UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Der Einfluss der Sprachprobenlänge auf die Reliabilität

Eine Reliabilitätsanalyse der mittleren Äusserungslänge und der *moving average type-token ratio* bei Spontansprachanalysen von Kindern mit und ohne Spracherwerbsstörungen

Masterarbeit, eingereicht bei der Philosophischen Fakultät
der Universität Freiburg (CH)
Studienprogramm Master of Arts in Sonderpädagogik,
Option Logopädie

Lena Graf
aus Birrwil AG
Matrikel-Nr.: 17-497-322

Betreuer: Prof. Dr. phil. Erich Hartmann

Datum: 24. April 2024

https://doi.org/10.51363/unifr.lma.2024.010

Abstract

Zweck: Diese Arbeit verfolgt innerhalb der Reliabilitätsanalyse zwei Ziele: Erstens soll die instrumentelle Reliabilität der MLU und der MATTR in den gesamten Sprachproben untersucht werden. Zweitens soll der Frage nachgegangen werden, wie kurz Sprachproben sein können, um reliable Werte zu ergeben.
Methode: Narrative Spontansprachproben von 49 Schweizerdeutsch sprechenden Kindern mit und ohne Spracherwerbsstörungen im Alter von 4 - 7 Jahren wurden gesammelt, transkribiert und automatisiert bezüglich der MLU und der MATTR analysiert.
Resultate: Die Analyse mittels RMANOVA ergab keine signifikanten Unterschiede zwischen den Sprachprobenlängen. Die Spearman-Brown Korrelationskoeffizienten lagen für die MLU ab 30 Äusserungen und mehr auf einem akzeptablen Niveau. Für die MATTR wurde die erwartete Schwelle der relativen Reliabilität von 0.9 nicht erreicht. Der Korrelationskoeffizient stagnierte hier ab 105 Wörtern bei circa 0.87, was knapp unter dieser Grenze liegt. Eine genügende absolute Reliabilität auf individueller Ebene ergaben sich durch die Analyse von B-A Plots ab 30 Äusserungen für die MLU und 70 Wörtern für die MATTR.
Fazit: Für Sprachproben, die in der schweizerdeutschen Sprache mittels *personal narratives* bei Kindern im Alter von 4 - 7 Jahren mit und ohne Spracherwerbsstörungen erhoben wurden, ist für die MLU eine Mindestlänge von 30 Äusserungen und für die MATTR von 105 Wörtern empfehlenswert. Die gewählte Länge der Sprachprobe sollte sich an den Anforderungen an die Messgenauigkeit orientieren.

Inhaltsverzeichnis

1 Einleitung **8**

2 Die Spontansprachanalyse **10**

2.1 Stärken und Schwächen der LSA 11

2.1.1 Vorteile 11

2.1.2 Nachteile 12

2.2 Aktuelle Anwendung 13

2.3 Elizitationsmethoden 15

2.4 Automatisierte Spontansprachanalyse 17

2.5 Logopädische Indikatoren 18

2.5.1 Indikatoren im DigiSpon 1 18

2.5.2 Auswahlkriterien 19

2.5.3 Indikator Äusserungslänge 20

2.5.4 Indikator lexikalische Vielfalt 22

2.6 Gütekriterien der Testevaluation 26

2.7 Reliabilität 27

2.7.1 Schätzmethoden 27

2.7.2 Reliabilität von kurzen Sprachproben 31

2.7.3 Bisherige Forschungsbefunde 32

3 Fragestellungen **38**

4 Daten und Methoden **39**

4.1 Datenerhebung 39

4.2 Stichprobe 41

4.3 Transkription und Analyse 42

4.4 Statistische Analyse 44

4.4.1 Instrumentelle Reliabilität 44

4.4.2 Reliabilität von kurzen Sprachproben 47

5 Resultate **53**

5.1 Instrumentelle Reliabilität 53

5.2 Reliabililtät von kurzen Sprachproben 55

5.2.1 RMANOVA für die MLU 57

5.2.2 RMANOVA für die MATTR 59

5.2.3 Spearman-Brown Korrelationskoeffizient für die MLU 61

5.2.4 Spearman-Brown Korrelationskoeffizient für die MATTR 62

5.2.5 Bland-Altman Plots für die MLU 63

5.2.6 Bland-Altman Plots für die MATTR 66

6 Diskussion **68**

7 Limitationen und zukünftige Forschung **76**

8 Anhang **86**

Weitere Resultate der statistischen Analyse 86

R Code 89

Ehrenwörtliche Erklärung 109

Abbildungsverzeichnis

1 Histogramm der randomisierten Testhalbierungsreliabilitäten der MLU . . 54
2 Histogramm der randomisierten Testhalbierungsreliabilitäten der MATTR mit Überschneidungen der Fenster 54
3 Histogramm der randomisierten Testhalbierungsreliabilitäten der MATTR ohne Überschneidung der Fenster 55
4 Q-Q Plots zur Verteilung der MLU-Werte 57
5 Boxplots der MLU-Werte gruppiert nach Sprachprobenlänge 58
6 Q-Q Plots zur Verteilung der MATTR-Werte 59
7 Boxplots der MATTR-Werte gruppiert nach Sprachprobenlänge 60
8 Entwicklung der Spearman-Brown Koeffizienten mit zunehmender Sprachprobenlänge für die MLU 62
9 Entwicklung der Spearman-Brown Koeffizienten mit zunehmender Sprachprobenlänge für die MATTR 63
10 Bland-Altman Plot für den Vergleich der MLU-Werte von 10 Äusserungen und der gesamten Sprachprobe 64
11 Bland-Altman Plot für den Vergleich der MLU-Werte von 20 Äusserungen und der gesamten Sprachprobe 65
12 Bland-Altman Plot für den Vergleich der MLU-Werte von 30 Äusserungen und der gesamten Sprachprobe 65
13 Bland-Altman Plot für den Vergleich der MATTR-Werte von 35 Wörtern und der gesamten Sprachprobe 67
14 Bland-Altman Plot für den Vergleich der MATTR-Werte von 70 Wörtern und der gesamten Sprachprobe 67
15 Bland-Altman Plot für den Vergleich der MATTR-Werte von 105 Wörtern und der gesamten Sprachprobe 68
16 Bland-Altman Plot für den Vergleich der MLU-Werte von 40 Äusserungen und der gesamten Sprachprobe 86
17 Bland-Altman Plot für den Vergleich der MLU-Werte von 50 Äusserungen und der gesamten Sprachprobe 87
18 Bland-Altman Plot für den Vergleich der MLU-Werte von 30 Äusserungen und der gesamten Sprachprobe mit ± 1.75 Standardabweichungen 87
19 Bland-Altman Plot für den Vergleich der MLU-Werte von 30 Äusserungen und der gesamten Sprachprobe mit ± 1.65 Standardabweichungen 88
20 Bland-Altman Plot für den Vergleich der MATTR-Werte von 140 Wörtern und der gesamten Sprachprobe 88
21 Bland-Altman Plot für den Vergleich der MATTR-Werte von 175 Wörtern und der gesamten Sprachprobe 89

22 Bland-Altman Plot für den Vergleich der MATTR-Werte von 70 Wörtern und der gesamten Sprachprobe mit ± 1.75 Standardabweichungen 89

Tabellenverzeichnis

1 Längen der analysierten Sprachproben in Minuten, Äusserungen und Wörtern 42
2 Schätzwerte für die instrumentelle Reliabilität der MLU und der MATTR . 53
3 Schätzwerte für die instrumentelle Reliabilität der MATTR ohne Überschneidung der Fenster 55
4 Deskriptive Statistiken der MLU und der MATTR bei kurzen Spontansprachproben 56
5 Mauchly's Test für Spherizität für die MLU-Werte 58
6 Resultate des RMANOVA-Gesamtmodells für die MLU mit der Greenhouse-Geisser Korrektur 58
7 Paarweise Vergleiche zwischen den Äusserungslängen für die MLU mit der Bonferroni-Korrektur 59
8 Mauchly's Test für Spherizität für die MATTR-Werte 60
9 Resultate des RMANOVA-Gesamtmodells für die MATTR mit der Greenhouse-Geisser Korrektur 60
10 Paarweise Vergleiche zwischen den Äusserungslängen für die MATTR mit der Bonferroni-Korrektur 61
11 Spearman-Brown Korrelationskoeffizienten für den Vergleich der MLU-Werte der gekürzten Sprachproben mit denjenigen der Gesamtsprachprobe 61
12 Spearman-Brown Korrelationskoeffizienten für den Vergleich der MATTR-Werte der gekürzten Sprachproben mit denjenigen der Gesamtsprachprobe 62
13 Übersicht über die MLU-Werte bei den verschiedenen Sprachprobenlängen 86
14 Übersicht über die MATTR-Werte bei den verschiedenen Sprachprobenlängen 86

Danksagung

Die vorliegende Arbeit wurde im Rahmen des Projekts «Digital unterstützte Spontansprachanalyse - DigiSpon 1» erarbeitet und umgesetzt. Das Forschungsprojekt ist eine Kooperation der HfH Zürich (Institut für Sprache und Kommunikation; Susanne Kempe-Preti, Sonja Schäli), der Universität Freiburg (CH, Institut für Sonderpädagogik; Julia Winkes, Ramona Rüegg, Lena Graf), der Universität Zürich (Institut für Computerlinguistik; Sarah Ebling, Anja Ryser) und der PH Bern (Institut Primarstufe; Pascale Schaller). Ziel ist die Entwicklung und Evaluation einer Software für (semi-) automatische Spracherkennung und Analyse von Kindersprache (Kempe Preti, 2023).

An dieser Stelle möchte ich mich herzlich bei Julia Winkes bedanken, die mir die Möglichkeit geboten hat, meine Masterarbeit in diesem Projekt zu schreiben und mir stets unterstützend zur Seite stand. Bei meinem Betreuer Erich Hartmann möchte ich mich für das wertvolle Feedback bedanken. Für Fragen zur Statistik und Methodik nahm sich Verena Hoffmann Zeit und brauchte mich in meiner Arbeit so enorm weiter - vielen Dank! Zu guter Letzt möchte ich mich bei allen Teammitgliedern für die spannende und die gewinnbringende Zusammenarbeit im DigiSpon 1 - Projekt bedanken. Insbesondere sei an dieser Stelle Anja Ryser erwähnt, die das Tool im letzten Jahr so weit brachte, dass die Datensätze, die ich in dieser Arbeit brauchte, automatisch ausgegeben werden konnten.

1 Einleitung

Eine zentrale Grundlage der logopädischen Arbeit liegt in der Diagnostik von sprachlichen Fähigkeiten. Zu diesem Zweck wurden unzählige Verfahren entwickelt und evaluiert. Neben den standardisierten Testverfahren entwickelte sich dabei die Spontansprachanalyse (fortan auch LSA; Language Sample Analysis genannt) heraus. Bei einer LSA wird die spontane Sprachproduktion der Klient:innen analysiert. Dadurch soll sichergestellt werden, dass mit der logopädischen Diagnostik diejenigen sprachlichen Fähigkeiten erhoben werden, die den Menschen mit sprachlichen Einschränkungen im Alltag tatsächlich zur Verfügung stehen. Die LSA wird in der Sprachtherapie, der Linguistik, der Psychologie und anderen verwandten Bereichen angewandt, um die Sprachentwicklung, die Kommunikationsfähigkeiten und andere Aspekte der Sprache zu untersuchen und daraus gezielte therapeutische Interventionen zu planen. Zu diesem Zweck werden Sprachproben von Klient:innen einerseits hinsichtlich verschiedener sprachsystematischer Indikatoren wie beispielsweise der mittlere Äusserungslänge, der Anzahl verschiedener Wörter oder der Satzkomplexität untersucht. Andererseits können unter anderem Aussagen auf der Ebene der Pragmatik wie plötzliche Themenwechsel, das Turn-Taking oder die Perspektivenübernahme gemacht werden. Zusätzlich hat die LSA einen hohen Stellenwert in Forschung:

«Language sample analysis (LSA) is invaluable to describe and understand child language use and development for clinical purposes and research.» (Lüdtke, Bornman et al., 2023, S. 1).

Trotz der von verschiedenen Autor:innen betonten Wichtigkeit scheint die Spontansprachanalyse in der logopädischen Praxis noch nicht adäquat angewandt zu werden (Pavelko et al., 2016). Eine Ursache dafür liegt im grossen Zeitaufwand der für die Transkription und Auswertung nötig ist (Liu et al., 2023). Eine Möglichkeit, diese Schwelle der Anwendung zu senken, bestünde in der Verkürzung der Spontansprachproben. Es stellten sich daher schon verschiedene Forschende die Frage, wie lang eine Spontansprachprobe sein muss, respektive wie kurz sie sein kann. Verschiedene Studien proklamierten über lange Zeit die These, dass für valide und reliable Messungen ein Minimum von 50 Äusserungen transkribiert und analysiert werden müssen (Casby, 2011; Eisenberg et al., 2001; Paul, 2007; Pavelko et al., 2020). Pavelko et al. (2020) wendeten in ihrem Artikel, indem sie reliable Werte für Sprachproben mit 25 Äusserungen für Kinder mit einer typischen Sprachentwicklung fanden, dagegen unter anderem folgendes ein:

«... it is possible that the procedures used in the current study, which only included children developing typically, may not transfer when obtaining language samples from children with LI[1]. ... however, future research could explore whether the procedures used in this study can be used to accurately, validly, and reliably assess the language skills of children with LI.» (Pavelko et al., 2020, S. 787)

[1]LI steht hier als Abkürzung für *language impairment*, zu Deutsch Sprachentwicklungsstörung

Verstärkt wird die Relevanz der Frage, wie kurz die Sprachproben sein können durch folgende Begebenheit: Seit die Schweiz das Behindertengleichstellungsgesetz (Bundesrat der Schweizerischen Eidgenossenschaft, 2004) ratifiziert hat, ist die inklusive Schule eines der brennendsten und aktuellsten Themen im Bildungssektor. In diesem Kontext wurde unter anderem das *Response-to-Intervention*-Modell entwickelt (Justice, 2006). Ein wesentlicher Bestandteil dieses Konzeptes ist die regelmässige Fortschrittsdiagnostik. Dazu werden effektive und effiziente Tools benötigt. Für den sprachlichen Bereich fehlen diese aber noch weitgehend (Ukrainetz, 2006). Die Spontansprachanalyse hat das Potenzial, diese Lücke zu füllen (Heilmann et al., 2010), auch da im Gegensatz zu standardisierten Test kein testspezifischer Lerneffekt die Resultate verzerren kann (Gallagher & Hoover, 2020; Roberts et al., 2022). Um sicherzustellen, dass die Diagnostik alle 2-3 Wochen durchgeführt werden kann, muss das Diagnostiktool innerhalb weniger Minuten durchführbar sein. Dies erfordert eine kurze Sprachprobe.

Für Kinder mit Spacherwerbsstörungen finden sich unter anderem bei Heilmann et al. (2013) erste Hinweise darauf, dass reliable Werte für einige Indikatoren bereits bei kürzeren Spontansprachproben (3 Minuten) erreicht werden. Bisher wurde jedoch keine Studie publiziert, die dieser Frage in der schweizerdeutschen Sprache nachgeht. Aus diesen Gründen wird in dieser Arbeit die folgenden Fragestellung untersucht:

«Wie lange muss eine Spontansprachprobe bei Kindern mit und ohne Spracherwerbsstörungen im Alter von 4 - 7 Jahren sein, um reliable Werte zu erhalten?»

Um die erwähnte Fragestellung zu beantworten, wird in einem ersten Schritt die theoretische Grundlage der LSA aufgezeigt. Dabei wird zunächst eine Begriffsklärung vorgenommen und anschliessend die Vor- und Nachteile dieser im diagnostischen Prozess und die aktuelle Verwendung in der Praxis diskutiert. Um die Datenerhebung der dieser Arbeit zu Grunde liegenden Sprachproben einordnen zu können, werden anschliessend verschiedene Elizitationsmethoden vorgestellt. In einem nächsten Schritt wird ein kurzer Einblick in die Möglichkeiten und den aktuellen Stand bezüglich der (semi-) automatisierten Transkription und Analyse von Spontansprachanalysen gegeben. Dann wird der Auswahlprozess der untersuchten logopädischen Indikatoren beschrieben und der Forschungsstand zu den ausgewählten Indikatoren (MLU und MATTR) beleuchtet. In einem weiteren Schritt wird das theoretische Konzept der Reliabilität beleuchtet, um darauf aufbauend relevante Forschungsbefunde bezüglicher der Fragestellung zu diskutieren. Aus all diesen Grundlagen werden anschliessend die Fragestellungen und Unterfragen der Arbeit abgeleitet.

Im empirischen Teil dieser Arbeit wird zunächst die Stichprobe, die den Analysen zugrunde liegt deskriptiv beschrieben und die Datenerhebung detailliert beleuchtet. Zusätzlich wird die Transkription und die Analyse der Sprachproben betrachtet. Um die vorliegenden Daten analysieren und interpretieren zu können, wird die vorgenommene

statistische Analyse detailliert beschrieben und diskutiert. Anschliessend werden die gefundenen Resultate präsentiert, diskutiert und mit der bisherigen Forschung in Verbindung gebracht. Zum Schluss werden die Limitationen dieser Arbeit aufgezeigt und ein Ausblick auf mögliche zukünftige Forschung gegeben.

2 Die Spontansprachanalyse

Im folgenden Abschnitt wird erläutert, was eine Spontansprachanalyse ist und wofür sie im logopädischen Alltag eingesetzt wird. Es werden Vor- und Nachteile der LSA diskutiert und eingeordnet.

Die Spontansprachanalyse ist eine diagnostische Methode zur Einschätzung und Bewertung der mündlichen Sprache. Der Einsatz von Spontansprachanalysen im diagnostischen Prozess wird von diversen Autor:innen empfohlen und als *Best Practice* (Scott et al., 2022) oder *Goldstandard* (Charest & Skoczylas, 2019; Ebert & Scott, 2014; Escobedo et al., 2023) bezeichnet. Die spezifische Stärke der Spontansprachanalyse liegt in der Abbildung der alltäglichen kommunikativen Fähigkeiten der untersuchten Person (Overton et al., 2021).

Die Spontansprachanalyse kann mit Menschen einer sehr breiten Altersspanne durchgeführt werden und deckt damit einen sehr grossen Teil des logopädischen Klientels ab, welches aufgrund sprachlicher Schwierigkeiten behandelt wird. Um die Leserlichkeit des folgenden Textes zu erhöhen, wird darauf verzichtet, immer alle davon zu nennen. Stattdessen wird der Fokus sprachlich auf die Kinder gelegt, da diese im Fokus dieser Arbeit stehen. Natürlich lassen sich aber sehr viele der besprochenen Themen auch auf erwachsene Personen übertragen.

Um die sprachlichen Fähigkeiten eines Kindes mit Hilfe einer Spontansprachanalyse zu evaluieren, müssen zunächst einmal Sprachdaten dieses Kindes gesammelt und aufgezeichnet werden. Dazu werden verschiedene Elizitationsmethoden angewandt, die vom Freispiel bis zum Erklären von komplexen Regeln reichen (siehe Kapitel 2.3). Laut einer gemeinhin geltenden Empfehlung müssen diese Sprachdaten mindestens 50 Äusserungen umfassen. In einem weiteren Schritt werden die gesammelten Äusserungen des Kindes nun transkribiert. Dazu werden sie möglichst lauttreu niedergeschrieben. Die lauttreue Verschriftlichung ist zur Analyse der Phonetik-Phonologie unumgänglich. Anschliessend werden die transkribierten Äusserungen des Kindes anhand verschiedener Indikatoren analysiert, die Auskunft über die sprachlichen oder kommunikativen Fähigkeiten des Kindes geben (siehe Kapitel 2.5).

2.1 Stärken und Schwächen der LSA

In der Literatur werden diverse Stärken und Schwächen der LSA beschrieben. In ihrer Summe erklären diese Vor-und Nachteile die aktuelle Verwendung von Spontansprachanalysen in der logopädischen Praxis. Daher werden zunächst die Vorteile und die Nachteile beschrieben, um anschliessend auf die aktuelle Nutzung der LSA einzugehen.

2.1.1 Vorteile

Um zunächst grundlegend zu klären, weshalb die LSA sowohl in der Ausbildung als auch in der logopädischen Praxis und der Forschung einen hohen Stellenwert hat, werden einige Vorteile ins Feld geführt.

Alltagsnähe und hohe ökologische Validität: Ein erster Vorteil der LSA ist, dass durch die Analyse von spontanen Äusserungen geprüft werden kann, über welche sprachlichen Möglichkeiten Klient:innen in realen Alltagssituationen verfügen (Heilmann et al., 2010; Hewitt et al., 2005; Wilder & Redmond, 2022). Dadurch wird eine hohe ökologische Validität erreicht (Charest & Skoczylas, 2019). Das heisst, dass die sprachlichen Fähigkeiten, die ein Kind bei einer LSA zeigt, eine valide Aussage über seine sprachlichen Alltagsfähigkeiten machen. Daher können aufgrund der LSA Therapieziele gezielt abgeleitet werden (Ebert & Scott, 2014; Lundine, 2020; Nippold et al., 2014; Spencer et al., 2020). Dies ist insbesondere im Sinne der *International Classification of Functioning, Disability and Health* (ICF)[2] sehr wünschenswert, da die LSA auch eine Aussage über die sprachliche Partizipation und Teilhabe eines Kindes treffen kann.

Breite Anwendbarkeit: Die LSA kann bei den verschiedensten Störungsbildern und Altersklassen eingesetzt werden (Channell et al., 2018; Costanza-Smith, 2010). So beispielsweise auch bei Menschen mit Down-Syndrom (Thurman et al., 2021) oder einer Autismus-Spektrumsstörung (MacFarlane et al., 2023). Auch eignet sich die Methode für mehrsprachige Kinder (Ebert & Pham, 2017; Heilmann & Westerveld, 2013) und für Sprecher:innen von Dialekten (Ebert & Pham, 2017; Heilmann & Westerveld, 2013; Stockman, 1996), für die es kaum standardisierte Testverfahren gibt.

Keine Testsituation nötig: Standardisierte Tests mit Kindern, die Auffälligkeiten im Verhalten zeigen, durchzuführen, kann eine grosse Herausforderung sein. Mittels Spontansprachanalysen lässt der Sprachstand in diesen Fällen deutlich einfacher einschätzen als mit standardisierten Tests (Wilder & Redmond, 2022). Einen ähnlichen Vorteil bietet die LSA bei jungen Kindern im Vorschulalter (circa 2 bis 4 Jahre). Da viele der standardisierten Tests mit jungen Kindern, unter anderem aufgrund von motivationalen Hindernissen, kaum durchführbar sind, fällt der Spontansprachanalyse eine essentielle Rolle im diagnostischen Prozess zu (Eisenberg et al., 2001; Lüdtke, Ehlert et al., 2023). Verstärkt wird

[2]Das Klassifikationssystem der World Health Organization (WHO)

dies dadurch, dass die LSA bereits bei Kindern mit 2-Wort-Äusserungen eingesetzt werden kann (De Anda et al., 2023).

Analyse auf allen Sprachebenen: Durch die alltagsnahe Form der LSA ist es möglich, spezifische Indikatoren auf den Ebenen der expressiven Sprache (Wortschatz, Grammatik, Aussprache) zu analysieren. Darauf zielen auch sehr viele standardisierte Tests ab. Zusätzlich besteht die Option die Pragmatik und Kommunikation in den Fokus zu nehmen. Dies spiegelt sich beispielsweise insbesondere in der Forschung zum Einsatz der LSA bei Kinder mit einer Autismus-Spektrumsstörung (MacFarlane et al., 2023). Dies ist in standardisierten Verfahren oft schwer erreichbar.

2.1.2 Nachteile

Neben oder mit diesen Vorteilen bringt die LSA auch Nachteile mit sich. Einige davon sind im folgenden aufgeführt.

Begrenzte Generalisierbarkeit der Sprachproben: Einer der Nachteile der Spontansprachanalyse liegt darin, dass möglicherweise innerhalb der erhobenen Sprachprobe nicht die maximalen sprachlichen Fähigkeiten des getesteten Kindes abgebildet werden. Kinder scheinen beispielsweise je nach Form der Erhebung und besprochenem Thema unterschiedlich komplexe Sätze zu bilden (Westerveld & Vidler, 2016). Diese Schwäche der LSA findet sich auch innerhalb der Wortschatzbeurteilung. Wenn sich ein Kind in einem Thema sehr wohl fühlt und viele Wörter kennt, wird womöglich verpasst, dass in anderen semantischen Feldern viele Wörter fehlen (Kapantzoglou et al., 2019). Wenn das Ziel der Diagnostik also ist, die maximalen sprachlichen Fähigkeiten eines Kindes zu evaluieren, ist die Spontansprachanalyse womöglich nicht die optimale Wahl.

Fehlende psychometrische Normen: Aktuell ist die Spontansprachanalyse noch kaum nach psychometrischen Testnormen evaluiert. In den allermeisten Sprachen - auch in der schweizerdeutschen - fehlt es an Vergleichsdaten aus der Grundgesamtheit. Dadurch können auch keine statistischen Normen berechnet werden. Verschiedene Forschungsprojekte, die sich mit der Automatisierung der Transkription und Analyse von Sprachproben befassen, versuchen dies aktuell anzugehen (siehe Kapitel 2.4). Die dünne Studienlage stellt zusätzlich ein Fragezeichen hinter die oben erwähnte Empfehlung für die Verwendung der LSA bei Dialektsprechenden. Aufgrund der hohen Anzahl Variablen, die zu Einschätzung dieser aktuell noch ausgewertet werden müssten, scheint die Anwendung von standardisierten Test aktuell klinisch noch praktikabler (Ramos et al., 2022). Auch dies könnte sich in Zukunft durch den Einbezug von künstlicher Intelligenz ändern.

Spontan ist nicht gleich spontan: Spontansprachproben werden mit verschiedenen Techniken, sogenannten Elizitationsmethoden, erhoben. Der Name «Spontansprachanalyse» impliziert, dass die analysierte Sprache spontan produziert wurde. Dies trifft genau genommen aber nur auf Sprachdaten zu, die in unstrukturierten und unbeobachteten

Situationen erhoben wurden (also beispielsweise zu Hause während des Alltags). Eine Strukturierung des Erhebungskontextes (also beispielsweise eine gemeinsame Bilderbuchbetrachtung), wie es in der Praxis oftmals gemacht wird, erhöht zwar in einem gewissen Masse die Vergleichbarkeit der Daten, führt aber zu weniger Spontaneität und Natürlichkeit (Lüdtke, Ehlert et al., 2023). Dieses Dilemma lässt sich kaum auflösen.

Grosser Zeitaufwand: Ein letzter Nachteil der LSA besteht im grossen Zeitaufwand, den die Durchführung einer LSA bedeutet. Insbesondere die Transkription der aufgenommenen Sprachprobe braucht ein Vielfaches der Länge der Aufnahme. Im logopädischen Alltag ist die Aufwendung dieser Zeit kaum realistisch. Der grosse Zeitaufwand wird wird im nächsten Abschnitt nochmals vertieft besprochen.

Diese Nachteile zeigen, dass die Spontansprachanalyse die standardisierte Tests nicht ersetzen kann und soll. Sie dient als Ergänzung dieser und führt zu einer umfassenden Beschreibung der sprachlichen Fähigkeiten eines Kindes.

2.2 Aktuelle Anwendung

Nachdem nun die Vor- und Nachteile der Spontansprachanalyse beleuchtet wurden, soll der Blick auf die aktuelle Anwendung der LSA in der logopädischen Praxis gerichtet werden. Die Studienlage bezüglich der Anwendung von Spontansprachanalysen in der logopädischen Praxis ist uneinheitlich. Während viele der befragten logopädischen Fachpersonen in den Studien von Kemp und Klee (1997), Bawayan und Brown (2022) (beide in den USA) und Westerveld und Claessen (2014) (Österreich) angaben, regelmässig Spontansprachanalysen durchzuführen, deuten andere Studienresultate (Klatte et al., 2022; Pavelko et al., 2016) auf eine deutlich seltenere Anwendung der LSA hin. Wie auch Bawayan und Brown (2022) anführen, erklären sich diese Unterschiede neben den unterschiedlichen Abrechungssystemen in den verschiedenen Ländern vermutlich vor allem durch die Selbstselektion der Studienteilnehmenden (also der Logopäd:innen). Möglicherweise nehmen Logopäd:innen, die häufig Spontansprachanalysen durchführen, eher an einer Befragung zu diesem Thema teil als andere. Gleichzeitig nehmen Logopäd:innen, die selten Spontansprachanalysen durchführen möglicherweise eher an Befragungen teil, die die Hindernisse diesbezüglich evaluieren. Klatte et al. (2022) untersuchten in ihrer Studie, die in Holland durchgeführt wurde, die Hindernisse, die Logopäd:innen von der Anwendung der LSA im Alltag abhalten. Dabei fanden sie insbesondere die folgenden Hindernisse:

- Die Fachpersonen gaben an, dass ihnen das Wissen und die Fähigkeiten zu Umsetzung der LSA fehlen.
- Die LSA wird teilweise nicht als Inhalt der professionellen Rolle verstanden.
- Die logopädischen Fachpersonen haben kaum Erfahrung mit der LSA.

- Sie haben auch Zweifel, ob sich die aufgewendete Zeit im Vergleich zum Endresultat lohnt.

- Teilweise werden die Kosten nicht übernommen, und dies bei einem gleichzeitig sehr grossen Zeitaufwand.

Vergleichbare Resultate finden sich ebenfalls bei Pavelko et al. (2016), wobei die limitierten Ressourcen in dieser Studie am häufigsten genannt wurden[3]. Dass Spontansprachanalysen sehr zeitaufwändig sind, spiegelte sich in denjenigen erwähnten Studien, in denen die Logopäd:innen angaben die LSA häufig anzuwenden: Eine Mehrheit der Logopäd:innen gaben an, die LSA meist in Form von Echtzeit-Transkriptionen durchzuführen (Bawayan & Brown, 2022; Kemp & Klee, 1997; Pavelko et al., 2016; Westerveld & Claessen, 2014). Bei Echtzeit-Transkriptionen werden die Äusserungen des Kindes direkt während der Interaktion notiert. Das heisst, es wird weder eine Tonband- noch eine Videoaufnahme für die Durchführung der LSA verwendet. Diese Praxis wird von vielen Forschenden nicht empfohlen, weil es dabei gehäuft zu Ungenauigkeiten kommen kann (Evans & Miller, 1999; Heilmann et al., 2010; Nippold et al., 2014). Viele der befragten logopädischen Fachpersonen (42 %) gaben in der Befragung von Pavelko et al. (2016) an, zwischen 26 und 50 Äusserungen eines Kindes zu analysieren[4], womit vermutlich viele unter der empfohlenen Menge von mindestens 50 Äusserungen bleiben.

