research assistant to form the basis for the stimuli. The non-manipulated "fluent" versions were altered to create a version that included FPs (*uh* and *um*) before six of 17–18 sentences (depending on the story). A third version included coughs before the same six sentences. Native English-speaking participants (n=72) heard each story in a different condition and were asked afterwards to retell the story in their own words. The focus of their experiments was not on retelling the stories word-for-word, but rather on whether the subjects remembered all the important events (plot points). The results of Fraundorf & Watson (2011a) showed that the FP condition was better recalled than the fluent condition, while the cough condition impaired recall. Based on these results, it seems that FPs are beneficial for the recall of information not only in their immediate vicinity but also on a discourse level.

Noteworthy here is that the stimuli used by Fraundorf & Watson (2011a) are not entirely fluent in a phonetic sense.[2] Although they do not include any FPs, they do contain syllable lengthenings, which can also be assigned to the class of disfluencies. An informal inspection showed that these lengthenings occur very frequently in each story (17 per story on average). Furthermore, the FPs in the stimuli are much longer (mean 1,314 ms for *uh*, 1,441 ms for *um*) than what we observe in spontaneous speech (see Chapter 4) and what is reported in the literature (de Leeuw, 2007; Shriberg, 2001). This may be due to differences in the tasks the speakers were performing and the cognitive load involved.

6.1.3 Hypotheses

Based on the literature reviewed above, we hypothesise that FPs facilitate the memorisation of discourse in English, as described by Fraundorf & Watson (2011a). Due to the fact that the recall effect has not been reported for German or a language other than English (Chen et al., 2022b; Bosker et al., 2014), we do not have any predictions for German as an experimental language. Furthermore, we expect that silent phases (instead of FPs) also lead to a positive effect on memory, as they affect word

[2]We kindly thank Scott Fraundorf and Duane Watson for providing us with the stimuli of their study.

processing in a similar way (Brennan & Schober, 2001; Corley & Hartsuiker, 2003, 2011).

In Experiment 2, we test whether the beneficial effect of FPs is also present when using a list recall paradigm (i.e., lists of twelve items from the same semantic field (e.g. fruits) with FPs before two of the items) that primarily focuses on short-term memory. As there is no latency between hearing the list and reproducing it apart from the time the audio is playing, this paradigm differs from the memory test employed by Corley et al. (2007) and Collard et al. (2008) who found a longer-term beneficial effect of FPs on the memorisation of words.

6.2 Experiment 1a: German

Experiment 1a was designed to assess whether the recall effect of FPs is also present in German.

Methods

Stimulus recording

The short passages of *Alice in Wonderland* (Carroll, 1916) provided by Fraundorf & Watson (2011a) were translated to German and recorded by a female native German-speaking research assistant using an H1 Zoom recorder with a clip-on microphone in a quiet room. The stories were memorised and retold in a way that mimicked the original retellings from Fraundorf & Watson (2011a), rather than read aloud by the speaker. A fluent version, not containing lengthenings nor FPs, was created by concatenating the most fluent sentences produced by the speaker to form the entire story. Concatenation was not noticeable. Two manipulated versions were created: one included FPs before six of the 16-18 sentences (depending on the story) and the other substituted long silences for the FPs. The sentences that were manipulated were the same as those used by Fraundorf & Watson (2011a) in their experiment. The duration of the long silences was also matched to the duration of the FPs. The condition with long silence was created to investigate whether unfilled pauses have the same effect on recall as FPs. Six FP tokens were chosen from a disfluent trial of the speaker retelling the stories. Three tokens of each FP type (*uh* and *um*) were chosen that differed in their duration: 371–606 ms for *uh* (mean 482 ms) and 611-784 ms for *um* (mean 679 ms) (Table 6.1). These FPs were spliced into the fluent version of the stories (into the existing silent phases, before six of the sentences) to create the FP condition. The order in which the FP tokens appeared in each story differed. Minor changes in intensity were applied to fit the FP to the surrounding speech material. To create the long silence condition, quiet passages from the same speech recordings were taken and spliced into the fluent version of the stories in the same position as the FPs in the previous condition.

Table 6.1: Comparison of stimuli durations

Mean (SD) FP duration of stimuli in ms in Experiments 1a and 1b reported here compared to stimuli by Fraundorf & Watson (2011a).

FP type	*uh*	*um*
Exp 1a (FP condition)	482 (112)	679 (59)
Exp 1b (short FP condition)	571 (77)	726 (74)
Fraundorf & Watson (2011)	1314 (252)	1441 (290)

Experimental set-up

The experiment was set up in Labvanced (Finger et al., 2017). Nine different lists were created containing each of the three stories in a different condition (fluent condition, FP condition, long silence condition) and by randomising the order in which the stories and conditions were presented to the subjects. The experiment included a practice passage that was not evaluated in the analyses in order to familiarise the participants with the task.

Subjects

45 subjects (mean age 31.2 yrs; age range 18–63 yrs; 19 females, 25 males, 1 non-binary) with German as their native language were recruited and paid £2.80 via Prolific (2014). They were randomly assigned to a list. Subjects heard each story in a different condition once and were asked after each story to retell it in their own words while recording themselves. The experiment took approximately 20 minutes. Recordings have a 48 kHz sampling rate. It is expected that recording hardware and audio equipment differ from subject to subject, indicated by the varying quality of the recordings.

Scoring of answers

Each story consisted of 14 plot points which included events that were important to the outcome of each story. The answers of each participant were checked as to whether or not these events were mentioned in the retellings. The scoring of participants' retellings was done by two native German annotators (author and a student assistant). To check inter-rater agreement on the plot points, 20 % of the data were scored by both annotators, resulting in 94.7 % agreement. Where annotators disagreed, the first annotator's (author) scores were used. Scoring was done based on guidelines developed by both annotators. The difficulty was that one plot point possibly contained more than one single piece of information. For example, one plot point of one story read: *The White Rabbit runs by and drops his fan and gloves.* Had the subject only mentioned one event (e.g. the rabbit running by) and not the other (e.g. the rabbit dropping his fan and gloves), the question was whether this plot

Table 6.2: Contrast coding of Experiments 1a & b

Contrast coding matrix of statistical models of Experiment 1a (German) and Experiment 1b (English), which differ in experimental conditions.

Experiment 1a	fluent	FP	long silence
Experiment 1b	fluent	long FP	short FP
Contrast1 (C1)	2/3	-1/3	-1/3
Contrast2 (C2)	0	1/2	-1/2

point would be counted as correctly recalled or not. The annotation guidelines clearly stated which details of each story had to be mentioned in order for this plot point to be coded as correct. In the example above, the subject would not receive the point for the plot point in question, as the second part was considered more important for the outcome of the story. Annotations were done in a binary fashion (0 = not recalled, 1 = recalled) with no partial points given.

Results

Analyses were performed in R (Version 4.1.3) (R Core Team, 2022) by fitting Generalised Linear Mixed-Effects Models using the lme4 package (Bates et al., 2015) (Version 1.1-28). Estimates, standard errors, and z-values were reported. We extracted confidence intervals by using the confint function of the stats package (Version 4.1.3) (R Core Team, 2022). All models were run using the bobyqa optimizer and iterations were increased to $2{\cdot}10^5$ to avoid convergence issues. The random effects structure was determined using a PCA analysis as reported in Bates et al. (2015). Models were compared using the Akaike Information Criterion (AIC) (Akaike, 1998); models with the lowest AIC are preferred. Contrast coding was used for the levels of manipulations in the following way. The first contrast (C1) compared the fluent condition (2/3) against the other two conditions (FP condition and long silence condition; each -1/3), which is why the coded value for fluent was the same as the absolute sum of the values of the other two conditions. The second contrast (C2) only compared the FP condition (1/2) to the long silence condition (-1/2), which was achieved by setting the value of the fluent condition to equal 0 (cf. Table 6.2).

A first analysis was done on the full data set using all plot points, target plot points, and “distractor" plot points that were not subject to manipulation in any condition. A best fit model was determined including a random intercept and slope for the subject and a random intercept for the plot point. The analysis revealed a main effect for C2, but no significant effect for C1 (see Table 6.3 for estimates). The significant effect for C2 suggests a difference between the long silence condition and the FP condition, in that the long silence condition improves recall compared to the FP condition. However, as there are 14 plot points in every story and six of them were manipulated only in two conditions, the full data set is imbalanced in this way.

Table 6.3: GLMM of Experiment 1a

Best fit models of the full and the target data set as well as their estimates for the statistical analysis of Experiment 1a for the German data set.

	Estimate	Std. Error	z-value	Pr(>|z|)	CI
Full data set:	*Answer∼C1+C2+(1+C1+C2|Subject)+(1|Plotpoint)*				
(Intercept)	0.94	0.25	3.70	<.001 ***	
C1	0.14	0.18	0.78	.44	[-0.22; 0.53]
C2	-0.46	0.19	-2.46	<.05 *	[-0.86; -0.09]
Target data set:	*Answer∼C1+C2+(1+C2|Subject)+(1|Story)+(1|Plotpoint)*				
(Intercept)	1.18	0.38	3.08	<.05 **	
C1	0.47	0.19	2.52	<.05 *	[0.1; 0.84]
C2	-0.50	0.26	-1.88	.06 .	[-1.06; 0.03]

To create a balanced data set, the full data set was reduced to only the target plot points, i.e., the potentially manipulated plot points in each story. This resulted in 810 observations (3 stories * 6 plot points * 45 participants). The model with the best fit for the target data set (including intercepts for subject, story and plot points, and a random slope for the subject regarding C2) showed a main effect for C1 and a trend for C2 (see Table 6.3 for estimates).

This suggests that the fluent condition improves recall of the plot points but the long silence and the FP condition impair recall (Figure 6.1). The trend of the second contrast suggests that the FP condition impairs recall more than the long silence condition.

Discussion

The results of Experiment 1a with our German data did not confirm the findings reported by Fraundorf & Watson (2011a). In both analyses of the full data set and the target data set, the FP condition's effect was worse than that of the other two conditions. Thus, recall seems best after the fluent stories and decreases the more salient the manipulation becomes, i.e., an FP was a more profound manipulation than the insertion of a short silence.

6.3 Experiment 1b: English

Experiment 1b is also a partial replication of Fraundorf & Watson (2011a); however, compared to Experiment 1a, it is closer to the original as it uses the same stimuli, so the language and the speaker are identical. This experiment was conducted in order to look into the different results between the previous experiment and the one by Fraundorf & Watson (2011a).

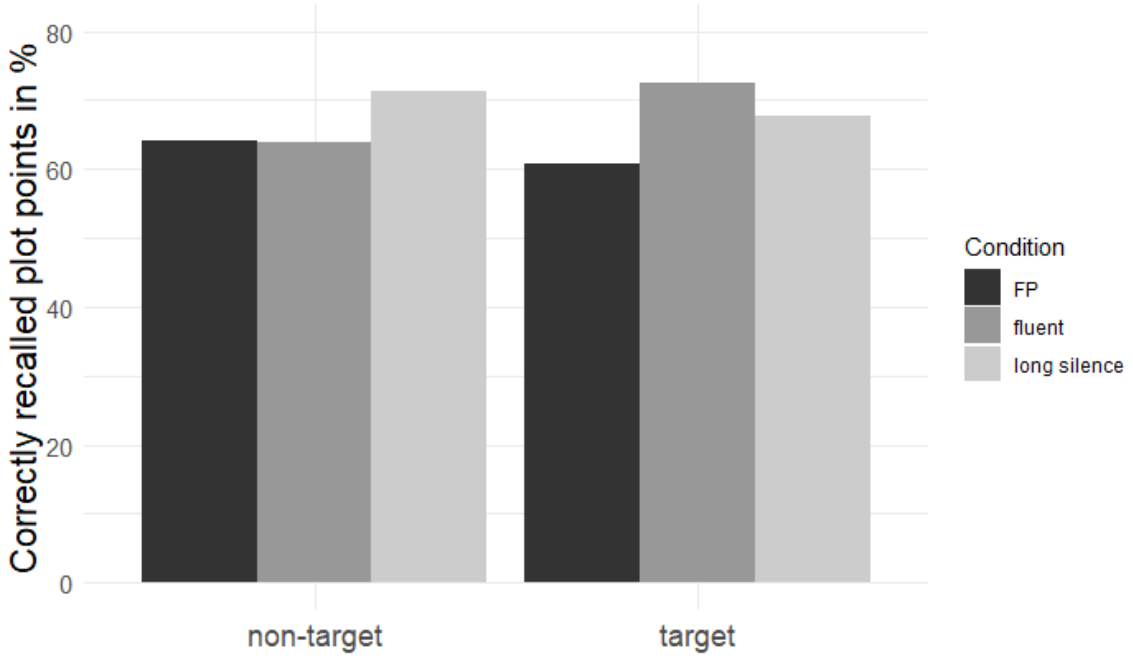

Figure 6.1: Experiment 1a: Recalled plot points

Correctly recalled plot points per condition for items. Target items are the potentially manipulated plot points (Note: target items in the fluent condition are not manipulated but still included as a reference).

Methods

Stimuli

The original recordings of Fraundorf & Watson (2011a), which were used for this experiment, were made by a female research assistant with a North American accent, in a similar way as reported for Experiment 1a. The speaker memorised the plot points and then retold them from memory rather than reading directly from the script. The fluent version of each story did not contain FPs, but disfluencies like syllable lengthenings were present in the recordings.[3] We used the stimuli from the fluent and FP conditions from Experiment 1 (Fraundorf & Watson, 2011a) while creating a new condition based on the FP condition. As reported in Table 6.1, the FP stimuli had a mean duration of 1,314 ms (sd = 252 ms) for *uh* and a mean duration of 1,440 ms (sd = 290 ms) for *um*, which is longer than usually reported (see Chapter 4 and also (de Leeuw, 2007)). For English, German, and Dutch, de Leeuw (2007) reports mean values ranging from 317 ms to 379 ms for *uh*, and from 457 ms to 611 ms for *um* in spontaneous speech. For the German data reported before (Chapter 4), mean values of 382 ms for *uh* and 559 ms for *um* were found. Adding two standard deviations to these mean values would just exceed the 1 s mark for the vocalic-nasal FPs, but not the vocalic FP, illustrating the exceptional length used in Fraundorf & Watson (2011a). As the FPs used in Experiment 1a were substantially shorter (mean = 482 ms, sd = 112 ms for *uh*; mean = 679 ms, sd = 59 ms for *um*), a third condition was created where the FPs and the surrounding silence were manually shortened by half their duration, resulting in FPs with a mean duration of 571 ms (sd = 77 ms) and 726 ms (sd = 74 ms), respectively. This will be referred to as the short FP condition while the unchanged FP condition will be called the long FP condition hereafter. Since the stimuli in 1a and 1b are recorded by two different speakers, the speaking rate (including pauses) differs across the two experiments, with the German speaker producing faster speech (4.3–4.5 syllables/second) than the English speaker (3.0–3.2 syllables/second). These measures were taken from the fluent conditions.

Experimental set-up

The experiment was set up in Labvanced (Finger et al., 2017). Nine different lists were created containing a combination of each story with the different conditions (fluent, long FP, and short FP). The order of stimuli within the lists was fixed such that all subjects who saw a particular list encountered the stimuli in the same order. The experiment included a practice passage to familiarise the participants with the task.

In the experiment, 58 subjects (mean age 31.4 yrs; age range 18—57 yrs; 35 females, 23 males) with English as their native language were recruited and paid £2.80 via Prolific (2014). None of these participants took part in Experiment 1a. They were

[3]For more details on the recording procedure, see Fraundorf & Watson (2011a).

Table 6.4: GLMM of Experiment 1b

Best fit models of the full and the target data set as well as their estimates for the statistical analysis of Experiment 1b for the English data set.

	Estimate	Std. Error	z-value	Pr(>| z|)	CI
Full data set:	*Answer∼C1+C2+(1+C1+C2||Subject)+(1|Story)+(1+C1+C2||Plotpoint)*				
(Intercept)	0.46	0.22	2.12	<.05 *	
C1	0.12	0.15	0.83	.41	[-0.175; 0.425]
C2	0.14	0.16	0.89	.38	[-0.19; 0.48]
Target data set:	*Answer∼C1+C2+(1+C1+C2|Subject)+(1| Story/Plotpoint)*				
(Intercept)	0.56	0.31	1.81	.07 .	
C1	0.07	0.19	0.37	.72	[-0.3; 0.46]
C2	0.19	0.23	0.84	.4	[-0.27; 0.67]

randomly assigned to each list. Subjects heard each story once and were asked after each story to retell the story in their own words while recording themselves. The experiment took approximately 20 minutes. Recordings have a 48 kHz sampling rate. As in Experiment 1a, the quality of the recordings was highly variable with respect to background noise and recording hardware.

The scoring was performed as described under Experiment 1a. Annotators' plot point agreement for 20 % of the data reached 94.8 %. Where annotators disagreed, the first annotator's (author) scores were used.

Results

A statistical analysis was done as reported under Experiment 1a. The contrasts were coded as follows: for C1, the fluent condition received the value 2/3, while the other two conditions (long FP condition, short FP condition) were coded as -1/3 each (see Table 6.2). For the second contrast (C2), the fluent condition was set to 0, the longer FP condition received the value 1/2, and the value for the short FP condition was -1/2. For the full data set (i.e., all plot points including distractors), we used a random effects structure, including random intercepts for the subject, story, and plot points. Additionally, we included a random slope for the subject and plot points excluding their correlation term. We did not find a significant effect on any contrast when including them in the statistical model. When reducing the data set to targets only (i.e., the six manipulated plot points), significant results were not found. Here, we included a random intercept and slope for the subject and a random intercept for plot points nested under the story (see Table 6.4 for the best fit model).

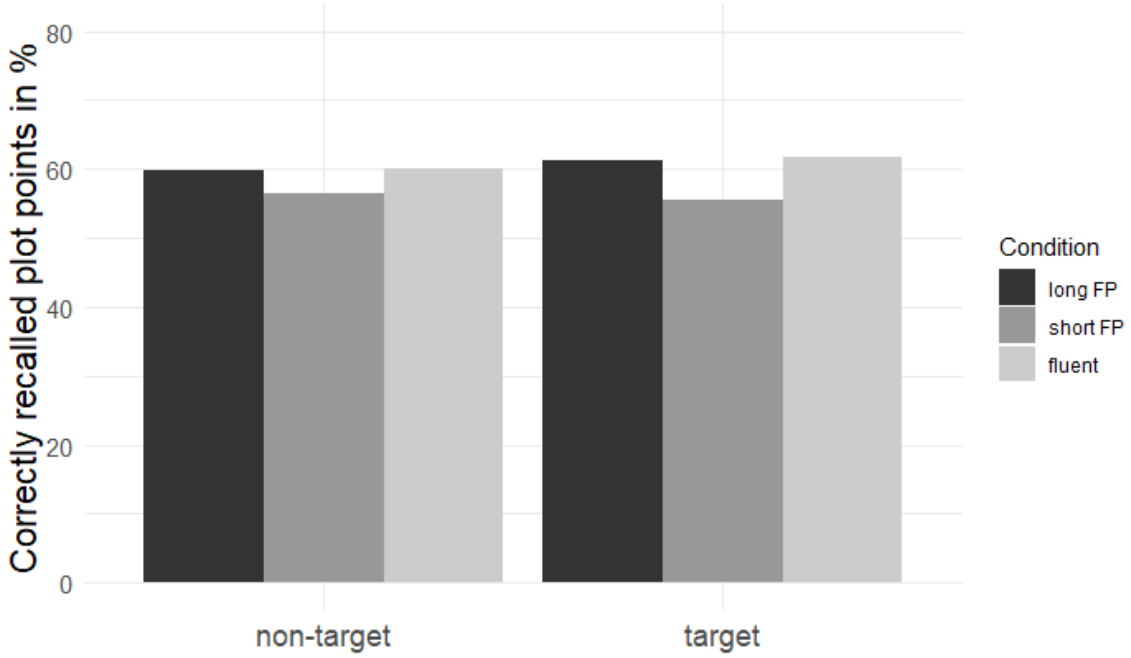

Figure 6.2: Experiment 1b: Recalled plot points

Correctly recalled plot points per condition for items. Target items are the potentially manipulated plot points (Note: target items in the fluent condition are not manipulated but still included as a reference).

Discussion

With Experiment 1b, it was not possible to replicate the beneficial recall effect of FPs found in Fraundorf & Watson (2011a) nor could we confirm the opposite findings of Experiment 1a– that the fluent version led to better recall. This is surprising as the exact same stimuli from the original study (Fraundorf & Watson, 2011a) were used for Experiment 1b. Possible reasons for these results will be reviewed in the general discussion.

6.4 Experiment 2

The previous experiments used short stories as stimuli and subjects were asked to recall the stories by explaining the important plot points. As the results from Fraundorf & Watson (2011a) could not be confirmed in Experiments 1a and 1b, we aimed to simplify the design by exerting more control over the stimuli. We chose a list design of twelve items that included FPs in one condition (= disfluent) and none in the other condition (= fluent). They were presented aurally to the subjects in a web-based experiment. Participants had to memorise the lists and then reproduce them by typing the items into an answer box. Another component of the experiment included a digit span task that was intended to assess a subject's memory capacity. A pretest was conducted to generate and norm the stimuli used in the experiment (section 4.2).

Digit span

It is assumed that overall memory capacity has an effect on the memorisation of list items. The digit span was designed to resemble the main experiment as closely as possible, which is why the stimuli of the digit span were also presented to the subjects aurally. This test included 25 stimuli of random digit sequences, including the digits from one to nine (no digit appeared twice in the same sequence). Stimuli were of different lengths ranging from a four-digit sequence (e.g. 5921) to an eight-digit sequence (e.g. 17568932). Each length appeared five times in the set of stimuli (see Table 6.5 for an overview of stimuli). Stimuli were presented randomly to the subjects.

The aim was to acquire one score for each participant that represented the participant's memory capacity, which we refer to as memory score. It represents the longest digit span stimulus length the subject was comfortably able to recall. We defined the threshold for the memory score to be a correct recall of 80% of the stimuli per category, i.e., per stimulus length. In other words, the memory score is the longest stimulus length for which the subject was able to score four out of five trials correctly. An example of the procedure is given in the last column of Table 6.5. The example subject scored at least four times correctly on three stimulus lengths (i.e., 4, 5 and 7),

Table 6.5: Stimuli digit span

Overview of the digit span stimuli, including an example subject to emphasise the scoring method of the memory score: the values denote the number of correctly recalled trials per stimuli length. The memory score per subject is determined by the longest stimuli length with a score of at least 4 out of 5.

Stimuli example	Stimuli length	Number of stimuli	Subject example
5921	4	5	5/5
41865	5	5	4/5
972841	6	5	3/5
1365284	7	5	4/5
17568932	8	5	2/5
		total = 25	memory score = 7

and as seven is the longest stimuli length, the subject's memory score is seven (even though they scored only three times correctly on the previous stimulus length). We also calculated the memory score with thresholds of three and five but these showed ceiling or floor effects, respectively. The current threshold (= 4/5) best approximates a normal distribution.

Stimulus collection

A pretest was conducted for the collection of stimuli. The aim was to generate English words for different semantic categories and establish their relative frequencies within each category. We presented seven common categories (zoo animals, body parts, clothing, fruit, furniture, musical instruments, and vegetables). The task was then to name as many items in each category as the subject could think of by typing them in the answer box. The experiment was set up using Labvanced (Finger et al., 2017). Every subject was presented with each category one, after the other. Subjects saw the category name (e.g. zoo animals); their task was then to type as many items they could think of in this category (e.g., lion, giraffe, elephant, etc.) in an answer box on the screen. The task was limited to 90 seconds. After the timer ran out, the subject was led to the next category or, after all categories were concluded, to a survey collecting information about the subject. Participants were randomly assigned to one of seven experimental lists, which differed only in the order in which the categories were presented. 35 subjects (mean age 29 yrs, sd = 10.1 yrs, age range 18–63 yrs; 22 females, 12 males, 1 non-binary person) were recruited and paid for their participation via Prolific (Prolific, 2014). All subjects indicated English as their native language; however, they were from different English-speaking regions (UK, US, Canada, South Africa, Australia). The majority of subjects were from the UK (n=25). The collected data were manually checked for spelling mistakes. A frequency count was performed to rank the items per category according to their frequency of occurrence. Items with

Table 6.6: List experiment: Example stimuli list

The target items in the fourth and ninth position (here: lemur *and* lion*) were the items to be manipulated with a vocalic-nasal FP in the disfluent condition. These items were swapped in half of the data to control for frequency effects. All other items remained in their positions. The lists in the other five categories had the target items in the same position and the same fixed order of high- and low-frequency items.*

Position	Frequency	Example list
1	low	donkey
2	low	ostrich
3	high	elephant
4	target: low	(um) lemur
5	low	peacock
6	high	tiger
7	high	penguin
8	low	alpaca
9	target: high	(um) lion
10	low	otter
11	high	monkey
12	high	giraffe

a frequency count of 20 or more (i.e., 20 or more participants listed this item) were considered “high-frequency", and items with a frequency count of one to four (i.e., one to four participants listed this item) were considered “low-frequency". The furniture category was excluded because there were not sufficient high-frequency items that fit the requirements. Instead, this category was used as a practice list. For each of the other six categories, a list with six high- and six low-frequency items was created which was then used as a stimulus set for the main experiment. For manipulation, an FP token was inserted before the fourth and ninth item, which was always one high- and one low-frequency item (see Table 6.6). In order to avoid the primacy and recency effects (Murdock, 1962), we chose not to manipulate the first and last three items, which is why the fourth and ninth items were selected. The frequency of the target item was balanced such that in half of the lists the fourth item was low frequency while in the other half the ninth item was low frequency. All remaining items were kept in their assigned positions and thus the frequency of the items was not changed (e.g., the item in the first position was a low-frequency item in all lists, while the item in the twelfth position was always a high-frequency one).

Methods

Stimulus recording

The lists that were created as described in the previous section were recorded in a quiet room using Audacity by a male speaker of American English in a fluent manner. He used a SteelSeries Arctis 7P headset microphone and a sampling rate of 44.1 kHz. An FP token (*um*) was selected from the same speaker from a recording of spontaneous speech with the same recording set-up. The selected token was 560 ms long, and the fluent lists of items ranged between 18 and 20 seconds.

Experimental set-up

The experiment was again created using Labvanced (Finger et al., 2017). Each task consisted of listening to one audio file, i.e., one list of a single category, and afterwards typing the answer in a box that appeared on the screen after the audio finished playing. The task was to listen to each list, memorise as many items as possible, and when prompted, type them in an answer box on the screen. Each subject listened to every category. Participants were randomly assigned to one of twelve experimental lists which differed in the order of categories. Manipulation, i.e., insertion of FPs, occurred in every other category. In between tasks, subjects could take a break. Before the list experiment, subjects also completed the digit span task described previously, and afterwards, they were asked to fill out a short survey on biographical information.

Subjects

73 subjects (42 female, 31 male; mean age 33, sd = 11.2, age range 19–69) were recruited and paid £2.80 for their participation via Prolific (2014). Participants indicated different English accents (British, American, Australian, South African, Irish) but all confirmed English as their native language. The experiment took approximately 20 minutes.

Statistical analysis

The statistical analysis was done using Generalised Linear Mixed Models (GLMMs) from the lme4 package (Bates et al., 2015) in R (R Core Team, 2022). The dependent variable was the binary answer ("recalled", "not recalled") of whether or not the subject recalled an item. Five independent variables were included in the modelling: the manipulation of each item (did an FP occur before the item or not; manipulated vs. non-manipulated), the semantic category, item frequency (high vs. low), item position in the list, and the memory score of the subject. The best fit model was determined using forward selection, i.e., adding each factor into a model with a random intercept per subject and item and comparing the results using the AIC (Akaike, 1998) to a

Table 6.7: GLMM of the full data set for Experiment 2

Estimates of the best-fit model for the full data analysis in Experiment 2. The reference level for the categorical variable "frequency" is the level "high".

	Estimate	Std. Error	z value	Pr(>| z |)
(Intercept)	-2.11	0.53	-4.02	<.001 ***
item order	0.23	0.05	4.19	<.001 ***
digitspan	0.35	0.08	4.36	<.001 ***
frequency	-0.83	0.79	-1.05	0.29
item order:digitspan	-0.02	0.01	-2.24	<.05 *
item order:frequency	0.28	0.11	2.49	<.05 *
digitspan:frequency	0.07	0.11	0.58	.56
item order:digitspan:frequency	-0.02	0.02	-1.42	.16

null model with only random effects. After the best first factor was determined, the second and third factors were chosen in a similar fashion. Models were run using the bobyqa optimizer and iterations were increased to $2{\cdot}10^5$ to avoid convergence issues.

Results

The best fit model was the following:
$glmer(Answer \sim item.order * memory.score * frequency +$
$(1 \mid Subject) + (1 \mid Item), data = full.data, family = binomial).$

The model revealed a main effect for item position as well as for the subjects' memory score (Table 6.7). Item frequency, however, was only significant in the interaction with item position, which may be due to the experimental design, as the items in each position were always of the same frequency, apart from the target item in positions 4 and 9 (which were exchanged for half of the data). Furthermore, it should be noted that in the previous analysis, a large set of non-manipulated items (n=4380) was compared to a small set of manipulated items (n=876). Due to this imbalanced data set, we decided to reduce the data set to only the target items in positions 4 and 9 and regard all other items as distractors. In reducing the data, we are left with a new data set that only includes potentially manipulated items in the fourth and ninth positions of the lists. These items were also controlled for frequency: in half of the data set, the high-frequency item occurred in the fourth position (and the low-frequency item in the ninth), and for the other half, the positions were changed. This was done to attribute possible effects to the frequency of the item and not only to the position in which the item occurred. Statistical modelling was done in the same way as described above for the full data set, and the best fit model was the following:
$glmer(Answer \sim item.order + memory.score +$
$(1 \mid Subject) + (1 \mid Item), data = target.data, family = binomial).$

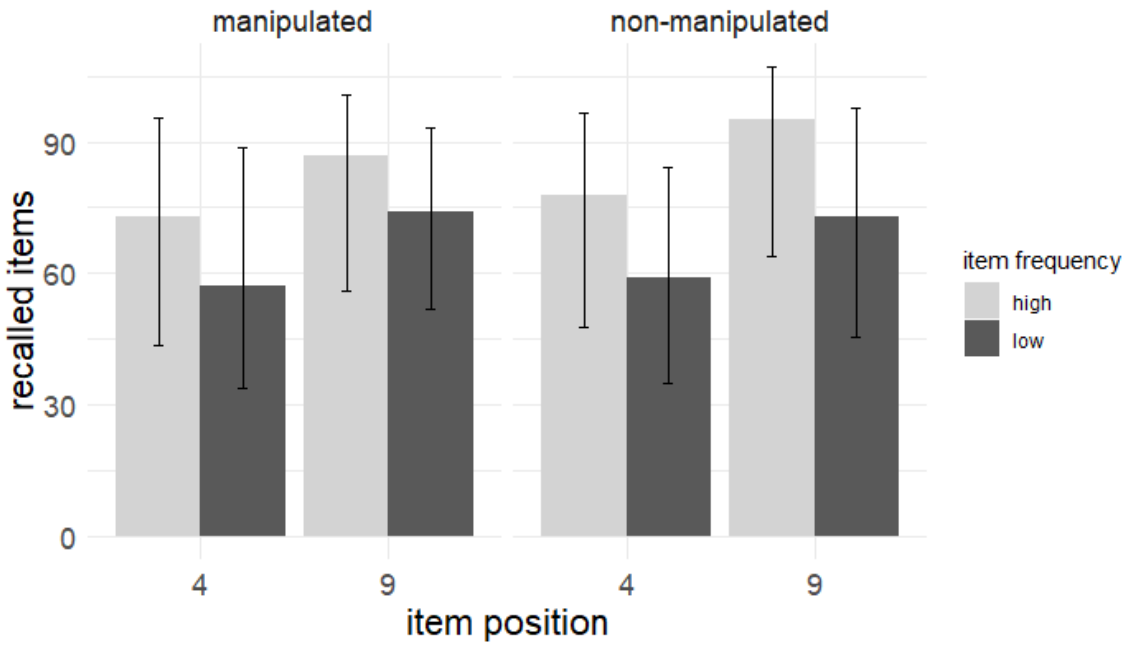

Figure 6.3: Experiment 2: correctly recalled items
Correctly recalled items per manipulation (manipulated = disfluent, non-manipulated = fluent) and item frequency. The error bars represent the standard deviation between subjects.

The model again revealed a main effect for item position ($\beta = 0.15$, SE = 0.03, z value = 4.72, $\Pr(>|z|) < .001$, CI = [0.09; 0.22]) and the subjects' memory score ($\beta = 0.32$, SE = 0.08, z value = 4.26, $\Pr(>|z|) < .001$, CI = [0.17; 0.48]); item frequency is not included in the best fit model and thus not a significant factor anymore. Moreover, the factor manipulation (i.e., whether or not an FP was present before the item) was not a significant factor in either of the two best fit models. In sum, we did not find a beneficial recall effect in our experimental data.

Discussion

Contrary to the results reported in the literature, a significant effect resulting from the inclusion of FPs on the recall of list items was not found in our data. Recollection was influenced by the position of the items, in that items in the ninth position were better recalled than those in the fourth position, which is in line with the frequently reported recency effect (Murdock, 1962). Item frequency was not a significant factor, which is in contrast to our hypothesis and to findings by Poirier & Saint-Aubin (1996). However, the main effect of memory score suggests that subjects who performed well in the digit span also performed well in the memorisation of lists.