Zusammenfassend kann also folgendes festgehalten werden: Spontansprachanalysen weisen diverse Vor- und Nachteile auf. Sie haben aber eine grosse Stärke, die standardisierte Tests kaum leisten können. Sie geben Auskunft darüber, welche sprachlichen Kompetenzen dem Kind im Alltag zur Verfügung stehen. Unter anderem aufgrund des grossen zeitlichen Aufwands werden Spontansprachanalysen dennoch entweder zu wenig oder im Vergleich zur empfohlenen Mindestanzahl von 50 Äusserungen zu kurz gehalten angewandt. Diese Befunde bringen uns zurück zur eingangs gestellten Frage, wie lange eine Sprachprobe sein muss, um reliable Werte zu generieren. Bislang ist nicht abschliessend geklärt, inwiefern kürzere Sprachproben zu reliablen oder eben nicht reliablen Werten führen. Wie später in dieser Arbeit vertieft ausgeführt wird, gibt es einige Evidenz für die Reliabilität von kurzen Spontansprachproben (siehe Kapitel 2.7.3). Die vorliegende Arbeit trägt einen Teil zur Beantwortung dieser Frage bei, indem einerseits zum ersten Mal Daten der schweizerdeutschen Sprache evaluiert werden und andererseits eine breite Auswahl an Schätzmethoden für die Reliabilität angewandt wurden.

[3]Keine der genannten Studien zur Anwendung der Spontansprachanalyse wurde in der Schweiz durchgeführt. Daher sind Aspekte wie die finanzielle Abdeckung nicht zwingend auf den hiesigen Kontext übertragbar. Insbesondere der Aspekt der fehlenden Zeit und Ressourcen für die LSA dürfte auf die Schweiz aber sicherlich zutreffen.

[4]35 % gaben an, 51 -100 Äusserungen zu sammeln. In Bezug auf die Dauer der Spontansprachproben in Minuten gaben 27 % an, 1 bis 5 Minuten aufzunehmen. 46 % wählen die Option 6 bis 10 Minuten. Die restlichen Antwortgebenden wählten eine kürzere der eine längere Option.

Auch durch die Digitalisierung und Automatisierung der LSA könnten einige der Anwendungshindernisse (insbesondere der zeitliche Aufwand für die Transkription) geschmälert werden (siehe Kapitel 2.4). Dies wiederum würde zu einer höheren Wahrscheinlichkeit führen, dass die LSA qualitativ hochwertig in der Praxis angewandt wird. An dieser Hoffnung knüpft das Projekt DigiSpon 1 für die schweizerdeutsche Sprache an, in dessen Rahmen die hier vorliegende Arbeit entstanden ist.

2.3 Elizitationsmethoden

Um die im DigiSpon 1 angewendete Elizitationsmethode einordnen zu können, werden im folgenden Abschnitt die verschiedenen Möglichkeiten aufgezeigt, mit welchen eine LSA elizitiert werden kann. In der Vergangenheit wurden die verschiedensten Methoden vorgestellt, welche im Folgenden besprochen werden. Zu beachten ist, dass bei vielen dieser Methoden die Äusserungen des Kindes durch die Elizitation nicht im engeren Sinne natürlich respektive spontan sind - der Begriff Spontansprachanalyse ist in dieser Hinsicht durchaus irreführend, wird aber in der logopädischen Praxis auch für elizitierte Sprachproben verwendet[5]. Im Folgenden findet sich eine Übersicht über die verschiedene Kategorien von Methoden[6]. Eine detaillierte Zusammenfassung der Studienlage der Elizitationsmethoden, also wie sich die Elizitationsmethode auf die gesammelten Sprachdaten auswirkt, findet sich in Voniati et al. (2021).

Freispiel: Wenn eine Sprachprobe von kleinen Kindern (ungefähr bis zum Alter von 4 Jahren) erhoben werden soll, wird meist eine Freispielsequenz zur Erhebung der Sprache verwendet. Dabei werden dem Kind verschiedene Spielmaterialien angeboten, welche dann in einem gemeinsamen Spiel verwendet werden. Die logopädische Fachperson wendet motivierende Techniken (wie z.B. das Wiederholen von Äusserungen des Kindes) an, um möglichst viele Äusserungen des Kindes zu evozieren. Die Erhebung der spontanen Sprache während des Freispiels ist zusammen mit der dialogischen Erhebungsmethode am nächsten an einer natürlichen Unterhaltung (Westerveld, 2019).

Dialog (Conversational): Bei der dialogischen Elizitationsmethode führt die logopädische Fachperson ein interviewartiges Gespräch mit dem Kind durch. Dabei ist es laut Westerveld (2019) wichtig, dass ein Thema aufgegriffen wird, welches das Kind interessiert. Dadurch kann mit dieser Methode eine Fähigkeit geprüft werden, welche in jedem Alter im Alltag sehr wichtig ist: Ein Gespräch führen und sich auf das Gegenüber einzustellen. Zusätzlich ist es dank der dialogischen Struktur möglich, die pragmatischen Kompetenzen zu beurteilen.

[5]Treffender ist der im englischen Sprachraum verwendete Begriff "Language Sample Analysis", da dieser nicht enger definiert, wie die Sprachproben erhoben werden.

[6]In der Literatur finden sich keine einheitlichen Namen für diese Kategorien. Die hier vorgestellten Kategorien orientieren sich an oft verwendeten Überbegriffen.

Geschichten erzählen / nacherzählen: Bei der Methode «Geschichten erzählen» wird meist mit Bildergeschichten gearbeitet. Dabei gibt es verschiedene Optionen: Entweder beginnt die logopädische Fachperson, indem sie die Geschichte zuerst erzählt und das Kind nacherzählen soll (englisch *retell* genannt) oder das Kind beginnt damit, die Geschichte frei zu erzählen (englisch *tell* genannt). Sofern immer wieder eine ähnliche Geschichte verwendet wird, führt das Geschichten erzählen zu einer hohen Vergleichbarkeit der Daten. Verstärkt wird dies dadurch, dass sich das Verhalten der erhebenden Person bei der Elizitationsmethode ziemlich einfach standardisieren lässt (Channell et al., 2018). Einige Studien zeigen, dass durch das Geschichten erzählen längere Äusserungen evoziert werden als mittels Freispiel und Dialog (Southwood & Russell, 2004; Wofford et al., 2022). Eine Gefahr der Verwendung von Nacherzählungen ist, dass die Gedächtnisleistungen einen verzerrenden Einfluss auf die Resultate haben können (Lüdtke, Ehlert et al., 2023).

Narrativ: Narrative Sprachproben bestehen entweder aus fiktiven Geschichten oder aus persönlichen Berichten des Kindes. Von Erlebtem zu erzählen gehört für die allermeisten Kinder zum alltäglichen Sprachgebrauch und ist daher sehr nahe an ihrer Lebensrealität (Westerveld, 2019). Die narrativen Erzählungen können beispielsweise über eine Frage oder ein Bild evoziert werden.

Erklärender Diskurs (Expository discourse): Die Methode des *expository discourse* besteht darin, das Kind einen Sachverhalten wie beispielsweise die Spielregeln seines Lieblingsspiels erklären zu lassen. Diese Methode scheint sich insbesondere dann anzubieten, wenn komplexe Satzstrukturen evoziert werden sollen (Westerveld & Moran, 2011).

Es bestehen aktuell keine einheitlichen Empfehlungen, welche Elizitationsmethode verwendet werden sollte. Bei der Auswahl sollten sicherlich Faktoren wie das Alter des Kindes und das Ziel der LSA einbezogen werden. Zusammenfassend kann die Spannweite der Elizitationsmethoden und deren Implikationen so beschrieben werden:

> «The more natural (unstructured) the elicitation context the more representative the sampled language, but also the longer it takes to collect (and then transcribe and code) a desired amount of language to calculate LSA measures, because the frequency of targeted structures (e.g., complex language) may be much lower in natural communication.» (Lüdtke, Ehlert et al., 2023, S. 41)

Für das DigiSpon-Projekt wurde in Anlehnung an Westerveld und Gillon (2002) ein Leitfaden für die erhebenden Personen erstellt. Dies sollte dem Ziel dienen, der Erhebungssituation einen standardisierten Rahmen zu geben, da viele verschiedene Personen die Erhebungen durchführten. Im Kapitel 4.1 wird das Vorgehen der Datenerhebung im DigiSpon 1-Projekt detailliert beschrieben.

2.4 Automatisierte Spontansprachanalyse

Die vorliegende Arbeit entstand im Projekt DigiSpon 1. Dieses beschäftigt sich mit der (semi-) automatischen Transkription und Analyse von schweizerdeutschen Spontansprachproben. Die automatisierte Transkription von schweizerdeutschen Sprachdaten soll mit Hilfe künstlicher Intelligenz (KI) ermöglicht werden. Durch die kommerzielle Bekanntheit, die diese im letzten Jahr insbesondere durch Programme wie ChatGPT erhalten hat, ist die KI endgültigen in unserem Alltag angekommen. Es ist dringend angezeigt, die Möglichkeiten, die solche Technologien für die logopädische Arbeit eröffnen, wissenschaftlich genauer zu beleuchten und nützliche Tools daraus zu generieren. Ein erster Schritt in diese Richtung versucht das DigiSpon 1 -Projekt zu gehen. Die Vorteile, die die Anwendung der automatisierten Transkription und Analyse bringen könnte, sind unter anderem folgende:

«Additionally, automated transcription and coding would strengthen the applicability of LSA and if long-dated, it could even be expanded to fully develop its potential in providing insights into multilingual language learning and how it is used within communicative interaction in natural settings, including children and their caregivers, peers, teachers and other communication partners.» (Lüdtke, Ehlert et al., 2023, S. 41)

Für den englischen und den standarddeutschen Sprachraum gibt es bereits Forschungsprojekte, die die automatisierte Transkription und Analyse von Sprachproben elaborieren. DigiSpon 1 untersucht nun die Anwendungsmöglichkeiten in der schweizerdeutschen Sprache. Insbesondere im englischen Sprachraum arbeiten Forschende schon lange an der automatisierten Transkription und Analyse von Spontansprachproben. Eine in vielen wissenschaftlichen Publikationen verwendete Software dazu ist SALT (Systematic Analysis of Language Transcripts der University of Wisconsin-Madison[7]). Im deutschen Sprachraum arbeiten Forschende der Leibniz Universität Hannover seit einigen Jahren an einer Software namens TALC (Tools for Analyzing Language and Communication[8]). Für eine aktuelle Übersicht der vorhandenen Software wird auf Lüdtke, Bornman et al. (2023) verwiesen. Mehrere Faktoren erschweren die Entwicklung einer entsprechenden Software in der schweizerdeutschen Sprache (Ryser, 2023):

i. Dialekt: Schweizerdeutsch ist eine kleinräumige Sprache, für die es wenig Datenmaterial gibt.

ii. Kindersprache: Kindersprache weist eine grössere Variabilität auf als Erwachsenensprache, was das Lernen des Algorithmus erschwert.

iii. Gestörter Spracherwerb: Durch die Auffälligkeiten erhöht sich die Variabilität zusätzlich und stellt spezielle Anforderungen an die Software.

[7] für mehr Informationen siehe: https://www.saltsoftware.com

[8] für mehr Informationen siehe: https://www.leibnizlab-communication.uni-hannover.de/de/forschung/projekte/talc

Eine ausführliche Diskussion der Herausforderungen im DigiSpon-Projekt, der technischen Umsetzung und ethischer Aspekte findet sich in Ryser (2023).

Aktuell ist das DigiSpon-Tool in der Lage einige Indikatoren automatisiert zu berechnen, wenn ein Transkript ins Programm geladen wird (mehr dazu im Kapitel 2.5). Die automatische Transkription ist aktuell noch in Arbeit, da zu diesem Zweck die Sprachdaten benötigt werden, die innerhalb des DigiSpon 1-Projektes gesammelt wurden. Leider konnten bis zum Abschluss dieser Arbeit noch nicht alle Sprachproben transkribiert werden und daher auch keine Aussage über die automatisierte Transkription des Tool getroffen werden.

2.5 Logopädische Indikatoren

Über die Zeit wurden verschiedenste Indikatoren für die Analyse von Sprachproben vorgestellt. Die Breite dieser Indikatoren reicht über alle linguistischen Sprachebenen von der Pragmatik bis zur Aussprache. Aus diesem Grund wird in diesem Kapitel zunächst aufgezeigt, welche Indikatoren aktuell mit dem DigiSpon 1-Tool analysiert werden können. Dies bereitet die Grundlage für die Diskussion im anschliessenden Teil, welche dieser Indikatoren für die Analysen in dieser Arbeit ausgewählt wurden. Dazu werden Kriterien präsentiert, anhand derer die Indikatoren verglichen werden. Auch soll dieses Kapitel für diese Indikatoren einen kurzen Forschungsüberblick bieten.

2.5.1 Indikatoren im DigiSpon 1

Zum Zeitpunkt der Konzeptualisierung dieser Arbeit konnte das DigiSpon 1-Tool folgende Indikatoren auswerten:

- Anzahl Äusserungen des Kindes (TNU[9])
- Mittlere Äusserungslänge (MLU[10])
- Anzahl Wörter (TNW[11])
- Anzahl verschiedener Wörter (NDW[12])
- Type-Token Ratio[13] (TTR)

[9]TNU = Total Number of Utterances; die Anzahl der Äusserungen werden addiert. Im Falle des DigiSpon 1-Tools entspricht eine Äusserung dem Gesagten zwischen zwei Sprecherwechseln, also von dem Zeitpunkt an, an dem das Kind die Sprecherrolle übernimmt bis es sie zum nächsten Mal abgibt.

[10]MLU = Mean Length of Utterance; wird im DigiSpon 1-Tool als Durchschnitt der verwendeten Wörter pro Äusserung berechnet.

[11]TNW = Total Number of Words; alle geäusserten Wörter des Kindes werden addiert.

[12]NDW = Number of Different Words; die Anzahl der voneinander verschiedenen Wörter werden gezählt. Wenn ein Wort bereits im Geäusserten vorkam, wird dies nicht noch einmal gezählt.

[13]Das Verhältnis der Anzahl verschiedener Wörter und der gesamten Anzahl Wörter

- Moving-Average Type-Token Ratio[14] (MATTR)

- Brùnet's Index[15]

- Honoré-Statistik[16]

Seither kamen laufend weitere Indikatoren dazu, welche allerdings aufgrund des zeitlichen Ablaufs nicht für die Auswahl der Indikatoren dieser Arbeit in Frage kommen:

- Verteilung der Wortarten

- Subjekt-Verb Kongruenz

- Falsche Pluralbildung

Die Auswertung von Indikatoren auf der pragmatisch-kommunikativen Ebene, dem Sprachverständnis und der Aussprache waren zum Abgabezeitpunkt dieser Arbeit noch nicht möglich.

2.5.2 Auswahlkriterien

Der Auswahl der Indikatoren wird im Folgenden detailliert besprochen, da nicht davon auszugehen ist, dass sich die gefundenen Resultate beim einen Indikator auf andere Indikatoren übertragen lassen. Dies liegt daran, dass verschiedene Indikatoren verschiedene Eigenschaften aufweisen:

«Evidence is growing that some measures, such as MLU, may be less influenced by sample length than measures that evaluate the linguistic content of the sample in more detail.» (Lüdtke, Ehlert et al., 2023, S. 39)

Ein Ziel innerhalb der Auswahl der Indikatoren war es, verschiedene sprachliche Ebenen abzudecken. Aufgrund der aktuell verfügbaren Indikatoren, die das Tool auswerten konnte, beschränkte sich die Auswahl auf die syntaktische Ebene respektive die Äusserungslänge und die semantisch-lexikalische Ebene. Auf ersterer stand ausschliesslich die MLU zur Auswahl. In Bezug auf die semantische Ebene war die Auswahl grösser. Da aber sowohl der Brùnet's Index als auch die Honoré-Statistik im logopädischen Alltag im Moment kaum oder gar keine Relevanz haben, wurden diese in der weiteren Diskussion nicht berücksichtigt und nur die anderen Indikatoren (TNW, TTR und MATTR) weiter verfolgt.

Charest et al. (2020) folgend ergeben sich bezüglich der Anwendung von logopädischen Indikatoren verschiedene Faktoren aus der wissenschaftlichen Literatur, die es zu berücksichtigen gilt:

[14]Die TTR wird über ein festgelegtes Fenster von Wörtern berechnet, welches immer ein Wort weiter verschoben wird. Der Durchschnitt all dieser Werte bildet die MATTR.

[15]Ebenfalls ein Indikator der lexikalischen Vielfalt, der auf Types und Tokens basiert

[16]Verwendet eine logarithmische Formel, um die Types und Tokens in ein Verhältnis zu stellen.

i. Wenn aus den gemessenen Indikatoren Wissen über spezifische Fähigkeiten des Kindes abgeleitet werden, sollten sie frei von Verzerrungen durch andere Faktoren sein. Diese Faktoren reichen von der Sprachprobenlänge bis hin zum Satzbau (Owen & Leonard, 2002). Wenn also beispielsweise die lexikalische Vielfalt als Indikator für die lexikalischen Fähigkeiten interpretiert werden soll, ist es wichtig, dass dieser Indikator nicht von anderen Faktoren wie der Satzbildungsfähigkeit beeinflusst wird. Ist dies nicht der Fall, wird womöglich etwas anderes - in diesem Fall die Satzbildungsfähigkeit anstatt der lexikalischen Fähigkeiten - gemessen.

ii. Für den sinnvollen Einsatz eines Indikators als Mass für sprachliche Fähigkeiten und den Fortschritt dieser, sollten die Indikatoren einen Entwicklungsverlauf über das Alter aufweisen (Channell et al., 2018; Eisenberg et al., 2001)

iii. Um einen Beitrag zur Identifikation einer Spracherwerbsstörung beitragen zu können, sollte der Indikator für die Charakteristiken der entsprechenden Population sensitiv sein, das heisst er sollte zwischen Kinder mit und ohne Spracherwerbsstörungen differenzieren können.

iv. Im Hinblick auf die evidenzbasierte Praxis sollten neben den Gruppenunterschieden (Punkt iii.) auch beachtet werden, inwiefern aus den Indikatoren der individuelle Förderbedarf abgeleitet werden kann.

Die Vor- und Nachteile der genannten Indikatoren der Äusserungslänge und der Semantik werden im Folgenden diskutiert, bevor ein Fazit zum Forschungsstand und den genannten Kriterien gezogen wird.

2.5.3 Indikator Äusserungslänge

Die mittlere Äusserungslänge ist ein viel verwendeter Indikator in der logopädischen Praxis und bezeichnet die durchschnittliche Länge der Äusserungen eines Kindes. Eisenberg et al. (2001) betonen, dass die MLU nicht als morpho-syntaktischer Indikator verstanden werden sollte, sondern als eine von mehreren Möglichkeiten, wie die Äusserungslänge eines Kindes gemessen werden kann. Längere Äusserungen sind nicht zwingend morpho-syntaktisch komplexer, wie diese Autor:innen an folgenden Beispielsätzen der Berechnung des MLU in Morphemen zeigen:

i. Want more cookies Mommy

ii. I want to go home

Beide Sätze bestehen aus 5 Morphemen[17] und haben damit dieselbe Äusserungslänge. Während der erste Satz aber ein einfacher, ungrammatikalischer Satz ist, ist der zweite, grammatikalisch korrekte Satz deutlich komplexer.

Die Berechnung des MLU ist in der Forschung nicht eindeutig definiert, wobei drei Faktoren unterschiedlich gehandhabt werden:

i. Segmentierung von Äusserungen: Unter Äusserungen können z.B. Sätze oder Sinneinheiten verstanden werden. Es kann aber auch nach Atempausen oder der fallenden Intonation am Ende einer Äusserung segmentiert werden[18].

ii. Analyseebene: Die Länge einer Äusserung kann entweder auf Wort- oder auf Morphemebene gezählt werden. Innerhalb der Zählung in Morphemen ergibt sich wiederum eine Diskussion, was alles als ein Morphem gezählt werden soll[19].

iii. Ausschluss von Äusserungen: Nicht immer werden alle Äusserungen in die Analyse der MLU miteinbezogen. So können beispielsweise unverständliche Äusserungen oder Ellipsen[20] ausgeschlossen werden.

Eine Übersicht, wie diese Faktoren in der Vergangenheit umgesetzte wurden, findet sich in Eisenberg et al. (2001). Diese Uneinheitlichkeit in der Definition der MLU erschwert die Vergleichbarkeit der Resultate der bisherigen Studien. Dazu kommt, dass die verwendete Berechnungsart in der Vergangenheit nicht immer klar beschrieben wurde.

Die MLU zeigt für eine breite Altersspanne (4 - 21 Jahre) eine signifikante Entwicklung mit zunehmendem Alter wie Channell et al. (2018) für den narrativen Kontext in der englischen Sprache zeigen konnten. Beim Alter von 18;6 Jahren[21] zeigte sich in den Daten ein Plateau, das heisst ab diesem Alter fand sich kein weiterer Zuwachs der MLU. Diverse weitere Untersuchungen fanden ebenfalls eine statistisch signifikantes Wachstum des MLU mit zunehmendem Alter (Heilmann et al., 2010; Miller, 1991; Owens & Pavelko, 2020). Ähnliche Resultate wurden auch mit anderen Elizitationsmethoden (konversationell, expository) gefunden (Rice et al., 2006; Rice et al., 2010). Für die schweizerdeutsche (und auch hochdeutsche) Sprache gibt es aktuell keine entsprechenden Befunde.

[17]Das Morphem bezeichnet die kleinste Spracheinheit, die eine grammatische Funktion hat. Ein Wort kann entweder aus einem einzelnen Morphem bestehen oder aus mehreren zusammengesetzt sein. So können beispielsweise Mehrzahlendungen oder Vorsilben Morpheme bilden, wobei ein Morphem nicht gleichbedeutend mit einer Silbe ist. Beispiele: i) Läufer in Silben «Läu-fer», in Morphemen «Läuf-er» ii) zerreissen in Silben «zer-rei-ssen», in Morphemen «zer-reiss-en»

[18]Teilweise werden diese Indikatoren etwas anders genannt. Wenn beispielsweise nach Satzstrukturen segmentiert wird, kann der Indikator auch «Mean Length of C-Units» (MLCU) genannt werden.

[19]Auch hier werden die Namen des Indikators teilweise leicht angepasst. So wird beispielsweise eine mittlere Äusserungslänge in Morphemen oft MLU-m genannt.

[20]In der Linguistik bezeichnen Ellipsen unvollständige Satzstrukturen, bei denen ein Satzteil oder ein Wort ausgelassen wird. Im Alltag werden Ellipsen häufig angewandt und werden in der mündlichen Sprache nicht als inkorrekt gewertet. In den folgenden Ellipsen wurde der weggelassene Teil der in eckige Klammern gesetzt: «Was [machen wir] nun?», «Erst [kommt] die Arbeit, dann das Vergnügen.», «Dieser Zug geht nach Bern. Nein, [dieser Zug geht] nach Zürich.»

[21]Diese Schreibweise notiert das Alter in folgender Art und Weise: Jahre;Monate.

Diverse Studien untersuchten die Spezifität und die Sensitivität der MLU entweder als Einzelindikator oder kombiniert mit anderen Indikatoren. Ohne andere Indikatoren zeigen diese Studien eine hohe Variablilität an Resultaten wie Ramos et al. (2022) in ihrem systematischen Review zeigen. Durch eine Kombination mit anderen Indikatoren konnte die Genauigkeit erhöht werden. So kamen beispielsweise Pavelko und Owens (2019a) mit dem «Sampling Utterances and Grammatical Analysis Revised (SUGAR) protocol» auf eine Sensitivität und Spezifität von jeweils 86 %. Gemäss diesen Resultaten sollte die MLU sicherlich nicht als einzelner Indikator zur Diagnostik einer Spracherwerbsstörung eingesetzt werden. Kombiniert werden könnte es beispielsweise mit dem *finite verb morphology composite*[22]. Dieser zeigt eine hohe diagnostische Genauigkeit für Kinder zwischen 4 und 6 Jahren (Ramos et al., 2022).

Um die MLU als Indikator einordnen zu können, werden zusammenfassend die Kriterien von Charest et al. (2020) beschrieben:

i. Es ist zum aktuellen Zeitpunkt noch nicht vollständig geklärt, wie stark die MLU von anderen Faktoren wie der Sprachprobenlänge beeinflusst wird. Auch, ob sie von anderen sprachlichen Fähigkeiten abhängig ist oder nicht, ist nicht abschliessend geklärt.

ii. Die MLU zeigt sich - zumindest in der englischen Sprache - als entwicklungssensitiv für das Alter der, in der vorliegenden Arbeit, untersuchten Kinder.

iii. Als Einzelindikator zur Unterscheidung zwischen Kindern mit und ohne Spracherwerbsstörung eignet sich die MLU nicht. In Kombination mit anderen Indikatoren ist dieses Kriterium eher gegeben.

iv. Die MLU kann einen Hinweis auf mögliche Therapieziele geben. Um ein spezifisches Ziel zu setzen, sind in vielen Fällen neben der MLU zusätzliche Informationen nötig.

Die MLU scheint demnach einige der Kriterien zu erfüllen. Andere wie beispielsweise die Entwicklungssensitivität in der schweizerdeutschen Sprache können aktuell nicht abschliessend beurteilt werden und erfordern weitere Forschung. Aufgrund des aktuellen Forschungsstandes scheint es auf jeden Fall lohnenswert zu sein, die MLU weiter zu untersuchen. Sie wurde daher in der vorliegenden Arbeit bezüglich der Reliabilität in schweizerdeutschen Sprachproben vertieft analysiert.

2.5.4 Indikator lexikalische Vielfalt

Die lexikalische Vielfalt (englisch *lexical diversity*) ist einer der zentralen Indikatoren der Semantik. Die Operationalisierung der lexikalischen Vielfalt erfolgt als ein Wert, der die Diversität der verwendeten Wörter über eine definierte Länge einer Sprachprobe misst.

[22]Der FVMC ist ein grammatikalischer Indikator, der die Markierung von Verben beurteilt.

Dieser Wert soll die Breite und den Abruf des aktiven Wortschatzes widerspiegeln (Charest et al., 2020).

Neben diversen anderen Arten, die lexikalische Vielfalt zu messen[23], wurden in der Forschung insbesondere die NDW (*number of different words*), die TTR (*type-token ratio*) und die MATTR (*moving average type-token ratio*) diskutiert. Die NDW bezeichnet die Anzahl verschiedener Wortstämme, die in einem Text von einer definierten Länge vorkommen. Die TTR teilt die Anzahl die Anzahl *word tokens* (Anzahl verschiedener Wörter) durch *word types* (Anzahl Wörter). Die TTR kann Werte zwischen 0 und 1 annehmen, wobei diese der Lesbarkeit halber oft statt in Dezimalzahlen in ganzen Zahlen dargestellt wird (also multipliziert mit 100). Je grösser diese Zahl, desto mehr verschiedene Wörter werden im analysierten Text verwendet. Die MATTR wird als durchschnittlicher TTR über eine Serie von Teilsprachproben bestimmter Länge (z.B. 50 Wörter) berechnet. In anderen Worten bedeutet dies, dass dieses Fenster von beispielsweise 50 Wörtern nach und nach über den gesamten Text verschoben wird (Wörter 1-50, Wörter 2-51, Wörter 3-52, und so weiter) und in diesen Fenstern die TTR berechnet wird. Zum Schluss wird aus diesen TTR-Werten der Mittelwert berechnet, was dann der MATTR entspricht (Charest et al., 2020). Die TTR wird als Option für den Indikator der semantischen Entwicklung in der vorliegenden Arbeit aus verschiedenen Gründen ausgeschlossen:

i. Die TTR ist abhängig von der Sprachprobenlänge, indem der TTR-Wert mit zunehmender Länge sinkt (Owen & Leonard, 2002). Je länger eine Sprachprobe ist, desto höher wird die Wahrscheinlichkeit, dass das nächste Wort bereits schon einmal verwendet wurde. Je häufiger Wörter wiederholt werden, desto tiefer wird der TTR-Wert. Dies liegt unter anderem daran, dass der Korpus einer Sprache der Zipf'schen Verteilung folgt. Diese besagt, dass wenige Wörter sehr hochfrequent in einem Korpus vorkommen, während viele Wörter sehr niederfrequent vorkommen. Es gibt also Wörter wie beispielsweise «der», die sehr häufig verwendet werden, während viele andere wie beispielsweise «Tannenzapfen» deutlich seltener auftauchen. Je häufiger ein Wort wie «der» in einer Sprachprobe verwendet wird, desto niedriger wird der TTR-Wert liegen.

ii. Die TTR bildet keinen Entwicklungsverlauf über das zunehmende Alter ab (Watkins et al., 1995).

iii. Diverse Studien zeigen, dass die TTR nicht zwischen Kindern mit und ohne Spracherwerbsstörung differenziert (Thordardottir & Weismer, 2001; Watkins et al., 1995). Yang et al. (2022) zeigten in ihrer Studie gar, dass die Anwendung des TTR viele

[23]Hier sei beispielhaft der *measure of lexical richness* erwähnt, der die relative Frequenz innerhalb des Korpus der verwendeten Wörter in die Messung einbezieht und dadurch der natürlichen Erwerbsreihenfolge in der Sprachentwicklung gerechter wird (Van Hout & Vermeer, 2007). Dieser und weitere Indikatoren werden an dieser Stelle nicht weiter betrachtet, da DigiSpon diese aktuell noch nicht berechnen kann.

Kinder mit verifizierten Spracherwerbsstörungen als Kinder mit typischem Spracherwerb und umgekehrt kategorisiert hätte.