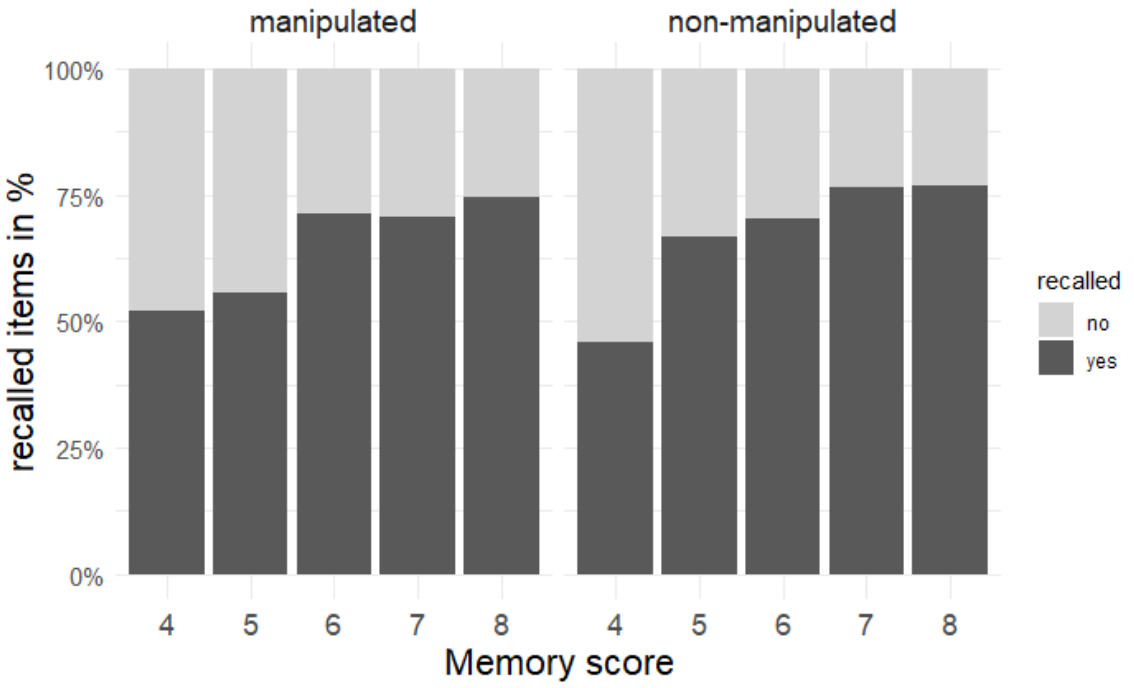

Figure 6.4: Experiment 2: recalled items as a function of memory score
Recollection of subjects categorised by their memory score acquired in the digit span task for manipulated (disfluent) and non-manipulated (fluent) target items separately.

6.5 General discussion

The three studies presented in this paper aimed at replicating the beneficial recall effect of filler particles (FPs). Experiment 1a and 1b were partial replication studies of Fraundorf & Watson (2011a), while Experiment 2 used a list recall paradigm to test the effect. Experiment 1a tested whether disfluent short stories were better recalled than fluent short stories and if there was a difference between the disfluent conditions using FPs or long silences. The results of this experiment, which was conducted in German, did not confirm the presence of a beneficial recall effect but rather found an inhibitory effect, i.e., the fluent version yielded the best recall and the FP condition the worst. The long silence condition tended to lead to a better recall than the FP condition, but this difference was not statistically significant.

While the experiment was very similar to the original one by Fraundorf & Watson (2011a), there were two noticeable differences: First, the study was conducted online, which may have led to participants being less engaged than they would have been in an on-site lab experiment. Second, the FPs in our study were shorter than the ones used in the original experiment. The recall effect may be sensitive to the FP duration, as shorter particles may not provide sufficient time to create the recall benefit or may not be salient enough to draw attention to upcoming material.

Experiment 1b included two FP conditions that differed in the length of the stimuli. The short FP condition matched the FP condition of Experiment 1a, while the long FP condition was the same as in the original study by Fraundorf & Watson (2011a) using the exact same stimuli. Neither a recall benefit nor impairment effect from FPs was found for this experiment as no factor reached statistical significance. The effect of FP duration has to be further examined. It is possible that Experiment 1b was underpowered with a sample size of 58 participants (compared to Fraundorf & Watson's (2011) 72 participants). The studies by Diachek & Brown-Schmidt (2022) suggest that the recall effect is rather small, which means a large data set is needed to obtain a robust result.

Experiment 2 tested the recall effect of FPs using a list recall paradigm. Subjects had to listen to and reproduce lists of twelve items in which two FPs had been inserted in the disfluent condition. Results did not show an effect from the FP on the immediately following item, nor the recollection of the entire list. A reason for this might be the artificiality of the task and/or the limited salience of the FPs.

Phonetic perspectives on filler particles

An important point from a phonetic perspective is that FPs are not all alike. This is in contrast to transcriptions of FPs. For instance, the word-like unit *um* can be very different depending on the surrounding silences and the phonetic characteristics of the FP as outlined in Chapter 4. Both the phonetic context and the phonetic characteristics of the FP may have a strong impact on its *appropriateness* (i.e., naturalness or artificiality) and on its *perceptual salience*.

In both experiments, we observed a mismatch between FPs in natural and experimental data. FPs "in the wild" are often directly connected with speech, be it before the FP, after the FP, or on both sides as discussed in Chapter 2. Of course, there are also cases of an FP occurring with a stretch of silence before and after. However, these cases do not necessarily represent the majority of the cases. In fact, in German corpora, the type speech-FP-speech is predominant (Belz et al., 2017). Furthermore, *um* tends to appear more often as silence-FP-silence, whereas *uh* appears more often as speech-FP-speech (Clark & Fox Tree, 2002). The same was found in Chapter 4 for German data. It could be shown that the vocalic FP occurs most often within speech, while the vocalic-nasal FP occurs most often within isolation, i.e., as silence-FP-silence.

One aim of discussing the phonetic details of stimuli is to raise awareness in research for the extreme phonetic variability of FPs and the unknown effects they have in listening experiments. For this reason, we advocate for a good description of the phonetic characteristics such as duration and surrounding silences, but also for combinations of FPs such as lengthening before the FP or the frequently observed and cliticised sequence *and uh*. All these factors contribute to the naturalness of FPs and their contextual fit, which may, in turn, influence the salience and recall benefit of

the specific token.

Experimental conditions and task demands

One question that arises from our study is why the effect of the FP on recall is inconsistent across experiments. Specifically, we found no effect of FPs on the recall of plot points in English (Experiment 1b) and an inhibiting effect of FPs on recall in German (Experiment 1a). These results contrast with Fraundorf & Watson (2011a), who found a benefit of FPs on recall in English. Moreover, we were also unable to replicate the recall effect previous studies have found (Fraundorf & Watson, 2011a; Corley et al., 2007; Collard et al., 2008; Diachek & Brown-Schmidt, 2022) using a simplified task of list recall. There are several potential explanations for these inconsistencies. Firstly, it is possible that FPs simply do not modulate memory in the recollection of information. However, while this could explain our results (for experiments in English), it is at odds with the results of Fraundorf & Watson (2011a) as well as earlier studies which suggest a facilitative effect of FPs on word recall (e.g., Corley et al., 2007; Collard et al., 2008).

Differences in experimental conditions due to different testing platforms (laboratory vs. online) may account for the discrepancy between the results of Fraundorf & Watson (2011a) ours. While a number of lab-based results have been replicated in online paradigms, it has been noted that web-based methods present the challenge of reduced control over the experimental environment, which may lead to greater variation in how participants respond (Gallant & Libben, 2019). A third consideration is differences in task demands across experiments in our study: the task of recalling items from a sequence is qualitatively different from remembering plot points based on a narrative discourse, and the two likely place different cognitive demands on the comprehender. In particular, the former is typically studied within the framework of serial memory (e.g., Burgess & Hitch, 1999) and is known to invoke organisational strategies such as chunking (Miller, 1956) or mnemonic techniques (Massen & Vaterrodt-Plünnecke, 2006), whereas the latter may rely more on semantic (long-term) memory (Burkhardt, 2007) and requires a global understanding of the narrative. This requires processing at multiple levels for the comprehender to establish coherent relations between statements within the discourse as well as between the discourse and their broader world knowledge (Graesser et al., 1997).

Given these differences, it is perhaps less surprising that FPs may produce dissimilar effects across the two tasks. Experiments investigating the effect of FPs on comprehension indicate that disfluencies impact response times in a word verification task but not in a question-answering task targeting discourse-level relations (Cevasco & van den Broek, 2016). Cevasco & van den Broek (2016) suggest that FPs have a different impact on comprehension at the surface level, where the exact wording of statements is processed, compared to the discourse level, where connections between statements are established to construct a larger meaningful representation. It is also

notable that other types of disfluency (e.g., perceptual disfluency in written language comprehension) have been found to facilitate learning and recall in some situations (e.g., memorising information from a list; Diemand-Yauman et al., 2011), but not in others (e.g., passage comprehension; Pieger et al., 2016). Our results highlight that the behaviour of FPs in comprehension may depend on various factors, including the task goals and, thereby, the cognitive demand on the listener. An investigation of how FPs influence recall across different task contexts is necessary to obtain a more complete understanding of the phenomenon.

6.6 Interim conclusion

Reviewing the results of our three experiments in light of previous work, we conclude that the recall effect of filler particles (FPs) is a complex phenomenon dependent on a multitude of factors. One decision every researcher must make concerns the trade-off between control and the naturalness of the data. In our experiments, we spliced naturally occurring FPs into fluent stories and lists while ensuring that no other disfluencies occurred in the stimuli. While this method is highly controlled, it does not reflect natural disfluent data. It is unusual in natural speech to have FPs as the only type of disfluency. Combinations of different types of disfluencies occur frequently (e.g., a lengthened syllable, plus a short silence, plus an FP). Other studies (Fraundorf & Watson, 2011a; Collard et al., 2008) used stimuli that not only included FPs but also lengthened syllables, which may reflect natural data more closely. Findings in these studies, however, cannot only be attributed to the occurrence of FPs but rather to the combination of disfluencies.

We conclude that the phonetic details of FPs should be taken into account when testing the recall effect of FPs as well as the cognitive demands of the task. While we made the first step in that direction with the studies reported here, there is still a necessity to compare the recall effect of FPs using different lengths in a large-scale study. Furthermore, how FPs may vary in different contexts should be examined. It is possible that when producing lists, FPs are much longer than in semi-spontaneous speech (which the short retellings of FPs tried to model). The form of the FP should always be appropriate to the task in order to keep the stimuli as natural as possible.

In the interest of open science, granting access to the audio stimulus materials is important. Furthermore, we see great potential for FP research in expanding to other (also non-Germanic) languages. Providing a suitable experimental paradigm would be a key element for reaching this goal. While the list recall paradigm used here is easy enough to replicate, it did not show any recall effects for English. The story design includes a more natural set-up. However, the analysis of the recollection of each plot point may not be as straightforward as should be the case for an experimental design used for cross-linguistic comparisons. For the development of a cross-linguistic test design, further research is needed.

Part III

Discussion and Conclusion

Chapter 7

General discussion

Spontaneous speech rarely lacks disfluencies. Repairs, lengthenings, false starts, and filler particles (FPs) are often the result of difficulties in the planning process. As such, they act as features inserted into the speech stream, buying the speaker time to plan their upcoming utterances. FPs are one type of disfluency and the time-buying function seems to be their main function next to attention-seeking, turn-taking and yielding, and discourse structuring (Clark & Fox Tree, 2002; Shriberg, 1994; Maclay & Osgood, 1959; Goodwin, 1981). Even though there is a debate in the academic field on this topic (see Chapter 2.2.1), most researchers seem to agree on the symptom hypothesis of FPs which means that the phenomenon is a result of processing trouble and not used as a signal by speakers (O'Connell & Kowal, 2005; Corley & Stewart, 2008). Nevertheless, listeners have learned to interpret FPs and deduct information from them, e.g., a speaker is likely to refer to a low-frequency word or a discourse-new referent when producing FPs (Bosker et al., 2014; Arnold et al., 2003, 2004).

The present dissertation set out to examine the phonetic characteristics of FPs in spontaneous speech in different languages and explore the importance of FPs' phonetic composition in a series of recall experiments. As collecting spontaneous speech and annotating it for disfluencies is highly time-consuming, not many studies exist that investigate disfluencies on a large scale. Another large-scale investigation on disfluencies was conducted by Shriberg (1994) who used three corpora containing data from over 100 native English speakers. Studies on languages other than English are scarcer making the Pool2010 corpus (Jessen et al., 2005), which investigated the FPs of 100 male speakers, an important contribution to the field. It includes normal spontaneous speech as well as Lombard speech which had not been used to examine FPs before. After examining FPs, their pause context, and voice and vowel quality in detail in a large-scale corpus containing native German speech, this thesis reports analyses on three more languages, including English, Spanish, and Arabic. While work on FPs in Spanish is scarce, work on FPs in Arabic is non-existent and forms a research gap that this thesis alone cannot fully close.

FPs are reported to improve the recall of subsequent speech material (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2014), however, a focus on phonetic detail in the speech stimuli is often lacking. FPs in the stimuli may be unnaturally long (Fraundorf & Watson, 2014), or they may occur in combination with other disfluencies (Corley et al., 2007; Collard et al., 2008), which is more natural, but the resultant effect cannot be attributed to solely one of the phenomena. The aim of the recall experiments in Chapter 6 was to replicate the recall effect of FPs using ones that match natural FPs in spontaneous speech more closely and to examine whether their length affects the recall benefit.

Annotation scheme

In order to annotate the data according to our needs, we developed an annotation scheme in Praat (Boersma & Weenink, 2022) for the Pool2010 corpus (Jessen et al., 2005) which is presented in detail in Chapter 3. This scheme consists of five disfluency tiers and one optional one for annotating vowel tokens in lexical words from each speaker, such as the corner vowels, which frequently act as reference points for FP vowels. The five disfluency tiers are dedicated to silent pauses, respiration noises and laughter, FPs and tongue clicks, and disfluency phenomena such as repetitions, repairs, lengthenings, and truncations. As the focus of this work is on the phonetic characteristics of FPs, a tier was added to segment each FP into vowel and nasal consonants, but also to annotate creaky voice and glottal pulses within the FP. The aim of presenting a new annotation scheme for disfluencies was not to propose its widespread use in the disfluency community, but to present the details of the scheme, so colleagues can choose features that may be useful for their own research. In many cases, the annotation of all the phenomena in our scheme is not necessary for the research question, and thus too time-consuming and costly to include. A detailed report of the annotation guidelines is nevertheless essential in every study that involves speech data.

German

In the first study, the phonetic characteristics of FPs were investigated using a large corpus of semi-spontaneous speech including 100 male native German speakers in two conditions: a normal condition and a Lombard speech condition. The features investigated were the frequency distribution of typical FPs as well as glottal FPs and tongue clicks, the duration and pause context, and the voice and vowel quality. It was found that the phenomena that are often ignored in disfluency research, glottal FPs and tongue clicks, are equally as frequent as the typical FPs *uh, um,* and *hm* and should therefore be included in future disfluency research. When considering the frequency of these typical FPs for the German Pool2010 corpus (Jessen et al., 2005), the vocalic FP is used the most often, followed by the vocalic-nasal FP and the nasal FP, which is used rarely. The predominance of the vocalic FP is surprising

as Wieling et al. (2016) and de Leeuw (2007) report a preference for the vocalic-nasal FP in German and other Germanic languages. Wieling et al. (2016) argue that there has been a change from the vocalic FP to the vocalic-nasal FP, which has been led by women of the younger generation. This could be why we find a predominance of the vocalic FP, as the corpus was compiled in 2001 and includes only male speakers. The favour of a specific type of disfluency could also be related to the frequency of syllable types. As German slightly favours closed syllables, a predominance of the vocalic-nasal FP would be expected (Delattre & Olsen, 1969). It is possible, however, that the favour of closed syllables in German is not strong enough as the relation is only 60 % to 40 %. Furthermore, it becomes apparent that the pause context influences the duration of FPs in that FPs produced in isolation are the longest and FPs produced within speech are the shortest. FPs produced with pauses on only one side fall in between the two extremes, but IPU-final FPs are longer than IPU-initial FPs. Both types of typical FPs behave in the same way which is why we term this phenomenon a duration hierarchy. The same pattern could be found for the vocalic FP in Hungarian (Gósy & Silber-Varod, 2021). This may be due to the final lengthening phenomenon and/or the fact that the speaker needs more preparation time at the end of one utterance to plan the next than at the beginning of a new utterance that is potentially already planned (Lindblom, 1968). In terms of voice quality, creaky voice and glottal pulses are very frequent at the beginning of FPs, but quite rare at the end of FPs which confirms findings by Belz (2021) and Cataldo et al. (2019).

The vowel quality of the FPs that include a vowel portion, i.e., *uh* and *um*, largely overlaps between the two types and spreads over a large area of the central vowel space. This speaks for the hypothesis that FPs do not have a target. Any change in the vowel quality of an FP would not lead to a difference in meaning as would be the case for lexical vowels. Alternating the vowel quality in lexical items could lead to a different lexeme and could thus entail misunderstandings. This is not the case when producing FPs which explains the high variation of the FP vowel quality.