Aufgrund der erhöhten Wahrscheinlichkeit für die Wiederholung von Wörtern mit zunehmender Testlänge, ist auch die NDW als Summe der verschiedenen Wörter abhängig von der Textlänge. Die NDW muss daher über eine klar definierte Länge einer Sprachprobe gemessen werden, um vergleichbar zu sein (Charest et al., 2020). Es wurden vorgeschlagen, dazu eine bestimmte Anzahl Äusserungen oder Wörter der zu verwenden[24]. Charest et al. (2020) verglichen in ihrer Studie unter anderem diese beiden Methoden und die MATTR in Bezug auf ihre Eignung als logopädische Indikatoren. Die stärksten Resultate fanden die Autor:innen, wie auch andere vor ihnen für die NDW, berechnet anhand von einer festgelegten Anzahl Äusserungen. Diese Berechnungsart ist entwicklungssensitiv (Charest et al., 2020; Miller, 1991) und differenziert zwischen Kindern mit und ohne Sprachentwicklungsstörungen (Charest et al., 2020; Heilmann et al., 2010; Hewitt et al., 2005; Watkins et al., 1995). Sowohl Miller (1991) als auch Charest et al. (2020) zeigten allerdings, dass diese Werte sehr stark mit der MLU korrelieren. Dies ist vermutlich darauf zurückzuführen, dass längere Äusserungen mit einer hohen Wahrscheinlichkeit auch mehr unterschiedliche Wörter beinhalten und daher auch die Vielfalt grösser wird. Verschiedene Autoren kamen zum Schluss, dass mit diesen beiden Indikatoren nahezu identische Konzepte der produktiven Sprache gemessen werden (Charest & Skoczylas, 2019).

Bezüglich der NDW für eine bestimmte Anzahl Wörter fanden sowohl Watkins et al. (1995) als auch Charest et al. (2020) und Yang et al. (2022), dass der Indikator entwicklungssensitiv ist, wenn auch mit einer kleinen Effektstärke. Kaum eine Untersuchung fand einen signifikanten Unterschied im NDW für eine bestimmte Anzahl Wörter zwischen Kindern mit und ohne Sprachentwicklungsstörung (Charest & Skoczylas, 2019). Die NDW wird in dieser Studie nicht zur Verwendung empfohlen. Aufgrund dieser Befunde wird dieser Indikator auch in dieser Arbeit nicht weiter verfolgt.

Eine weitere Methode, um dem Problem der Abhängigkeit von der Sprachprobenlänge zu begegnen, ist die MATTR (*moving average type-token ratio*). Für diesen Indikator gibt es aktuell nur wenige, sich teilweise widersprechende Studien.Wu et al. (2019) untersuchten die MATTR mit Mandarin sprechenden Kindern. Sie fanden keine Unterschiede zwischen 3- und 4-jährigen Kindern, aber signifikante Gruppenunterschiede zwischen Kindern mit und ohne Spracherwerbsstörung. Charest und Skoczylas (2019) fanden keine Gruppenunterschiede zwischen Kindern mit und ohne Spracherwerbsstörungen. In der bereits erwähnten Studie von Charest et al. (2020) zeigte sich die MATTR zwar als entwicklungssensitiv aber als nicht oder kaum statistisch differenzierend zwischen Kindern mit und ohne Spracherwerbsstörung. Eine Studie von Kapantzoglou et al. (2019), in der

[24]Daneben wurden auch andere Längenmasse wie zum Beispiel ein Sample vorgeschlagen. Diese Einheiten passen aber insbesondere auf narrative Sprachproben und nicht auf die Art von Spontansprachproben, die in DigiSpon erhoben wurden. Sie werden daher auch nicht weiter ausgeführt.

vier lexikalische Indikatoren (MTLD[25], MATTR, TTR, D[26] und HD-D[27]) verglichen wurden, kam zum Schluss, dass die MATTR unter diesen Indikatoren die beste Messvariante für die lexikalische Diversität darstellt. Dies, weil die MATTR den grössten Anteil an Varianz der lexikalischen Diversität abbilden konnte.

Insgesamt scheint keine der vorgelegten Option ein perfekter logopädischer Indikator zu sein. Eine Verwendung der NDW nach einer bestimmten Anzahl Äusserungen ist für die vorliegende Arbeit nicht sinnvoll, da die Gefahr bestehen würde, ein und das selbe Konzept zwei Mal auf die Reliabilität zu prüfen. Im direkten Vergleich von Charest et al. (2020) zeigte die MATTR unter den restlichen Indikatoren die besten Resultate. Daher werden die vorgestellten Kriterien für diesen Indikator zusammengefasst:

Auch die MATTR wird zusammenfassend nach den Kriterien von Charest et al. (2020) beschrieben:

i. Die MATTR scheint - im Gegensatz zu anderen Indikatoren der semantischen Vielfalt - kaum von der Äusserungslänge beeinflusst. Auch korreliert sie weniger stark mit der MLU.

ii. Die MATTR zeigt sich - zumindest in der englischen Sprache - als entwicklungssensitiv für das Alter der, in der vorliegenden Arbeit, untersuchten Kinder.

iii. Als Einzelindikator zur Unterscheidung zwischen Kindern mit und ohne Spracherwerbsstörung eignet sich die MATTR nicht.

iv. Aus der MATTR kann insofern ein Therapieziel abgeleitet werden, als dass das Verhältnis von Types und Tokens auf Lücken im Wortschatz hindeuten kann. Um ein spezifisches Therapieziel abzuleiten sind aber weitere Informationen notwendig.

Auch die MATTR zeigt bezüglich dieser Kriterien einige Schwächen. Um diesen Indikator ohne Vorbehalte in der logopädischen Praxis einsetzen zu können, ist weitere Forschung nötig. Dies insbesondere bezüglich der Entwicklungssensitivität in der schweizerdeutschen Sprache und der Unterscheidungsfähigkeit zwischen Kindern mit und ohne Spracherwerbsstörungen. Auch scheint es sinnvoll, die Entwicklung von neuen Indikatoren, möglicherweise auch solchen, die nicht auf einer *type-token ratio* basieren, in der Forschung weiter zu verfolgen. Innerhalb der Indikatoren, die das DigiSpon-Tool aktuell

[25]MTLD steht für *measure of textuel, lexical diversity* und bildet die durchschnittliche Anzahl aufeinander folgender Wörter ab, innerhalb derer die TTR aufrecht erhalten wird, bevor die TTR aufgrund der Textlänge absinkt. Weitere Infos finden sich in Kapantzoglou et al. (2019).

[26]D bezeichnet einen Indikator, der entwickelt wurde, um die Schwächen der TTR anzugehen. Um den Indikator D zu berechnen, wird analysiert, wie schnell die TTR mit zunehmender Sprachprobenlänge abnimmt. Wenn D einen niedrigen Wert ausweist, deutet dies auf einen schnellen Abfall des TTR im Laufe des Textes hin, was auf einen begrenzten Wortschatz hindeutet. Für mehr Details siehe zum Beispiel Kapantzoglou et al. (2019).

[27]HD-D ist eine Version des Indikators D, der auf einer anderen Verteilung basiert. Weitere Infos finden sich ebenfalls in Kapantzoglou et al. (2019).

analysieren kann, scheint die Analyse des MATTR für die vorliegende Arbeit dennoch am sinnvollsten.

2.6 Gütekriterien der Testevaluation

Die Reliabilität, die Validität und die Objektivität bilden die drei Gütekriterien der Testevaluation. In der bisherigen Forschung wurden die verschiedensten Methodiken verwendet, um die Reliabilität von Spontansprachanalysen zu bestimmen. Um diese einordnen und kritisch betrachten zu können, wird in einem ersten Schritt die Reliabilität als theoretisches Konzept betrachtet und in ein Verhältnis zu den anderen Gütekriterien gestellt.

In der Literatur finden sich verschiedene Definition der Reliabilität. Eine davon lautet: «Test or assessment data are reliable to the degree to which they can be replicated or reproduced.» (Downing & Yudkowsky, 2009, S. 59)

Für die Einordnung der vorliegenden Arbeit leistet die folgende Erläuterung zum Konzept der Reliabilität einen wichtigen Beitrag:

«Die Reliabilität wird auch als Messgenauigkeit bezeichnet. Der Begriff «Messgenauigkeit» verleitet jedoch zu einer Fehlinterpretation. In der Umgangssprache wird man beispielsweise von einer Waage, die sehr genau misst, auch annehmen, dass sie das Gewicht richtig anzeigt. Dies wäre aber eine Fehlinterpretation des Reliabilitätskonzepts. Etwas sehr reliabel zu messen bedeutet lediglich, dass die Messung kaum durch unsystematische Fehler (Messfehler) gestört wird. Ob die Waage wirklich das Gewicht oder der Intelligenztest wirklich die Intelligenz misst, ist eine Frage der Validität.» (Schmidt-Atzert & Amelang, 2012, S. 137)

Für die Einordnung der Resultate der Reliabilität von Spontansprachanalysen in der vorliegenden Arbeit ist diese Erläuterung essentiell, um Missverständnisse zu vermeiden. Die nachfolgende Analyse wird eine Aussage darüber machen, wie stark die Messung der Indikatoren durch zufällige Fehler gestört ist. Die Frage der Validität - also ob die Indikatoren in dieser Erhebungsform tatsächlich die intendierten sprachlichen Fähigkeiten messen, sollte Gegenstand einer zukünftigen Forschungsarbeit sein, da sie eine wesentliche Voraussetzung für die Güte eines diagnostischen Verfahrens ist. Dazu könnte beispielsweise ein Vergleich mit einem standardisierten, bereits evaluierten Testverfahren erfolgen.

Das dritte der Testkriterien bildet die Objektivität. Die Objektivität geht der Frage nach, inwiefern die Messbedingungen standardisiert wurden und damit vergleichbar sind. Die drei genannten Gütekriterien diagnostischer Verfahren - Reliabilität, Validität und Objektivität - bedingen einander gegenseitig (Moosbrugger & Kelava, 2020) und sollten daher alle untersucht werden. In der vorliegenden Arbeit liegt der Fokus auf der Reliabilität. Die Objektivität und die Validität könnten zukünftig in weiteren Studien untersucht

werden. Daher wird im nächsten Teil dieser Arbeit die Reliabilität genauer beschrieben und auf den Forschungsstand der Spontansprachanalyse bezogen.

2.7 Reliabilität

Das Gütekriterium der Reliabilität kann durch diverse Schätzverfahren beschrieben und quantifiziert werden. Die Resultate dieser Schätzverfahren tragen unterschiedliche Kennwerte zu einem Test bei. Diese Erkenntnisse sind vergleichbar mit den Einzelteilen eines Puzzles - für ein umfassendes Verständnis eines diagnostischen Verfahrens sind sie alle unverzichtbar und nicht untereinander austauschbar (Schmidt-Atzert & Amelang, 2012). Im folgenden wird ein Überblick über die verschiedenen Aspekte der Reliabilität gegeben. In der vorliegenden Arbeit können und werden aufgrund der vorliegenden Daten nicht alle davon untersucht werden. Sie zu beschreiben ist jedoch relevant, um die in dieser Arbeit vorgenommenen Analysen als Einzelteile im Puzzle der Reliabilität einzubetten.

2.7.1 Schätzmethoden

Um die unterschiedlichen Schätzmethoden der Reliabilität zu veranschaulichen, wird ein fiktives Beispiel eines Sprachverständnistests durch das folgende Kapitel führen. Nehmen wir an, dieser Test beinhaltet 16 Items, die verschiedene Kompetenzen im Bereich des Sprachverständnisses abdecken. Jeweils vier Items testen folgende Kompetenzen: das Verständnis von Einzelwörtern, einfachen Sätze, Präpositionen und Negationen. Die nachfolgend beschriebenen Schätzmethoden werden jeweils auf diesen fiktiven Test angewandt, um deren Beitrag zum Gesamtbild der Reliabilität zu veranschaulichen.

Testwiederholungsreliabilität Beim Verfahren der Testwiederholungsreliabilität (englisch: *retest-reliability* genannt) wird ein diagnostischer Test mit jeder Person einer Stichprobe zwei Mal durchgeführt. Der Kennwert der Testwiederholungsreliabilität ergibt sich aus der Korrelation der beiden Durchführungen. Die zeitliche Distanz der Durchführungen spielt dabei eine essentielle Rolle, da die Reliabilität durch die Stabilitätseigenschaften des Merkmals beeinflusst werden kann (Schmidt-Atzert & Amelang, 2012). Dabei gibt es Merkmale wie beispielsweise die Intelligenz, die deutlich stabiler sind als andere. Insbesondere können entweder Übungs- und Erinnerungseffekte oder tatsächlich veränderte Persönlichkeitsmerkmale einen Einfluss auf die Testwerte haben (Moosbrugger & Kelava, 2020). Wenn eine Testwiederholungsreliabilität eines stabilen Merkmals innert einiger Tage durchgeführt wird, sind hohe Reliabilitätskoeffizienten zu erwarten. Abweichungen von einer perfekten Reliabilität wären entsprechend entweder der Testkonstruktion oder Faktoren wie der Aufmerksamkeit zuzuschreiben.

In Bezug auf den fiktiven Sprachverständnistest wäre folgendes Vorgehen zur Erhebung der Testwiederholungsreliabilität denkbar: Eine Kindergartenklasse bildet die Stichprobe anhand derer der Test überprüft werden soll. Nun wird der Test mit jedem Kind

der Klasse zwei Mal mit einem Abstand von 1-2 Wochen durchgeführt. Der Datensatz besteht also pro Kind aus zwei Testresultaten für jedes Item. Werden diese beiden Testresultate verglichen, ergibt sich daraus die Testwiederholungsreliabilität. Eine Herausforderung dieses Verfahrens liegt darin, dass die Kinder der Klasse eventuell bei der ersten Durchführung des Tests spezifische Strukturen oder Wörter neu erlernt haben, die sie aufgrund dieses Lerneffektes in der zweiten Durchführung nun adäquater lösen können. Zusätzlich könnten die Kinder in der Zeit zwischen den Testdurchführungen Fortschritte in ihrer Sprachentwicklung gemacht haben. Die Veränderungen zwischen den Durchführungen müssen daher genau betrachtet und analysiert werden. Diese Form der Reliabilität gibt uns Auskunft darüber, wie stabil ein Test ein Merkmal über die Zeit misst oder ob es sehr stark von anderen Faktoren wie der Motivation abhängt.

Paralleltest-Reliabilität Bei der Paralleltestreliabilität wird ein zweiter, paralleler Test zum originalen Test erstellt. Beide Testversionen werden mit derselben Stichprobe durchgeführt. Durch die Durchführung von zwei parallelen Testverfahren können einige der sogenannten reliabilitätsverändernden Faktoren eliminiert werden. Reliabilitätsverändernde Faktoren sind beispielsweise die fortlaufende Entwicklung oder Lerneffekte, aber auch die Aufmerksamkeit und Motivation. Mithilfe der Paralleltestreliabilität können insbesondere die ersten beiden (Lerneffekte und die fortlaufende Entwicklung) kontrolliert werden, da zwischen den Durchführungen weniger Zeit verstreichen muss. Die zwei erstellten Testversionen sollten aus unterschiedlichen Items bestehen, aber dasselbe Konzept messen. Die Paralleltestreliabilität wird bestimmt, indem die Resultate der beiden Testversionen miteinander verglichen werden. Wenn die beiden Versionen parallel sind, ist davon auszugehen, dass bei einer empirischer Überprüfung dieselben Verteilungen der Testwerte feststellbar sind und entsprechend eine hohe Reliabilität resultiert (Schmidt-Atzert & Amelang, 2012). In der Praxis wird diese Schätzmethode aufgrund des enormen Konstruktionsaufwandes nur selten angewendet (Moosbrugger & Kelava, 2020).

In Bezug auf den fiktiven Sprachverständnistest müsste zur Bestimmung der Paralleltestreliabilität ein zweiter Test mit wiederum 16 Items erstellt werden. Auch dieser müsste die vier Kompetenzen (Verständnis von Einzelwörtern, einfachen Sätzen, Präpositionen und Negationen) testen und dies auf einem äquivalenten Niveau wie der originale Test. Mit jedem Kind aus der Kindergartenklasse, die unsere Stichprobe bildet, würden beide Testversionen durchgeführt (hier möglichst ohne zeitlichen Unterschied). Der Datensatz besteht also für jedes Kind aus den Resultaten der beiden Testversionen. Um die Paralleltestreliabilität zu berechnen, würden die Resultate der beiden Testversionen verglichen. Ergäbe diese Berechnung eine hohe Paralleltestreliabilität, würde dies darauf hinweisen, dass die beiden Testversionen das Sprachverständnis beide in einer ähnlichen Art und Weise und damit stabil messen.

Testhalbierungsreliabilität Die Testhalbierungsreliabilität (englisch: *splithalf-reliability*) dient der Überprüfung der instrumentellen Reliabilität. Die instrumentelle Reliabilität untersucht die effektive Messgenauigkeit eines Tests, indem die Einflüsse der reliabilitätsverändernden Faktoren (insbesondere Schwankungen in der Stimmung, Motivation, Aufmerksamkeit und so weiter) möglichst eliminiert werden. Zur Evaluation der Testhalbierungsreliabilität wird der untersuchte Test nur einmal mit der Stichprobe durchgeführt. Daher können die Schwankungen der reliabilitätsverändernden Faktoren auf diese eine Erhebung minimiert werden. Im Vergleich dazu können sie sich bei der Testwiederholungsreliabilität beispielsweise abhängig von der Tagesform deutlich stärker unterscheiden (Schmidt-Atzert & Amelang, 2012). Während die ersten beiden Schätzmethoden (Testwiederholung und Paralleltest) Informationen bezüglich der praktischen Umsetzung der Tests liefern, liegt der Fokus der Testhalbierungsreliabilität auf der Konsistenz der Items innerhalb des Tests. Zur Bestimmung der Testhalbierungsreliabilität werden die Items eines diagnostischen Verfahrens in zwei Hälften aufgeteilt und die Korrelation zwischen den Testhälften bestimmt. Voraussetzung dafür ist, dass die Items homogen sind und eine Aufteilung in zwei parallele Hälften erlauben (Schmidt-Atzert & Amelang, 2012). Parallel bedeutet in diesem Kontext, dass die Hälften gleich viele Items enthalten und die Items der beiden Hälften vergleichbar sind. Die Korrelation der Testhalbierungsreliabilität fällt in der Regel niedriger aus, als dies bei der Gesamtreliabilität der Fall wäre. Dies liegt daran, dass Korrelationskoeffizienten unter anderem auch von der Anzahl Items abhängen und durch die Aufteilung der Items auf zwei Testhälften weniger Items in die Berechnung einfliessen. Es wird daher eine Korrekturformel (oftmals die Spearman-Brown-Formel) eingesetzt, mit der der Korrelationskoeffizient auf die ursprüngliche Testlänge hochgerechnet werden kann (Moosbrugger & Kelava, 2020).

Um die Testhalbierungsreliabilität für den fiktiven Sprachverständnistest zu berechnen, müsste folgendes getan werden: Mit jedem Kind der Kindergartenklasse würde der Sprachverständnistest ein Mal durchgeführt. Der Datensatz besteht bei dieser Schätzmethode also nicht aus zwei Sets, die verglichen werden können. Um dennoch einen Vergleich zu ziehen, würden für die Testhalbierungsreliabilität die 16 Items in zwei Hälften geteilt. Beispielsweise könnten aus jeder getesteten Kompetenz zwei Items der ersten Hälfte und zwei Items in die zweiten Hälfte zugeteilt werden. Diese beiden Hälften wären parallel. Würden jedoch alle Items des Verständnisses von Einzelwörtern und einfachen Sätzen der einen Hälfte zugeteilt und die anderen beiden Kompetenzen der anderen Hälfte, wären diese nicht parallel, weil sie unterschiedliche Kompetenzstufen abbilden. Wenn nun also zwei parallele Hälften der Items gebildet wurden, könnte daraus die Testhalbierungsreliabilität berechnet werden. Da aber aufgrund dieses Vorgehens nur noch 8 Items miteinander verglichen würden und nicht mehr je 16, würde der Korrelationskoeffizient aufgrund der geringeren Itemzahl tiefer ausfallen. Daher würde der errechnete Korrelationskoeffizient mittels der Spearman-Brown Formel auf die gesamte Länge des Tests hochgerechnet wer-

den. Wenn dieser Korrelationskoeffizient hoch ist, würde dies auf eine hohe instrumentelle Reliabilität hinweisen. Das heisst, dass die Items das zu Grunde liegende Konzept konsistent messen.

Interne Konsistenz Das Konzept der internen Konsistenz basiert auf der Überlegung, dass ein Test nicht nur in zwei Hälften, sondern in viele kleine Teile zerlegt werden kann. In der Berechnung werden entsprechend alle einzelnen Items wie einzelne Test behandelt, die miteinander verglichen werden (Brotz & Döring, 2006). Auch mit der internen Konsistenz kann eine Aussage über die instrumentelle Reliabilität gemacht werden, da die reliabilitätsverändernden Faktoren bei dieser Erhebungsform ebenfalls kontrolliert werden. Zur Schätzung der internen Konsistenz wird der Test nur einmal mit einer Stichprobe durchgeführt. Der Datensatz besteht also wie bei der Testhalbierungsreliabilität aus einem Set an Resultaten. Eine mögliche spezifische Berechnungsvariante der internen Konsistenz ist das Cronbach's Alpha (Schmidt-Atzert & Amelang, 2012).

Um die interne Konsistenz des Sprachverständnistests zu berechnen, würde der Test gleich wie bei der Testhalbierungsreliabilität mit jedem Kind der Kindergartenklasse nur ein Mal durchgeführt. Anstatt die Items nun in zwei Hälften aufzuteilen, werden nun alle Items miteinander verglichen. Mathematisch kann dazu beispielsweise das Cronbach's Alpha verwendet werden, welches Werte zwischen 0 und 1 annehmen Kann. Je näher das Cronbach's Alpha bei 1 liegt, stärker deutet dies auf eine hohe Korrelation der einzelnen Items untereinander hin.

Interrater-Reliabilität Um zu überprüfen, inwiefern die Bewertungen von verschiedenen Prüfenden übereinstimmen, kann eine Interrater-Reliabilität berechnet werden (Schmidt-Atzert & Amelang, 2012). Dazu werden die erhobenen Tests von zwei oder mehr verschiedenen Prüfenden beurteilt. Der Korrelationskoeffizient entsteht durch den Vergleich der verschiedenen Beurteilungen.

In Bezug auf den fiktiven Sprachverständnistest würde der Test jedem Kind der Kindergartenklasse ein Mal durchgeführt. Die Durchführungen könnten gefilmt werden. Aufgrund dieser Videoaufzeichnungen könnten nun verschiedene Prüfende die Leistungen der Kinder beurteilen. Das Datenset bestünde bei dieser Form der Reliabilität pro Kind also aus verschiedenen Sets, die jeweils eine Beurteilung durch eine prüfende Person darstellen. Um die Interrater-Reliabilität zu bestimmen, würden diese Resultate miteinander verglichen. Je ähnlicher die Resultate der verschiedenen Prüfenden sind, desto höher ist die Interrater-Reliabilität.

All diese verschiedenen Schätzmethoden tragen zum Gesamtbild der Reliabilität eines Tests bei. Für die vorliegende Arbeit wurde mit jedem Kind eine Sprachprobe erhoben und einmalig durch das DigiSpon-Tool bezüglich der Indikatoren analysiert. Die Schätzmethoden der Testwiederholungsreliabilität, der Paralleltestreliabilität und der Interraterreliabilität können daher im Rahmen dieser Arbeit nicht untersucht werden. Die

instrumentelle Reliabilität (Testhalbierungsreliabilität und die interne Konsistenz) wurde aber untersucht. Die Fragestellung dieser Arbeit versucht zu klären, wie lange eine Sprachprobe sein muss, um reliable Werte zu erhalten. Die bisher vorgestellten Schätzverfahren beziehen sich auf die Reliabilität des gesamten Tests. Um die Fragestellung zu beantworten, wurden die Sprachproben gekürzt. Auch um die Reliabilität dieser gekürzten Sprachproben zu bestimmen, können verschiedene Schätzmethoden verwendet werden. Diese sollen im nächsten Abschnitt diskutiert werden.

2.7.2 Reliabilität von kurzen Sprachproben

Zur Schätzung der Reliabilität von kurzen Spontansprachproben werden in der Literatur diverse der vorgestellten Schätzmethoden verwendet. Eine Unterscheidung, die in der wissenschaftlichen Diskussion in den letzten Jahren an Bedeutung gewonnen hat, ist diejenige zwischen der relativen und der absoluten Reliabilität.

Die relative Reliabilität misst die Konstanz, in der eine Person ihre Rangposition bei wiederholten Messungen behält (Wilder & Redmond, 2022). Eine hohe relative Reliabilität kommt also dann zu Stande, wenn die relative Position der einzelnen Proband:innen zwischen verschiedenen Messzeitpunkten konstant bleibt. Um beim fiktiven Beispiel des Sprachverständnistests zu bleiben, bedeutet eine perfekte relative Reliabilität folgendes: Mit allen Kindern wird der Sprachverständnistest einmal durchgeführt. Aufgrund der Resultate des gesamten Tests kann eine Rangliste der Kinder erstellt werden, die nach der Anzahl korrekt gelöster Items geordnet wird. Anschliessend wird der Test beispielsweise um die Hälfte gekürzt und die Testwerte der Kinder anhand dieser Items neu berechnet. Nun wird erneut eine Rangliste der Kinder erstellt. Bei einer perfekten relativen Reliabilität sind die beiden Ranglisten identisch. Je mehr Kinder innerhalb dieser Ranglisten die Plätze getauscht haben, desto tiefer wird die relative Reliabilität sinken. Eine Möglichkeit der Schätzung der relativen Reliabilität ist die Berechnung von Korrelationskoeffizienten. Häufig verwendet werden zum Beispiel Pearson-Korrelation, Cronbach's Alpha[28] oder Spearman-Brown-Korrelation. Der Korrelationskoeffizient wird mit r gekennzeichnet.

Die absolute Reliabilität hingegen gibt an, wie gross die Variation innerhalb einer Person in den Punktwerten über verschiedene Messungen hinweg ist. In Bezug auf den fiktiven Sprachverständnistest würde also wiederum das Resultat des gesamten Tests und einer verkürzten Version für jedes Kind der Stichprobe berechnet. Nun steht allerdings nicht mehr die Rangfolge der Kinder im Fokus sondern die mögliche Differenz zwischen

[28]Das Cronbach's Alpha ist eine weit verbreitete mathematische Formel, die sich verschiedentlich einsetzten lässt. In die Berechnung fliessen verschiedene Variablen ein (Moosbrugger & Kelava, 2020). Weil das Cronbach's Alpha als Werkzeug verstanden werden kann, um zu überprüfen, wie sich verschiedene Items zueinander verhalten, wird es sowohl zur Berechnung der internen Konsistenz eines Tests als auch der relativen Reliabilität von kurzen Sprachproben verwendet. Für die mathematische Berechnung bildet für Letzteres die gesamte Sprachprobe ein Item und die kurze Sprachprobe ein zweites Item. Aus diesen beiden Item kann entsprechend das Cronbach's Alpha berechnet werden.

den beiden Resultaten (des gesamten und des gekürzten Tests). Es wird also untersucht, ob sich die Resultate der Kinder absolut verändern. Auf Gruppenebene können die Mittelwerte der beiden Testresultate unter Einbezug der Varianzen mit Verfahren wie dem linearen Modell (zum Beispiel ANOVA, RMANOVA oder *Mixed-Model Random Effects Analysis*) oder auch der Generalisierungstheorie (G-Theorie) berechnet werden. Die G-Theorie untersucht den Einfluss verschiedener Varianzquellen auf die Messung und zerlegt die Gesamtvarianz in einzelne Varianzkomponenten (Moosbrugger & Kelava, 2020). Diese wird insbesondere für die Analyse der Interrater-Reliabilität eingesetzt (Revelle & Condon, 2019). Auch die *Coefficient of Variance (CV)* zählt zu den Messmethoden der absoluten Reliabilität. Dabei wird die Standardabweichung in Bezug zum Mittelwert gesetzt und im Endeffekt als Prozentzahl ohne Masseinheit dargestellt. Um die absolute Reliabilität auf individueller Ebene zu betrachten, können beispielsweise Bland-Altman Plots (B-A Plots) verwendet werden (Pavelko et al., 2020). In dieser grafischen Vergleichsmethode von zwei Messverfahren (oder eben kurzen und langen Sprachproben) werden die Differenzen der Resultate eingetragen. So können diese genauer betrachtet und beurteilt werden.

All diese Methoden, die Reliabilität von gekürzten Sprachproben zu beurteilen, werden im Kapitel 4.4 nochmals aufgenommen und diskutiert. Grundsätzlich wären all dies Möglichkeiten, sich der Fragestellung dieser Arbeit anzunähern. Welche davon tatsächlich verwendet wurden und weshalb, wird im genannten Kapitel beschrieben. Zunächst soll nun aber der aktuelle Forschungsstand bezüglich der Reliabilität von Spontansprachproben aufgezeigt werden.

2.7.3 Bisherige Forschungsbefunde

Für die vorliegende Arbeit sind insbesondere die Testhalbierungsreliabilität respektive die interne Konsistenz und die Reliabilität von kurzen Sprachproben relevant. Auf publizierte Studien zu diesen Themenfeldern wird im nächsten Abschnitt eingegangen. Die allermeisten Studien stammen dabei aus dem englischen Sprachraum. Wenn im folgenden nicht spezifisch auf die Erhebungssprache der Kinder hingewiesen wird, stammt die Studie aus dem Englischen. Da auch andere Schätzmethoden zum Gesamtverständnis der Reliabilität sehr wichtig sind, werden auch diese zum Schluss dieses Kapitels kurz besprochen.