A comparison of the Lombard condition with the normal condition shows that the FP types *uh, um*, and *hm* decrease in frequency in the Lombard condition while the rate of tongue clicks increases. The explanation for this is only tentative and would assume the signal hypothesis rather than the symptom hypothesis (see Chapter 2.2.1). The signal hypothesis assumes that speakers use FPs intentionally to signal an upcoming delay to the listener. FPs under this hypothesis are therefore listener-oriented. The speaker's awareness of the listener in the Lombard condition, due to the increased noise, could be reduced to the degree that the speaker also decreases the rate of FPs. This would explain the reduced number of FPs but not the increased number of tongue clicks. The vowel quality of FPs differs in the sense that the vocal tract area used in the Lombard condition is significantly smaller than that in the normal condition. The reasons for this effect are speculative, but the higher muscle tension could play a role (Wohlert & Hammen, 2000). Additionally, an increase of the

first formant is detectable which can be explained by the increased jaw opening in the Lombard condition, which is brought about by the higher intensity (Van Summers et al., 1988).

Speaker-specificity was analysed by looking at a subset of 12 speakers from the Pool2010 corpus (Jessen et al., 2005). As the two conditions are of the mismatch type (Alexander et al., 2005), differences from one condition to the other can either stem from a large within-speaker variation or be due to the difference in conditions. The data investigated here, suggests that the between- and within-speaker variation is quite large. Taking only FPs as a disfluency feature in a forensic analysis into account will not be beneficial, based on the findings here. Other work, however, suggests that combining FPs with other disfluency features, such as lengthenings, repetitions, repairs, and silent pauses, could lead to a speaker-specific disfluency profile that could be added as a feature in a forensic phonetic case report (McDougall et al., 2019; Braun et al., 2023; Braun & Elsässer, 2023).

English and Spanish

Chapter 5.2 focused on two parameters of FPs, namely their frequency distribution and the vowel quality of the vocalic part of the particle. These parameters were examined using two languages other than German: English and Spanish. The subcorpus of the Diapix-FL corpus (Cooke et al., 2013) used here includes conversational speech from 20 women, 10 native English speakers and 10 native Spanish speakers. Participants produced speech in their native language as well as their second language, which was again either Spanish or English. Analyses of the frequency distribution of FPs in English and Spanish native speech reveal a language-specific preference for FP type. The English group prefers the vocalic-nasal FP *um* and the Spanish group prefers the vocalic FP *uh*. It is hypothesised that this result is connected to the preferred type of syllable structure in each language. Closed syllables are more common in English (Crystal & House, 1990) which may promote the production of the vocalic-nasal FP as a closed syllable rather than the open syllable in the vocalic FP. Spanish favours open syllables and restricts the coda in the word-final position to include only certain phonemes under which is the alveolar nasal consonant /n/ but not the bilabial nasal consonant /m/ (Gabriel, 2022). The syllable in the vocalic-nasal FP consisting of a vowel and the bilabial nasal is highly unnatural for native Spanish speakers and thus, the preference for the vocalic FP *uh* is expected. However, native Spanish speakers produce some vocalic-nasal FPs in their L1 and this rate increases when the group speaks English as their L2.

In accordance with findings by Cenoz (2000), the preferred FP type of each speaker group also transfers to their L2 even though the L2 proficiency is quite high for both groups. Increasing proficiency is expected to aid in eliminating a foreign accent in the L2, but it seems that the preferred type of FP is not affected (Flege, 1995). The vowel quality of all L1 FPs was compared to the vowel quality of lexical vowels from L1

speech, for the English and Spanish native speech separately. Results show that the vowel of the native English FP is a low central vowel similar to the English /ʌ/ while a close-mid unrounded front vowel /e/ is produced in Spanish. This vowel quality in FPs is quite unique as most languages report a central vowel quality close to [ə] or slightly more open, similar to [ɐ] as reported in Chapter 2.2. And if the vowel quality takes the form of a front vowel, the open-mid vowel [ɛ] seems more frequent (Belz, 2021; Cao & Mok, 2023). Thus, the use of the close-mid front vowel /e/ in FPs in native Spanish is unexpected.

As outlined in Chapter 2.2.2, it is suspected that the vowel inventory of a language affects the vowel quality of FPs but more specifically, the existence of a central vowel in the investigated language. The previously investigated languages in this work, German and English, both produce schwa regularly in unstressed syllables even though the status as its own phoneme is debated (Hirschfeld & Wallraff, 2002). Spanish does not have a central vowel, neither in its vowel inventory nor in unstressed syllables as it does not reduce syllables like German and English do (Gabriel, 2022). So why did the Spanish FP *uh* emerge as [eː] and not as [aː], [oː], or [uː]? The open vowel phoneme /a/ and the rounded close-mid back vowel phoneme /o/ are Spanish lexical words (meaning: the preposition *at* and the conjunction *or* in English) which leaves only /e/ and /u/ as possible candidates. The choice between the two is then influenced by the frequency of these two phonemes. Guirao & García Jurado (1990) report a frequency of 15 % for the front vowel /e/ and only 2.8 % for the rounded back vowel /u/ for American Spanish in a frequency count of all phonemes of the language. The accent of Spanish is not expected to have a major effect on the frequency of these two phonemes. The choice of vowel quality in Spanish FPs can be explained quite well by the vowel inventory and the non-existence of schwa, as well as the vowels' frequency distribution. It would be highly interesting and beneficial for the field to investigate the vowel quality of FPs in other languages that do not use schwa as a phoneme or an allophone.

Comparisons of L1 and L2 speech of native English and Spanish speakers show that while we see a clear-cut separation of English and Spanish FP vowel spaces in the speakers' native speech, vowel spaces in the second language overlap to a considerable degree. The L2 FP vowel quality of native English and Spanish speakers exhibits a high degree of overlap and a bimodal distribution for both speaker groups. This suggests that speakers approximate the vowel quality of their foreign language but are not fully able to adopt it. Similar trends have been shown in Muhlack (2020a). Analyses of speaker-specific differences show three types of speakers regarding their foreign FP adaptation. Most speakers approximate the foreign FP vowel quality but this group also shows a high degree of within-speaker differences. This is category 1: approximation with high variation. The other two categories can be viewed as the end points of the adaptation process: no adaptation and clear separation. Only a few speakers can be assigned to these two classes in the analysis of the Diapix-FL corpus (Cooke et al., 2013). Reasons for a clear separation of the FPs in the two languages

may be, besides L2 proficiency, a speaker's stance towards the foreign language, the regularity of interaction with native speakers of the language, the length of stay in a country in which the foreign language is spoken, media consumption in the L2, and teaching context such as learning the L2 in the country where it is spoken or learning it from a native speaker compared to learning it from a second language speaker. This information was not available for the current data set. However, it would be beneficial to include in future data sets that aim to investigate FP adaptation in L2 speech.

Advanced L2 learners seem to be able to perceive a difference between their native FP vowel quality and the FPs of their foreign language. This ability to perceive a difference between a native category and an L2 category is a prerequisite in producing the foreign category according to Flege's Speech Learning Model (SLM) (Flege, 1995). Even though the SLM is applied to phoneme categories of a native and a foreign language, this can also be applied to the FP vowel quality. It may be the case that speakers form a category for their native FP, and once they reach a certain proficiency in their second language, they establish an L2 FP category given that they perceive (consciously or subconsciously) a difference between the native FP and the FP in the target language. It is plausible that speakers refine this FP category with increasing proficiency, i.e., approach the vowel quality more closely or reduce the variation. Speakers who use their native FP vowel quality in their L2 are thought to have not formed an L2 category for the FP, and subsequently, use their L1 FP in their foreign language. This may be due to the perceptual similarity between the two types making the speaker unable to perceive an acoustic difference. Speakers that separate clearly between native and foreign L2 FPs are then expected to have perceived a difference between the two FPs and successfully formed different categories for the two FP vowels; consequently, the foreign FP category has been refined sufficiently so that it resembles that of a native speaker of the foreign language.

Arabic

A corpus (Ibrahim et al., 2020) of Egyptian Arabic speech by 7 speakers (4 male, 3 female) was subsequently used to examine the frequency distribution of several disfluency phenomena as well as the vowel quality of Arabic FPs (Chapter 5.3). The previous analyses on speaker-specificity in Chapter 4 revealed high within- and between-speaker variations regarding the frequency distribution of FPs and tongue clicks. Other literature suggests a speaker-specific disfluency profile beyond FPs (Braun & Rosin, 2015; McDougall & Duckworth, 2018), which is why the scope of investigated phenomena was expanded for this data set to also include silent pauses, repetitions, lengthenings, and lexical FPs. The corpus includes two spontaneous speech tasks which are hypothesised to differ in cognitive load. The first is a *Daily life task* in which participants talk about their everyday life, e.g., what they do on a regular day; the second task, a *Map task* requires speakers to describe the directions from a pop-

ular sight in the city to the university using a map as a visual aid. Results show that speakers are quite consistent in their disfluency pattern across the two tasks and that these speakers produce a much higher rate of silent pauses than any other disfluency phenomena. The most frequent disfluency phenomenon after the silent pause is the vocalic FP *uh*. This finding supports the hypothesis that the most frequent syllable type affects the preferred FP type. Syllable structures in Modern Standard Arabic are reported to occur in the following frequencies: CV in about 50 %, another open syllable type (CVV) about 17 %, and the CVC type only 24 %.

The comparison of the vowel quality of Arabic FPs and the Modern Standard Arabic corner vowels shows that the FP occupies the central vowel space centring around the schwa position. Vowel reduction plays an important role in Arabic phonology as each vowel can be reduced in unstressed positions, then taking the form of a schwa (Embarki, 2013). This can also be seen in our data in the high variation across the vowel space of the corner vowels. Due to the high degree of vowel reduction in Arabic, schwa as an allophone is very familiar to speakers of Modern Standard Arabic and also quite frequent. The hypothesis that the existence of schwa in a language influences the vowel quality of the FP can be confirmed for this data set of Modern Standard Arabic.

Investigations into the vowel quality across the reported languages reveal that the Arabic FPs overlap to a high degree with the German FPs. English overlaps partially with both these data sets but more so with the Arabic FP vowel quality. The Spanish FPs show a quite unique vowel quality compared to the other three languages. They overlap only minimally with the Arabic data set. The position of the Spanish FP vowel in the vowel quadrilateral is not central as is the case for German, English, and Arabic FPs; it takes the form of an unrounded close-mid front vowel which overlaps to a high degree with the Spanish phoneme /e/. The reasons for this uniqueness have been outlined above.

Recall

While Chapters 4 and 5 presented analyses of the production data, Chapter 6 examined a hypothesised function of FPs. Several studies suggest that FPs have a beneficial effect on the recall of information that follows an FP (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2011a). The studies were conducted in English and studies in other languages are sparse and do not support the reported effect (Chen et al., 2022a; Bosker et al., 2014). Psycholinguistic studies often miss out on reporting how the phonetic details of their stimuli, e.g., the duration of FPs and their pause context in this case, could influence the effect. Corley et al. (2007) and Collard et al. (2008) use FPs that occur near lengthenings and do not report the duration of the particle. On the other hand, Fraundorf & Watson (2011a) report the duration of the particle but this duration ($>$ 1s) exceeds the duration of FPs observed in natural data by far. The salience of the FP in the stimuli of Fraundorf & Watson (2011a)

could have affected the recall of participants to a higher degree as the FP length was quite unnatural. This led us to conduct a series of experiments examining the recall effect in combination with the length of the FP in the stimuli. Experiment 1a was a partial replication study of Fraundorf & Watson (2011a) conducted in German with a fluent condition, a long silence condition, and a filler particle condition. Our FPs in the stimuli stories matched naturally observed FPs more closely, their duration ranged from 480-730 ms. Participants listened to three short stories from *Alice in Wonderland* (Carroll, 1916), each in a different condition, and it was hypothesised that the important plot points from the story including FPs would be better recalled by participants than after the fluent condition, as was shown in Fraundorf & Watson (2011a). Results of Experiment 1a showed no such effect, but the contrary. The fluent condition was significantly better recalled than the long silence and FP condition. A marginal effect even suggests that the long silence condition was better recalled than the FP condition, which was unexpected. As this experiment was performed in a different language than the original and, due to the COVID-19 pandemic, conducted online with less control over the participants' equipment and concentration level, we decided to take a step back and eliminate the language difference.

Experiment 1b was then conducted in English, and thanks to Scott Fraundorf and Duane Watson, we were able to work with the exact stimuli from the original study. The question for this experiment was then whether we could replicate the reported recall effect in a web-based setup. Furthermore, we hoped to explore the influence of the FP duration. We did this by taking the fluent condition and the FP condition from the original experiment and manipulating the FP condition (= long FP condition) to create a short FP condition. The procedure and analyses of this experiment remained the same as in the previous experiment with a different group of participants. The results of this experiment showed no effect of condition at all, neither a beneficial effect of FPs nor an inhibitory effect. Visual examination of the results showed that the long FP condition tended to be as well recalled as the fluent condition. But again, this effect was not significant. It may be possible, as the size of the recall effect was small, that an increase in participants would lead to the tendency reaching significance (Diachek & Brown-Schmidt, 2022) which would suggest that the duration of FPs has an effect on the recall effect. However, as the difference is not significant, no such influence can be confirmed.

As Experiments 1a and 1b failed to replicate the reported recall effect of FPs using the story paradigm, we created a new simpler paradigm aimed at finding the recall effect. It was hypothesised that if FPs have a beneficial effect on the recollection of the material that follows one, this should also show in a simple recollection of lists. In Experiment 2, participants heard six lists of different topics (e.g., zoo animals, body parts, vegetables, etc.) and three of these lists contained two FPs before two items. The participants were asked after each auditory presentation of the list to recall as many items as possible by writing them down in an answer box. Results showed that the position of the items in the list and the general memory score of the participant,

as determined by a digit span test, influenced the recollection of items in the list but not the occurrence of an FP in the list or before the target item.

However, this new list paradigm also failed to replicate the beneficial recall effect of FPs. Reasons for this could be the difference in task between our experiments and the ones conducted in Corley et al. (2007); Collard et al. (2008) and Diachek & Brown-Schmidt (2022). These experiments asked participants, after an initial EEG experiment, to identify the words that occurred in the previous experiment, which possibly placed different cognitive demands on the participants than recalling a short story or a twelve-item list. The story paradigm requires participants to understand the narrative globally and possibly calls for a better long-term memory (Burkhardt, 2007), whereas the list paradigm requires a better short-term memory and/or organisational strategies such as chunking or mnemonic techniques (Miller, 1956; Massen & Vaterrodt-Plünnecke, 2006). However, this does not explain why neither of our experiments showed a recall effect while Fraundorf & Watson (2011a) found one. Another possibility for the discrepancy is the mode of the experiment. The original experiment tested participants in the lab while our experiments were conducted as web-based experiments. This comes with the caveat that we had less control over participants' equipment, concentration level, and general surroundings. Web-based methods have been reported to lead to a higher variation in participants' answers which could affect the outcome of the study (Gallant & Libben, 2019).

Another limitation of our recall experiments is that we failed to calculate their power based on the effect size from previous studies. It may be the case that, due to the small effect size, a large number of participants is needed to find the investigated effect. Diachek & Brown-Schmidt (2022) recently investigated the recall effect using a similar paradigm as Corley et al. (2007) and Collard et al. (2008), and they included power analyses for each of their experiments. They conducted four experiments with 110, 161, 200, and 293 participants, respectively, basing each sample size on the previous experiment and a new power analysis, using Corley et al. (2007) as a starting point. They were able to conclude that different forms of disfluencies (pauses, FPs, repetitions) do have a beneficial effect on the recollection of spoken material and that this effect is position-dependent and only occurs at the end of sentences (Diachek & Brown-Schmidt, 2022). The large sample size of Diachek & Brown-Schmidt's (2022) experiments suggests that our recall experiments were highly underpowered with each sample size coming short of 100 participants (45, 58, and 73 participants) The danger of an underpowered experiment is that any obtained results are unreliable and cannot be taken at face value. Especially when the effect size is so small, as was the case with the recall effect; finding the effect with a small sample size (i.e., too few participants) is very unlikely. Thus, any further experiments examining this effect, possibly in combination with the phonetic composition of FPs, which is still a research gap, will need to conduct power analyses beforehand and then recruit the required number of participants.

Limitations and future research

In the previous section, the results of the empirical studies were summarised and put into the context of disfluency research. Due to the scope of this work, some aspects were omitted from the studies which can be considered in future research. A brief discussion of the limitations of this work will be given in this section, along with an outlook on future research topics that logically arise from the results and limitations of the work presented here.

One aspect of FPs has been omitted intentionally in Chapters 4 and 5 from the start, which is the function of FPs. As assigning function may be a highly subjective matter and closely connected with the phonetic characteristics of FPs, this part was not considered for this work but is an important topic to be considered. The danger here is that the annotators assigning a function to the FPs could be influenced by the FP type, their duration, or other characteristics which would make the analysis of FPs and their function problematic. A way of escaping this problem would be a context analysis in which the FP is omitted by another researcher first. In this way, the annotators assign the function to the context, i.e., which function arises from the utterance itself without being influenced by the FPs. To the best of my knowledge, this has not been done before. Usually, multiple annotators are consulted for annotating the function, and their scores are then compared in an annotator agreement analysis. This must also be done in a context analysis to verify the method, and if this shows a particularly low annotator agreement, we may conclude that the function is assigned mainly based on the FPs themselves.