Instrumentelle Reliabilität Cole et al. (1989) untersuchten Freispiel-Sprachproben von Kindern (n = 10) mit Entwicklungsauffälligkeiten im Alter von 4;6 - 6;8, die mittels SALT analysiert wurden. Es wurde die Testhalbierungsreliabilität berechnet. Dabei ergab sich bei der Berechnung eines t-Tests für die MLU und die lexikalische und morphologische Produktion eine hohe Reliabilität (kein signifikanter Unterschied zwischen den Hälften, $r = 0.94$). Für die Type-Token-Ratio fanden McKee (2000) bei Kindern im Alter von

27 - 33 Monaten eine akzeptable Korrelation[29] ($r = 0.76$). Eine weitere Untersuchung der Testhalbierungsreliabilität stellte Guo et al. (2019) anhand von narrativen Sprachproben von Kindern im Alter zwischen 4 und 9 Jahren mit typischer Sprachentwicklung und Spracherwerbsstörungen an. Für die *percentage of grammatical utterances* (PGU[30]) fanden sie über alle Altersklassen eine hohe Testhalbierungsreliabilität (rs > 0.63, ps < 0.001). Guo et al. (2021) untersuchten in der selben Altersspanne mit und ohne Spracherwerbsstörungen die *clausal density*[31] bei narrativen Sprachproben mittels Pearson's Korrelationskoeffizient. Es zeigte sich eine starke Reliabilität über alle Altersklassen (rs > 0.54, ps < 0.001).

Insgesamt liegen für die instrumentellen Reliabilität bei Spontansprachanalysen nur wenige Studien vor. In allen bestehenden Studien zeigt sich über verschiedene Elizitationsmethoden, Altersgruppen und Indikatoren hinweg generell eine gute Testhalbierungsreliabilität. Das Cronbach's Alpha wurde in keiner gefundenen Studie zur Bestimmung der instrumentellen Reliabilität verwendet. Wenn es eingesetzt wurde, dann für den Vergleich verschiedener Sprachprobenlängen (siehe im nächsten Abschnitt).

Sprachprobenlänge Bezüglich der Frage, wie lange eine Spontansprachprobe sein muss, wurden vergleichsweise viele Studien publiziert. Es besteht allerdings eine grosse Herausforderung bezüglich all dieser Studien:

«Studies focused on the length/size of the language sample vary greatly e.g., according to the unit of comparison that is used (e.g., utterances, minutes, tokens, number of elicitation materials), the measures that are calculated (e.g., MLU, D, TTR, FVMC, TAPS, TNW, TDW[32]), the sample sizes that are used in the comparisons (e.g., > 200 vs. 100 utterances, 100 vs. 50/25/12 utterances), the position of the subsamples (e.g., beginning/middle/end of the transcript, transcript split into two halves, even/uneven number of tokens), and the statistical method used to assess measurement reliability (e.g., relative vs. absolute reliability). This diversity complicates clustering and synthesis of results on this component of the LSA process.» (Lüdtke, Ehlert et al., 2023, S. 33)

Zusätzlich scheint es auch keinen Konsens bezüglich der minimalen Höhe der Reliabilität zu geben. Einige Autor:innen (Gavin & Giles, 1996; Guo & Eisenberg, 2015; Heilmann et al., 2010) sprechen aber ab $r > 0.7$ von einer ausreichenden Reliabilität,

[29] In dieser Studie wurde der Pearson's Korrelationskoeffizient berechnet. Die Interpretation der Höhe dieses Korrelationskoeffizienten unterscheidet sich stark voneinander. In der Studie von Guo et al. (2019) wird der Koeffizient folgendermassen interpretiert: 0 - 0.1 entspricht einem schwachen Zusammenhang, 0.1 - 0.5 einem mittleren Zusammenhang und ab 0.5 wird von einem starken Zusammenhang gesprochen.

[30] Bei diesem Indikator wird wird aus den grammatikalisch korrekten und inkorrekten Sätzen ein Prozentsatz gebildet.

[31] Ein Mass für die Verwendung komplexer Satzstrukturen.

[32] Dies sind Abkürzungen für verschiedene Indikatoren. Der Vollständigkeit halber werden sie an dieser Stelle aufgeschlüsselt: D = Instrument zur Messung der Wortschatzdiversität, das auf der Wahrscheinlichkeit neuer Wörter basiert; FVMC = finite verb morphology composite, ein syntaktisch-morphologischer Indikator; TAPS = A Language Processing Skills Assessment, ein standardisiertes Vorgehen zur Analyse von LSA; TNW = total number of words

andere erst ab $r > 0.9$ und dies teilweise innerhalb eines Schätzverfahrens wie die Ausführungen zur Pearson's Korrelation gezeigt haben. Auch dies erschwert den Vergleich zwischen den Studien.

Um diesen Abschnitt zu strukturieren, werden zunächst Studien besprochen, die Kinder mit typischer Sprachentwicklung untersuchten und anschliessend diejenigen, die entweder Kinder mit und ohne oder nur solche mit Sprachentwicklungsstörungen einbezogen. Zudem werden zunächst Studien besprochen, die einen in dieser Arbeit analysierten Indikator untersuchten.

Gavin und Giles (1996) publizierten eine der ersten Studien mit dieser Fragestellung und untersuchten je zwei Freispiel-Sprachproben von Kindern im Alter von 2 - 3 Jahren mit typischer Sprachentwicklung. Die Sprachproben umfassten insgesamt 20 Minuten respektive 175 Äusserungen pro Kind. Mit Hilfe der SALT-Software wurden verschiedene Indikatoren analysiert: TNW, NDW, MLU-m und MSL[33]. Die Autor:innen fanden reliable Werte für die MLU-m ab 75 Äusserungen ($r = 0.74$). Ab 100 Äusserungen erreichten alle Indikatoren (TNW, NDW und MSL) eine adäquate Reliabilität.

Zu gänzlichen anderen Resultaten gelangten Heilmann et al. (2010), die in ihrer Studie 7-minütige dialogische und narrative Sprachproben von Kindern mit typischer Sprachentwicklung (n = 231) im Alter von 2;8 - 13;3 untersuchten. Die Analyse der Daten mittels *repeated measures analysis of variance* (RMANOVA) zeigte sowohl für den Vergleich von 1-minütigen und 3-minütigen Sprachproben mit der gesamten Sprachprobe keine signifikanten Unterschiede für die MLU-m ($p = 0.54$, $n^2 < 0.01$). Auch fanden die Autor:innen keine signifikanten Interaktionen zwischen der Sprachprobenlänge, der Elizitationsmethode und dem Alter der Kinder. Laut den Autor:innen führten die unterschiedlichen Sprachprobenlängen zu maximal 1 % der Variation. Um eine Aussage über die relative Reliabilität treffen zu können, wurde das Cronbach's Alpha berechnet. Dabei ergab sich eine grosse Spannweite von Werten (αs = 0.56 - 0.79), wobei die Werte für die MLU-m $\alpha > 0.7$) und für die NDW ($\alpha > 0.81$) bereits ab einer Sprachprobenlänge von einer Minute reliabel waren. Die absolute Reliabilität wurde mittels *Coefficients of Variation (CV)* berechnet. Generell stellten die Autor:innen fest, dass die Variabilität stieg, je kürzer die Sprachproben waren. Dabei wurden zwischen zwischen 3 und 7 Minuten moderate Anstiege (C_{var} = 0.03 - 0.07) und zwischen 1 und 7 Minuten grössere Unterschiede (C_{var} = 0.04 - 0.14) gefunden. Die Autoren schlussfolgern, dass die Unterschiede zwischen 3 und 7 Minuten klinisch nicht relevant sein dürften und somit auch 3-minütige Sprachproben verwendet werden können, um die MLU-m zu berechnen.

Eine weitere Untersuchung des MLU-m findet sich bei Guo und Eisenberg (2015), die die SALT-Analyse bei Freispiel-Spontansprachproben von Kindern mit typischer Sprach-

[33]NDW = number of different words, hierbei werden die Anzahl verschiedener Wortstämme gezählt; MSL = mean syntactic length, ein Indikator für die Äusserungslänge, bei dem Antworten auf ja/nein-Fragen ausgeschlossen werden

entwicklung zwischen 3;0 und 3;11 (n = 60) anwendeten. Sie verglichen kürzere (1, 3, 5, 7 und 10 Minuten) mit 22-minütigen Sprachproben. Die berechneten RMANOVA ergaben keine signifikanten Unterschiede für die MLU-m zwischen den Sprachprobenlängen ($F < 1.15$, $ps > 0.29$, $n^2 < 0.019$). Die relative Reliabilität erreichte bei einer Untersuchung mit dem Pearson's Korrelationskoeffizient bei 10 Minuten sehr hohe Werte ($r = 0.92$, $p < 0.01$, bei 7 Min: $r = 0.88$, $p < 0.01$)[34]. Neben der MLU-m untersuchten die Autor:innen die NDW und TNW. Bei der NDW unterschied sich nur die 10-minütige Sprachprobe nicht signifikant vom längsten Sample. Bei der TNW unterschied sich nur die 3-minütige Probe signifikant von der längsten Sprachprobe, was auf eine erhöhte Variabilität bei kürzeren Sprachproben hinweisen könnte. Beide erlangten mit 10-minütigen Aufnahmen hohe Korrelationskoeffizienten.

Eine neuere Untersuchung der Reliabilität der MLU wurde von Pavelko et al. (2020) veröffentlicht. In dieser Studie wurden Spontansprachproben von 50 und 25 Äusserungen bei Kindern mit typischer Sprachentwicklung im Alter von 3;2 – 7;10 Jahren (n = 220) verglichen. Die Analyse der Spontansprachproben erfolgte mittels SUGAR-Methode (Sampling Utterances and Grammatical Analysis Revised[35]). Die Autor:innen berechneten eine *mixed-model* Analyse, welche keine signifikanten Unterschiede zwischen den beiden Sprachprobenlängen ergab ($t(219) = -0.47$, $p = 0.64$). Für die MLU ergab sich bei 25 Äusserungen eine signifikant grössere *within-subject random* Variabilität als bei 50 Äusserungen, deren Effektgrösse aber sehr klein war. Zusätzlich analysierten die Autoren die Sprachproben mittels Bland-Altman Plots. Die *Limits Of Agreement* (LOA[36]) wurden mit einer Standardabweichung von ± 1.96 berechnet. 95 % der Werte fielen in den Bereich des Konfidenzintervalls, was ein klinisch akzeptables Resultat darstellt. Die Autor:innen kommen zum Schluss, dass mit Sprachproben von 25 Äusserungen reliable Werte erreicht werden.

Zusammenfassend wurden für die Analyse der MLU(-m) bei Kindern mit einer typischen Sprachentwicklung in der englischen Sprache folgende Sprachprobenlängen empfohlen:

- Gavin und Giles (1996): 75 Äusserungen
- Heilmann et al. (2010): 3 Minuten
- Guo und Eisenberg (2015): 10 Minuten
- Pavelko et al. (2020): 25 Äusserungen

[34] Anders als bei Guo et al. (2019) werden in dieser Studie erst Korrelationskoeffizienten von über 0.9 als starker Zusammenhang betrachtet.

[35] Für eine detaillierte Erklärung der SUGAR-Methode siehe Pavelko und Owens (2019b)

[36] Die LOA entsprechen den Grenzen des Konfidenzintervalls um den Mittelwert der Differenzen der Resultate der beiden im Plot verglichenen Testlängen. Genauer wird dies im Kapitel 4.4 beschrieben.

Dass sich die Verlässlichkeit von Spontansprachindikatoren zwischen verschiedenen Sprachen unterscheiden kann, zeigt die Studie von Tomas und Dorofeeva (2019), die sich mit der MLU in der russischen Sprache befasste. Dazu wurden von 27 Kinder im Alter von 2;9 - 5;7 Jahren mit typischer Sprachentwicklung Sprachproben während einer Freispielsequenz aufgezeichnet. Die Studie untersuchte verschiedene Formen des MLU mittels Pearson's Korrelationskoeffizienten und Bland-Altman Plots. Letztere zeigte für kurze Sprachproben (25 und 50 Äusserungen) einen klaren Bias. Die Autor:innen empfehlen daher für die russische Sprache mindestens 100 Äusserungen zur Analyse des MLU.

Eine der ersten Untersuchungen des MLU bei Kindern mit Spracherwerbsstörungen führten Casby (2011) durch. Dazu verwendeten sie Sprachproben von 10 Kindern im Alter von 3;0 - 11;8 Jahren aus der CHILDES-Datenbank[37]. In ihrer Analyse unterteilten sie die Sprachproben (150 Äusserungen) in verschiedene Teilproben: Jeweils vom Anfang, der Mitte und dem Ende der Sprachprobe wurden 10 und 20 Äusserungen entnommen. Zusätzlich wurden drei quasi-zufällige Sprachproben aus der Gesamtstichprobe in der Länge von 10, 20 und 50 Äusserungen gezogen. Die Analyse mit ANOVA ergab keine signifikanten Unterschiede zwischen den kürzeren und der langen Sprachprobe. Die berechneten Pearson's Korrelationskoeffizienten ergaben unterschiedliche Werte, je nach Sprachprobenlänge und wo die Äusserungen aus der Sprachprobe entnommen wurden ($r = 0.52$ - 0.94). Sie lagen aber alle im Bereich eines starken Zusammenhangs[38].

Zwei Jahre später untersuchten Heilmann et al. (2013) die Reliabilität erneut anhand von 10-minütigen Interview-Sprachproben. Dazu wurden 20 Kinder mit Risiko für oder diagnostizierter Spracherwerbsstörung im Kindergartenalter befragt. Analysiert wurden mittels SALT verschiedene Indikatoren, unter anderem die mittlere Länge eines Turns und die NDW. Die Autor:innen zeigten mit der *Generalizability-Theory*, dass die Länge der Sprachproben einen bescheidenen Einfluss auf die Indikatoren hat und kurze, dreiminütige Spontansprachproben zu reliablen Ergebnissen führen können.

Ebenfalls verschiedene Indikatoren untersuchten Wilder und Redmond (2022), welche Sprachproben im Freispiel von Kindern mit und ohne Spracherwerbsstörung (n = 42) erhoben und zwischen kurzen (1, 3, 5, 7, 10 Minuten) und einer langen Sprachprobe (20 Minuten) verglichen. Die RMANOVA ergab nur für die NDW einen signifikanten Unterschied zwischen den verschiedenen Längen. Für alle andern (NTU, WPM, MLU-m, percentage of maze words, errors and omissions, PGU, subordination index[39]) ergaben sich keine signifikanten Unterschiede. Die Koeffizienten der Spearman-Brown Analyse zeigten sehr unterschiedliche Resultate bei Kindern mit und ohne Spracherwerbsstörung, der Länge

[37] Child Language Data Exchange System, eine Datenbank, die dem Austausch von kindlichen Sprachdaten im englischen Sprachraum dient (MacWhinney & Fromm, 2022)

[38] Hier wiederum wurde dieselbe Interpretation des Pearson's Korrelationskoeffizienten angewandt, wie bei Guo et al. (2019), eine starker Zusammenhang beginnt also bei $r > 0.5$

[39] Die bisher noch nicht erläuterten Abkürzungen stehen für folgende Indikatoren: NTU = number of total utterances; WPM = words per minute.

der Sprachprobe und den Indikatoren. Einige Indikatoren zeigten bereits bei 3-minütigen Sprachproben Werte über 0.9, viele erreichten diesen Wert ab 7 Minuten. Aufgrund dieser Resultate und der Analyse von B-A-Plots, deren LOA mit ± 1.00 Standardabweichung berechnet wurden, kamen die Autor:innen zum Schluss, dass 7 minütige Sprachproben für reliable Werte verwendet werden sollten.

Zusammenfassend wurden für die Analyse der MLU(-m) bei Kindern mit einer Spracherwerbsstörung in der englischen Sprache folgende Sprachprobenlängen empfohlen:

- Casby (2011): 10 Äusserungen
- Heilmann et al. (2013): 3 Minuten
- Wilder und Redmond (2022): 7 Minuten

Insgesamt zeigen sich grosse Unterschiede in den Empfehlungen der Sprachprobenlängen zur Berechnung der MLU. Obwohl die Ergebnisse alles andere als eindeutig sind, empfehlen doch einige der besprochenen Studien Sprachprobenlängen, die kürzer sind als die bis anhin geforderten 50 Äusserungen. Die Hypothese, dass die MLU - zumindest in der englischen Sprache - bereits bei kurzen Sprachproben zu reliablen Resultaten zu führen kann, schient im Moment aufgrund der Resultate weder falsifizier- noch verifizierbar. Und dies sowohl bei Kindern mit als auch bei Kindern ohne Spracherwerbsstörung. Noch weniger gibt die aktuelle Studienlage eindeutige Antworten auf die Frage, wie kurz die Sprachproben genau sein können.

Zur MATTR wurde keine Studien gefunden, die sich mit der Länge der Sprachproben auseinandersetzte. Die dürftige Studienlage kombiniert mit der beschriebenen Praxishandhabung der Spontansprachanalyse, nämlich dass viele Logopädinnen weniger als 50 Äusserungen eines Kindes transkribieren und auswerten, zeigt die Relevanz der vorliegenden Arbeit.

Einige Studien befassten sich in der Vergangenheit mit Indikatoren, die in dieser Arbeit keine vertiefte Rolle spielen. Daher werden sie hier nur kurz erwähnt. Sowohl Leonard et al. (2017) als auch Tommerdahl und Kilpatrick (2013) untersuchten in ihren Studien morphologische Indikatoren. McDaniel und Brady (2022) analysierten verschiedene *vocal measures* bei Kindern mit Autismus-Spektrumsstörung. Die Untersuchung von van Severen et al. (2012) befasst sich mit dem phonologischen Inventar von Kindern mit typischer Sprachentwicklung in der holländischen Sprache. Eine weitere Studie befasste sich mit der Reliabilität der *Percentage of Grammatical Utterances* (Eisenberg & Guo, 2015).

Testwiederholungsreliabilität Die Testwiederholungsreliabilität wurde in diversen Studien untersucht (Cole et al., 1989; Gavin & Giles, 1996; Heilmann et al., 2013; Heilmann et al., 2008; MacFarlane et al., 2023; McKee, 2000; Soleymani et al., 2016; Thurman et al., 2021; Tommerdahl & Kilpatrick, 2013). Da diese Form der Reliabilität nicht im Fokus

dieser Arbeit steht, werden die Resultate dieser Studien im Folgenden kurz zusammengefasst: Insgesamt scheint die Testwiederholungsreliabilität über verschiedene Elizitationsmethoden und mit unterschiedlichen Klientel reliabel zu sein. Insbesondere für den MLU und die NDW liegen relativ konsistente Befunde vor. Die MATTR wurde bisher in keiner gefundenen Studie bezüglich der Testwiederholungsreliabilität untersucht. Die meisten Befunde beziehen sich auf die Altersspanne, die auch im DigiSpon1-Projekt im Fokus steht (Kindergartenalter). Obwohl die Studienlage sicherlich noch ausbaufähig ist, kann davon ausgegangen werden, dass die Testwiederholungsreliabilität für Kinder im Kindergartenalter insbesondere für die MLU und die NDW gut ist.

3 Fragestellungen

Die bisher beschriebenen Studien und der theoretische Hintergrund bilden die Grundlage für die Fragestellungen dieser Arbeit. Die im Folgenden formulierten Teilanalysen dieser Arbeit wurden aus diversen Gründen anstatt als Hypothesen als Unterfragen formuliert. Einerseits lassen sich weder aus dem aktuellen Forschungsstand noch aus der Theorie klar erwartbare Resultate begründen. Andererseits weisen die getätigten Analysen einen eher explorativen Charakter auf. Um die Stringenz der Arbeit zu wahren, wurde auch eine inferenzstatistische Analyse als Frage anstatt als Hypothese formuliert.

Da bisher keine gefundene Studie die instrumentelle Reliabilität in der schweizerdeutschen Sprache untersucht hat, soll dies in einem ersten Schritt gemacht werde. Dazu wurden die Testhalbierungsreliabilität und die interne Konsistenz der beiden Indikatoren berechnet. Eine hohe instrumentelle Reliabilität bildet Voraussetzung für die weiteren Analysen, da nur so davon ausgegangen werden kann, dass die Indikatoren innerhalb der gesamten Sprachprobe reliabel sind. Wenn die einzelnen Items hoch miteinander korrelieren, ist davon auszugehen, dass der Indikator ein stabiles Mass für die jeweiligen Teilfähigkeiten eines Kindes ist. Dies führt zu folgender Fragestellung:

Unterfrage 1: Wie hoch ist die instrumentelle Reliabilität der MLU und der MATTR bei den Sprachproben?

Zur Beantwortung der übergeordneten Frage dieser Arbeit, wie lange Sprachproben sein müssen, um reliable Werte zu generieren, beziehen sich die folgenden Fragen auf die Reliabilität von kurzen Sprachproben. Entsprechend wurden für die Analysen, die diese Fragen beantworten sollen, aus den gesamten Sprachproben kleinere Samples entnommen, welche nun mit den gesamten Sprachproben verglichen wurden. Die zweite grosse Unterfrage lautet entsprechend:

Unterfrage 2: Führen kürzere Sprachproben zu vergleichbaren Resultaten wie die gesamten Sprachproben?

Um diese Frage zu beantworten, wurden verschiedene Teilanalysen durchgeführt. Eine detaillierte Diskussion der angewandten statistischen Methoden findet sich im Kapitel 4.4. Zum einen soll mittels einer inferenzstatistischen Methode untersucht werden, ob sich kurze Sprachproben signifikant von längeren Unterscheiden. Durch diese Analyse wird die absolute Reliabilität der kurzen Sprachproben auf Gruppenebene beleuchtet. Sollten diese einen signifikanten Unterschied ergeben, würde dies gegen die Verwendung der gekürzten Sprachproben sprechen. Die Frage lautet also:

Unterfrage 2.1: Wie zeigt sich die absolute Reliabilität von kurzen Spontansprachanalysen auf der Gruppenebene?

Anschliessend wurde die relative Reliabilität der Indikatoren anhand eines Korrelationskoeffizienten geschätzt, um eine weitere Kennzahl der Reliabilität abbilden zu können. Falls diese Analyse hohe Korrelationen ergeben, zeigt dies, dass die Rangfolge der Resultate der Kinder zwischen langen und kurzen Sprachproben vergleichbar bleibt. Die Frage lautet:

Unterfrage 2.2: Wie hoch ist die relative Reliabilität von kurzen Spontansprachproben?

Um zum Schluss ein Licht auf die absolute Reliabilität auf der individuellen Ebene werfen zu können, wurden die Differenzen der Resultate der beiden Indikatoren zwischen den kürzeren und den längeren Sprachproben grafisch aufbereitet. So können systematische Verzerrungen der Resultate erkannt werden.

Unterfrage 2.3: Wie zeigt sich die absolute Reliabilität von kurzen Spontansprachanalysen auf der individuellen Ebene?

4 Daten und Methoden

In den folgenden Abschnitten wird zunächst die Datenerhebung und anschliessend die daraus resultierte Stichprobe beschrieben. Ein weiterer Abschnitt widmet sich der vorgenommenen Transkription der Auswertung dieser. Ein Fokus liegt hier auf der genauen Beschreibung der Berechnung der Indikatoren, da es hierfür aktuell keine Standards gibt und eine genaue Beschreibung so die Vergleichbarkeit der verschiedenen Studien erhöhen kann. In einem weiteren Abschnitt soll beschrieben werden, wie die gewonnen Daten statistisch bearbeitet wurden.

4.1 Datenerhebung

In diesem Kapitel wird zunächst die Frage beantwortet, welche Kinder zur Zielgruppe der Datenerhebung gehörten. Anschliessend soll auf die Art und Weise der Datenerhebung eingegangen werden.

Die Zielgruppe der Datenerhebung wurde durch folgende Einschlusskriterien definiert:

- Kinder zwischen 4 und 7 Jahren (Kindergartenalter)

- Kinder mit und ohne Spracherwerbsstörungen

Da die Durchführung der gewählten Elizitationsmethode und der Transkription nicht mit allen Kindern durchführbar ist, wurde folgende Ausschlusskriterien festgelegt:

- Das Kind kann noch keine 3-Wort-Äusserungen bilden.

- Das Kind zeigt eine stark unverständliche Aussprache.

Die Einschlusskriterien wurden definiert, damit die Sprachdaten untereinander vergleichbar sind. Es wurden Sprachproben von Kindern mit und ohne Spracherwerbsstörungen erhoben, damit das entwickelte Tool für die vorgegebenen Altersspanne verschiedene sprachliche Kompetenzen zu transkribieren lernen kann. Eine unverständliche Aussprache (einer schweren phonetisch-phonologischen Störung entsprechend) wurde als Auschlusskriterium festgelegt, da dies für die automatische Transkription eine starke Erschwernis bedeutet hätte. Bei Kindern mit einem Sprachstand von produktiven Zweiwort-Äusserungen wäre die gewählte Elizitationsmethode in der Durchführung kaum machbar gewesen, da mit einer solch kurzen Äusserungslänge noch sehr wenig erzählt werden kann.

Die Datenerhebung erfolgte über zwei verschiedene Strategien: Zum einen wurden in der Praxis tätige Logopäd:innen angefragt an der Datenerhebung teilzunehmen. 20 Logopädinnen erklärten sich bereit, mit Kindern, die bei ihnen in der Therapie waren, Sprachproben aufzunehmen. Zum anderen wurden Kindergartenlehrpersonen angefragt, ob in ihrer Klasse Aufnahmen gemacht werden können. In diesem Fall wurden die Sprachproben von Kindern mit und ohne Spracherwerbsstörungen durch Mitarbeiterinnen des DigiSpon-Projekts erhoben, die für diesen Zweck in die Kindergärten fuhren. Für alle Sprachproben wurde das Einverständnis der Eltern und der erhebenden Personen eingeholt. Je nach Wunsch der Eltern wurde entweder eine Video- oder eine Tonbandaufnahme erstellt. Die Datenerhebung fand zwischen Oktober 2023 und März 2024 in Regel- und Sprachkeilkindergärten statt.

Die im DigiSpon-1 Projekt gesammelten Daten dienten verschiedenen Zwecken. Einerseits sollte das im Projekt entwickelte Tool mit den Sprachdaten trainiert werden, um eine möglichst gute automatische Transkription der schweizerdeutschen Sprache zu erreichen. Dazu sollten die Daten einerseits nicht zu ähnlich aber auch nicht zu verschieden sein. Daher wurde in einem ersten Schritt entschieden, sich auf Kinder im Kindergartenalter zu beschränken und somit die Varianz der Sprachentwicklung in einem gewissen Rahmen zu halten. Da das Tool in einem zukünftigen Endprodukt insbesondere Aufnahmen von Kindern mit einer Spracherwerbsstörung transkribieren soll, lag ein Fokus der

Erhebungen auf diesen Kindern. Es sind aber auch Kinder ohne Spracherwerbsstörungen im Datensatz vorhanden. Diese beiden Faktoren tragen zur Heterogenität der gesammelten Sprachdaten bei. Aus diesen Überlegungen und den Erkenntnissen, die im Kapitel 2.3 zu den Elizitationsmethoden zusammengefasst sind, wurde die Entscheidung gefällt, die Erhebungen mit einem Leitfaden zu strukturieren. Dadurch sind die erhobenen Daten zwar nicht mehr wirklich spontan, aber dafür vergleichbarer und vermutlich etwas homogener. Dies schien ebenfalls sinnvoller Weg zu sein, da sehr viele verschiedene Personen an den Erhebungen beteiligt waren.

Die gewählte Elizitationsmethode wurde an die *personal narratives*, eine von Westerveld (2011) entwickelte Elizitationsmethode, angelehnt. Dabei werden die Kinder durch Erzählanregungen ermutigt, von eigenen Erlebnissen zu berichten. Die Erzählanregung bestanden in diesem Fall aus Bildersets von je 10 Fotos. Das Ziel war, Sprachproben von circa 10 Minuten zu erhalten. Wenn ein Kind nicht auf die angebotenen Erzählanregungen ansprach, durfte das Vorgehen so angepasst werden, dass trotzdem eine Sprachprobe gewonnen werden konnte. Es wurden zwei Bildersets (A und B) erstellt. Welche der Versionen verwendet wurde, wurde durch den Wurf einer Münze randomisiert. Falls ein Kind sowohl Schweizerdeutsch als auch Hochdeutsch sprach, wurden Erhebungen in beiden Sprachen durchgeführt. In diesem Fall sollte jeweils ein Bilderset für eine Sprache verwendet werden. Die Zuteilung dieser sollte ebenfalls via Münzwurf getroffen werden.

Zusätzlich wurden über einen Fragebogen, den die Eltern der Kinder ausfüllten, zusätzlich detailliertere Informationen zu den Kindern und ihren familiären Hintergründen eingeholt. Diese Angaben umfassten folgende Bereiche:

- Geburtstag
- Geschlecht
- Klasse
- Sprachen und Dialekte, die das Kind spricht
- *Age of Onset* der Zweit- / Drittsprache
- Familiensprache, Sprache des Elternteils 1, Sprache des Elternteils 2
- Besucht das Kind die Logopädie
- Allfällig bestehende Diagnosen (Lern- und Sprachstörungen)
- Beruf der Eltern

4.2 Stichprobe

In dieser Arbeit wurde eine Teilstichprobe der erhobenen Daten des DigiSpon 1-Projektes verwendet, da zum Zeitpunkt dieser Arbeit noch nicht alle Sprachproben transkribiert

wurden. Der gesamte Datensatz umfasst 115 Spontansprachproben. Die Stichprobe, die dieser Arbeit zu Grunde liegt, umfasst n = 50 Sprachproben. Eine dieser Sprachproben wurde von den Analysen ausgeschlossen, weil sie weniger als 50 Äusserungen beinhaltet. Die finale Stichprobe umfasste daher 49 Sprachproben von Kindern im Alter von 4;5 bis 7;6 Jahren, wobei der Median bei 5;7 Jahren lag[40]. Von den 49 Kindern waren 28 männlich und 21 weiblich. Laut den Angeben der Eltern wachsen 25 der Kinder mehrsprachig und 24 einsprachig (verschiedene Dialekte der schweizerdeutschen Sprache) auf. Auch wurden die Eltern gefragt, ob ihr Kind in logopädische Therapie erhält oder eine Diagnose vorliegt. Bei 25 Kindern gaben die Eltern an, dass diese in der Logopädie seinen. Sechs davon haben laut Aussagen der Eltern Schwierigkeiten in der Phonetik-Phonologie, 10 eine schwere Sprachentwicklungsstörung. Die Auffälligkeiten der anderen Kinder wurden nicht genauer beschrieben.

In der analysierten Teilstichprobe von n = 49 wurde bei 24 Sprachproben die Version A und bei 25 Sprachproben die Version B verwendet. Die kürzeste so erhobene Sprachprobe umfasste 7 Minuten, während die längste 16 Minuten dauerte. Der Mittelwert liegt bei 11.35 Minuten, der Median bei 10 Minuten. Die Entsprechungen in Äusserungen und Wörtern zur Dauer in Minuten sind in Tabelle 1 ersichtlich. Etwas mehr als die Hälfte der analysierten Transkripte wurden von zwei Mitarbeitenden des DigiSpon-Projekts erhoben (n = 26), die restlichen wurden von Logopädinnen aus der Praxis. Diese führten die Erhebungen mit jeweils einem bis drei Kindern durch.

Einheit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Minuten	7.00	10.00	10.00	11.35	14.00	16.00
Äusserungen	52.00	70.00	92.00	90.22	102.00	194.00
Wörter	198.00	448.00	545.00	565.12	674.00	975.00

Tabelle 1: Längen der analysierten Sprachproben in Minuten, Äusserungen und Wörtern

4.3 Transkription und Analyse

Die Transkription der Sprachproben folgte einem Transkritionsleitfaden, welcher innerhalb des DigiSpon 1-Projektes auf Basis der Dieth-Dialäktschrift entwickelt wurde. Es wurden stets sowohl die Äusserungen des Kindes als auch diejenigen der Fachperson transkribiert und mit automatisierten Zeitstempeln versehen. Namen wurden aus Gründen des Datenschutzes nicht transkribiert. Wichtig bezüglich Analyse der MLU ist, dass Verständigungssignale (wie beispielsweise mhm, aha, ja) nicht transkribiert wurden, wenn diese während des Redebeitrags des Gegenübers geäussert wurden. Wenn das Kind also noch nicht fertig gesprochen hatte, wurde die Äusserung aufgrund eines solchen Verständigungssignales der Fachperson nicht unterbrochen. Ein Sprecherwechsel wurde dann transkribiert, wenn das Gegenüber mehr als nur ein Verständigungssignal äusserte, also

[40] Aufgrund eines fehlenden Wertes (bei einem Kind wurde kein Geburtstag abgegeben) basieren die Angaben des Alters auf n = 48 Proband:innen

beispielsweise etwas kommentierte oder eine Frage stellte. Die angewandten Regeln bezüglich der Verschriftlichung des Dialekts, der Handhabung von Pausen und so weiter werden nicht weiter vertieft, da diese keine Auswirkungen auf die vorliegende Fragestellung haben.

Jede Sprachprobe wurde mit Hilfe der Transkriptionssoftware f4 von einer ersten Person transkribiert und anschliessend von einer zweiten Person kontrolliert. Diese Arbeit führten diverse Logopädiestudierende der Universität Freiburg (CH) und der HfH Zürich aus. Die Auswertung der so erstellten Transkripte erfolgte durch das im Projekt DigiSpon 1 programmierte Tool zur automatisierten Analyse von LSA. Da insbesondere die MLU in den bisherigen Studien sehr unterschiedlich berechnet wurde und es keine allgemeingültigen Standards gibt, wird im folgenden die Berechnung der beiden Indikatoren, wie sie das DigiSpon-Tool vornimmt, genauer beschrieben.

In einem ersten Schritt wurden die Transkripte bereinigt, indem Kommentare in Klammern, also beispielsweise, wenn etwas flüsternd gesprochen wurde, entfernt wurden. Anschliessend wurden die Äusserungen der Untersuchungsleiterin (die Logopädin oder die DigiSpon-Mitarbeiterin) aus der Analyse entfernt. Als Segmentierungseinheit der MLU diente der Sprecherwechsel. Alles, was ein Kind äusserte ohne von der Untersuchungsleiterin unterbrochen zu werden, gilt also als eine Äusserung. Es wurden keine Äusserungen ausgeschlossen. In einem nächsten Schritt wurden die Äusserungen in Wörter unterteilt, was die Analyseebene der MLU in diese Arbeit darstellt. Es wurde also nicht wie in anderen Studien in Morphemen, sondern in Wörtern gezählt. Die Anzahl Wörter wurden in einem letzten Schritt durch die Anzahl Äusserungen geteilt, um die durchschnittliche Anzahl Wörter pro Äusserung, also die mittlere Äusserungslänge, zu berechnen. Das Tool generierte sowohl einen Datensatz, indem die Rohwerte der MLU pro Äusserung (also die absolute Zahl an Wörtern) eingetragen waren als auch eines, das die MLU-Werte der gesamten Sprachprobe enthielt.

Zur Berechnung der MATTR wurden ebenfalls zunächst die Transkripte bereinigt und nur die Äusserungen des Kindes zur Analyse behalten. Dann wurden die kindlichen Äusserungen in Wörter unterteilt. In einem nächsten Schritt unterteilte das Tool diese Wörter in die vorgegebene Fenstergrösse. Im DigiSpon-Tool kann die Fenstergrösse aktuell vor der Analyse beliebig festgelegt werden. Es besteht aktuell in der Literatur kein Konsens zur Frage, wie gross das Fenster für die Berechnung des MATTR gesetzt werden sollte. Eine bestehende Empfehlung lautet folgendermassen: «it is likely that using a larger window allows MATTR to estimate lexical diversity scores more reliably» (Kapantzoglou et al., 2019, S. 79). Um im Gegensatz zur TTR nicht von der Sprachprobenlänge abzuhängen, muss das Fenster aber auch klein genug sein, um verschiedene TTR-Werte berechnen zu können. Im DigiSpon-Tool besteht eine Voreinstellung der Fenstergrösse von 15 Wörtern. Möglicherweise wären die Resultate der MATTR, der Empfehlung von Kapantzoglou et al. (2019) folgend, reliabler, wenn grössere Fenster gewählt würden. Da

diese Arbeit aber spezifisch die Analyse des DigiSpon-Tools untersuchen will, wurde die Fenstergrösse von 15 für diese Arbeit übernommen. Das Tool bildete also aus jeweils 15 Wörtern die Analyseeinheit der MATTR und unterteilt die Sprachdaten des Kindes entsprechend in die Wörter 1-15, dann die Wörter 2-16, dann 3-17 und so weiter, bis das letzte Wort erreicht wurde. Entsprechend ist die Anzahl Fenster bei einer Fenstergrösse von 15 Wörtern, um 15 kleiner, als die absolute Anzahl Wörter in der Sprachprobe. Innerhalb dieser Einheiten wurden die Wörter anschliessend die Anzahl *types*, also der verschiedenen Wörter gebildet. Aus der Anzahl *types* und der Anzahl *tokens*, die in diesem Fall immer 15 entspricht, wurde dann das Verhältnis der beiden für jedes Fenster berechnet. Dies entspricht dem TTR-Wert. In einem letzten Schritt wurde der Durchschnitt TTR-Werte berechnet und somit der MATTR-Wert gebildet. Um die Werte übersichtlicher zu gestalten, wurden die MATTR-Werte, die zu diesem Zeitpunkt zwischen 0 und 1 zu liegen kamen, mit dem Faktor 100 multipliziert. Auch für diese Analyse generierte das Tool einerseits eine Tabelle mit den TTR-Werten für jedes einzelne Fenster (die Rohwerte) und andererseits eine Tabelle mit den MATTR-Werten für die gesamten Sprachproben.

4.4 Statistische Analyse

Nachdem nun die Datenerhebung und die daraus gewonnen Daten betrachtet wurden, stellt sich die Frage, wie diese statistisch analysiert werden sollen und weshalb. Diesen Fragen widmet sich das folgende Kapitel. Da keine Untersuchung in der schweizerdeutschen Sprache zur Frage der Reliabilität von Spontansprachproben gefunden wurde, kann sich die vorliegende Arbeit nicht auf Vorwissen abstützen. Es können weder gesicherte Annahmen zur instrumentellen Reliabilität noch zur Reliabilität von kurzen Sprachproben gemacht werden. Entsprechend soll in dieser Arbeit beides untersucht werden. Dazu werden im folgenden Abschnitt zunächst die verwendeten statistischen Methoden zur Schätzung der instrumentellen Reliabilität diskutiert. In einem zweiten Schritt werden die verschiedenen Aspekte der Reliabilität von kurzen Sprachproben diskutiert und entsprechend statistische Methoden zur Analyse ausgewählt.

4.4.1 Instrumentelle Reliabilität

Im ersten Teil der Analyse soll geklärt werden, inwiefern die gesamten Sprachproben reliabel sind. Es stellt sich also die Frage, wie exakt die MLU und die MATTR bei Kindern im Alter von 4 - 7 Jahren mit und ohne Spracherwerbsstörungen in der schweizerdeutschen Sprache durch die vorgestellte Elizitationsmethode gemessen werden. Zu diesem Zweck soll in einem ersten Teil die instrumentelle Reliabilität bestimmt werden. Dieser Analyse dient die, vom Tool ausgegebene Tabelle, die die Rohwerte der beiden Indikatoren beinhaltet. Für die Analyse der MLU flossen entsprechend die absoluten Anzahlen der Wörter pro Äusserung in die Analyse ein. Die Untersuchung der MATTR basierte hier auf den

TTR-Werten der einzelnen Fenster. Diese Daten bildeten die Items für die Analyse der instrumentellen Reliabilität.

Es gibt verschiedene statistische Methoden, mit denen die instrumentelle Reliabilität untersucht werden kann. Eine davon ist das Cronbach's Alpha, welches sehr häufig in Studien zum Einsatz kommt. Dennoch wird das Cronbach's Alpha von diversen Autor:innen als ungenauer Schätzwert für die interne Konsistenz eines Tests kritisiert. Revelle und Condon (2019) führen dies auf die restriktiven Annahmen zurück, die für die Anwendung des Cronbach's Alpha erfüllt sein müssen. Ist eine davon verletzt, führt die Berechnung je nach Verletzung zu einer Unter- oder einer Überschätzung der Reliabilität. Hinzu kommt, dass der Alphakoeffizient abhängig ist von der Anzahl Items. Je mehr Items ein Test enthält, desto höher wird der Koeffizient ausfallen (Schmidt-Atzert & Amelang, 2012). Da das Cronbach's Alpha aber oft verwendet wird, erhöht seine Berechnung die Vergleichbarkeit mit anderen Studien. Aus diesem Grund wurde das Cronbach's Alpha in dieser Arbeit zur Analyse verwendet. Aufgrund der genannten Kritik wurde aber zusätzlich die Testhalbierungsreliabilität berechnet.

Es gibt viele verschiedene Strategien, wie ein diagnostisches Verfahren halbiert werden kann, um anschliessend die Testhalbierungsreliabilität zu berechnen (Schmidt-Atzert & Amelang, 2012). Eine Möglichkeit ist es, die erste Hälfte der Items mit der zweiten Hälfte zu vergleichen. Dies setzt eine hohe Vergleichbarkeit der beiden Hälften voraus und wäre zum Beispiel dann keine Option, wenn die Items nach Schwierigkeit geordnet wären (Schmidt-Atzert & Amelang, 2012). Bei Spontansprachanalysen ist dies grundsätzlich nicht der Fall. Es ist aber theoretisch denkbar, dass ein Kind eine «Anlaufzeit» braucht, um mit der Aufgabe und der Situation vertraut zu werden und daher erst in der zweiten Hälfte seine sprachlichen Fähigkeiten vollumfänglich präsentiert. Die Trennung der Items in der Hälfte der Spontansprachprobe wäre dann keine gute Option, weil in der ersten Hälfte der Sprachprobe möglicherweise vermehrt zurückhaltende Äusserungen zu hören wären.

Eine andere Option ist es, das Verfahren nach geraden und ungeraden Items aufzuteilen. Dieses Vorgehen ist bei diagnostischen Verfahren empfehlenswert, die nach Schwierigkeit geordnet sind oder bei denen keine spezifische Reihenfolge erkennbar ist (Schmidt-Atzert & Amelang, 2012). Die Daten der Sprachproben sind nicht nach Schwierigkeit geordnet, wie dies teilweise bei klassischen Leistungstests der Fall ist. Es kann aber, aus dem bereits genannten Grund, dass gewisse Kinder möglicherweise eine «Anlaufzeit» brauchen, auch nicht davon ausgegangen werden, dass keine spezifische Reihenfolge erkennbar wäre. Auch diese Aufteilung erscheint daher nicht optimal.

Eine weitere Möglichkeit wäre es, den Test nach Itemkennwerten wie der Itemschwierigkeit zu unterteilen (Schmidt-Atzert & Amelang, 2012). Da die Spontansprachanalyse

keine spezifischen Aufgaben vorgibt, die von den Proband:innen entweder gelöst werden kann oder nicht gelöst werden kann, fällt diese Vorgehensweise als Option weg.

Zu guter Letzt bleibt die Option der randomisierten Halbierung der Items in zwei Hälften (Schmidt-Atzert & Amelang, 2012). Dabei werden die Items zufällig der einen oder der anderen Hälfte zugeteilt. Die einzige Bedingung ist, dass am Schluss in beiden Hälften gleich viele Items sind. Durch eine mehrfache Wiederholung dieses Vorgehens kann die Verteilung der Testhalbierungsreliabilitäten aufgezeigt werden, die einen umfassenden Einblick in die Resultate dieser Schätzmethoden gibt. Aufgrund der so erhobenen Reliabilitätswerte kann zusätzlich ein Konfidenzintervall für die Testhalbierungsreliabilität berechnet werden. Insgesamt erscheint die randomisierte Halbierung der Spontansprachprobe am naheliegendsten, da nicht anzunehmen ist, dass eine spezifische Ordnung der Items vorliegt, aber auch nicht mit absoluter Sicherheit davon ausgegangen werden kann, dass die erste Hälfte der Spontansprachprobe mit der zweiten vergleichbar ist. In den Resultaten werden daher einerseits die minimale, die durchschnittliche und die maximale Reliabilität ausgewiesen und andererseits die Verteilung der Koeffizienten in einem Histogramm präsentiert. Die randomisierte Einteilung der Items kann in enorm vielen verschiedenen Varianten durchgeführt werden. Um dies in einem sinnvollen Rahmen zu halten, wurden die Aufteilungen auf 10'000 dieser Varianten begrenzt.

Eine zu klärende Frage ist, wie hoch die minimale Höhe der Reliabilität sein muss, um ein reliables Resultat zu widerspiegeln. Die Reliabilität ist in der Regel ein Kennwert, der sich zwischen 0 und 1 bewegt. Je näher der Wert bei 1 liegt, desto höher ist die Reliabilität des Diagnostikmaterials. Da die Reliabilität von verschiedenen Faktoren (z.B. Homo- beziehungsweise Heterogenität des Konstrukts sowie der untersuchten Population) abhängt, gibt es keine eindeutige Schwelle, die die Mindesthöhe einer guten Reliabilität angibt. Laut Moosbrugger und Kelava (2020) setzte sich vor längerer Zeit die Konvention durch, dass für heterogene Konstrukte eine Reliabilität von 0.7 angemessen und eine Reliabilität zwischen 0.8 und 0.9 für homogene Konstrukte hoch genug sei. Für Leistungstests wurden Reliabilitätswerte von über 0.9 als sehr gut definiert. Die Autoren kritisieren allerdings, dass diese Faustregeln aufgrund der fehlenden theoretischen Grundlage nicht überbewertet werden sollen und im Einzelfall möglicherweise andere Werte sinnvoll sind. Aufgrund dieser Unklarheit wird im Folgenden jeweils klar festgelegt, für welchen Test welche Werte als genügend reliabel bezeichnet werden. Für die Testhalbierungsreliabilität sollte die durchschnittliche Reliabilität mindestens bei 0.7 liegen. Im Idealfall würde diese zwischen 0.8 oder 0.9 liegen (Moosbrugger & Kelava, 2020, S. 330). Die Korrelationskoeffizienten des Cronbach's Alpha werden Price (2016) folgend, so interpretiert:

- $\alpha = 0.65$ - 0.7 minimal akzeptabel
- $\alpha = 0.7$ - 0.8 respektabel
- $\alpha \geq 0.8$ sehr gut

Da beide Analysen keine fehlenden Werte zulassen, wurde die Länge der Sprachproben für die Analyse der MLU auf 52 Äusserungen festgelegt. Einerseits entspricht dies der kürzesten Sprachprobe, die im Datensatz vorhanden ist[41] und andererseits liegt diese Länge über der empfohlenen Mindestanzahl von 50 Äusserungen für eine Sprachprobe. Zur Analyse der MATTR-Werte wurden die Sprachproben auf 198 Wörter gekürzt (dies entspricht 183 Fenstern für die MATTR, da die Fenstergrösse auf 15 Wörter festgelegt wurde), was wiederum der kürzesten Sprachprobe im Datensatz entspricht.

Einschränkend ist zu beachten, dass sich die Fenster des MATTR überschneiden und nur schon aufgrund dessen eine hohe instrumentelle Reliabilität zu erwarten ist. Die Berechnung der instrumentellen Reliabilität mit überschneidenden Fenstern ist sinnvoll, weil der Indikator aus diesen Überschneidungen konzipiert ist. Gleichzeitig können die Werte dieser Analyse kaum mit denjenigen der MLU verglichen werden, bei welcher sich die Analyseeinheiten nicht überschneiden. Um einen vollständigen Überblick zu erhalten, wurde daher eine zweite Analyse der instrumentellen Reliabilität für die MATTR ohne Überschneidung der Fenster vorgenommen. Dieses Vorgehen führte zu einer starken Abnahme der Anzahl Items. Um dem entgegenzuwirken, wurde nochmals die zu diesem Zeitpunkt kürzeste Sprachprobe ausgeschlossen. So konnten 22 MATTR-Werte ohne überschneidende Fenster pro Sprachprobe in die Analyse einbezogen werden. Dieses Vorgehen wurde gewählt, weil die Anzahl Items einen Einfluss auf das Cronbach's Alpha hat. Auch wenn so einige Items mehr einbezogen werden konnten, ist dieser Einfluss nicht zu vernachlässigen.

Eine genügend hohe Reliabilität in diesen beiden Analysen bildet die Voraussetzung für die nachfolgenden Untersuchungen. Wenn die Schätzmethoden der internen Konsistenz und der instrumentellen Reliabilität hohe Reliabilitätswerte ergeben, ist davon auszugehen, dass der Indikator ein stabiles Mass für die jeweiligen Teilfähigkeiten eines Kindes ist. Dies ist wichtig, damit der Vergleich der kurzen Sprachproben mit der gesamten Sprachprobe sinnvoll ist. Nur wenn der Test insgesamt reliabel misst, ist es sinnvoll, den Test in kürzere Teile zu zerlegen und diese mit dem ganzen Test zu vergleichen.

4.4.2 Reliabilität von kurzen Sprachproben

Um die zweite Fragestellung zu beantworten wurden in einem weiteren Schritt die Werte der Indikatoren für die gesamten und für kürzere Sprachproben berechnet. Aus den zuvor verwendeten Rohdaten wurden nun also Durchschnittswerte gebildet. Dabei wurden die Indikatoren für die gesamten Sprachproben auf der Basis aller zur Verfügung stehenden Sprachdaten und nicht mehr nur auf der Basis von 52 Äusserungen respektive 198 Wörter berechnet. Dies führt hier nicht mehr zu einem Problem, da zur Überprüfung der Reliabilität von kurzen Sprachproben nicht mehr die Werte der einzelnen Items verwendet

[41] Nach Ausschluss der Sprachprobe, die weniger als 50 Äusserungen beinhaltete.

wurden, sondern die daraus resultierenden Werte für die Indikatoren. Dadurch kommt es nicht mehr zu fehlenden Werten. Es ist zu beachten, dass die Indikatoren der gesamten Sprachproben aufgrund der unterschiedlichen Längen der Sprachproben auf unterschiedlich vielen Items (für die MLU sind dies die Rohwerte pro Äusserungen, für die MLU die TTR-Werte pro Fenster) basieren. Die Sprachproben variieren in ihrer Länge von 52 bis 174 Äusserungen respektive 198 bis 961 Wörtern.

Die gekürzten Sprachproben begannen jeweils zu Beginn der Gesamtsprachprobe (das heisst, es werden nicht randomisiert Äusserungen aus der Sprachprobe entnommen). Diese Entscheidung wurde aufgrund folgender Überlegung getroffen: Für die Praxis stellt sich die Frage, wie lange die Sprachproben sein müssen, um reliable Werte zu erhalten. Wenn nun eine kürzere Sprachprobe zur Analyse verwendet wird, beginnt diese äquivalent zu einer langen Sprachprobe. Wenn zuvor eine Art «Vorgespräch» geführt werden müsste – und dann entsprechend die Sprachprobe aus der Mitte der Sprachprobe entnommen würde, würde dies nicht zu einer zeitlichen Erleichterung im logopädischen Alltag führen. Dasselbe gilt auch für randomisiert entnommene Äusserungen aus einer längeren Sprachprobe. Es stellt sich damit allerdings die Frage, ob Kinder bereits von Anfang an ihre sprachlichen Leistungen abrufen können, oder ob sie dazu länger Zeit brauchen (z.B. weil sie das Setting und die Aufgabe noch nicht kennen).

Für die MLU lag der Kürzung die Einheit der Äusserungen zu Grunde und es wurden fünf verschiedene Kurzsprachproben entnommen: 10, 20, 30, 40 und 50 Äusserungen. Für die MATTR lag die Einheit der Wörter der Kürzung zu Grunde, da dies die Berechnungsgrundlage des Indikators ist. Entsprechend der MLU wurden fünf gekürzte Sprachproben entnommen. Da die kürzeste Sprachprobe 183 Wörter enthält, wurden die Sprachproben nach einer Abrundung auf 35, 70, 105, 140 und 175 Wörter gekürzt. Für diese gekürzten Sprachproben wurden nun entsprechend der gesamten Sprachproben die Werte für die MLU und die MATTR berechnet. Diese Werte wurde anschliessend bezüglich der Reliabilität verglichen.

Aktuell besteht in der Forschungsgemeinschaft kein Konsens, welche Reliabilitätsmessverfahren angewendet werden sollen, um die Reliabilität von kurzen Sprachproben zu untersuchen (Bruton et al., 2000; Pavelko et al., 2020). Bruton et al. (2000) plädierten dafür, dass die Reliabilität optimalerweise stets mit mehr als einem Messverfahren berechnet werden soll. Dies liegt darin begründet, dass die verschiedenen Messverfahren jeweils unterschiedliche Aspekte der Reliabilität messen und somit ein umfassenderes Bild generiert wird, wenn mehrere Schätzmethoden angewandt werden. Pavelko et al. (2020) empfehlen dabei sowohl relative als auch absolute Messverfahren zu verwenden. Diese spezifische Anforderung kommt dadurch zu Stande, dass sich diese beiden Formen der Reliabilität stark unterscheiden aber beide nicht zu unterschätzen sind. Wenn nur eine davon evaluiert wird, kann dies zu einem verzerrten Bild der gesamten Reliabilität führen.

Pavelko et al. (2020) kritisieren an der relativen Reliabilität beispielsweise, dass sich mit ihr einzig zeigen lässt, ob die Ergebnisse einer längeren Sprachprobe mit denjenigen einer kürzeren Sprachprobe in Beziehung stehen. Sie erlaubt aber keine Aussage darüber, wie sehr sich die Resultate einer einzelnen Proband:in zwischen den Messungen unterscheiden. Diese Problematik verschärft sich, je heterogener die getestete Gruppe ist. Je grösser die Unterschiede zwischen den einzelnen Proband:innen sind, desto weiter auseinander können die Messpunkte einer einzelnen Proband:in liegen, ohne dass diese ihre Position in der Rangliste verlässt. Wie Wilder und Redmond (2022) argumentieren, führt dies bei der heterogenen Population der Kinder mit Spracherwerbsstörungen möglicherweise zu einer ungenauen Schätzung der Reliabilität. Auch ist es nicht möglich, systematische Fehler mit Korrelationskoeffizienten zu entdecken. Aus diesen Gründen wurde entschieden, die Reliabilität der gekürzten Sprachproben anhand einer Auswahl von Schätzmethoden zu untersuchen. Eine davon soll die absolute Reliabilität auf Gruppenebene untersuchen, eine die relative Reliabilität und eine Letzte die absolute Reliabilität auf individueller Ebene.

Schätzmethoden der absolute Reliabilität auf Gruppenebene: Zunächst wurde über eine inferenzstatistische Analyse die absolute Reliabilität auf Gruppenebene geschätzt. Auch dazu standen verschiedene Analysemöglichkeiten zur Auswahl. In den beschriebenen Studien im Kapitel 2.7.3 wurde oft eine Form der ANOVA verwendet. Pavelko et al. (2020) verwendeten eine *mixed-model* Analyse. Dies begründeten sie wie folgt:

«First, the mixed model correctly models the covariance structure of dependent variables that are correlated (i.e., avoids violating the assumption of statistical independence) that occurs when analyzing multiple repeated dependent measures. Second, the mixed model does not assume that all within-subject observations have equal variance/correlations.» (Pavelko et al., 2020, S. 781)

In anderen Worten liegt ein Vorteil der *mixed-model* Analyse darin, dass es die Tatsache berücksichtigt, dass die Resultate der Testungen miteinander korrelieren können. Das bedeutet, dass es die Annahme der Unabhängigkeit nicht verletzt. Ausserdem berücksichtigt eine *mixed-model* Analyse, dass nicht alle Resultate innerhalb eines Kindes die gleiche Streuung und Korrelationen haben müssen. Es ist allerdings zu beachten, dass dieser Studie eine vergleichsweise grosse Stichprobe (n = 220) zu Grunde lagen. Genau dafür eignet sich die *mixed-model* Analyse. Für kleinere Stichproben hingegen wird sie nicht empfohlen. Die Mindestanzahl an Observationen pro Kondition liegt bei 1600 wenn ein Merkmal wiederholt gemessen wurde (Brysbaert & Stevens, 2018, S. 16).

Übertragen auf die vorliegende Arbeit liegen 49 MLU-Werte (Observationen) pro Sprachprobenlänge (Kondition) und äquivalent viele MATTR-Werte vor. Da dies deutlich unter der minimalen Menge von 1600 Beobachtungen liegt, musste die *mixed-model* Analyse als Option ausgeschlossen werden. Gegen die Anwendung der RMANOVA spricht die Grösse der Stichprobe hingegen nicht. Da in der vorliegenden Arbeit die getesteten

Personen mit sich selbst verglichen werden, ist die RMANOVA zur Analyse besser geeignet als die ANOVA, die nicht für diese Abhängigkeiten kontrolliert. Entsprechend fiel der Entscheid auf die RMANOVA und in einem ersten Schritt wurden die Voraussetzungen dieser geprüft. Als Voraussetzungen der RMANOVA gelten die Normalverteilung der Daten und die Spherizität. Die Normalverteilung der Daten wurde mittels Q-Q Plots untersucht. Da die Stichprobe allerdings grösser als 40 ist, würde eine Nicht-Normalverteilung der Daten nicht ist Gewicht fallen (Park et al., 2009). Die Spherizität ist eine spezifische Annahme, die die RMANOVA macht, sobald mehr als zwei Stufen vorhanden sind. Der Test auf Spherizität überprüft, ob zwischen allen Messpaaren die Varianz der Differenzen gleich sind. Spherizität ist anders ausgedrückt dann gegeben, wenn zwischen allen Messpaaren Homoskedastizität gegeben ist. Ein für diesen Zweck anwendbarer Test ist der Mauchly-Test für Spherizität. Ergibt dieser ein signifikantes Resultat, ist die Spherizität der Daten nicht gegeben und die Analyse muss durch eine Einschränkung der Freiheitsgrade korrigiert werden, wie dies beispielsweise mit der Greenhouse-Geisser Korrektur gemacht werden kann. Wird diese Korrektur bei einem signifikanten Resultat für die Spherizität nicht angewandt, besteht das Risiko eines Fehlers 1. Art. Die Nullhypothese (also die Annahme, dass in der Population kein Unterschied besteht) wird also möglicherweise zurückgewiesen, obwohl sie wahr ist. Im Rahmen einer *post-hoc* Analyse wurden die paarweisen Vergleiche zwischen den einzelnen Sprachprobenlängen berechnet. Da in der vorliegenden Arbeit viele Vergleiche durchgeführt wurden, wurde hierzu die Bonferroni-Korrektur angewandt. Ohne diese Korrektur könnte aufgrund der sogenannten Alphafehlerkommulierung ein signifikantes Resultat gefunden werden, obwohl dieses in Wahrheit nicht vorhanden ist. Dies entspräche also einem falsch-positiven Resultat (Park et al., 2009). Falls in der Analyse mittels RMANOVA ein statistischer Unterschied gefunden würde, spräche dies gegen die Verwendung von kurzen Sprachproben. Die nach Sprachprobenlänge gruppierten Daten wurden zusätzlich in Boxplots zur visuellen Analyse abgebildet.

Eine weitere Möglichkeit, die absolute Reliabilität auf Gruppenebene zu untersuchen, ist die *coefficients of variation (CV)*. Diese wurde in der Vergangenheit zwar verwendet, bringt aber auch Herausforderungen mit sich: «The CV is calculated by dividing the standard deviation by the group mean. CV estimates are limited by the lack of a standard interpretation of CV values. Also, acceptable CV values may vary by measure.» (Wilder & Redmond, 2022, S. 1941) Weil es keinen Anhaltspunkt gibt, wie viel Varianz innerhalb einer Messung akzeptabel wäre, wurde diese Option nicht weiter verfolgt.