Another aspect not considered in this thesis is lexical FPs. However, languages and, especially, individuals could differ in their preference for non-lexical and lexical FPs. Widening the scope of this research to include lexical FPs would benefit the discussion of language- and speaker-specificity, e.g., Laserna et al. (2014) suggest that the preference and disfavouring of FPs (*uh, um*) and discourse markers (*I mean, you know, like*) may give information about the social class of a speaker or their personality.

In Chapter 4, a large corpus of native German speech was analysed, but this corpus only consisted of male speakers. As this corpus, the Pool2010 corpus (Jessen et al., 2005), was compiled as a reference corpus for forensic phonetic casework, i.e., to offer insights into the typicality of a feature, female speakers were not considered. A direct comparison of male and female speakers would, nevertheless, be necessary to confirm the suggestion of Wieling et al. (2016) that females prefer the vocalic-nasal type and thus, lead the language change from *uh* to *um*.

As suggested in previous literature (Braun & Rosin, 2015; McDougall & Duckworth, 2018; Braun et al., 2023; Braun & Elsässer, 2023), FPs alone may not be speaker-specific, but several disfluencies together may show speaker-specific patterns, as also observed in the Arabic corpus in Chapter 5. Different disfluencies such as repairs, repetitions, truncations, and lengthenings have been annotated in the Pool2010

corpus, but analyses including these features have not been conducted, yet. It would be highly beneficial to apply a statistical analysis, to investigate speaker-specificity, to the Pool2010 corpus data (Jessen et al., 2005) such as presented in Braun et al. (2023). They showed that eight speakers were differentiated based on their disfluency profile above chance.

Due to the lack of a full orthographic transcription of the Pool2010 corpus (Jessen et al., 2005), an analysis of the syntactic context of FPs has not been conducted. A detailed analysis of FPs in different dialogue turns and intonation phrases, however, can be found in Belz (2021).

The analysis of tongue clicks in Chapter 4 was limited to their frequency distribution. Further analyses could involve the analysis of context, e.g., whether they only occur at the beginning of utterances as an artefact of speech preparation gestures, such as swallowing and mouth opening, or whether they frequently occur in combination with respiration noises. Further analyses could determine the place of articulation and the differentiation of proper tongue clicks and percussives.

In Chapter 5, the frequency distribution and the vowel quality of the FPs were analysed. The first part dealt with these aspects in English and Spanish L1 and L2 speech. Further analyses using the data could include examining the pause context as reported in Chapter 4. However, direct comparisons of the Diapix-FL corpus (Cooke et al., 2013) and the German Pool2010 corpus (Jessen et al., 2005) could not be drawn, as the former corpus includes dialogue data and the latter includes (rather) monologic data which could especially affect the frequency and duration of pauses surrounding the FPs. It is hypothesised that dialogues inhibit the production of (long) pauses as the risk could be to lose one's turn when the interlocutor starts to speak (Maclay & Osgood, 1959; Clark & Fox Tree, 2002). A limitation of the data in the Diapix-FL corpus is the limited information available for the participants. Participants are roughly at the same L2 proficiency level, but they did not take a controlled test of L2 proficiency which would give a more detailed insight into the variation of this factor within the group. Proficiency is expected to influence the ability to adapt a native-like disfluency pattern and assimilate to the foreign FP vowel quality. Further information about the second language history would also be beneficial, such as the age of acquisition in which the language was learned and the duration of a possible extended stay in the country where the L2 is spoken. An open question in second language research is whether teaching students the foreign disfluency pattern (e.g., the preferred type and vowel quality of FPs) would be beneficial for acquiring a native-like accent in the L2.

A small corpus of Egyptian Arabic speech was analysed in terms of the frequency distribution of several disfluency phenomena and the vowel quality of the FP. Due to the small corpus size, reliable generalisations cannot be drawn for the population of Egyptian Arabic speakers. In order to be able to do this, a larger corpus is needed to verify the tendencies shown here, e.g., the speaker-specific disfluency patterns and the central FP vowel quality. Future corpora would benefit from a balanced set of

women and men as well as controlled collections of lexical vowels, as vowel reduction is frequent in Egyptian Arabic (Embarki, 2013). A word list could be compiled including the vowels of the language in a controlled phonetic context which then could be extracted as a reference for the FP vowel. Vowel tokens should be checked for possible vowel reductions which could then be either excluded or taken to form another category to be compared to the FP vowel. In the current study, lexical vowel tokens were taken from (rather fast) spontaneous speech, so the vowel tokens were very short which may have led to a higher degree of vowel reduction.

The cross-linguistic comparisons of FP vowels would furthermore benefit from a controlled collection of the phonemic vowel inventory of each language and the inclusion of any allophonic central vowels in the data. Comparison can then be made between the FP vowel and each vowel phoneme of the language. Furthermore, this full vowel inventory can then be used to normalise formant values per language or speaker. Whether different vowel inventories have an effect on vowel normalisation procedures is another topic for future research.

Chapter 6 looked at one hypothesised function of FPs, which is their beneficial recall effect. In three experiments, we were unable to replicate the effect which, as outlined above, may be due to the small sample size and, with it, the lack of experimental power. Future studies would benefit from a power analysis before gathering participants which would give insight into the sample size needed to find the investigated effect. A research gap, that we attempted to close, is the importance of the phonetic characteristics of FPs in combination with the recall effect. Controlled experimental studies are needed to examine whether longer FPs (> 1) have a different effect on the recollection of subsequent speech material than shorter FPs (approx. 500 ms).

Chapter 8

Conclusion

This thesis is situated in the area of disfluencies, focusing on the phonetic characteristics of filler particles (FPs), such as *uh* and *um*. While this phenomenon has received more attention in the last decade, detailed phonetic descriptions of FPs in different languages and cross-linguistic comparisons are sparse.

For the purpose of this study, a new annotation scheme has been developed that uses Praat TextGrids (Boersma & Weenink, 2022) as an annotation tool. With this scheme, not only can FPs be annotated, but also respiration noises and other disfluencies, such as repairs and truncations. When analysing the function of FPs, it is proposed to conduct a context analysis without being presented with the form of the FP, as this may lead to circularity of the analysis. The form of the FP may influence the annotator's preference for a specific function for the FP in question.

This work further provides a thorough analysis of FPs (including glottal FPs and tongue clicks) in a large corpus of German, investigating five phonetic features: frequency distribution, duration, pause context, voice quality, and vowel quality. Investigating these features in combination, using a corpus with 100 speakers, and taking two speech conditions– normal speech and Lombard speech– into account has not been done before. It is apparent that the duration of FPs is dependent on the pause context it appears in and that this applies to vocalic and vocalic-nasal FPs to the same extent. FPs that occur within two silent pauses are the longest, followed by FPs that occur utterance-finally, and FPs that occur utterance-initially. FPs that are produced within an utterance, which are more often vocalic rather than vocalic-nasal FPs, are the shortest FP type. This pattern can be called a duration hierarchy.

More surprising are the findings that the frequency of clicks increases in the Lombard condition while the frequency of the typical FPs (*uh, um, hm*) decreases in the Lombard condition. Explanations of this effect are speculative and a satisfying answer cannot be reached. A reason for the decrease in typical FPs could be the listener orientation of FPs, as, in the Lombard condition, participants were presented with white noise over headphones. This may have influenced the awareness of a listener

by the speaker which is why they (unconsciously) produced less typical FPs. The listener orientation of FPs, however, is highly debated in the disfluency community (O'Connell & Kowal, 2005; Corley & Stewart, 2008). The increase in clicks may be explained by the raised muscle tension in the condition. Lombard speech is characterised by a raised intensity which results from a higher vocal effort and thus higher muscle tension (Jessen et al., 2005; Wohlert & Hammen, 2000). This higher muscle tension may facilitate the production of clicks as they are produced by pressing the tongue to the palate which forms a small pocket of air (Clark et al., 2007). The release then produces an audible click noise.

Another unexpected finding is associated with the minimal production of creaky voiced portions within the FPs in the Lombard condition. As was just established, the Lombard condition is produced with higher muscle tension. There are two different types of creaky voice, one produced with lower subglottal pressure and one with higher subglottal pressure (Keating et al., 2015). As we encounter higher muscle tension in Lombard speech we would expect a higher amount of (constricted) creaky voice in this speech condition. However, creaky voice within the FPs in this condition decreases compared to the normal condition which does not adhere to the expectations. Furthermore, glottal FPs were investigated, which are FPs solely produced with creaky voice, i.e., they are sequences of glottal pulses. The Lombard condition does not have an effect on the frequency of these FPs, even though we would expect more glottal FPs with higher muscle tension and the constricted creaky voice type or fewer glottal FPs with the unconstricted creaky voice type.

Speaker-specificity of the phonetic features of FPs was investigated by comparing within-speaker variations to between-speaker variations, e.g., by comparing standard deviations of the features. It seems that both variations are quite large and that speakers do not exhibit a speaker-specific use of FPs, e.g., in their duration or production of creaky voice. However, the pattern of the disfluency features investigated (clicks, glottal FPs, and typical FPs) seems to show some speaker-specificity, and it may be worth investigating this using the full speaker set.

Another aspect of this thesis is the cross-linguistic comparison of the frequency distribution and vowel quality of FPs using a small set of languages and the hypotheses that can be formed from these results. While work on FPs in English and German, and to some extent Spanish, has been conducted before (e.g., Shriberg, 1994; Clark & Fox Tree, 2002; de Leeuw, 2007; Pätzold & Simpson, 1995; Belz, 2021; Erker & Bruso, 2017; Garciá-Amaya & Lang, 2020), disfluency research on Semitic languages, and Arabic, specifically, has been neglected at large. This work contributes to the filler particle research by comparing the preferred FP types of different languages and by investigating the vowel quality of FPs with the vowel inventory of each language. It was found that native English speakers prefer the vocalic-nasal FP *um* while native Spanish speakers prefer the vocalic type *uh* in both their native as well as a foreign language. The vowel quality was found to be a mid-central quality in German, English, and Arabic, only Spanish speakers do not produce a central vowel but a front

vowel in their FPs. Two hypotheses were formed that aim at predicting the preferred type of FPs and the vowel quality used in other languages. 1) The preferred type of FPs can be predicted by the most frequent syllable type of a language, i.e., either open (= *uh*) or closed (= *um*) syllables. 2) Languages that include a mid-central vowel (schwa) in their vowel inventory, either as a phoneme or allophone, use this vowel quality in their FPs. The question then emerges of which vowel quality is used when the language in question does not include a mid-central vowel in their vowel inventory. Spanish, as one candidate of a language that does not have a vowel similar to schwa in their inventory, uses the close-mid front vowel /e/ in their FPs. This vowel also happens to be the most frequent vowel phoneme in the language (Guirao & García Jurado, 1990). The systematic comparison of FP vowels with the corner vowels taken from lexical vowels by the same speaker is another benefit of this thesis. To shed more light on the relationship between schwa and the FP vowels, it is necessary to investigate more languages and categorise them as languages that include schwa and those that do not.

The last aspect considered in this thesis is a hypothesised function of FPs, namely the beneficial recall effect. Some studies in English have found that information that is preceded by an FP is more easily recalled than information that is not preceded by an FP (Corley et al., 2007; Collard et al., 2008; Fraundorf & Watson, 2011b; Diachek & Brown-Schmidt, 2022). The experimental series conducted here could not confirm this effect, neither in English nor German which shows that the effect may be more fragile than expected. This failure may be due to the low sample size of the studies and thus a low experimental power and/or because the experiments were conducted online and participant engagement could not be controlled. The experiments were designed to shed more light on the influence of the phonetic characteristics of the FPs on the recall effect, i.e., does the recall effect only occur with particularly long FPs or is it stronger when the FP is more salient? These questions could not be answered but introducing the phonetic make-up of FPs as an influencing factor into the discourse of the recall effect was an important step in advancing the investigation of said effect.

This work adds to the field of disfluency research by providing a new annotation scheme, which is suitable for annotating several disfluency phenomena. Furthermore, a detailed description of FPs in German is given, and a cross-linguistic comparison of FPs in some languages is enclosed. The results open up new lines of research, investigating FPs in combination with the syllable type and the existence of schwa in a language, which may shed more light on the predictability of the form of FPs. The influence of the form of FPs on a specific function and the recall effect, could not be deduced in this thesis, but including the area of phonetics in cognitive research is advantageous for future studies investigating the recall effect. In sum, this thesis advances our understanding of FPs and paves the way for future research in the area of disfluencies.

Appendices

Appendix A

Frequency of features per speaker

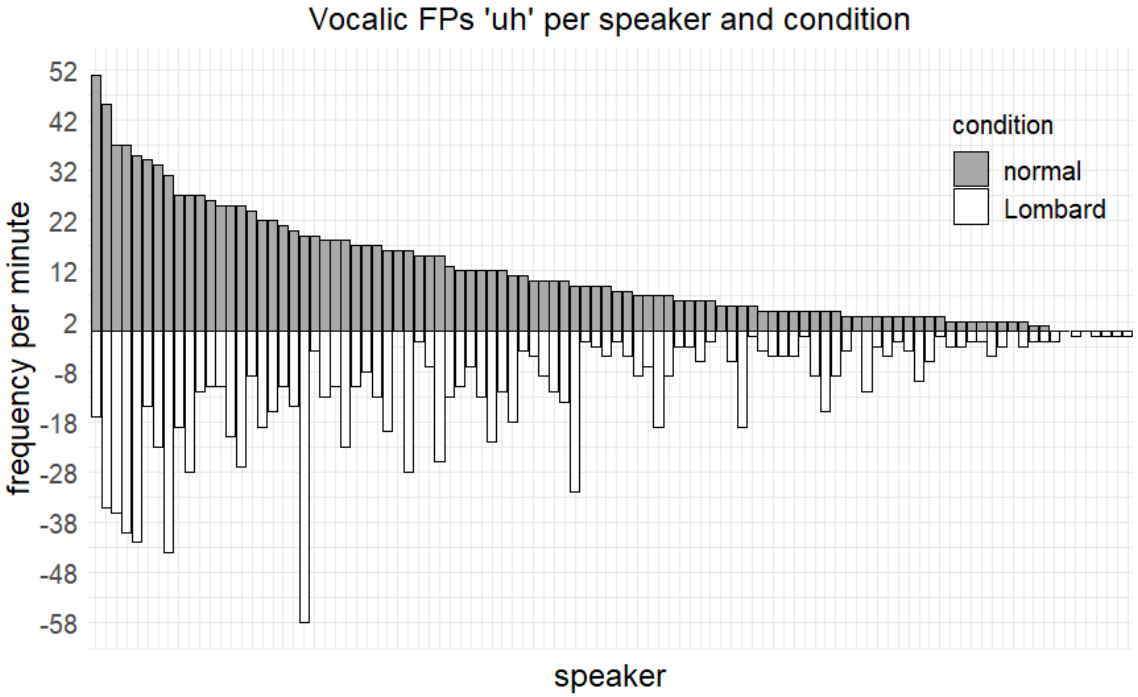

Figure 1: Bar plot per speaker for the vocalic FP
Frequency per minute of the vocalic FPs uh *per speaker as a function of condition. Negative values for the Lombard condition have to be understood as positive values.*

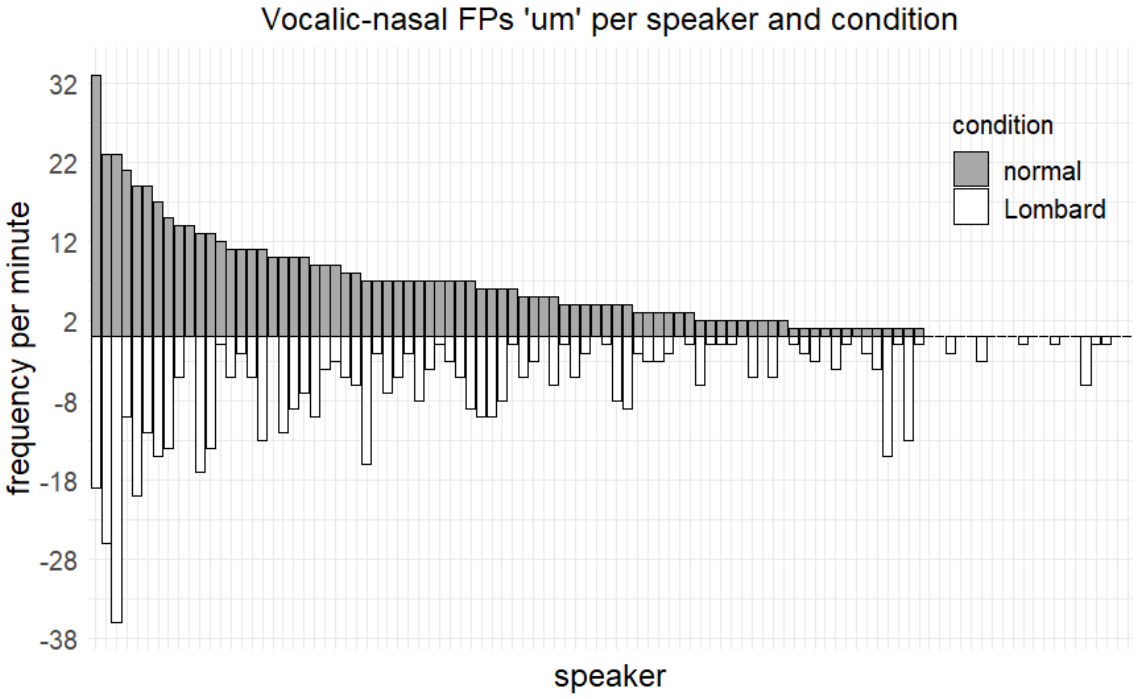

Figure 2: Bar plot per speaker for the vocalic-nasal FP
Frequency per minute of the vocalic-nasal FPs um *per speaker as a function of condition. Negative values for the Lombard condition have to be understood as positive values.*