Schätzmethoden der relativen Reliabilität: In einem zweiten Schritt wurde die relative Reliabilität der kürzeren Sprachproben mithilfe des Korrelationskoeffizienten geschätzt. Diese statistische Analyse gibt Auskunft darüber, ob die Rangfolge der Kinder (geordnet nach Höhe der MLU- respektive MATTR-Werten) in der Auswertung der Indikatoren gleich bleibt, wenn die gesamte Sprachprobe oder eine gekürzte Sprachprobe verwendet wird. Wenn die relative Reliabilität hoch ist, bedeutet dies also, dass das Kind mit dem

höchsten MLU-Wert in der Gesamtsprachprobe mit einer hohen Wahrscheinlichkeit auch in der gekürzten Sprachproben den höchsten MLU-Wert hat. Entsprechend verhält es sich auch mit allen anderen MLU-Werten der gesamten Reihenfolge. Da jeweils die gekürzte Sprachprobe mit der gesamten verglichen wurde, handelt es sich in dieser Analyse um eine sogenannte Zwei-Item-Skala. Um die relative Reliabilität einer solchen zu schätzen, empfehlen Eisinga et al. (2013) die Verwendung des Spearman-Brown Korrelationskoeffizienten r[42], da dieser am wenigsten Verzerrungen aufweist. Bisher besteht keine klare Grenze für die relative Reliabilität, nach welcher unakzeptable Werte von akzeptablen unterschieden werden. Daher orientierte sich die Berechnung in dieser Arbeit an anderen Autoren (Gavin & Giles, 1996; Guo & Eisenberg, 2015; Wilder & Redmond, 2022), welche $r \geq 0.90$ als Schwelle festlegten. Ein weiteres Argument für diese strenge Auslegung der Grenzwerte ist, dass wenn die Teilsprachproben mit der gesamten Sprachprobe verglichen werden, sich zwingend ein Teil der Sprachproben überschneiden und damit eine höhere Korrelation ergeben als wenn die Sprachproben komplett unabhängig wären.

Schätzmethoden der absolute Reliabilität auf individueller Ebene: Schliesslich wurden Bland-Altman Plots (B-A Plots) erstellt, um potenzielle systematische Verzerrungen und die Kongruenz der Resultate zwischen den Sprachprobenlängen zu untersuchen und somit einen Schätzung der absoluten Reliabilität auf der individuellen Ebene zu erhalten. Diese graphische Vergleichsmethode wurde für den Vergleich von zwei verschiedenen Messmethoden entwickelt und dient dazu, einen möglichen *bias* in den Daten zu erkennen und die Übereinstimmung der beiden Messmethoden zu überprüfen (Bland & Altman, 1999). Dabei wird einerseits die Übereinstimmung im Durchschnitt aufgezeigt und andererseits wird die Übereinstimmung der individuellen Werte quantifiziert (Pavelko et al., 2020), indem diese in Diagrammen grafisch dargestellt werden. Dazu wird auf der y-Achse die Differenz der Werte der beiden Messmethoden (im Falle dieser Arbeit beispielsweise die Differenz der MLU-Werte der Gesamtsprachprobe und der gekürzten Sprachproben) abgebildet. Auf der x-Achse finden sich die Durchschnitte der Resultate der beiden Messmethoden (Bland & Altman, 1999). Entsprechend wird pro Kind ein Datenpunkt im Diagramm abgebildet. Der Mittelwert der Differenzen der durch die beiden Methoden gemessenen Werte wird in der Mitte des Diagramms als horizontale Linie abgebildet. Wenn diese stark von 0 abweicht, deutet dies auf einen systematischen *Bias* hin. Würde diese Linie beispielsweise bei 10 liegen, lägen die Resultate der beiden Messmethoden im Mittelwert um den Wert 10 auseinander. Anschliessend werden zwei weitere horizontale Linien ins Diagramm eingefügt,welche die obere und die untere Grenze des 95 % Konfidenzintervalls abbilden, auch Grenzen der Übereinstimmung (englisch: *Limits of Agreement*; LOA) genannt. Da unter anderem Pavelko et al. (2020) für die Berechnung der LOA $\pm$ 1.96 Standardabweichungen (SD) vorschlagen, wird die obere LOA so berechnet: Mittelwert

[42]In dieser Studie vergleichen Eisinga et al. (2013) drei verschiedene Korrelationskoeffizienten in Bezug auf Zwei-Item-Skalen.

der Differenzen + (1.96 x Standardabweichung der Differenzen). Für die untere lautet die Berechnung entsprechend so: Mittelwert der Differenzen - (1.96 x Standardabweichung der Differenzen). Wenn 90 % der Daten innerhalb der LOA zu liegen kommen, entspricht dies einem klinisch akzeptablen Resultat (Giavarina, 2015).

Die vorgestellte Berechnung der LOA führt dazu, dass je nach Streuung der Daten die Grenzen der Übereinstimmung in Bezug auf die absoluten Werte der Indikatoren unterschiedliche Spannbreiten zulassen. Je grösser die Streuung der Daten, desto breiter ist die Spannbreite der Differenzen, die als klinisch akzeptabel betrachtet wird. Dementsprechend können die Differenzen der MLU- und MATTR-Werte grösser sein, je breiter die Streuung ist. Aktuell bestehen bezüglich der MLU in der vorliegenden Form (Segmentation nach Sprecherwechseln) keine standardisierten Normen, die angeben würden, ab wann ein Unterschied klinisch relevant ist. Im Gegensatz dazu kann dies auf Satzebene klarer definiert werden. Wenn die Segmentation der Äusserungen anhand von Sätzen vorgenommen worden wäre, läge die klinische Relevanz einer Differenz bei einem Wort, da dies in der Entwicklung der Satzkonstruktion von Kindern oft eine nächste erreichte Stufe darstellt (Clahsen, 1986). Da nun aber in der vorliegenden Arbeit mehr als nur ein Satz in einer Äusserung vorhanden sein kann, wird die Grenze dessen, was klinisch relevant ist, schwammig. Die Äusserungslänge schwankt im vorliegenden Datensatz aufgrund der Konzeption enorm. Der Datensatz enthält sehr viele Äusserungen mit nur einem Wort, aber auch Äusserungen mit bis zu 78 Wörtern. Wäre nach Sätzen segementiert worden, wäre eine solche Schwankung nicht möglich. Aus diesen Gründen muss für die Analyse der B-A Plots für die MLU wohl etwas mehr Spielraum eingeräumt werden. Dennoch kann insbesondere bei Kindern, die eher kurze Äusserungen bilden, nicht eine beliebig grosse Differenz in der MLU zugelassen werden. Um diese Überlegungen in die Bland-Altman Plots einzubinden, kann der gewählte Multiplikator von 1.96 SD's verkleinert werden. Damit wird dann zwar nicht mehr das standardisierte 95 % Intervall berechnet, es können aber verschiedene Szenarien gezeichnet werden und nach wie vor überprüft werden, wie viele der einzelnen Datenpunkte innerhalb des neuen Intervalls liegen. Dieses Vorgehen findet sich auch in der Studie von Wilder und Redmond (2022), die auch begründen, dass die Resultate dadurch an Ausdifferenzierung gewannen und bessere Vergleiche ermöglichten . Diese Option wird in der vorliegend Arbeit angewandt, wenn die Resultate der Plots mit 1.96 SD's keine eindeutigen Resultate ergeben sollten. Grundsätzlich werden LOA in der Höhe von ± 2.5 als klinisch akzeptabel definiert, da dies der nötigen Erweiterung aufgrund der Segmentation anhand der Sprecherwechsel entgegen kommt.

Dieselbe Frage der klinischen Akzeptanz stellt sich auch für die MATTR, wobei diese durch die unterschiedliche Setzung der Fenster erschwert wird. Diese führt zwischen den Studien zu unterschiedlich hohen Werten und es bleibt daher ungeklärt, welche Breite der LOA für die absolute Reliabilität klinisch akzeptabel sind. Für das in dieser Untersuchung festgelegte Fenster von 15 Wörtern gilt: Pro Wort, welches nicht von allen anderen

verschieden ist, sinkt der MATTR-Wert um 6.6 (da $1/15 * 100 = 6.6$). Für die Interpretation der B-A Plots in dieser Arbeit wurde angenommen, dass weder die obere noch die untere LOA diesen Wert überschreiten sollte (die Grenzen der Übereinstimmung dürfen also maximal bei - 6.6 und + 6.6 liegen).

Zusammenfassend wurden für diese Arbeit die folgenden Schätzmethoden der Reliabilität verwendet:

- instrumentelle Reliabilität: Cronbach's Alpha und Testhalbierungsreliabilität
- absolute Reliabilität auf Gruppenebene: RMANOVA
- relative Reliabilität: Spearman-Brown Korrelationskoeffizient
- absolute Reliabilität auf individueller Ebene: Bland-Altman Plots

Alle statistischen Analysen wurden mit Software R durchgeführt. Der erstellte Code findet sich im Anhang. Die verwendeten *packages* werden dort genannt.

5 Resultate

In den folgenden Abschnitten werden die Resultate der eben beschriebenen Analysen präsentiert. Grafiken und Ergebnistabellen, die nicht hier im Haupttext aufgeführt sind, finden sich im Anhang.

5.1 Instrumentelle Reliabilität

Die Resultate der Analyse der instrumentellen Reliabilität sind in Tabelle 2 ersichtlich. Die maximale Reliabilität liegt bei beiden Indikatoren über 0.9. Die minimale Reliabilität hingehen unterscheidet sich deutlich stärker zwischen den Indikatoren. Für die MLU erreicht die minimale Reliabilität einen Wert von 0.55 während dieser für die MATTR bei 0.92 liegt. Die durchschnittliche Reliabilität liegt für die MLU bei 0.78 und für die MATTR bei 0.97. Genau dieselben Werte ergab die Analyse des Cronbach's Alpha (MLU = 0.78, MATTR = 0.97).

Indikator	Maximale Reliabilität	Minimale Reliabilität	Durchschnittliche Reliabilität	Crohnbach's Alpha
MLU	0.92	0.58	0.78	0.78
MATTR	0.99	0.91	0.97	0.97

Tabelle 2: Schätzwerte für die instrumentelle Reliabilität der MLU und der MATTR

Durch die randomisierte Einteilung der Äusserungen in zwei verschiedene Hälften ergeben sich verschiedene Testhalbierungsreliabilitäten. Diese sind grafisch in der Abbildung 1 für die MLU ersichtlich. Die Werte streuen in einer Normalverteilung. Das Konfidenzintervall der Testhalbierungsreliabilitäten liegt zwischen 0.67 und 0.87.

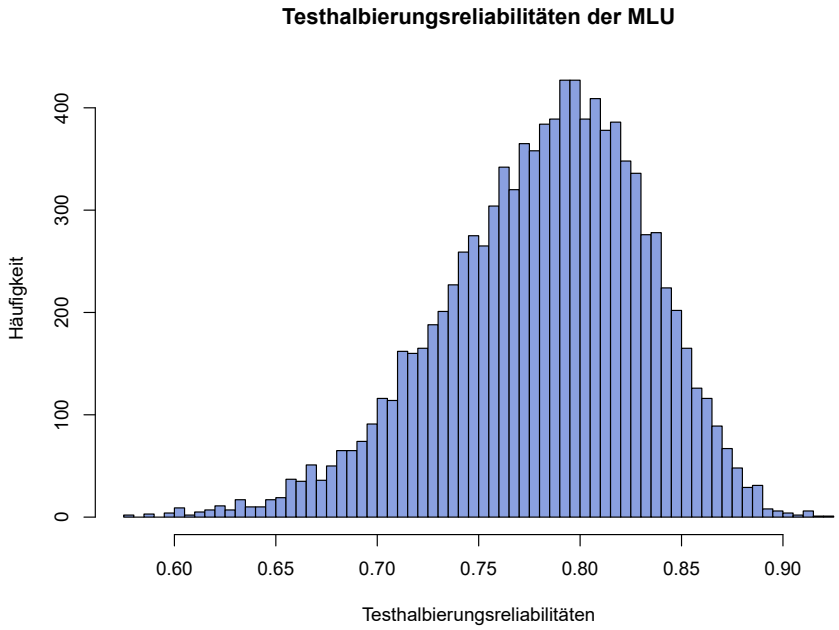

Abbildung 1: Histogramm der randomisierten Testhalbierungsreliabilitäten der MLU

In der Abbildung 2 sind die errechneten 10'000 Werte abgebildet, die aus den Testhalbierungsreliabilität bei der MATTR mit Überschneidungen der Fenster resultierten. Das Konfidenzintervall für diese Werte liegt zwischen 0.95 und 0.99.

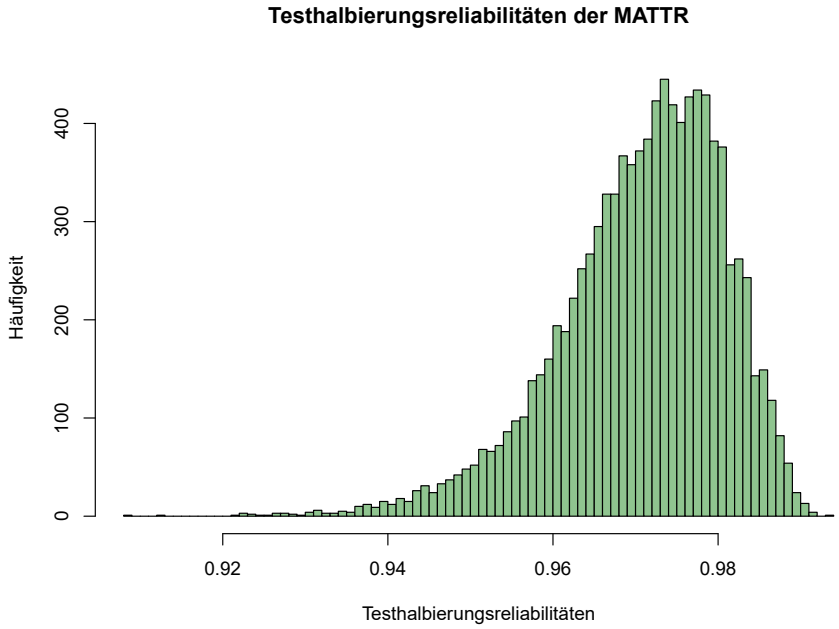

Abbildung 2: Histogramm der randomisierten Testhalbierungsreliabilitäten der MATTR mit Überschneidungen der Fenster

Die Reliabilitätswerte der zweiten Analyse für die MATTR ohne Überschneidung der Fenster, welche in der Tabelle 3 abgebildet sind, unterscheiden sich stark von der ersten. Die maximale Reliabilität liegt nun bei 0.86, was deutlich tiefer ist. Auch die minimale

(0.35) und die durchschnittliche Reliabilität (0.67) sinken deutlich. Das Cronbach's Alpha liegt bei 0.67.

Indikator	Maximale Reliabilität	Minimale Reliabilität	Durchschnittliche Reliabilität	Crohnbach's Alpha
MATTR	0.87	0.37	0.67	0.67

Tabelle 3: Schätzwerte für die instrumentelle Reliabilität der MATTR ohne Überschneidung der Fenster

Auch hier wurden die errechneten Werte in einem Histogramm dargestellt (siehe Abbildung 3). Auch hier zeigten die Daten eine Normalverteilung. Das Konfidenzintervall liegt zwischen 0.53 und 0.79 und ist somit das breiteste und niedrigste der berechneten Intervalle.

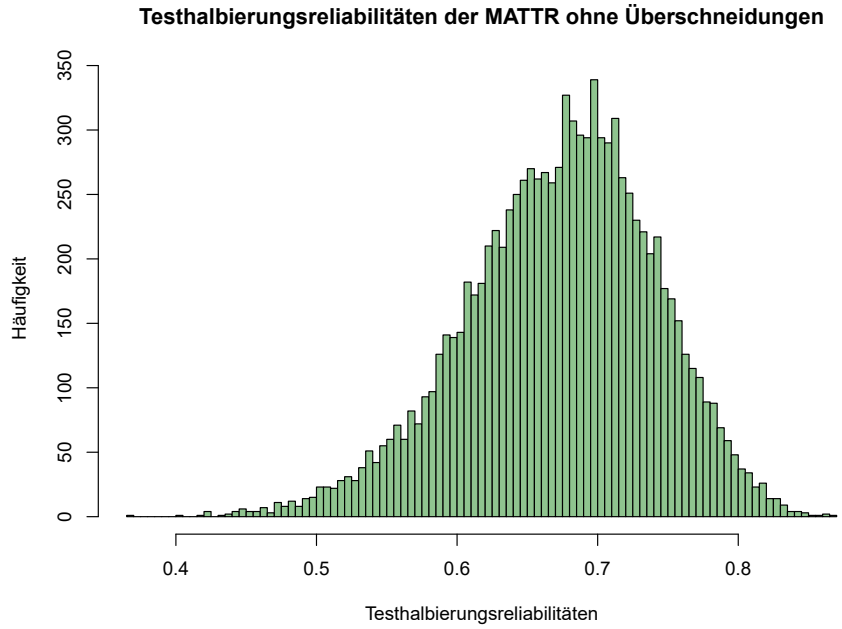

Abbildung 3: Histogramm der randomisierten Testhalbierungsreliabilitäten der MATTR ohne Überschneidung der Fenster

5.2 Reliabililtät von kurzen Sprachproben

Entsprechend des festgelegten Vorgehens wurde für die Evaluation der gekürzten Sprachproben nun nicht mehr mit den Rohwerten der MLU und der MATTR gerechnet. Diese wurden für die verschiedenen Sprachprobenlängen in die Form der entsprechenden Indikatoren gebracht. Das heisst, dass für alle Sprachprobenlängen die Durchschnitte der umfassten Rohwerte berechnet wurden. Deskriptiv zeigt sich in der Übersichtstabelle 4 zu den gekürzten Sprachproben folgendes Bild:

Sprachprobenlänge	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MLU 10 Äusserungen	1.20	3.70	5.50	6.23	7.50	19.50
MLU 20 Äusserungen	1.80	4.05	5.25	5.94	7.00	14.60
MLU 30 Äusserungen	1.97	4.43	5.67	6.11	7.23	13.50
MLU 40 Äusserungen	1.85	4.67	5.97	6.29	7.50	12.97
MLU 50 Äusserungen	2.34	4.40	5.96	6.36	7.96	12.70
MLU Gesamtsprachprobe	2.60	4.67	6.63	6.47	7.68	12.89
MATTR 35 Wörter	72.76	84.38	88.76	87.91	93.52	98.48
MATTR 70 Wörter	74.95	85.24	88.00	87.74	90.95	97.24
MATTR 105 Wörter	77.14	84.63	88.76	87.97	91.24	98.03
MATTR 140 Wörter	78.90	85.10	88.38	87.84	90.48	97.48
MATTR 175 Wörter	76.95	85.18	88.04	87.90	90.48	96.19
MATTR Gesamtsprachprobe	78.60	85.97	87.78	87.52	90.20	94.04

Tabelle 4: Deskriptive Statistiken der MLU und der MATTR bei kurzen Spontansprachproben

Die minimalen MLU-Werte steigen von der kürzesten Sprachprobe zur gesamten tendenziell an. So liegt der kleinste MLU-Wert bei 10 Äusserungen bei 1.2 und bei der gesamten Sprachprobe bei 2.6. Dies deutet darauf hin, dass Kinder, die nur sehr wenige Wörter pro Äusserung bilden, mit einer kurzen Sprachprobe tendenziell zusätzlich unterschätzt werden. Die durchschnittlichen MLU-Werte der verschiedenen Sprachprobenlängen zeigen hingegen keinen klaren Aufwärts- oder Abwärtstrend. Bei den maximalen Werten sinken die MLU-Werte tendenziell, je länger die Sprachproben sind. So liegt der maximale MLU-Wert bei 10 Äusserungen bei 19.5 und bei der gesamten Sprachprobe bei 12.89. Auch dies deutet darauf hin, dass einige Kinder mit einer Sprachprobe von 10 Äusserungen möglicherweise falsch eingeschätzt werden. Kinder, die bereits sehr lange Äusserungen bilden, werden mit einer Sprachprobenlänge von 10 Äusserungen möglicherweise zusätzlich überschätzt. Diese Hypothesen werden im folgenden mit den statistischen Schätzmethoden untersucht.

Für die minimalen Werte der MATTR zeigt sich ein unklareres Bild: Diese nehmen von 35 bis 140 Wörter klar zu (von 72.76 bis 78.90). Anschliessend sinkt der Wert zunächst bei 175 Wörter leicht ab auf 76.95 und steigt dann für die gesamte Sprachprobe wieder auf einen ähnlichen Wert (78.60) wie er bei 140 Wörtern hatte. Hier lässt sich auf den ersten Blick keine klare Vermutung zu möglichen Verzerrungen der Daten bei kurzen Sprachproben ziehen. Es gilt dennoch zu überprüfen, ob Kinder mit einem niedrigen MATTR-Wert mit einer kurzen Sprachprobe von 35 Wörtern möglicherweise tendenziell unterschätzt werden. Die Durchschnittswerte liegen für alle Sprachprobenlängen sehr nahe beieinander (87.52 - 87.97). Die maximalen MATTR-Werten verharren zwischen 35 und 140 auf einem vergleichbaren Niveau zwischen 97.24 und 98.48. Anschliessend fallen die Werte ab, bis sie für die gesamte Sprachprobe auf 94.04 liegen. Hier muss im Folgenden überprüft werden, ob Kinder mit hohen MATTR-Werten mit kurzen Sprachproben tendenziell überschätzt werden.

5.2.1 RMANOVA für die MLU

Zunächst sollen die Sprachproben mittels inferenzstatistischer Methoden verglichen werden. So kann die absolute Reliabilität auf Gruppenebene untersucht werden. Zunächst werden die Voraussetzungen der RMANOVA geprüft.

Die Normalverteilung der Daten ist insbesondere bei Stichproben, die kleiner sind als n = 40 von Bedeutung, weil dies zu falschen p-Werten führen kann. Obwohl die Stichprobe dieser Arbeit grösser als 40 ist, wurde die Normalverteilung überprüft, um sicherzugehen, dass dies kein Hindernis für die Analyse darstellt. Die Diagramme in Abbildung 4 zeigen zwar keine perfekte Normalverteilung aber auch kein komplett von der Normalverteilung abweichendes Bild. Aufgrund dessen, dass die Stichprobe etwas über den geforderten Grösse von n = 40 liegt, wurden keine Fälle ausgeschlossen und angenommen, dass diese Voraussetzung für die Analyse adäquat gegeben ist.

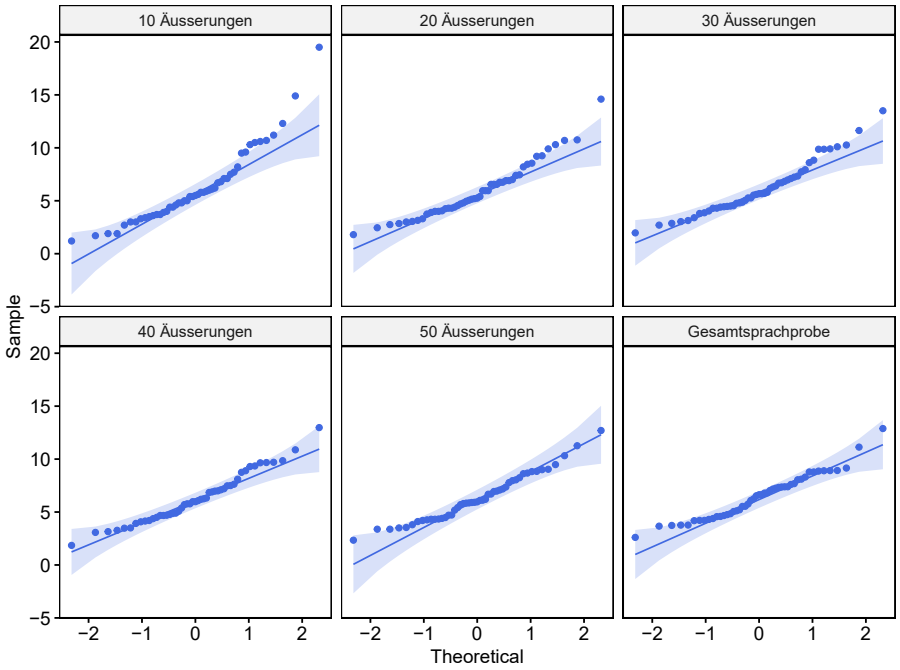

Abbildung 4: Q-Q Plots zur Verteilung der MLU-Werte

In den Boxplots der Abbildung 5 wurden die MLU-Werte gruppiert nach Sprachprobenlänge abgebildet. Die Streuung der Werte um den Median zeigen hohe Überschneidungen zwischen den Sprachprobenlängen. Auffallend ist dabei, dass sowohl bei 10 als auch bei 40 Äusserungen das untere und der obere Quartil deutlich weiter auseinander liegen als bei den anderen Sprachprobenlängen. Bei fast allen Sprachprobenlängen zeigen sich ein oder zwei Ausreisser gegen oben, wobei diese extremer ausfallen, je kürzer die Sprachprobe ist. Die exakten Werte der Mediane und Quartile der verschiedenen Sprachprobenlängen finden sich im Anhang in der Tabelle 13.

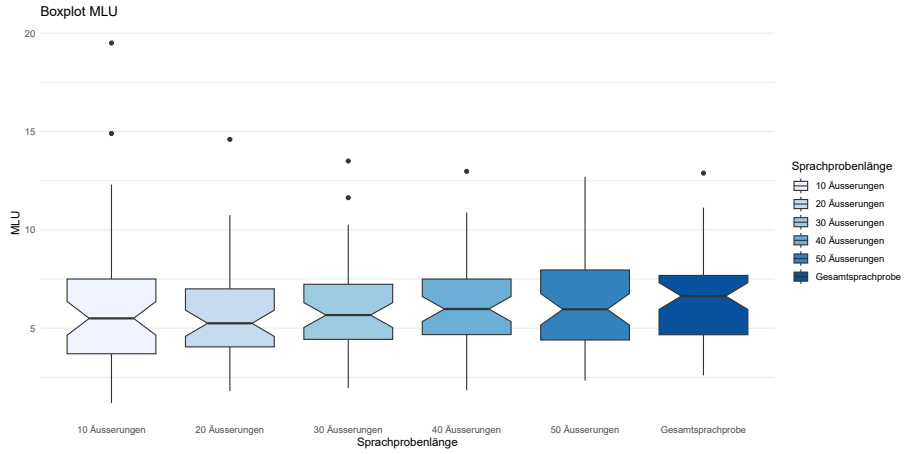

Abbildung 5: Boxplots der MLU-Werte gruppiert nach Sprachprobenlänge

Da der Mauchly's Test für Sphärizität ein signifikantes Resultat ergab (siehe Tabelle 5, p = 0.00), wurde die Greenhouse-Geisser Korrektur zur Berechnung des Modells verwendet.

Effect	W	p	p<.05
Sprachprobenlänge	0.00	0.00	*

Tabelle 5: Mauchly's Test für Sphärizität für die MLU-Werte

Das Gesamtmodell ergab mit der Greenhouse-Geisser Korrektur keinen signifikanten Effekt der Sprachprobenlängen auf die MLU (F(1.55, 74.37) = 1.53, p = 0.22). Die Freiheitsgrade zwischen den Gruppen betrugen 1.55 und innerhalb der Gruppen 74.37. Das generalisierte Eta-Quadrat (abgekürzt als «ges») für diesen Effekt betrug 0.00, was auf eine sehr geringe bis keine Effektstärke hinweist. Auch ohne die Greenhouse-Geisser Korrektur wäre das Gesamtmodell nicht signifikant geworden (F(5) = 1.53, p = 0.18), was zusätzlich darauf hindeutet, dass keine Effekt der Sprachprobenlänge auf die MLU vorliegt.

Effect	DFn	DFd	F	p	p<.05	ges
Sprachprobenlänge	1.55	74.37	1.53	0.22		0.00

Tabelle 6: Resultate des RMANOVA-Gesamtmodells für die MLU mit der Greenhouse-Geisser Korrektur

Die paarweisen Vergleiche zwischen den einzelnen Sprachprobenlängen, die mittels *post-hoc* Tests errechnet wurden, ergaben keine signifikanten Unterschiede für die MLU-Werte zwischen den Sprachprobenlängen. Die Werte wurden aufgrund der mehrfachen Vergleiche mit der Bonferroni-Korrektur berechnet. Ohne die Bonferroni-Korrektur würden einige Vergleiche signifikant, wobei die Effektstärken (berechnet mit Cohen's d) bei diesen Vergleichen klein waren (zwischen -0.34 und -0.3).

group1	group2	statistic	df	p	p.adj	effectsize
10 Äusserungen	20 Äusserungen	1.32	48.00	0.19	1.00	0.19
10 Äusserungen	30 Äusserungen	0.43	48.00	0.67	1.00	0.06
10 Äusserungen	40 Äusserungen	-0.20	48.00	0.84	1.00	-0.03
10 Äusserungen	50 Äusserungen	-0.36	48.00	0.72	1.00	-0.05
10 Äusserungen	Gesamtsprachprobe	-0.64	48.00	0.52	1.00	-0.09
20 Äusserungen	30 Äusserungen	-1.59	48.00	0.12	1.00	-0.23
20 Äusserungen	40 Äusserungen	-2.39	48.00	0.02	0.32	-0.34
20 Äusserungen	50 Äusserungen	-2.24	48.00	0.03	0.44	-0.32
20 Äusserungen	Gesamtsprachprobe	-2.36	48.00	0.02	0.33	-0.34
30 Äusserungen	40 Äusserungen	-2.07	48.00	0.04	0.66	-0.30
30 Äusserungen	50 Äusserungen	-1.78	48.00	0.08	1.00	-0.25
30 Äusserungen	Gesamtsprachprobe	-1.94	48.00	0.06	0.87	-0.28
40 Äusserungen	50 Äusserungen	-0.81	48.00	0.42	1.00	-0.12
40 Äusserungen	Gesamtsprachprobe	-1.27	48.00	0.21	1.00	-0.18
50 Äusserungen	Gesamtsprachprobe	-1.23	48.00	0.22	1.00	-0.18

Tabelle 7: Paarweise Vergleiche zwischen den Äusserungslängen für die MLU mit der Bonferroni-Korrektur

5.2.2 RMANOVA für die MATTR

Auch für die MATTR wurde die Normalverteilung trotz der genügend grossen Stichprobe von über 40 Sprachproben mittels Q-Q-Plot überprüft. Die Verteilung der Daten in Abbildung 6 zeigt eine nahezu optimale Normalverteilung. Die Voraussetzungen für die folgenden Analysen ist damit gegeben.