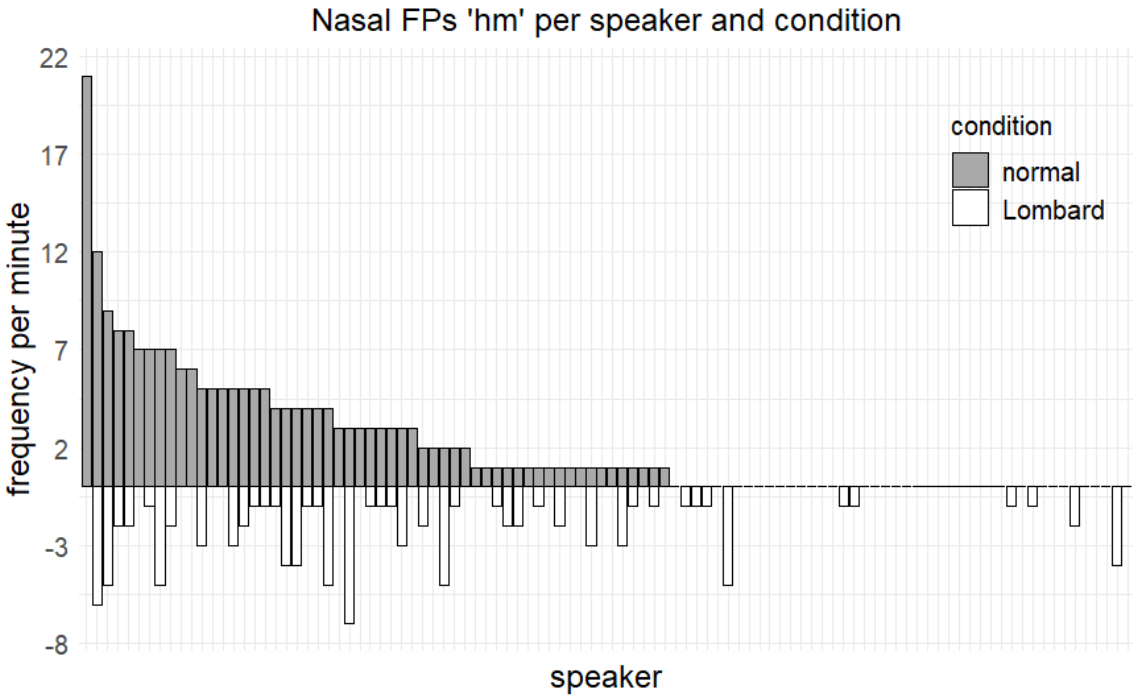

Figure 3: Bar plot per speaker for the nasal FP
Frequency per minute of the nasal FPs hm *per speaker as a function of condition. Negative values for the Lombard condition have to be understood as positive values.*

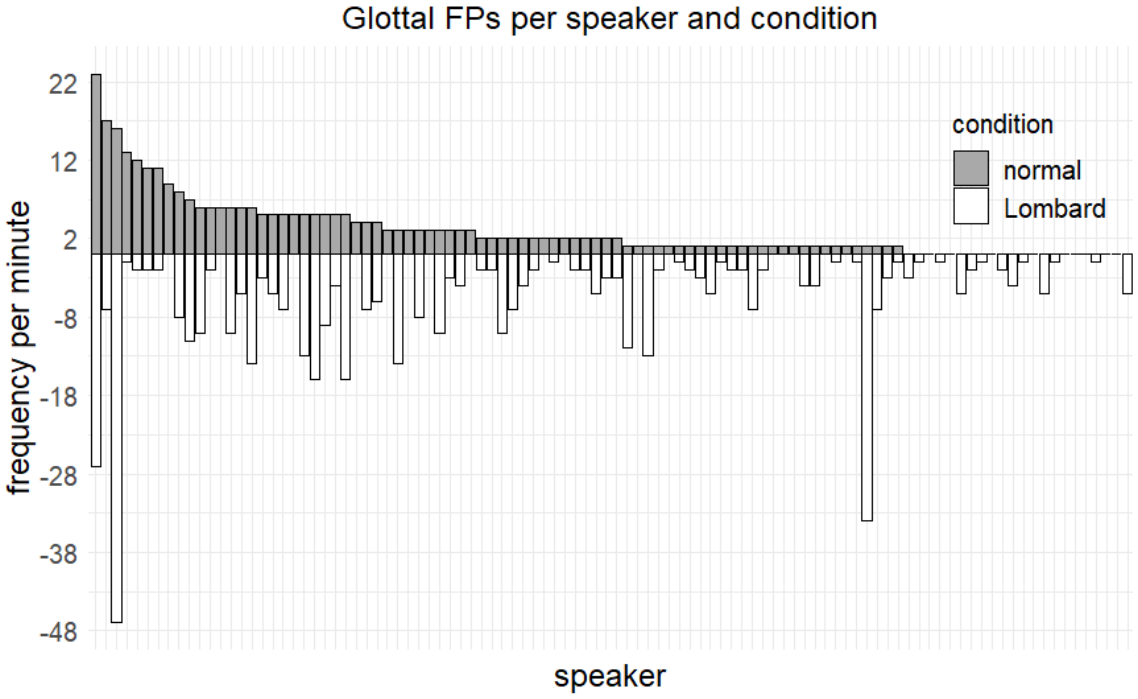

Figure 4: Bar plot per speaker for the glottal FP
Frequency per minute of the glottal FPs per speaker as a function of condition. Negative values for the Lombard condition have to be understood as positive values.

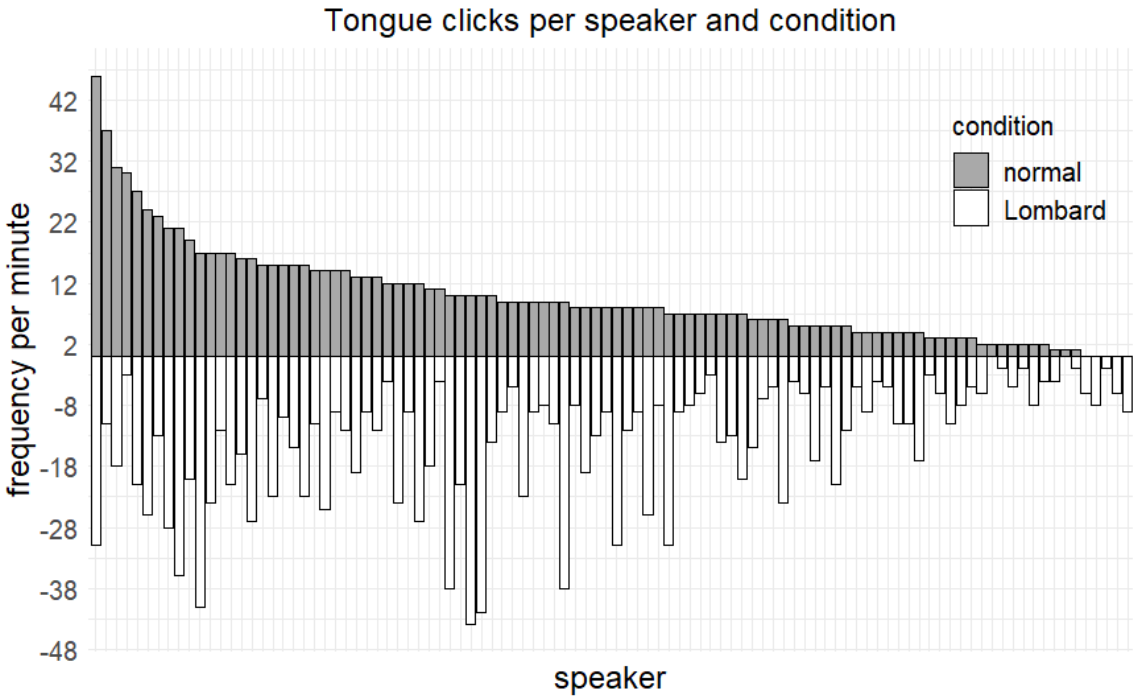

Figure 5: Bar plot per speaker for tongue clicks
Frequency per minute of the tongue clicks per speaker as a function of condition. Negative values for the Lombard condition have to be understood as positive values.

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
CI	confidence interval
EEG	Electroencephalography
FP	filler particle
f0	fundamental frequency
F1	first formant
F2	second formant
F3	third formant
GLMM	Generalised Linear Mixed Model
Hz	Hertz
IPU	inter-pausal unit
LADO	language analysis for the determination of origin
LM	Linear Model
LMM	Linear Mixed Model
L1	first/native language
L2	second/foreign language
SLM	Speech Learning Model
TTS	text-to-speech

Bibliography

Adams, M. R., & Hutchinson, J. (1974). The effects of three levels of auditory masking on selected vocal characteristics and the frequency of disfluency of adult stutterers. *Journal of Speech and Hearing*, *17*, 682–688.

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*, 3099–3107. URL: `https://doi.org/10.1121/1.1795335`.

Adell, J., Bonafonte, A., & Escudero, D. (2010). Synthesis of filled pauses based on a disfluent speech model. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. URL: `https://doi.org/10.1109/ICASSP.2010.5495136`.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers by Hirotugu Akaike* (pp. 199–213). New York: Springer.

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech, Language and the Law*, *12*, 214–234. URL: `https://doi.org/10.1558/sll.2005.12.2.214`.

Andreeva, B., Dimitrova, S., Gabriel, C., & Grünke, J. (2019). The intonation of Bulgarian Judeo-Spanish spontaneous speech. *International Congress of Phonetic Sciences (ICPhS), Melbourne*, (pp. 3827–3831).

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, *32*, 25–36. URL: `https://doi.org/10.1023/A:1021980931292`.

Arnold, J. E., Kam, C. L., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during refer-

ence comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*, 914–930. URL: https://doi.org/10.1037/0278-7393.33.5.914.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*, 578–582. URL: https://doi.org/10.1111/j.0956-7976.2004.00723.x.

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*, 441–455. URL: https://doi.org/10.1080/01690960903047122.

Barry, W. J. (1995). Schwa vs. schwa + /r/ in German. *Phonetica*, *52*, 228–235. URL: https://doi.org/10.1159/000262175.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*. URL: https://doi.org/10.18637/jss.v067.i01. arXiv:1406.5823.

Bates, S. A. R. (1995). *Towards a definition of schwa: an acoustic investigation of vowel reduction in English*. Ph.D. thesis University of Edinburgh.

Bateson, M. C. (2003). *Arabic Language Handbook*. Georgetown: Washington University Press.

Batliner, A., Kießling, A., Burger, S., & Nöth, E. (1995). Filled pauses in spontaneous speech. *International Congress of Phonetic Sciences (ICPhS), Stockholm*, (pp. 472–475).

Beattie, G. W., & Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*, 201–211.

Bellinghausen, C., Betz, S., Zahner, K., Sasdrich, A., Schröer, M., & Schröder, B. (2019). Disfluencies in German adult-and infant-directed speech. *SEFOS: 1st International Seminar on the Foundations of Speech. Breathing, Pausing and the Voice*, (pp. 44–46).

Belz, M. (2017). Glottal filled pauses in German. *Workshop on Disfluency in Spontaneous Speech (DiSS 2017), Stockholm*, (pp. 5–8).

Belz, M. (2018). Vowel quality of German äh and ähm in dialogue moves. *Phonetik und Phonologie im deutschsprachigen Raum (PP14), Wien*, (pp. 13–17).

Belz, M. (2019a). Dokumentation der Forschungsdaten GECO-FP v1: Phonetische Füllpartikelannotation basierend auf GECO v1.1.

Belz, M. (2019b). *GECO-FP*. Humboldt-Universität zu Berlin. URL: https://doi.org/10.18452/20794.

Belz, M. (2021). *Die Phonetik von äh und ähm: Akustische Variation von Füllpartikeln im Deutschen*. Berlin: Metzler. URL: https://doi.org/10.1007/978-3-662-62812-6.

Belz, M. (2023). Defining filler particles: A phonetic account of the terminology, form, and grammatical classification of "filled pauses". *Languages*, *8*. URL: https://doi.org/10.3390/languages8010057.

Belz, M., & Klapi, M. (2013). Pauses following fillers in L1 and L2 German Map Task Dialogues. *Workshop on Disfluency in Spontaneous Speech (DiSS 2013), Stockholm*, (pp. 9–12).

Belz, M., & Mooshammer, C. (2020). *Berlin Dialogue Corpus (BeDiaCo). (Version 1)*. Humboldt-Universität zu Berlin. URL: https://rs.cms.hu-berlin.de/phon.

Belz, M., & Reichel, U. D. (2015). Pitch Characteristics of Filled Pauses. *Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015)*, (pp. 1–4).

Belz, M., Sauer, S., Lüdeling, A., & Mooshammer, C. (2017). Fluently disfluent?: Pauses and repairs of advanced learners and native speakers of German. *International Journal of Learner Corpus Research*, *3*, 118–148. URL: https://doi.org/10.1075/ijlcr.3.2.02bel.

Belz, M., & Trouvain, J. (2019). Are 'silent' pauses always silent? *International Congress of Phonetic Sciences (ICPhS), Melbourne*, (pp. 2744–2748).

Betz, S. (2020). *Hesitations in Spoken Dialogue Systems*. Ph.D. thesis.

Betz, S., Vobe, J., Zarrieb, S., & Wagner, P. (2017). Increasing recall of lengthening detection via semi-automatic classification. *Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm*, (pp. 1084–1088). URL: https://doi.org/10.21437/Interspeech.2017-1528.

de Boer, M. M., & Heeren, W. F. L. (2020). Cross-linguistic filled pause realization: The acoustics of uh and um in native Dutch and non-native English. *Journal of the Acoustical Society of America*, *148*, 3612–3622. URL: https://doi.org/10.1121/10.0002871.

Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer. URL: http://www.praat.org.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147.

Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal*

of Memory and Language, *75*, 104–116. URL: https://doi.org/10.1016/j.jml.2014.05.004.

Böttcher, M., & Zellers, M. (2023). Hesitating with and without language heritage - Prosodic aspects of filler particles in the RUEG corpus. In *International Congress of Phonetic Sciences (ICPhS), Prague*.

Brand, C., & Götz, S. (2011). Fluency versus accuracy in advanced spoken learner language. *International Journal of Corpus Linguistics*, *16*, 255–275. URL: https://doi.org/10.1075/ijcl.16.2.05bra.

Braun, A., & Elsässer, N. (2023). Are there individual disfluency patterns? In *International Congress of Phonetic Sciences (ICPhS), Prague*.

Braun, A., Elsässer, N., & Willems, L. (2023). Disfluencies Revisited — Are They Speaker-Specific? *Languages*, *8*, 155. URL: https://doi.org/10.3390/languages8030155.

Braun, A., & Rosin, A. (2015). On the speaker-specificity of hesitation markers. In *International Congress of Phonetic Sciences (ICPhS), Glasgow*.

Brennan, S. E., & Schober, M. F. (2001). How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, *44*, 274–296. URL: https://doi.org/10.1006/jmla.2000.2753.

Brennan, S. E., & Williams, M. (1995). The feeling of anothers knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. URL: https://doi.org/10.1006/jmla.1995.1017.

Bühler, J. C., Schmid, S., & Maurer, U. (2017). Influence of dialect use on speech perception: A mismatch negativity study. *Language, Cognition and Neuroscience*, *32*, 757–775. URL: https://doi.org/10.1080/23273798.2016.1272704.

Bühler, K. (1965). *Sprachtheorie: Die Darstellungsfunktion der Sprache*. (2nd ed.). Stuttgart: Gustav Fischer Verlag.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581. URL: https://doi.org/10.1037/0033-295X.106.3.551.

Burkhardt, P. (2007). The P600 reflects cost of new information in discourse memory. *NeuroReport*, *18*, 1851–1854.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research*, *4*, 75–87. URL: https://doi.org/10.1007/BF01066991.

Candea, M., Vasilescu, I., & Adda-Decker, M. (2008). Inter- and intra-language acoustic analysis of autonomous fillers. In *Workshop on Disfluency in Spontaneous Speech Workshop (DiSS 2008), Aix-en-Provence* (pp. 47–52).

Cao, G. W., & Mok, P. (2023). The acoustics of cross-linguistic filled pauses in Cantonese-English-Mandarin trilingual speech. In *International Congress of Phonetic Sciences (ICPhS), Prague*.