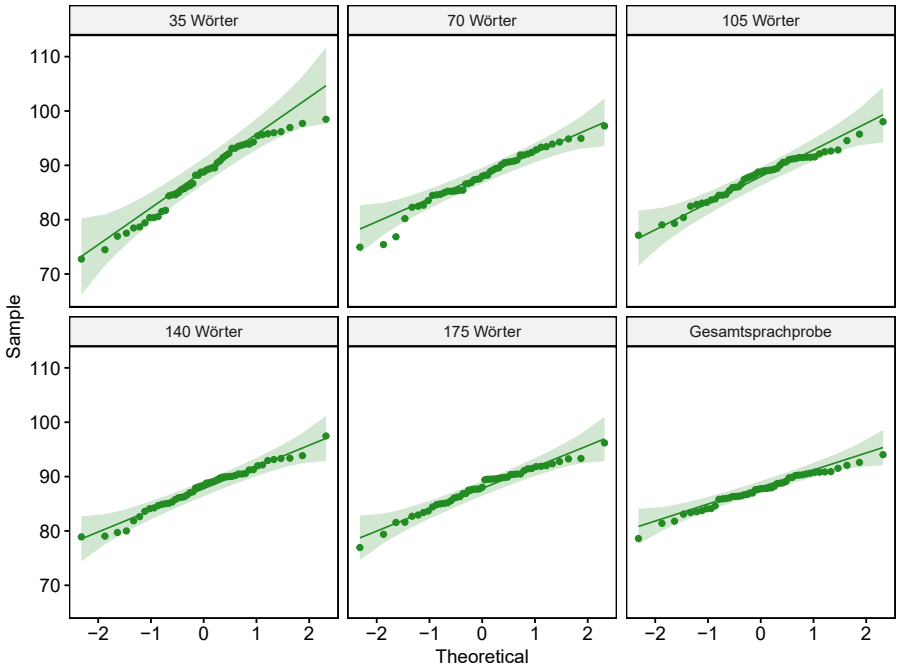

Abbildung 6: Q-Q Plots zur Verteilung der MATTR-Werte

Die Boxplots der MATTR gruppiert nach Sprachprobenlängen (siehe Abbildung 7) zeigen grosse Überschneidungen der Bereiche zwischen den oberen und unteren Quartilen. Deutlich zu erkennen ist, dass die Streuung der MATTR-Werte mit abnehmender Sprachprobenlänge zunimmt. Sowohl die Box als auch die Whiskar nehmen in der Län-

ge zu. Die Medianwerte scheinen mit abnehmender Sprachprobenlänge tendenziell leicht anzusteigen, wobei der Boxplot für 70 Wörter hier eine Ausnahme bildet. Gegen unten finden sich bei 70, 175 und den Gesamtsprachproben einzelne Ausreisser. Die exaktem Werte der Mediane und Quartile der verschiedenen Sprachprobenlängen finden sich im Anhang in der Tabelle 14.

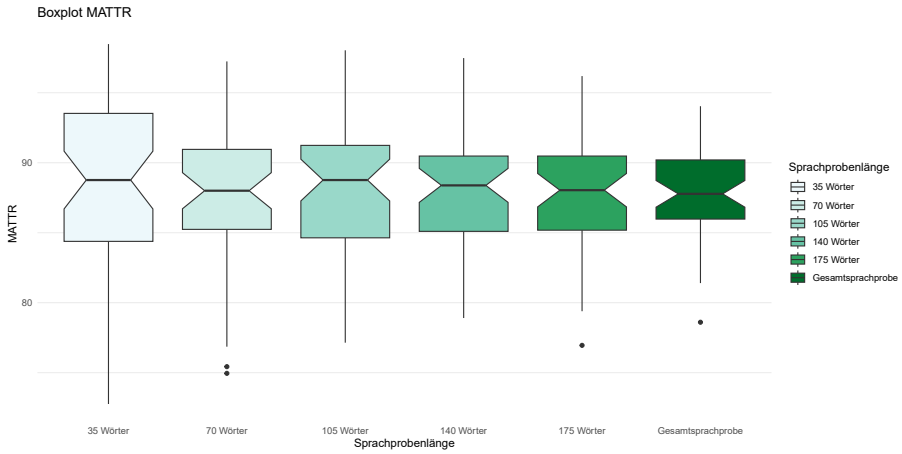

Abbildung 7: Boxplots der MATTR-Werte gruppiert nach Sprachprobenlänge

Da der Mauchly's Test für Spherizität ein signifikantes Resultat ergab (siehe Tabelle 8, p = 0.00), wurde zur die Greenhouse-Geisser Korrektur zur Berechnung des Modells verwendet.

Effect	W	p	p<.05
Sprachprobenlänge	0.01	0.00	*

Tabelle 8: Mauchly's Test für Spherizität für die MATTR-Werte

Das RMANOVA-Gesamtmodell ergab keinen signifikanten Effekt der Sprachprobenlängen auf die MATTR (F(1.90, 91.11) = 0.23, p = 0.78). Die Freiheitsgrade zwischen den Gruppen betrugen 1.90 und innerhalb der Gruppen 91.11. Das generalisierte Eta-Quadrat (ges) für diesen Effekt betrug 0.00, was auf eine sehr geringe bis keine Effektstärke hinweist. Gleiches hätte auch für das Gesamtmodell ohne Greenhouse-Geisser Korrektur gegolten (F(5) = 0.23, p = 0.95).

Effect	DFn	DFd	F	p	p<.05	ges
Sprachprobenlänge	1.90	91.11	0.23	0.78		0.00

Tabelle 9: Resultate des RMANOVA-Gesamtmodells für die MATTR mit der Greenhouse-Geisser Korrektur

Die paarweisen Vergleiche zwischen den einzelnen Sprachprobenlängen, die mittels *post-hoc* Tests errechnet wurden, ergaben sowohl mit als auch ohne die Bonferroni-Korrektur keine signifikanten Unterschiede für die MATTR-Werte zwischen den verschiedenen Sprachprobenlängen. Die Bonferroni-Korrektur wurde wiederum aufgrund der mehrfachen Vergleiche eingesetzt.

group1	group2	statistic	df	p	p.adj	effectsize
35 Wörter	70 Wörter	0.31	48.00	0.76	1.00	0.04
35 Wörter	105 Wörter	-0.09	48.00	0.93	1.00	-0.01
35 Wörter	140 Wörter	0.09	48.00	0.92	1.00	0.01
35 Wörter	175 Wörter	0.02	48.00	0.98	1.00	0.00
35 Wörter	Gesamtsprachprobe	0.50	48.00	0.62	1.00	0.07
70 Wörter	105 Wörter	-0.77	48.00	0.44	1.00	-0.11
70 Wörter	140 Wörter	-0.31	48.00	0.76	1.00	-0.04
70 Wörter	175 Wörter	-0.43	48.00	0.67	1.00	-0.06
70 Wörter	Gesamtsprachprobe	0.44	48.00	0.66	1.00	0.06
105 Wörter	140 Wörter	0.71	48.00	0.48	1.00	0.10
105 Wörter	175 Wörter	0.33	48.00	0.74	1.00	0.05
105 Wörter	Gesamtsprachprobe	1.11	48.00	0.27	1.00	0.16
140 Wörter	175 Wörter	-0.37	48.00	0.71	1.00	-0.05
140 Wörter	Gesamtsprachprobe	0.91	48.00	0.37	1.00	0.13
175 Wörter	Gesamtsprachprobe	1.11	48.00	0.27	1.00	0.16

Tabelle 10: Paarweise Vergleiche zwischen den Äusserungslängen für die MATTR mit der Bonferroni-Korrektur

5.2.3 Spearman-Brown Korrelationskoeffizient für die MLU

Der Korrelationskoeffizient der kürzesten Sprachprobe (10 Äusserungen) im Vergleich mit der Gesamtsprachprobe liegt unter der definierten Schwelle von 0.9 ($r = 0.81$). Mit 20 analysierten Äusserungen erreicht der Korrelationskoeffizient diese ebenfalls knapp noch nicht ($r = 0.88$). Für alle anderen Sprachprobenlängen liegen die Korrelationskoeffizienzen der MLU über dieser Schwelle, wobei sie weiterhin mit zunehmender Sprachprobenlänge grösser werden (siehe Tabelle 11). Um diesen Verlauf noch etwas genauer zu betrachten, wurden die Korrelationskoeffizienten in Abbildung 8 grafisch dargestellt.

Anzahl Äusserungen	Korrelationskoeffizient
10.00	0.81
20.00	0.88
30.00	0.92
40.00	0.95
50.00	0.98

Tabelle 11: Spearman-Brown Korrelationskoeffizienten für den Vergleich der MLU-Werte der gekürzten Sprachproben mit denjenigen der Gesamtsprachprobe

In der Abbildung 8 zeigt sich, dass die Korrelationskoeffizienten mit zunehmender Sprachprobenlänge zwar zunehmen, die Kurve aber abflacht. Die klarste Veränderung des Koeffizienten findet sich zwischen 10 und 20 Äusserungen. Im nächsten Zwischenschritt, zwischen 20 und 30 Äusserungen, steigt der Koeffizient deutlich weniger steil an und behält diese Steigung anschliessend in etwa bei.

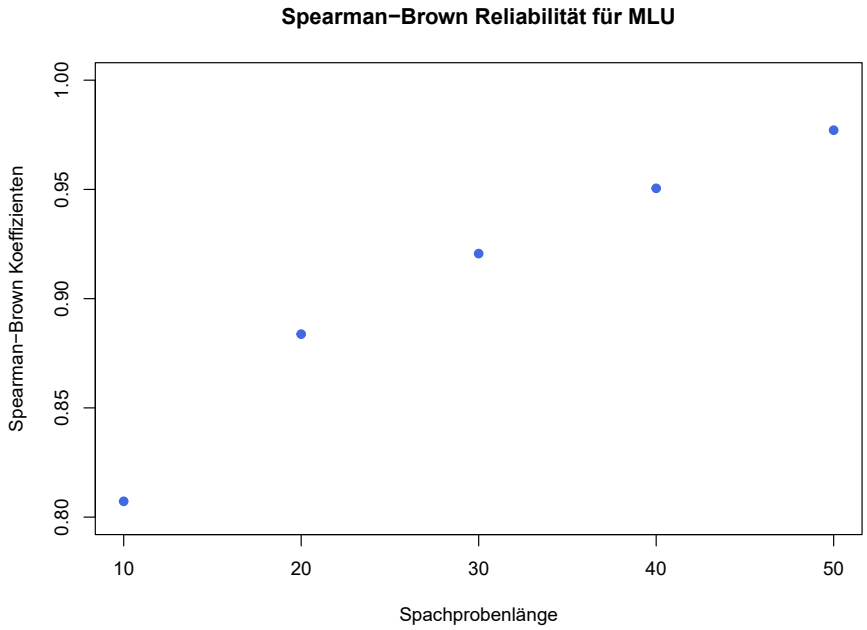

Abbildung 8: Entwicklung der Spearman-Brown Koeffizienten mit zunehmender Sprachprobenlänge für die MLU

5.2.4 Spearman-Brown Korrelationskoeffizient für die MATTR

In der Tabelle 12 zeigt sich, dass die Korrelationskoeffizienten für die MATTR ebenfalls mit abnehmender Sprachprobenlänge abnehmen. Diese liegen für alle Sprachprobenlängen (35, 70 und 105, 140 und 175 Wörter) unter der festgelegten Schwelle ($r > 0.9$). Ab der Sprachprobenlänge von 105 Wörtern liegt der Korrelationskoeffizient nur noch sehr knapp unter dieser Schwelle.

Anzahl Wörter	Korrelationskoeffizient
35.00	0.70
70.00	0.81
105.00	0.87
140.00	0.88
175.00	0.88

Tabelle 12: Spearman-Brown Korrelationskoeffizienten für den Vergleich der MATTR-Werte der gekürzten Sprachproben mit denjenigen der Gesamtsprachprobe

Auch für die MATTR wurde ein Diagramm erstellt, um den Verlauf der Korrelationskoeffizienten genauer betrachten zu können. In Abbildung 9 zeigt sich, dass die Korrelationskoeffizienten zu Beginn mit einer kleinen Zunahme der Sprachprobenlänge um 35 Wörter stark ansteigt (die Differenz beträgt 0.09). Dieser Anstieg setzt sich anschliessend verlangsamt bis zur Länge von 105 Wörtern fort. Anschliessend findet nahezu eine Stagnation des Korrelationskoeffizienten statt (Differenz zwischen 105 und 175 Wörtern beträgt 0.01).

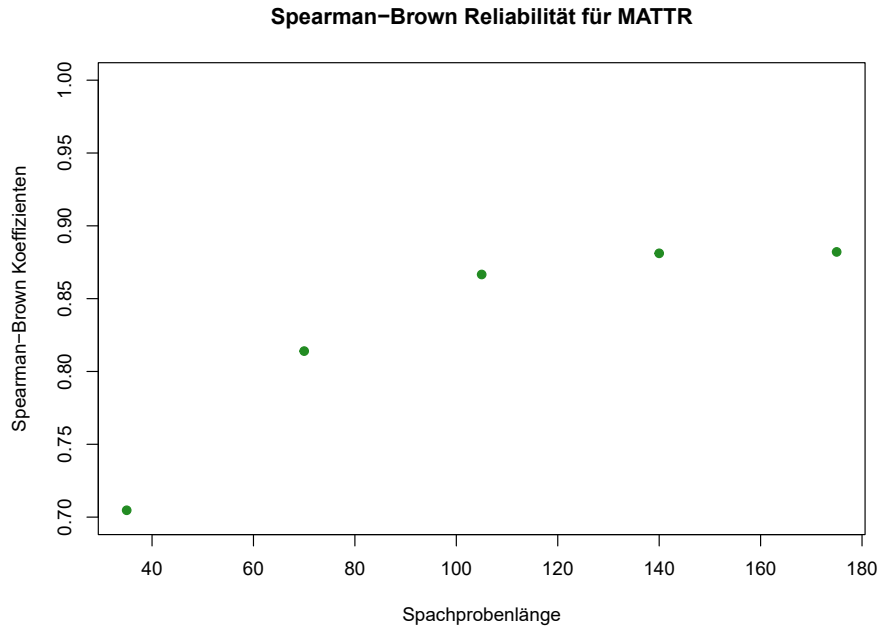

Abbildung 9: Entwicklung der Spearman-Brown Koeffizienten mit zunehmender Sprachprobenlänge für die MATTR

5.2.5 Bland-Altman Plots für die MLU

Im folgenden Abschnitt wird eine Auswahl der erstellten Bland-Altman Plots abgebildet und diskutiert. Einerseits leitet sich die Auswahl der Plots von den bisherigen Resultaten ab: Es zeigte sich, dass mittels inferenzstatistischer Verfahren kein Unterschied zwischen den Sprachprobenlängen gefunden wurde. Die Spearman-Brown Korrelationskoeffizienten erreichten aber nicht für alle Sprachprobenlängen Werte von über 0.9, wobei die Grenze bei 30 Äusserungen für die MLU lag und für die MATTR die Schwelle zwar nie überschritten wurde, die Korrelationskoeffizienten ihr aber ab 105 Wörtern sehr nahe kamen. Andererseits sind für die Praxis insbesondere die Resultate der drei kürzesten Sprachproben von Interesse, da die genannten Umfragen (siehe Kapitel 2.2) zeigten, dass mehr als die Hälfte der Logopäd:innen in der Praxis weniger als 50 Äusserungen transkribieren und die praktische Anwendbarkeit der LSA deutlich steigt, je kürzer die Sprachproben sein können. Daher werden die Plots der drei kürzesten Sprachproben im Folgenden genauer beschrieben. Die Kennwerte der anderen Sprachprobenlängen werden kurz beschrieben und die Plots finden sich im Anhang.

Die Abbildung 10 zeigt den erstellten Bland-Altman Plot für den Vergleich der MLU-Messungen bei 10 Äusserungen und der gesamten Sprachprobe. Die mittlere gestrichelte Linie zeigt die durchschnittliche Differenz der MLU-Werte dieser beiden Sprachprobenlängen und liegt bei -0.2. Das bedeutet, dass die MLU-Werte, wenn diese mit nur 10 Äusserungen berechnet werden, durchschnittlich um 0.2 tiefer liegen, als wenn mit der gesamten Sprachprobe gemessen wird. Das Diagramm zeigt deutlich, dass die Differenzen der MLU-Werte stark schwanken. Es gibt sowohl Kinder, die mit den ersten 10 Äusse-

rungen einen deutlich höheren MLU-Wert erreichen als mit der gesamten Sprachprobe als auch Kinder, die mit den ersten 10 Äusserungen einen deutlich tieferen MLU-Wert haben (Punkte über und unterhalb des Nullwerts der y-Achse). Die Differenzen der Mittelwerte scheinen nicht ganz zufällig um den Mittelwert zu streuen. Bei Kindern, die tiefe durchschnittliche MLU-Werte erreicht haben (etwa zwischen 2 und 6), scheint die Messung mit nur 10 Äusserungen die MLU in der Tendenz eher zu unterschätzen (hier fallen die Differenzen der Mittelwerte gehäuft in den negativen Bereich). Die Grenzen der Übereinstimmung (untere und obere gestrichelte Linie) liegen für diesen Vergleich bei -5.4 und 4.9, wobei die Differenz der Mittelwerte der MLU bei drei Kindern ausserhalb dieser Grenzen zu liegen kommen. Entsprechend liegen 94 % der Datenpunkte innerhalb der LOA, was einem klinisch akzeptablen Resultat entspricht (mindestens 90 % sollten innerhalb des Konfidenzintervalls liegen).

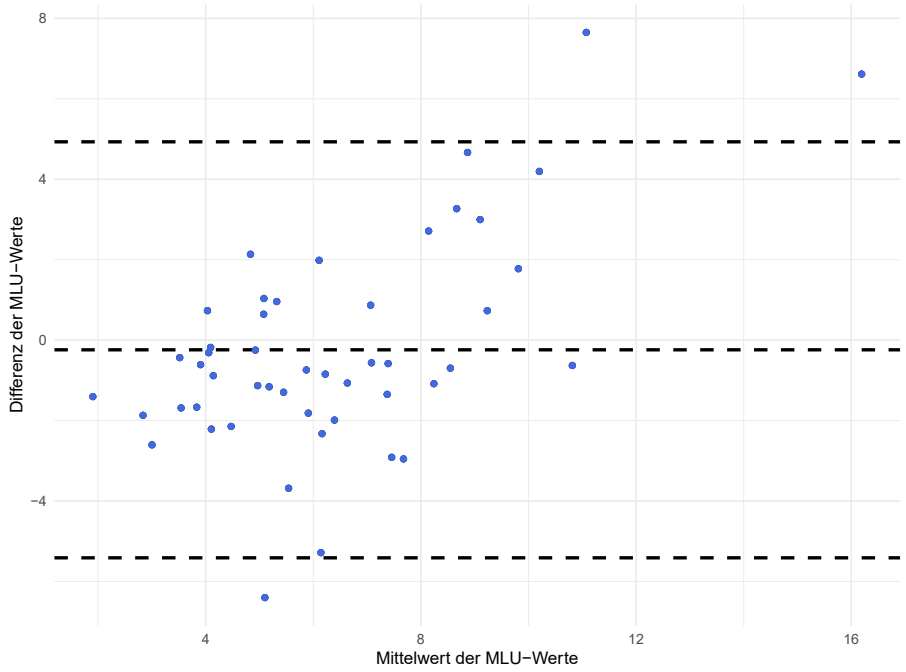

Abbildung 10: Bland-Altman Plot für den Vergleich der MLU-Werte von 10 Äusserungen und der gesamten Sprachprobe

In Abbildung 11 wurden die MLU-Werte von 20 Äusserungen mit den Gesamtsprachproben verglichen. Die mittlere Differenz der MLU-Werte liegt bei -0.5, was eine etwas grössere Differenz der Mittelwerte widerspiegelt als bei 10 Äusserungen. Dennoch ist aufgrund der geringen Abweichung nicht von einem grundlegenden Messunterschied auszugehen. Die Grenzen der Übereinstimmung liegen bei -3.7 und 2.5, was deutlich keiner ist als bei der kürzesten Sprachprobe in Abbildung 10. Es liegen wiederum drei Differenzwerte ausserhalb der LOA, was klinisch akzeptabel ist. Die Differenzen der MLU-Werte scheinen zufällig um den Mittelwert zu streuen. Die Tendenz, die MLU-Werte von Kindern zu überschätzen, die eine eher kurze Äusserungslänge aufweisen, scheint bei 20 Äusserungen weniger stark ausgeprägt zu sein.

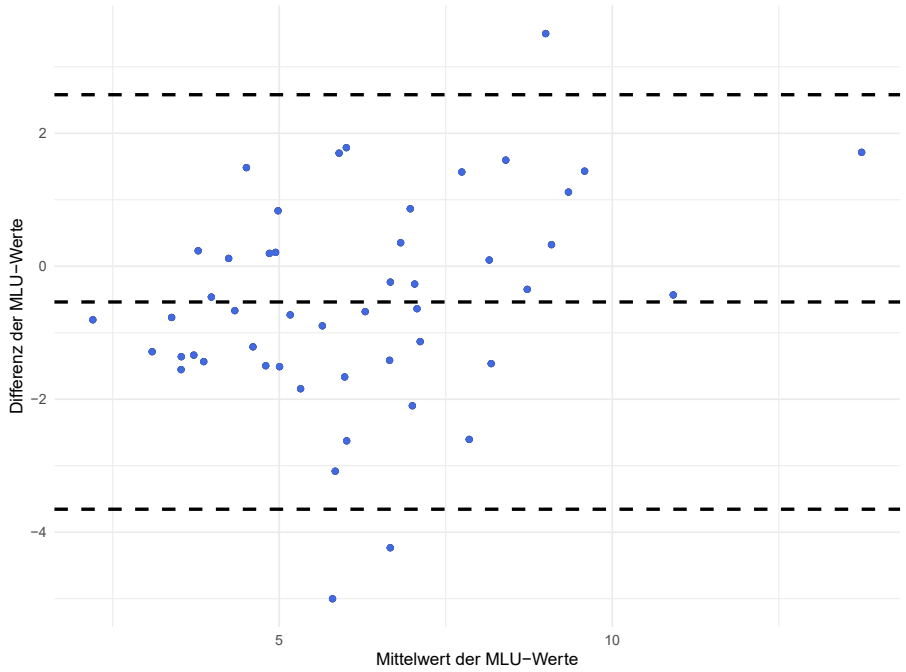

Abbildung 11: Bland-Altman Plot für den Vergleich der MLU-Werte von 20 Äusserungen und der gesamten Sprachprobe

Der B-A Plot für den Vergleich von 30 Äusserungen mit der gesamten Sprachprobe (Abbildung 12) zeigt wiederum eine etwas kleinere durchschnittliche Differenz der MLU-Werte von -0.4. Die Grenzen der Übereinstimmung liegen nochmals etwas näher zusammen bei -2.9 und 2.2, wobei nun vier Datenpunkte ausserhalb dieser Linien liegen. Dies entspricht weiterhin einem klinisch akzeptablen Wert (92 %). Es zeigt sich wie schon bereits bei 20 Äusserungen keine systematische Verteilung der Datenpunkte um den Mittelwert der Differenzen.

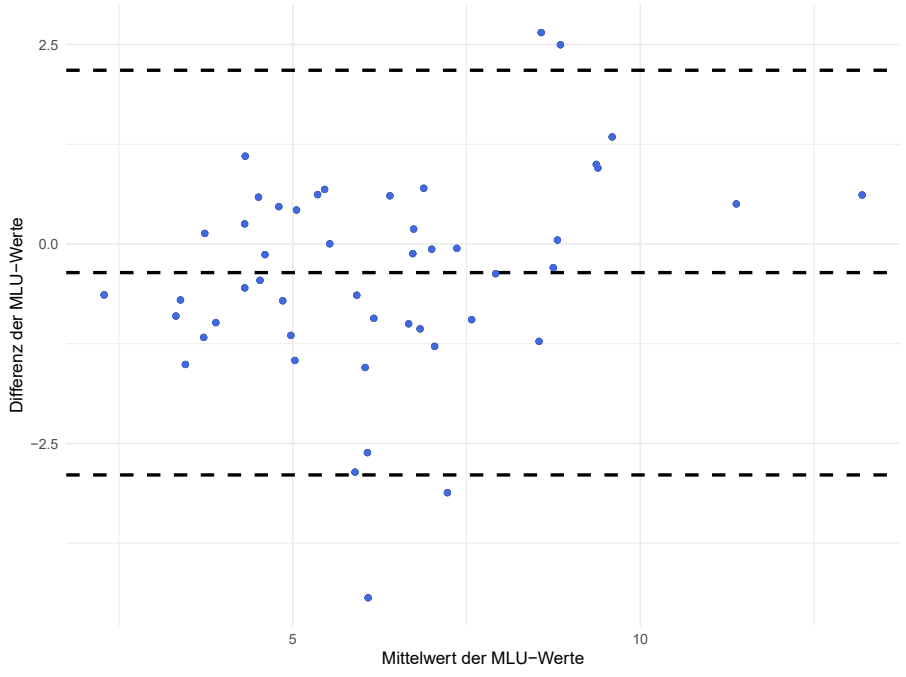

Abbildung 12: Bland-Altman Plot für den Vergleich der MLU-Werte von 30 Äusserungen und der gesamten Sprachprobe

Die Plots der anderen gekürzten Sprachproben befinden sich im Anhang (siehe Abbildungen 16 und 17) und werden hier kurz beschrieben. Die mittlere Differenz der MLU-Werte nimmt mit zunehmender Sprachprobenlänge ab und liegt bei 40 Äusserungen bei -0.2. Die Grenzen der Übereinstimmung befinden sich für 40 Äusserungen bei -2.1 und 1.8. 92 % der Daten liegen innerhalb dieser Grenzen. Die Differenzen der Mittelwerte streuen sowohl für 40 als auch für 50 Äusserungen zufällig um den Mittelwert der Differenzen. Bei 50 Äusserungen liegt die durchschnittliche Differenz der Mittelwerte bei -0.1 und die Grenzen der Übereinstimmung bei -1.4 und 1.2. 90 % der Daten liegen zwischen diesen Grenzen, was knapp in Rahmen der klinischen Akzeptanz liegt.

5.2.6 Bland-Altman Plots für die MATTR

Entsprechend der besprochenen Plots für die MLU und aufgrund der beschriebenen Überlegungen finden sich im Folgenden die B-A Plots für die drei kürzesten Sprachproben. In Bezug auf die MATTR entspricht dies 35, 70 und 105 Wörtern. Die anderen Diagramme werden kurz besprochen und sind im Anhang einsehbar.

Die Abbildung 13 zeigt den Vergleich der MATTR-Werte von 35 Wörtern und den gesamten Sprachproben. Der Mittelwert der Differenzen der MATTR-Werte liegt bei 0.4. Wenn die MATTR also mit 35 Wörtern berechnet wird, liegt dieser Wert durchschnittlich um 0.4 höher als wenn die MATTR mit der gesamten Sprachprobe berechnet wird. Die Differenzen der MATTR-Werte scheinen nicht zufällig um den Mittelwert zu streuen. Je tiefer respektive je höher der Mittelwert der beiden MATTR-Werte ist, desto grösser scheint die Differenz der MATTR-Werte zu sein. Bei Kindern, die eher wenig verschiedene Wörter verwenden (also einen tieferen MATTR-Wert haben), führt die Kürzung der Sprachproben tendenziell zu einer Unterschätzung der MATTR, während sie bei Kindern, die viele verschiedene Wörter verwenden, eher zu einer Überschätzung führt. Die untere Grenze der Übereinstimmung liegt bei -10.5, die obere bei 11.3. 98 % der Daten liegen innerhalb dieses breiten Feldes.

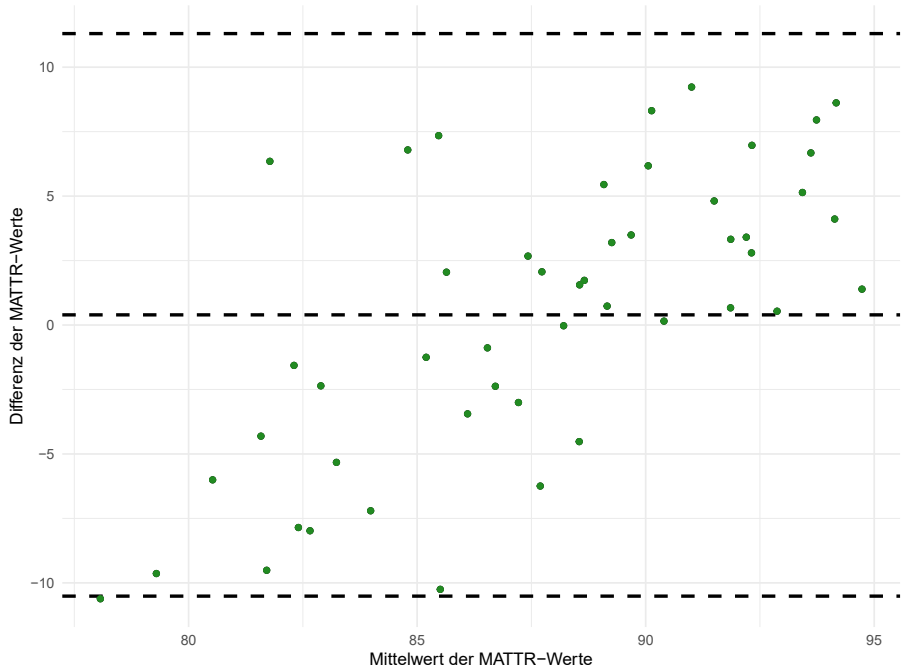

Abbildung 13: Bland-Altman Plot für den Vergleich der MATTR-Werte von 35 Wörtern und der gesamten Sprachprobe

Der Durchschnitt der Differenzen für den Vergleich von 70 Wörtern und den gesamten Sprachproben liegt bei 0.2 (siehe Abbildung 14). Die Grenzen der Übereinstimmung liegen bei -6.8 und 7.2. 94 % der Daten liegen innerhalb dieser Grenzen. Die Über- und Unterschätzung von Kindern mit einem hohen respektive einem tiefen MATTR-Wert scheint im Vergleich zur Abbildung 13 für 35 Wörter deutlich abzunehmen.