Carroll, L. (1916). *Alice's Adventures in Wonderland*. (Rev. ed. ed.). New York: Sam'l Gabriel Sons Company. URL: `https://www.gutenberg.org/files/11/11-h/11-h.htm`.

Cataldo, V., Schettino, L., Savy, R., Poggi, I., Origlia, A., Ansani, A., Sessa, I., & Chiera, A. (2019). Phonetic and functional features of pauses, and concurrent gestures, in tourist guides' speech. *Audio Archives at the Crossroads of Speech Sciences, Digital Humanities and Digital Heritage*, *6*, 205–231.

Cenoz, J. (2000). Pauses and hesitation phenomena in second language production. *ITL - International Journal of Applied Linguistics*, *127-128*, 53–69. URL: `https://doi.org/10.1075/itl.127-128.03cen`.

Cevasco, J., & van den Broek, P. (2016). The effect of filled pauses on the processing of the surface form and the establishment of causal connections during the comprehension of spoken expository discourse. *Cognitive Processing*, *17*, 185–194. URL: `https://doi.org/10.1007/s10339-016-0755-8`.

Chen, X., Liesenfeld, A. M., Li, S., & Yao, Y. (2022a). Effects of disfluent machine speech on memory recall in human-machine interaction. In S. Warchhold, D. Duran, I. Gessinger, & E. Raveh (Eds.), *Human Perspectives on Spoken Human-Machine Interaction* (pp. 52–57). Freiburg i. Breigau: Freiburg Institute for Advanced Studies. URL: `https://doi.org/10.6094/UNIFR/223818`.

Chen, X., Liesenfeld, A. M., Li, S., & Yao, Y. (2022b). Effects of filled pauses on memory recall in human-robot interaction in Mandarin Chinese. *Engineering Psychology and Cognitive Ergonomics. HCII 2022. Lecture Notes in Computer Science*, (pp. 3–17).

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111.

Clark, J., Yallop, C., & Fletcher, J. (2007). *An Introduction to Phonetics and Phonology*. (3rd ed.). Malden, Mass.: Blackwell Publishing.

Cohen Priva, U., & Strand, E. (2023). Schwa's duration and acoustic position in American English. *Journal of Phonetics*, *96*. URL: `https://doi.org/10.1016/j.wocn.2022.101198`.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research*, *14*, 565–577.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*, 696–702. URL: https://doi.org/10.1037/0278-7393.34.3.696.

Cooke, M., Garcia Lecumberri, M. L., & Wester, M. (2013). *DiapixFL*. LISTA Consortium: (i) Language and Speech Lab, Universidad del Pais Vasco, Spain and Ikerbasque, Spain; (ii) CSTR, University of Edinburgh, UK; (iii) KTH Royal Institute of Technology, Sweden; (iv) Institute of Computer Science, FORT. URL: https://doi.org/10.7488/ds/139.

Corley, M., & Hartsuiker, R. J. (2003). Hesitation in speech can. . . um. . . help a listener understand. *Annual Conference of the Cognitive Science Society*, (pp. 276–281).

Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS ONE*, *6*. URL: https://doi.org/10.1371/journal.pone.0019792.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668. URL: https://doi.org/10.1016/j.cognition.2006.10.010.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*, 589–602. URL: https://doi.org/10.1111/j.1749-818X.2008.00068.x.

Council of Europe (2011). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Strasbourg: Cambridge University Press. URL: https://rm.coe.int/1680459f97.

Crible, L., Didirková, I., Dodane, C., & Kosmala, L. (2022). Towards an inclusive system for the annotation of (dis)fluency in typical and atypical speech. *Clinical Linguistics and Phonetics*, (pp. 1–18). URL: https://doi.org/10.1080/02699206.2022.2126330.

Crystal, D., & Davy, D. (1976). *Advanced Conversational English*. (1st ed.). London: Longman.

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, *88*, 101–112. URL: https://doi.org/10.1121/1.399955.

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2019). Expectation and locality effects in the prediction of disfluent fillers and repairs in English speech. *Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 103–109).

Dammalapati, S., Rajkumar, R., & Agarwal, S. (2021). Effects of duration, locality, and surprisal in speech disfluency prediction in English spontaneous speech. *Proceedings of the Society for Computation in Linguistics*, (pp. 91–101).

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*, 223–243. URL: https://doi.org/10.1017/S0142716413000210.

Delattre, P., & Olsen, C. (1969). Syllabic features and phonic impression in English, German, French and Spanish. *Lingua*, *22*, 160–175. URL: https://doi.org/10.1016/0024-3841(69)90051-5.

Diachek, E., & Brown-Schmidt, S. (2022). The effect of disfluency on memory for what was said. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*, 1306–1324. URL: https://doi.org/10.1037/xlm0001156.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition*, *118*, 111–115. URL: https://doi.org/10.1016/j.cognition.2010.09.012.

Ditko, T., & Stauch, S. (2023). *Rhetorik und Redekunst für Dummies*. Weinheim: Wiley.

Eklund, R. (2004). *Disfluency in Swedish human – human and human – machine travel booking dialogues*. Ph.D. thesis Linköping University, Sweden. URL: https://doi.org/10.13140/RG.2.1.3015.0882.

Elmers, M., O'Mahony, J., & Székely, É. (2023). Synthesis after a couple PINTs: Investigating the role of pause-internal phonetic particles in speech synthesis and perception. In *Annual Conference of the International Speech Communication Association (INTERSPEECH), Dublin*.

Embarki, M. (2013). Phonetics. In J. Owens (Ed.), *The Oxford Handbook of Arabic Linguistics* (pp. 23–44). Oxford, New York: Oxford University Press.

Erker, D., & Bruso, J. (2017). Uh, bueno, em...: Filled pauses as a site of contact-induced change in Boston Spanish. *Language Variation and Change*, *29*, 205–244. URL: https://doi.org/10.1017/S0954394517000102.

Erker, D., & Vidal-Covas, L.-A. M. (2022). What we say when we say nothing at all: Clues to contact-induced language change in Spanish conversational pause-fillers. *Estudios del Observatorio/Observatorio Studies*, *2949*. URL: https://doi.org/10.15427/or080-09/2022en.

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. *International Conference on Computational Social Science*, (pp. 2016–2018).

Fisher, J. L., & Harris, M. B. (1973). Effect of note taking and review on recall. *Journal of Educational Psychology*, *65*, 321–325. URL: https://doi.org/10.1037/h0035640.

Flege, J. E. (1995). Second Language Speech Learning: Theory, Findings, and Problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). Timonium, MD: York Press.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*, 709–738. URL: https://doi.org/10.1006/jmla.1995.1032.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory Cognition*, *29*, 320–326. URL: https://doi.org/10.3758/BF03194926.

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, *34*, 37–55.

Fraser, H. (2012). Language analysis for the determination of origin (LADO). *The Encyclopedia of Applied Linguistics*, (pp. 2005–2007). URL: https://doi.org/10.1002/9781405198431.wbeal0597.

Fraundorf, S. H., & Watson, D. G. (2011a). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*, 161–175. URL: https://doi.org/10.1016/j.jml.2011.03.004.

Fraundorf, S. H., & Watson, D. G. (2011b). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*, 161–175. URL: https://doi.org/10.1016/j.jml.2011.03.004.

Fraundorf, S. H., & Watson, D. G. (2014). Alice's adventures in um-derland: Psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience*, *29*, 1083–1096. URL: https://doi.org/10.1080/01690965.2013.832785.

Fuchs, S., & Rochet-Capellan, A. (2021). The respiratory foundations of spoken language. *Annual Review of Linguistics*, *7*, 1–18. URL: https://doi.org/10.1146/annurev-linguistics-031720-103907.

Gabriel, C. (2022). Phonetik und Phonologie des Spanischen. In R. Klabunde, W. Mihatsch, & S. Dipper (Eds.), *Linguistik im Sprachvergleich: Germanistik, Romanistik, Anglistik* (pp. 27–48). Berlin: J.B. Metzler. URL: https://doi.org/10.1007/978-3-662-62806-5.

Gallant, J., & Libben, G. (2019). No lab, no problem: Designing lexical comprehension and production experiments using PsychoPy3. *The Mental Lexicon*, *14*, 152–168. URL: https://doi.org/10.1075/ml.00002.gal.

Garciá-Amaya, L., & Lang, S. (2020). Filled pauses are susceptible to cross-language phonetic influence. *Studies in Second Language Acquisition*, (pp. 1–29). URL: https://doi.org/10.1017/S0272263120000169.

Garcia Lecumberri, M. L., Cooke, M., & Wester, M. (2017). A bi-directional task-based corpus of learners' conversational speech. *International Journal of Learner Corpus Research*, *3*, 175–195. URL: https://doi.org/10.1075/ijlcr.3.2.04gar.

Garnier, M., Bailly, L., Dohen, M., Welby, P., & Lœvenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: Global effects on the utterance. *Annual Conference of the International Speech Communication Association (INTERSPEECH), Pittsburgh*, (pp. 2246–2249). URL: https://doi.org/10.21437/interspeech.2006-323.

Gerstenberg, A., Fuchs, S., Kairet, J. M., Frankenberg, C., & Schröder, J. (2018). A cross-linguistic, longitudinal case study of pauses and interpausal units in spontaneous speech corpora of older speakers of German and French. *International Conference on Speech Prosody, Poznań*, (pp. 211–215). URL: https://doi.org/10.21437/SpeechProsody.2018-43.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *AU-16*, 78–80. URL: https://doi.org/10.1109/TAU.1968.1161953.

Gessinger, I. (2022). *Phonetic Accomodation of Human Interlocutors in the Context of Human-Computer Interaction*. Ph.D. thesis.

Giannini, A. (2003). Hesitation Phenomena in Spontaneous Italian, . (pp. 2653–2656).

Gick, B. (2002). An X-ray investigation of pharyngeal constriction in American English schwa. *Phonetica*, *59*, 38–48. URL: https://doi.org/10.1159/000056204.

Gick, B., Wilson, I., Koch, K., & Cook, C. (2004). Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica*, *61*, 220–233. URL: https://doi.org/10.1159/000084159.

Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference? In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics: A Mutualistic Entente* (pp. 119–149). Berlin, Heidelberg, New York: Mouton de Gruyter. URL: https://doi.org/10.1515/9783110199024.119.

Godfrey, J. J., & Holliman, E. (1997). Switchboard-1 Release 2.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, *18*, 293–307. URL: https://doi.org/10.1558/ijsll.v18i2.293.

Gold, E., French, P., & Harrison, P. (2013). Clicking behavior as a possible speaker discriminant in English. *Journal of the International Phonetic Association*, *43*, 339–349. URL: https://doi.org/10.1017/S0025100313000248.

Goldman-Eisler, F. (1957). Speech Production and Language Statistics. *Canadian Journal of Chemistry*, *35*, 1578.

Goldman-Eisler, F. (1958). Speech Production and the Predictability of Words in Context. *Quarterly Journal of Experimental Psychology*, *10*, 96–106. URL: https://doi.org/10.1080/17470215808416261.

Goldman-Eisler, F. (1961). A comparative study of two hesitation phenomnea. *Language and Speech*, *4*, 18–26. URL: https://doi.org/10.1016/S0099-6963(27)90007-6.

Goodwin, C. (1981). *Conversation Organization: Interaction Between Speakers and Hearers*. New York: Academic Press.

Gósy, M. (2004). The manifold function of Schwa. *Grazer Linguistische Studien*, *62*, 15–26.

Gósy, M., Gyarmathy, D., & Beke, A. (2017). Phonetic analysis of filled pauses based on a Hungarian-English learner corpus. *International Journal of Learner Corpus Research*, *3*, 149–174. URL: https://doi.org/10.1075/ijlcr.3.2.03gos.

Gósy, M., & Silber-Varod, V. (2021). Attached filled pauses: Occurrences and durations. *Workshop on Disfluency in Spontaneous Speech (DiSS 2021), Paris*, (pp. 71–76).

Götz, S. (2013). *Fluency in Native and Nonnative English Speech*. Amsterdam, Philadelphia: John Benjamins Publishing Company. URL: https://doi.org/10.1075/scl.53.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psycholinguistics*, *48*, 163–189. URL: https://doi.org/10.1016/B978-0-12-407794-2.00053-5.

Grawunder, S., Oertel, M., & Schwarze, C. (2014). Politeness, culture, and speaking task - paralinguistic prosodic behavior of speakers from Austria and Germany. *International Conference on Speech Prosody*, (pp. 159–163). URL: `https://doi.org/10.21437/speechprosody.2014-20`.

Guirao, M., & García Jurado, M. A. (1990). Frequency of occurence of phonemes in American Spanish. *Revue québécoise de linguitique*, *19*, 135–149.

Gully, A. J., Foulkes, P., French, P., Harrison, P., & Hughes, V. (2019). The Lombard effect in MRI Noise. *International Congress of Phonetic Sciences (ICPhS), Melbourne*, (pp. 800–804).

Hamaker, J., Zeng, Y., & Picone, J. (1998). Rules and Guidelines for Transcription and Segmentation of the SWITCHBOARD Large Vocabulary Conversational Speech Recognition Corpus. URL: `https://www.isip.piconepress.com/projects/switchboard/doc/transcription_guidelines/transcription_guidelines.pdf`.

Hamdi, R., Ghazali, S., & Barkat-Defradas, M. (2005). Syllable structure in spoken Arabic: A comparative investigation. *Annual Conference of the International Speech Communication Association (INTERSPEECH), Lisbon*, (pp. 2245–2248).

Harrington, L., Rhodes, R., & Hughes, V. (2021). Style variability in disfluency analysis for forensic speaker comparison. *International Journal of Speech Language and the Law*, *28*, 31–58. URL: `https://doi.org/10.1558/ijsll.20214`.

Hay, J., Podlubny, R., Drager, K., & McAuliffe, M. (2017). Car-talk: Location-specific speech production and perception. *Journal of Phonetics*, *65*, 94–109. URL: `https://doi.org/10.1016/j.wocn.2017.06.005`.

Hirschfeld, U., & Wallraff, U. (2002). Untersuchungen zum schwa im Deutschen. In A. Braun, & H. R. Masthoff (Eds.), *Phonetics and its Applications: Festschrift for Jens-Peter Köster on the occasion of his 60th Birthday* (pp. 493–505). Stuttgart: Franz Steiner Verlag.

Hualde, J. I., & Ortiz de Urbina, J. (2003). *A Grammar of Basque*. Berlin, New York: Mouton de Gruyter. URL: `https://doi.org/10.1515/9783110895285`.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, *23*, 99–132. URL: `https://doi.org/10.1558/ijsll.v23i1.29874`.

Ibrahim, O., Asadi, H., Kassem, E., & Dellwo, V. (2020). Arabic speech rhythm corpus: Read and spontaneous speaking styles. *International Conference on Language Resources and Evaluation (LREC)*, (pp. 5337–5342).

Ibrahim, O., Yuen, I., van Os, M., Andreeva, B., & Möbius, B. (2022). The combined effects of contextual predictability and noise on the acoustic realisation of German syllables. *The Journal of the Acoustical Society of America*, *152*, 911–920. URL: https://doi.org/10.1121/10.0013413.

IPDS (2006). Video Task Scenario: Lindenstraße – The Kiel Corpus of Spontaneous Speech, Volume 4, DVD. *Institut für Phonetik und Digitale Sprachsignal- verarbeitung Universität Kiel*, .

Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science and Justice*, *47*, 50–67. URL: https://doi.org/10.1016/j.scijus.2007.03.003.

Jessen, M. (2012). *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs*. München: Lincom.

Jessen, M. (2018). Forensic voice comparison. In J. Visconti, & M. Rathert (Eds.), *Handbook of Communication in the Legal Sphere* (pp. 219–255). Berlin: Mouton de Gruyter.

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech Language and the Law*, *12*, 174–213. URL: https://doi.org/10.1558/sll.2005.12.2.174.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*, 385–390. URL: https://doi.org/10.3758/BRM.41.2.385.

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *International Congress of Phonetic Sciences (ICPhS), Glasgow*, (pp. 2–7).

Kelley, M. C., & Tucker, B. V. (2020). A comparison of four vowel overlap measures. *The Journal of the Acoustical Society of America*, *147*, 137–145. URL: https://doi.org/10.1121/10.0000494.

Kiesling, S., Dilley, L., & Raymond, W. D. (2006). The Variation in Conversation (ViC) Project: Creation of the Buckeye Corpus of Conversational Speech. URL: http://buckeyecorpus.osu.edu/BuckeyeCorpusmanual.pdf.

Kirchhübel, C., & Brown, G. (2022). Spoofed speech from the perspective of a forensic phonetician. In *Annual Conference of the International Speech Communication Association (INTERSPEECH), Incheon* (pp. 1308–1312). URL: https://doi.org/10.21437/Interspeech.2022-661.

Kjellmer, G. (2003). Hesitation. In defence of ER and ERM. *English Studies*, *84*, 170–198. URL: https://doi.org/10.1076/enst.84.2.170.14903.

Klug, K., & König, M. (2012). Untersuchung zur sprecherspezifischen Verwendung von Häsitationspartikeln anhand der Parameter Grundfrequenz und Vokalqualität. In U. Hirschfeld, & B. Neuber (Eds.), *Erforschung und Optimierung der Callcenterkommunikation* (pp. 175–193). Berlin: Frank Timme.

Kohler, K. J. (1994). Glottal stops and glottalization in German: Data and theory of connected speech processes. *Phonetica*, *51*, 38–51. URL: https://doi.org/10.1159/000261957.

Koopmans-van Beinum, F. J. (1994). What's in a schwa? *Phonetica*, *51*, 68–79. URL: https://doi.org/10.1159/000261959.

Kowal, S., O'Connell, D. C., Forbush, K., Higgins, M., Clarke, L., & D'Anna, K. (1997). Interplay of literacy and orality in inaugural rhetoric. *Journal of Psycholinguistic Research*, *26*, 1–31. URL: https://doi.org/10.1023/A:1025043620499.

Krech, E. M. (1968). *Sprechwissenschaftlich-phonetische Untersuchungen zum Gebrauch des Glottisschlageinsatzes in der allgemeinen deutschen Hochlautung*. Basel New York: Kager.

Künzel, H. J. (1987). *Sprechererkennung. Grundzüge forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik Verlag.

Kuzla, C., & Ernestus, M. (2011). Prosodic conditioning of phonetic detail in German plosives. *Journal of Phonetics*, *39*, 143–155. URL: https://doi.org/10.1016/j.wocn.2011.01.001.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. URL: https://doi.org/10.18637/jss.v082.i13.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, *2*, 205–254. URL: https://doi.org/10.4324/9781315683737-13.

Laserna, C. M., Seih, Y.-T., & Pennebaker, J. W. (2014). Um... Who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, *33*, 328–338. URL: https://doi.org/10.1177/0261927X14526993.

Laufer, A. (2019). The origin of the IPA schwa. *International Congress of Phonetic Sciences (ICPhS), Melbourne*, (pp. 1908–1911).

de Leeuw, E. (2007). Hesitation markers in English, German, and Dutch. *Journal of Germanic Linguistics*, *19*, 85–114. URL: https://doi.org/10.1017/S1470542707000049.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104.

Li, X. (2020). Click-initiated self-repair in changing the sequential trajectory of actions-in-progress. *Research on Language and Social Interaction*, *53*, 90–117. URL: https://doi.org/10.1080/08351813.2020.1712959.

Li, X., Ishi, C. T., Fu, C., & Hayashi, R. (2022). Prosodic and Voice Quality Analyses of Filled Pauses in Japanese Spontaneous Conversation by Chinese learners and Japanese Native Speakers, . (pp. 550–554). URL: https://doi.org/10.21437/speechprosody.2022-112.

Lickley, R. J. (2015). *Fluency and Disfluency*. Wiley-Blackwell. URL: https://doi.org/10.1002/9781118584156.ch20.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*, 783–783. URL: https://doi.org/10.1121/1.2142410.

Lindblom, B. E. F. (1968). Temporal organization of syllable production. *Speech Transmission Lab. Quarterly Progress Status Report. Stockholm: KTH Department of Speech, Music, and Hearing*, *9*, 1–5.

Lo, J. J. (2020). Between äh(m) and euh(m): The distribution and realization of filled pauses in the speech of German-French simultaneous bilinguals. *Language and Speech*, *63*, 746–768. URL: https://doi.org/10.1177/0023830919890068.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*, 606–608. URL: https://doi.org/10.1121/1.1912396.

Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales des maladies de l'oreille et du larynx*, *37*, 101–119.

Lomotey, C. F. (2021). Fillers in academic discourse: An analysis of lectures from a public university in Ghana. *European Journal of Applied Linguistics Studies*, *3*, 98–122. URL: https://doi.org/10.46827/ejals.v3i2.248.

Maclay, H., & Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *WORD*, *15*, 19–44. URL: https://doi.org/10.1080/00437956.1959.11659682.

Maddieson, I. (2005). Vowel quality inventories. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The World Atlas of Language Structures* (pp. 14–15). Oxford, New York: Oxford University Press.

Mahl, G. F. (1956). Disturbances and silences in the patient's speech in psychotherapy. *Journal of Abnormal and Social Psychology*, *53*, 1–15. URL: https://doi.org/10.1037/h0047552.

Martin, N. (2019). *Vor Publikum sprechen: Tipps und Tricks gegen die Angst vor Vorträgen*. 50Minuten.de.

Massen, C., & Vaterrodt-Plünnecke, B. (2006). The role of proactive interference in mnemonic techniques. *Memory*, *14*, 189–196. URL: https://doi.org/10.1080/09658210544000042.

McDougall, K., & Duckworth, M. (2017). Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication*, *95*, 16–27.

McDougall, K., & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *International Journal of Speech, Language and the Law*, *25*, 205–230. URL: https://doi.org/10.1558/IJSLL.37241.

McDougall, K., Rhodes, R., Duckworth, M., French, P., & Kirchhübel, C. (2019). Application of the 'TOFFA' framework to the analysis of disfluencies in forensic phonetic casework. *International Conference of Phonetic Sciences (ICPhS), Melbourne*, .

Meinhold, G., & Stock, E. (1980). *Phonologie der deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut Leipzig.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychology Review*, *63*, 81–97. URL: https://doi.org/10.1037/h0043158.

Muhlack, B. (2020a). L1 and L2 production of non-lexical hesitation particles of German and English native speakers. *Workshop on Laughter and Other Non-Verbal Vocalisations, Bielefeld*, (pp. 44–47).

Muhlack, B. (2020b). The vowel quality of non-lexical hesitation particles in German and English L1 and L2 speech [Poster]. In *Phonetik und Phonologie im deutschsprachigen Raum (PP16), Trier*.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488. URL: https://doi.org/10.1037/h0045106.

Niebuhr, O., & Fischer, K. (2019). Do not hesitate! - Unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance. In *Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz* (pp. 544–548). URL: https://doi.org/10.21437/Interspeech.2019-1194.

Nordström, P.-E. (1977). Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics*, *5*, 81–92. URL: https://doi.org/10.1016/s0095-4470(19)31115-5.

O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, *34*, 555–576. URL: https://doi.org/10.1007/s10936-005-9164-3.

O'Connell, D. C., & Kowal, S. (2008). *Communicating with one another: toward a psychology of spontaneous spoken discourse*. New York: Springer.

Ogden, R. (2013). Clicks and percussives in English conversation. *Journal of the International Phonetic Association*, *43*, 299–320. URL: https://doi.org/10.1017/S0025100313000224.

Ogden, R. (2020). Audibly not saying something with clicks. *Research on Language and Social Interaction*, *53*, 66–89. URL: https://doi.org/10.1080/08351813.2020.1712960.

Oostendorp, M. V. (2014). Schwa in phonological theory. *The Second Glot International State-of-the-Article Book*, (pp. 431–462). URL: https://doi.org/10.1515/9783110890952.431.

Pätzold, M., & Simpson, A. (1995). An acoustic analysis of hesitation particles in German. *International Congress of Phonetic Sciences (ICPhS), Stockholm*, (pp. 512–515).

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency - Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction*, *44*, 31–40. URL: https://doi.org/10.1016/j.learninstruc.2016.01.012.

Pistor, T. (2017). Prosodische Universalien bei Diskurspartikeln. *Zeitschrift für Dialektologie und Linguistik*, *84*, 46–76.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W. D., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech. URL: www.buckeyecorpus.osu.edu.

Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology*, *50*, 408–412. URL: https://doi.org/10.1037/1196-1961.50.4.408.

Prolific (2014). Prolific. London, UK. Accessed: Nov 2020 - Nov 2021. URL: https://www.prolific.co.

Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, *35*, 353–362. URL: https://doi.org/10.1016/j.wocn.2006.09.001.

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, . URL: https://www.r-project.org/.

Reitbrecht, S. (2017). *Häsitationsphänomene in der Fremdsprache Deutsch und ihre Bedeutung für die Sprechwirkung*. Berlin: Frank Timme.

Riazantseva, A. (2001). Second lanuage proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*, 497–526. URL: https://doi.org/10.1017/s027226310100403x.

Roach, P. J. (2009). *English Phonetics and Phonology: A practical course*. (4th ed.). Cambridge, New York, Melbourne: Cambridge University Press.

Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor Francis.

Rose, R., & Watanabe, M. (2019). A crosslinguistic corpus study of silent and filled pauses: When do speakers use filled pauses to fill pauses? *International Congress of Phonetic Sciences (ICPhS), Graz*, (pp. 2615–2619).

Rose, R. L. (2017). A Comparison of Form and Temporal Characteristics of Filled Pauses in L1 Japanese and L2 English. *Journal of the Phonetic Society of Japan*, *21*, 33–40.

Rose, R. L. (2020). Fluidity: Real-time feedback on acoustic measures of second language speech fluency. *International Conference on Speech Prosody, Tokyo*, (pp. 774–778). URL: https://doi.org/10.21437/SpeechProsody.2020-158.

Russo, M., & Barry, W. J. (2008). Measuring rhythm. A quantified analysis of Southern Italian dialects stress time parameters. *Language Design*, *2*, 315–322.

Sadanobu, T., & Takubo, Y. (1993). The discourse management function of fillers: A case of "eeto" and "ano(o)". *International Symposium on Spoken Dialogue*, (pp. 271–274).

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*, 362–367. URL: https://doi.org/10.1037/0022-3514.60.3.362.

Schmidt, J. E. (2001). Bausteine der Intonation? *Germanistische Linguistik*, *157-158*, 9–32.

Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America*, *85*, 295–312. URL: https://doi.org/10.1121/1.397737.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). Major trends in vowel system inventories. *Journal of Phonetics*, *25*, 233–253. URL: https://doi.org/10.1006/jpho.1997.0044.

Schweitzer, A., & Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. *Annual Conference of the International Speech Communication Association (INTERSPEECH), Lyon*, (pp. 525–529). URL: https://doi.org/10.21437/interspeech.2013-148.

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. New York, London: Routledge.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*, 379–423. URL: https://doi.org/10.1016/s0016-0032(23)90506-5.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis. URL: ftp://130.107.33.205/pub/papers/shriberg-thesis.pdf.

Shriberg, E. (2001). To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*, 153–169. URL: https://doi.org/10.1017/S0025100301001128.

Shriberg, E. E. (1999). Phonetic consequences of speech disfluency. *International Congress of Phonetic Sciences (ICPhS), San Francisco*, (pp. 619–622).

Shriberg, E. E., & Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, *50*, 172–179. URL: https://doi.org/10.1159/000261937.

Silverman, D. (2011). Schwa. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 1–15). Malden, Mass.: John Wiley Sons. URL: https://doi.org/10.1002/9781444335262.wbctp0026.

Simeonova, R. K. (1989). *Die Segmentsysteme des Deutschen und des Bulgarischen: Eine kontrastive phonetisch-phonologische Studie*. München: Verlag Otto Sagner. URL: https://doi.org/10.3726/b12213.

Šimko, J., Beňuš, Š., & Vainio, M. (2016). Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue. *The Journal of the Acoustical Society of America*, *139*, 151–162. URL: https://doi.org/10.1121/1.4939495.

Simpson, A. P. (2007). Acoustic and auditory correlates of non-pulmonic sound production in German. *Journal of the International Phonetic Association*, *37*, 173–182. URL: https://doi.org/10.1017/S0025100307002927.

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass*, *3*, 236–264. URL: https://doi.org/10.1111/j.1749-818X.2008.00112.x.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25–38. URL: https://doi.org/10.1006/jmla.1993.1002.

Staróbole Juste, F., & Furquim De Andrade, C. R. (2011). Speech disfluency types of fluent and stuttering individuals: Age effects. *Folia Phoniatrica et Logopaedica*, *63*, 57–64. URL: https://doi.org/10.1159/000319913.

Sugiura, K. (2015). *Production of English Schwa By Japanese Speakers*. Ph.D. thesis Kwansei Gakuin University.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*, 485–496.

Székely, É., Eje Henter, G., Beskow, J., & Gustafson, J. (2019). How to train your fillers: uh and um in spontaneous speech synthesis, . (pp. 245–250). URL: https://doi.org/10.21437/ssw.2019-44.

Tamaoka, K., & Makioka, S. (2004). Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments, and Computers*, *36*, 531–547. URL: https://doi.org/10.3758/BF03195600.

Temple, L. (2000). Second language learner speech production. *Studia Linguistica*, *54*, 288–297. URL: https://doi.org/10.1111/1467-9582.00068.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, *85*, 1699–1707.

Trebon, T. (2022). Effect of hesitation sound phonetic quality on perception of language fluency and accentedness. *Oregon Undergraduate Research Journal*, *20*, 1–18. URL: https://doi.org/10.5399/uo/ourj/20.1.3.

Trouvain, J. (2004). *Tempo variation in speech production. Implications for speech synthesis.*. Phonus 8, Phonetics, Saarbrücken: Saarland University.

Trouvain, J., & Belz, M. (2019). Zur Annotation nicht-verbaler Vokalisierungen in Korpora gesprochener Sprache. *Conference Elektronische Sprachsignalverarbeitung (ESSV), Dresden*, (pp. 280–287).

Trouvain, J., & Malisz, Z. (2016). Inter-speech clicks in an Interspeech keynote. *Annual Conference of the International Speech Communication Association (Interspeech), San Francisco*, (pp. 1397–1401). URL: https://doi.org/10.21437/Interspeech.2016-1064.

Trouvain, J., & Truong, K. P. (2012). Comparing non-verbal vocalisations in conversational speech corpora. *International Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, (pp. 36–39).

Trouvain, J., & Werner, R. (2020). Comparing annotations of non-verbal vocalisations in speech corpora. *Workshop on Laughter and Other Non-Verbal Vocalisations, Bielefeld*, (pp. 69–72).

Trouvain, J., & Werner, R. (2022). A phonetic view on annotating speech pauses and pause-internal phonetic particles. In C. Schwarze, & S. Grawunder (Eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion* (pp. 55–73). Tübingen: Narr.

Tuomainen, O., Taschenberger, L., Rosen, S., & Hazan, V. (2021). Speech modifications in interactive speech: Effects of age, sex and noise type. *Philosophical Transactions of the Royal Society B*, *377*. URL: `https://doi.org/10.1098/rstb.2020.0398`.

Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, *84*, 917–928. URL: `https://doi.org/10.1121/1.396660`.

Vasilescu, I., & Adda-Decker, M. (2007). A cross-language study of acoustic and prosodic characteristics of vocalic hesitations. In *Fundamentals of verbal and nonverbal communication and the biometric issue*.

Vasilescu, I., Nemoto, R., & Adda-Decker, M. (2007). Vocalic hesitations vs vocalic systems: a cross-language comparison. *International Congress of Phonetic Sciences (ICPhS), Saarbrücken*, (pp. 1101–1104).

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics Cognition*, *14*, 129–182. URL: `https://doi.org/10.1075/pc.14.1.08war`.

Watanabe, M., Shirahata, Y., Rose, R., & Maekawa, K. (2021). How do speakers pause and hesitate in English and Japanese? - A comparison using parallel corpora of English and Japanese presentation speeches. *Conference of the Oriental COCOSDA*, (pp. 164–167).

Werner, R., Fuchs, S., Trouvain, J., Kürbis, S., Möbius, B., & Birkholz, P. (2023). Acoustics of breath noises in human speech: Descriptive and three-dimensional modeling approaches. *Journal of Speech, Language, and Hearing Research*, (pp. 1–15).

Wester, M., García Lecumberri, M. L., & Cooke, M. (2014). DIAPIX-FL: A symmetric corpus of problem-solving dialogues in first and second languages. *Annual Conference of the International Speech Communication Association (INTERSPEECH), Singapore*, (pp. 509–513). URL: `https://doi.org/10.21437/interspeech.2014-126`.

Whalen, D. H., Chen, W.-R., Shadle, C. H., & Fulop, S. A. (2022). Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986). *The Journal of the Acoustical Society of America*, *152*, 933–941. URL: `https://doi.org/10.1121/10.0013410`.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., Francois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Seidel, D. R. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*, 1686. URL: https://doi.org/10.21105/joss.01686.

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, *6*, 199–234. URL: https://doi.org/10.1163/22105832-00602001.

Wiese, R. (1984). Language production in foreign and native languages: Same or different? In H. W. Dechert, D. Möhle, & M. Raupach (Eds.), *Second Language Productions* (pp. 11–25). Tübingen: Gunter Narr Verlag.

Wohlert, A. B., & Hammen, V. L. (2000). Lip muscle activity related to speech rate and loudness. *Journal of Speech, Language, and Hearing Research*, *43*, 1229–1239. URL: https://doi.org/10.1044/jslhr.4305.1229.

Zámečník, J. (2019). *Disfluency Prediction in Natural Spoken Language*. Ph.D. thesis Albert-Ludwigs-Universität Freiburg i.Br.

Zayats, V., Tran, T., Wright, R., Mansfield, C., & Ostendorf, M. (2019). Disfluencies and human speech transcription errors. *Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz*, (pp. 3088–3092).

Zellers, M. (2022). An overview of discourse clicks in Central Swedish. *Annual Conference of the International Speech Communication Association (Interspeech), Incheon*, (pp. 3423–3427). URL: https://doi.org/10.21437/interspeech.2022-583.