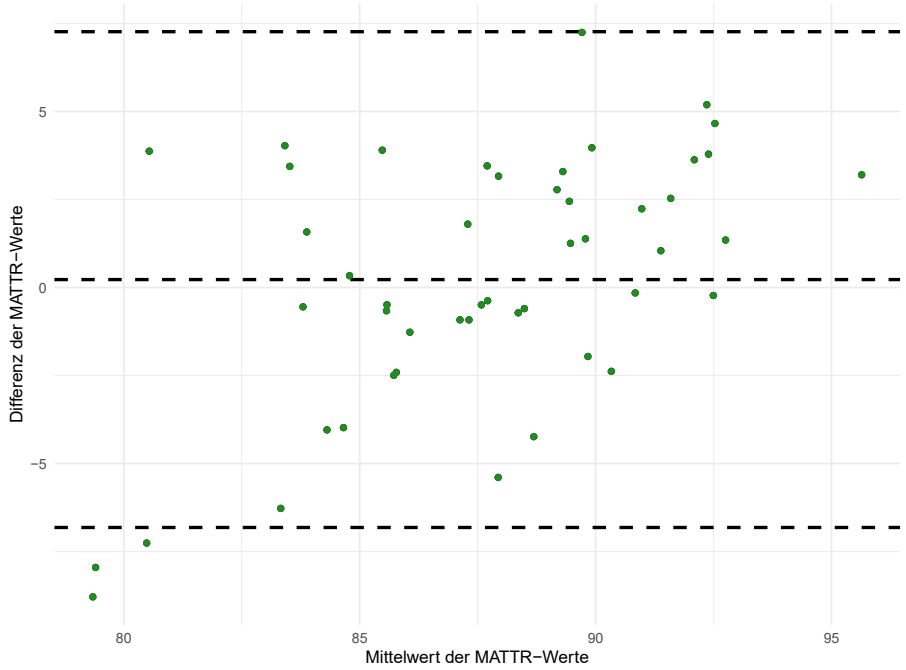

Abbildung 14: Bland-Altman Plot für den Vergleich der MATTR-Werte von 70 Wörtern und der gesamten Sprachprobe

Der B-A Plot für den Vergleich von 105 Wörtern und den gesamten Sprachproben (Abbildung 15) zeigt eine mittlere Differenz der MATTR-Werte von 0.5. Die Grenzen der Übereinstimmung liegen bei -5.1 und 6.0, wobei drei Datenpunkte nicht in diesem Bereich liegen (94 % innerhalb der Grenzen). Die Datenpunkte streuen zufällig um die durchschnittliche Differenz.

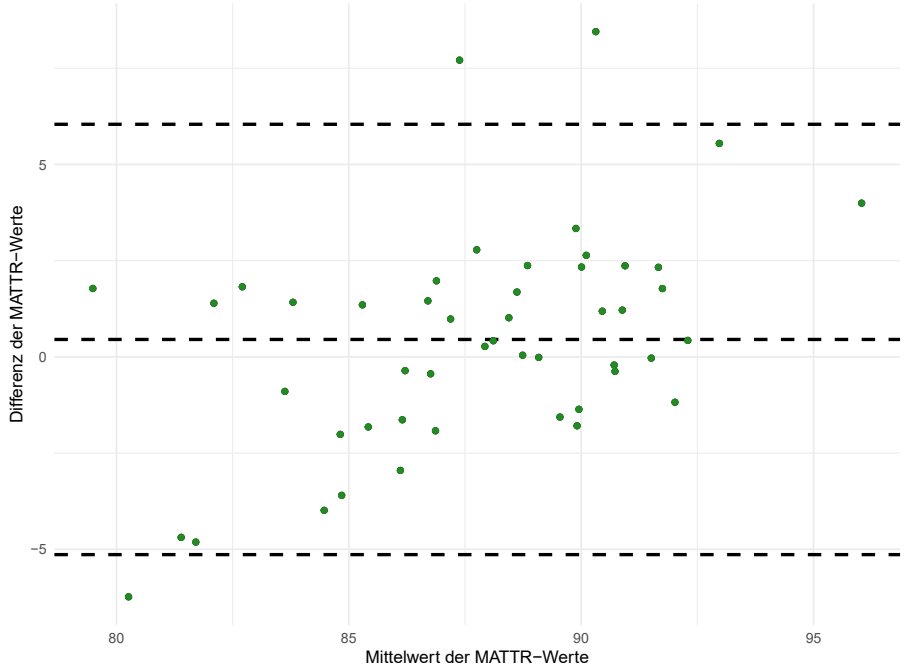

Abbildung 15: Bland-Altman Plot für den Vergleich der MATTR-Werte von 105 Wörtern und der gesamten Sprachprobe

Die durchschnittliche Differenz der MATTR-Werte beim Vergleich von 140 Wörtern und den gesamten Sprachproben (siehe Abbildung 20 im Anhang) liegt bei 0.4. Die Werte scheinen zufällig um diesen Mittelwert verteilt. Die untere Grenze der Übereinstimmung liegt bei -4.7, die obere bei 5.3. Hier liegen 96 % Der Daten innerhalb der Grenzen. Die mittlere Differenz für den Vergleich von 175 Wörtern und den gesamten Sprachproben liegt bei 0.4. Die Grenzen der Übereinstimmung liegen bei -4.3 und 5.1, 94 % der Daten liegen dazwischen (siehe Abbildung 21 im Anhang).

6 Diskussion

In diesem Abschnitt sollen die zu Beginn der Arbeit gestellten Fragen anhand der gefundenen Resultate beantwortet werden. Ein Bezug zur bisherigen Forschung wird hergestellt. Zur Beantwortung der zweiten Fragestellung wird zunächst auf die Unterfragen eingegangen, um anschliessend zusammenfassend die Frage nach der Mindestlänge einer Sprachprobe beantworten zu können.

Unterfrage 1: Wie hoch ist die instrumentelle Reliabilität der MLU und der MATTR bei den Sprachproben?

Das Cronbach's Alpha lag für die MLU bei 0.78, was der Definition von Price (2016) folgend, einem respektablen Wert entspricht. Das Konfidenzintervall der Testhalbierungsreliabilitäten zeigte sich als relativ breit (0.67 - 0.87). Ein Wert von 0.67 ist knapp als nicht genügend reliabel einzuschätzen, eine von 0.87 als hoch. Die Verteilung der Reliabilitätswerte im Histogramm (Abbildung 1) zeigt, dass die allermeisten berechneten Werte über 0.7 und somit in akzeptablen Bereich liegen. Für die nachfolgenden Analysen wird daher von einer genügenden instrumentellen Reliabilität ausgegangen.

Um diesen Wert genauer einschätzen zu können, wäre eine erneute Durchführung der Analyse mit der gesamten Sprachprobe spannend. Grundsätzlich müsste das Konfidenzintervall kleiner werden, je mehr Datenpunkte vorhanden sind. Dementsprechend wäre es interessant zu sehen, ob das Konfidenzintervall mit mehr Sprachproben weiterhin die Grenze von 0.7 umfassen würde oder nicht. Dadurch würde die Hypothese der instrumentellen Reliabilität an Robustheit gewinnen.

Im Vergleich zur Studie von Cole et al. (1989), der einzigen, die die MLU ebenfalls mittels der Testhalbierungsreliabilität untersuchte und dabei einen Wert von 0.94 fand, sind diese Werte niedrig. Einen Anteil an diesem Unterschied hat möglicherweise die unterschiedliche Berechnung der MLU. Während in der Studie von Cole et al. (1989) die einzelnen Äusserungen anhand der Intonation am Ende einer Äusserung und bei Pausen ab 2 Sekunden segmentiert wurden, geschah dies in der vorliegenden Arbeit anhand von Sprecherwechseln. Wenn ein Kind länger von einem Erlebnis erzählte, ohne dabei vom Gegenüber unterbrochen zu werden, kann so eine sehr viel längere Äusserung entstehen als mit den von Cole et al. (1989) angewandten Kriterien. Dies führte dazu, dass die gemessenen MLU-Werte bei Cole et al. (1989) vermutlich deutlich näher beieinander lagen. Grosse Unterschiede in den einzelnen Werten können dazu führen, dass sowohl das Cronbach's Alpha als auch die Testhalbierungsreliabilitäten sinken.

Für die MATTR-Werte mit überschneidenden Fenstern ergab sich ein sehr hoher Wert für das Cronbach's Alpha ($\alpha = 0.97$). Auch die Testhalbierungsreliabilitäten waren sehr hoch und das Konfidenzintervall lag zwischen 0.95 und 0.99. Wie im Kapitel 4.4 beschrieben, waren für die MATTR mit überschneidenden Fenstern hohe Werte zu erwarten, da den Items jeweils ein grosser gemeinsamer Anteil an äquivalenten Wörtern zu Grunde liegt. Dennoch sind diese Werte ein wichtiger Hinweis auf die instrumentelle Reliabilität der MATTR, da genau diese Überschneidungen dem zugrundeliegenden Konzept des Indikators inhärent sind. Entsprechend waren auch die deutlich tieferen Werte der zweiten Analyse ohne Überschneidungen tendenziell erwartbar. Das Cronbach's Alpha ($\alpha = 0.67$) liegt dennoch auch für diese Berechnungsart im minimal akzeptablen Bereich. Die durchschnittliche Testhalbierungsreliabilität ($r = 0.67$) liegt leicht unter der geforderten Schwelle. Das Histogramm (Abbildung 3) zeigt die breiteste Streuung der Testhalbierungsreliabilität. Dies spiegelt sich auch im breiten Konfidenzintervall (0.53 - 0.79).

Diese Analyse zeigt deutlich, dass der Indikator ohne die Überschneidungen der Fenster instrumentell nicht reliabel ist, da einzig das Cronbach's Alpha knapp im akzeptablen Bereich liegt. Dies ist ein Hinweis darauf, dass die Überschneidungen der MATTR-Fenster aus Sicht der Reliabilität sehr sinnvoll sind. Da diese in dieser Arbeit auch so berechnet wurden, wird für die nachfolgenden Analysen von einer genügenden instrumentellen Reliabilität ausgegangen.

In Bezug auf die erste Fragestellung wird also für beide Indikatoren von einer genügenden instrumentellen Reliabilität ausgegangen. In Bezug auf die MLU wäre aufgrund der Resultate eine Veränderung in der Segmentation der Äusserungen sinnvoll, da dies klarere Werte der instrumentellen Reliabilität erwarten lassen würde und die Vergleichbarkeit mit anderen Studien ermöglichen würde. Auch in Bezug auf den logopädischen Alltag hätte eine Segmentation nach Intonation oder nach Sinneinheiten eine grössere Aussagekraft, da beispielsweise das Stufenmodell von Clahsen (1986) zu jeder Stufe Äusserungslängen zuordnet, deren Segmentation nicht auf Sprecherwechseln basieren.

Unterfrage 2.1: Wie zeigt sich die absolute Reliabilität von kurzen Spontansprachanalysen auf der Gruppenebene?

Die Resultate der RMANOVA keine signifikanten Effekt der Sprachprobenlänge auf die MLU und die MATTR. Auch die *post-hoc* Tests ergaben keine signifikanten Unterschiede zwischen den verschiedenen Längen.

In Bezug auf die MLU waren diese Resultate im Gegensatz zu der MATTR erst nach der Bonferroni-Korrektur eindeutig. Ohne die Korrektur ergaben sich für vier Vergleiche signifikante Unterschiede: 20 und 40 Äusserungen, 20 und 50 Äusserungen, 20 Äusserungen und die gesamte Sprachprobe und 30 und 40 Äusserungen. Dies könnte darauf hindeuten, dass die absolute Reliabilität auf Gruppenebene bei 20 Äusserungen etwas geringer ist als bei den anderen Sprachprobenlängen. Da mit der Korrektur aber keiner der Vergleiche mehr zu signifikanten Resultaten führte, darf dieses Ergebnis wohl nicht zu stark überbewertet werden. Zumal die signifikanten Resultate ohne die Bonferroni-Korrektur auch falsch positiven Ergebnissen entsprechen könnten.

Diese Resultate decken sich mit der Studie von Guo und Eisenberg (2015), bei der für die MLU-m keine signifikanten Unterschiede zwischen den Sprachprobenlängen (1, 3, 7 und 10 Minuten verglichen mit 22 Minuten) gefunden wurden. Auch Wilder und Redmond (2022), Heilmann et al. (2010) und Casby (2011) fanden in ihren Untersuchung keine signifikanten Unterschiede zwischen den Sprachprobenlänge. Eine weitere inferenzstatistische Untersuchung findet sich bei Pavelko et al. (2020), welche eine lineare *mixed-model* Analyse für den Vergleich der MLU bei 25 und 50 Äusserungen durchführten. Auch diese ergab keinen signifikanten Effekt der Sprachprobenlängen.

Die Resultate der inferenzstatistischen Analyse des Einflusses der Sprachprobenlängen auf die MLU (mit unterschiedlichen Berechnungsvarianten) scheint also über verschiedene Sprachen (Englisch und Schweizerdeutsch), verschiedene Elizitationsmethoden (Freispiel, dialogisch und *personal narratives*), Altersgruppen (2;8 - 13;3 Jahre) und Kinder mit und ohne Spracherwerbsstörungen stabil zu sein und keine Unterschiede zu Tage zu führen.

Die vorliegende Arbeit zeigte erstmals, dass auch für die MATTR in der schweizerdeutschen Sprache keine signifikanten Unterschiede zwischen den Sprachprobenlängen bestehen.

Unterfrage 2.2: Wie hoch ist die relative Reliabilität von kurzen Spontansprachproben?

Die zuvor festgelegte Schwelle für eine genügende relative Reliabilität wurde bei der MLU mit 30 Äusserungen überschritten ($r = 0.92$). Die beiden kürzeren Sprachproben (10 und 20 Äusserungen) können aufgrund dieser Analyse nicht zur Anwendung empfohlen werden.

In der Analyse von Wilder und Redmond (2022) lagen die Spearman-Brown Korrelationskoeffizienten für die MLU bei den zwei kürzesten Sprachproben ebenfalls unter der Schwelle von 0.9. Sowohl bei Kindern mit und ohne Spracherwerbsstörungen wurde die Schwelle bei einer Länge von 7 Minuten überschritten. Noch längere Sprachproben (10 Minuten) wurden in der Untersuchung von Guo und Eisenberg (2015) benötigt, um diese Schwelle für die MLUC-m zu überschreiten ($r = 0.93$). In dieser Untersuchung wurde der Pearson's Korrelationskoeffizient verwendet. Heilmann et al. (2010) untersuchten die relative Reliabilität der kürzeren Sprachproben mit dem Cronbach's Alpha. Für die MLU fanden sie sowohl für die Länge von 1 Minute als auch für 3 Minuten minimal akzeptable Werte (ausser für die narrative Elizitationsmethode bei Kindern zwischen 6;0 und 13;3 Jahren) Für die dialogische Elizitationsmethode erreichten alle Werte ein akzeptables Niveau. Casby (2011) fanden für die ersten 10 Äusserungen der Sprachprobe einen Pearson's Korrelationskoeffizienten von $r = 0.86$, was auf eine hohe relative Reliabilität hinweist. Insgesamt zeigt sich bezüglich der relativen Reliabilität ein weniger klares Bild und die benötigte Sprachprobenlänge unterscheidet sich stark.

Für die MATTR wurde die festgelegte Schwelle von 0.9 für keine der Äusserungslängen überschritten. Interessanterweise zeigte sich für diesen Indikator bei 0.88 eine Stagnation der relativen Reliabilität. Dies erstaunt insbesondere im Vergleich zur vorhergegangenen Analyse mittels RMANOVA, bei der die Resultate der MATTR deutlicher darauf hinwiesen, dass keine Unterschiede zwischen den Sprachprobenlängen bestehen, während dies bei der MLU erst mit der Bonferroni-Korrektur eindeutig zu Tage trat. Ebenfalls konfrontiert mit einem Korrelationskoeffizienten knapp unter dieser Schwelle waren Guo und Eisenberg (2015):

«For instance, NDW/m and MLCUm from the 7-min conversational samples both correlated with those from the standard sample at a level of .88 ... , which was just slightly below the benchmark level. Does this mean that NDW/m and MLUm from 7-min samples are clinically unreliable? We do not have a clear answer for this question. ... clinicians will vary in the level of correlation that they consider desirable for interpreting LSA measures and, accordingly, will vary in decisions about sample length. If one believes that an incremental difference in correlation coefficients (i.e., .88 vs. .90) does not translate into clinically significant distinctions, collecting 7-min conversational samples (or approximately 63 utterances) to compute TNW, NDW, and MLCUm might be even more feasible in the clinical setting.» (Guo & Eisenberg, 2015, S. 150)

Die Frage, ob ein Korrelationskoeffizient von 0.87 (105 Wörter) oder 0.88 (140 und 175 Wörter) bedeutet, dass die Werte klinisch unreliabel sind, stellt sich auch in dieser Arbeit. Hier jedoch bezieht sich diese Frage statt auf die MLU auf die MATTR. Wie auch Guo und Eisenberg (2015) kann diese Arbeit keine klare Antwort auf diese Frage liefern. Übertragbar ist sicherlich die These, dass das Erheben und Auswerten von 105 Wörtern in Praxis machbarer scheint, als wenn mindestens 183 Wörter benötigt werden.

Da die verschiedenen Untersuchungen unterschiedliche Schätzmethoden für die relative Reliabilität verwendeten, bleiben die Resultate nur bedingt vergleichbar. Das Bild bleibt unklarer als bei der inferenzstatistischen Analyse. Generell scheint es längere Sprachproben zu brauchen, um eine akzeptable relative Reliabilität zu erreichen im Vergleich zu einer akzeptablen absoluten Reliabilität auf der Gruppenebene. Die vorliegenden Analysen deuten für die schweizerdeutsche Sprache darauf hin, dass für die MLU mindestens 30 Äusserungen verwendet werden sollten. Für die MATTR scheint eine relative Reliabilität von 0.9 kaum mit gekürzten Sprachproben erreichbar zu sein. Wenn allerdings davon ausgegangen wird, dass ein Korrelationskoeffizient von 0.87 nicht zu klinisch relevanten Veränderungen führt, kann eine Sprachprobe von 105 Wörtern verwendet werden.

Unterfrage 2.3: Wie zeigt sich die absolute Reliabilität von kurzen Spontansprachanalysen auf der individuellen Ebene?

Grundsätzlich kann für die MLU festgehalten werden, dass keine der Analysen einen für die Praxis relevanten *bias* im Sinne einer systematischen Verzerrung der durchschnittlich gemessenen Werte zutage führte. Die durchschnittlichen Differenzen der Messungen lag zwischen - 0.1 und - 0.5.

Dennoch spricht auch die Analyse mittels B-A Plots gegen die Verwendung von 10 Äusserungen zur Erhebung der MLU. Die Daten scheinen bei so kurzen Sprachproben einer Verzerrung zu unterliegen (regelmässige Verteilung der Datenpunkte). Ausserdem lässt das Konfidenzintervall hier eine grosse Spannbreite an Differenzen zwischen den beiden gemessenen MLU-Werten zu. Auch wenn die MLU in der vorliegenden Arbeit innerhalb eines Sprecherwechsels berechnet wurden, ist eine Differenz von nahezu minus oder plus

5 Wörtern eine grosse Spannbreite. Ein Kind könnte also in der gesamten Sprachprobe eine MLU von 7 Wörtern und in der gekürzten von 2 aufweisen und würde noch immer zwischen die Grenzen der Übereinstimmung fallen. In der logopädischen Praxis wäre eine solche Differenz klinisch relevant. Diese Schwierigkeit der Festlegung einer akzeptablen Breite der LOA in den B-A Plots diskutierten auch Wilder und Redmond (2022). Eine von ihnen vorgeschlagene Lösungsoption ist die Berechnung der LOA mit ± 1.00 Standardabweichung. Dies führt dazu, dass die Grenzen der Übereinstimmung enger zusammenrücken und so eine konservativere Schätzung der Reliabilität besteht. Wenn bei einer solchen Berechnung noch immer 90 % der Daten innerhalb der LOA liegen, gilt auch diese konservativere Schätzung als klinisch reliabel. Die Festlegung auf ± 1.00 SD's ergab sich aus den Daten der Studie.

In den Diagrammen der längeren Sprachproben (20 - 50 Äusserungen) nahm die Breite der LOA stetig ab. Auch konnten keine systematischen Verzerrungen in der Verteilung der Punktwolken mehr erkannt werden. Der B-A Plot für 30 Äusserungen zeigte die Grenzen der Übereinstimmung bei - 2.9 und 2.2. Die obere LOA lag damit unter der festgelegten Schwelle von 2.5, während die untere weiterhin leicht darüber lag. Um die MLU bei 30 Äusserungen genauer zu untersuchen, wurde ein weiterer Plot erstellt. Dabei wurde die Spannbreite der LOA mittels Senkung der LOA auf ± 1.75 SD's verkleinert, um eine restriktivere Schätzung zu erhalten (Diagramm siehe Anhang, Abbildung 18). Der Multiplikator der Standardabweichungen wurde auf diesen Wert gesetzt, weil dabei (im Vergleich zu noch kleineren SD's) noch 90 % der Fälle zwischen der unteren und der oberen Grenze der Übereinstimmung zu liegen kamen. Dieses Diagramm zeigte eine untere LOA von -2.6 und eine obere von 1.9, was sehr knapp nicht im definierten Rahmen der akzeptablen Spannbreite liegt. Ebenfalls im Anhang findet sich ein Diagramm, welches mit 1.65 SD's berechnet wurde und wegen eines Datenpunktes nicht mehr im klinisch akzeptablen Bereich liegt (Abbildung 19). In diesem Diagramm lagen aber beide LOA unter der Schwelle von ± 2.5. Mit diesen beiden zusätzlichen Diagrammen kann gezeigt werden, dass die festgelegten Grenzen der absoluten Reliabilität auf der individuellen Ebene mit einer Sprachprobenlänge von 30 Äusserungen nahezu erreicht werden.

Ab einer Sprachprobenlänge von 40 Äusserungen lagen die LOA im Bereich der gesetzten Grenzen. Hier wird die absolute Reliabilität auf der individuellen Ebene definitiv genügend erreicht. In der Praxis wäre eine minimale Sprachprobenlänge von 30 Äusserungen deutlich machbarer als eine minimale Sprachprobenlänge von 40 Äusserungen. Insbesondere, wenn diese im Rahmen einer Verlaufsdiagnostik durchgeführt werden sollte. Wie eben diskutiert kommen die Resultate der MLU bei 30 Äusserungen den gesetzten Grenzen der akzeptablen Differenzen sehr nahe. Es ist daher davon auszugehen, dass eine Kürzung der Sprachproben auf 30 Äusserungen nur mit einem kleinen Verlust bezüglich der absoluten Reliabilität auf der individuellen Ebene einhergeht. Es bleibt je nach Zweck

der Erhebungen abzuwägen, inwiefern ein solcher Verlust an Reliabilität der Daten im akzeptablen Rahmen liegt oder nicht.

Auch bezüglich der MATTR kann grundsätzlich festgehalten werden, dass bei keinem der Diagramme ein *bias* gefunden wurde. Die Mittelwerte der Differenzen liegen zwischen 0.2 und 0.5.

Wie bei der MLU zeigte sich auch für die MATTR bei der kürzesten untersuchten Sprachprobe (35 Wörter) eine Verzerrung der Punktwolke und eine grosse Spannbreite zwischen der unteren und der oberen LOA (- 10.5 und 11.3). Diese Unterschiede entsprechen in der Praxis klar klinisch relevanten Unterschieden. Auch hier kann eine Verwendung einer derart kurzen Sprachprobe nicht empfohlen werden.

Eine Sprachprobe von 70 Wörtern kam den definierten Grenzwerten der LOA von ± 6.6 deutlich näher. Die Spannbreite der LOA betrug hier noch -6.8 und 7.2. Auch hier wurde entsprechend ein weiteres Diagramm mit einem geringeren Multiplikator der Standardabweichung generiert, Das Diagramm mit dem Multiplikator von 1.75 findet sich im Anhang (Abbildung 22). Durch diese Veränderung kamen die LOA bei -6.1 und 6.5 zu liegen und 90 % der Daten lagen zwischen den Grenzen der Übereinstimmung. Aufgrund dessen, dass auch diese restriktivere Schätzung klinisch akzeptabel ist, scheint die absolute Reliabilität der MATTR bei einer Sprachprobenlänge von 70 Wörtern genügend hoch zu sein.

Ab einer Sprachprobenlänge von 105 zeigten die Diagramme eindeutig eine genügende absolute Reliabilität auf der individuellen Ebene.

Unterfrage 2: Führen kürzere Sprachproben zu vergleichbaren Resultaten wie die gesamten Sprachproben?

Die Resultate der durchgeführten Analysen verdeutlichen, weshalb es relevant ist, die Reliabilität von kurzen Sprachproben nicht bloss anhand einer davon zu untersuchen. Durch die Beleuchtung unterschiedlicher Facetten der Reliabilität kamen auch unterschiedliche Resultate zu Stande. Wäre beispielsweise nur die RMANOVA durchgeführt worden, hätte dies zu einer stark veränderten Interpretation geführt. In diesem Aspekt führen die Sprachproben zu sehr ähnlichen Resultaten wie die gesamten Sprachproben - bei der relativen Reliabilität fällt die Antwort durchaus anders aus. Im Folgenden soll zunächst auf die MLU und anschliessend auf die MATTR eingegangen werden.

Um auf die eingangs gestellte Frage, wie lange eine Spontansprachprobe bei Kindern mit und ohne Spracherwerbsstörungen im Alter von 4 - 7 Jahren sein muss, um reliable Werte zu erhalten, zurückzukommen, lässt sich zusammenfassend folgendes sagen: Mittels der inferenzstatistischen Analyse fanden sich bei der MLU keine statistischen Unterschiede, hier zeigten sich bereits 10 Äusserungen als reliabel. Eine genügend hohe relative Reliabilität wurde bei 30 Äusserungen erreicht. Die Analyse der absoluten Reliabilität

mittels Bland-Altman Plots ergab für alle Sprachproben klinisch akzeptables Resultat, da immer mehr als 90 % der Datenpunkte innerhalb des Konfidenzintervalls lagen. Erst ab 30 Äusserungen konnten die zugelassenen Differenzen der MLU (die Grenzen der Übereinstimmung) knapp als klein genug erachtet werden (festgelegt auf ± 2.5). Insgesamt scheint für die MLU mit einer Segmentation nach Sprecherwechseln eine Mindestlänge der Sprachproben von 30 Äusserungen empfehlenswert zu sein. Dies zu Erheben sollte mit den meisten Kindern innerhalb von circa 4 Minuten möglich sein.

Für die MATTR präsentiert sich die folgende Lage: die inferenzstatistische Untersuchung (ebenfalls RMANOVA) ergab auch hier keine signifikanten Unterschiede zwischen den Sprachproben. Eine genügende Reliabilität wird laut dieser Analyse demnach bereits bei 35 Wörtern erreicht. Die relative Reliabilität hingegen wurde mit keiner der gekürzten Sprachproben hoch genug. Die relative Reliabilität scheint also mit gekürzten Sprachproben für die MATTR nicht erreicht zu werden. Die absolute Reliabilität auf individueller Ebene wurde ab einer Sprachprobenlänge von 70 Wörtern als erreicht interpretiert. Ein detaillierter Blick auf die Spearman-Brown-Koeffizienten zeigt, dass die relative Reliabilität ab 105 Wörtern mit 0.87 mehr oder weniger knapp unter der Schwelle von 0.9 stagniert. Wenn also davon ausgegangen wird, dass sich eine Differenz von 0.03 nicht klinisch relevant auf die Resultate der MATTR auswirken, liegt die Empfehlung der minimalen Sprachprobenlänge für die Berechnung der MATTR bei 105 Wörtern. Ansonsten sollte an der Medianlänge der hier analysierten Sprachproben (545 Wörter) festgehalten werden, was innerhalb von 10 Minuten mit der vorgestellten Elizitationsmethode erreicht werden kann.

Sowohl die zeitlichen Empfehlungen für die MLU als auch diejenigen für die MATTR lassen sich nicht auf alle Kinder übertragen. Bei zurückhaltenden Kindern wird mehr Zeit nötig sein, als bei sprechfreudigen Kindern. Hier sollten sich die Logopäd:innen an der Anzahl Äusserungen orientieren.

Die Resultate dieser Arbeit reihen sich ein, in eine Reihe von neueren Untersuchungen, die sich mit der Frage der Reliabilität von kürzeren Sprachproben befasst haben. Einige dieser Studien kamen zum Schluss, dass wohl auch weniger als 50 Äusserungen zu reliablen Ergebnissen führen, andere nicht:

«So far, several studies have demonstrated that estimates are rather controversial, and there is no general agreement about language sample size.»(Voniati et al., 2021, S. 29)

Aufgrund der unterschiedlichen Konzeptionen der Studien - die einen untersuchen die Anzahl Minuten, die anderen die Anzahl Äusserungen und so weiter - bleibt die Vergleichbarkeit leider erschwert und es ist herausfordernd, übergreifende Antworten zu finden. Die eingangs gestellte Frage, wie lange Spontansprachproben sein müssen, um reliable Werte zu erhalten, kann auf Basis der vorliegenden Daten und Analysen folgendermassen beantwortet werden: Für die Sprachproben in der schweizerdeutschen Sprache, erhoben mittels

personal narratives bei Kindern im Alter von 4-7 Jahren mit und ohne Spracherwerbsstörungen scheinen 30 Äusserungen für die MLU und 105 Wörter für die MATTR zu reliablen Resultaten zu führen. Sollen beide Indikatoren in einer Sprachprobe analysiert werden, reicht dementsprechend bei den meisten Kindern eine 4-minütige Aufnahme, um reliable Werte zu generieren. Diese kurze Dauer liegt auch für die regelmässige Verlaufsdiagnostik definitiv im machbaren Bereich.

7 Limitationen und zukünftige Forschung

Zum Abschluss dieser Arbeit sollen nun zunächst die Limitationen dieser Arbeit diskutiert werden. In einem nächsten Schritt wird ein Ausblick auf mögliche zukünftige Forschung gegeben. Dabei werden sowohl mögliche Folgeuntersuchungen dieser Arbeit diskutiert, als auch mögliche Weiterführungen für das Projekt und die übergreifende LSA-Forschungen angesprochen.

Diese Arbeit untersuchte die Auswirkungen der Sprachprobenlänge auf einen Indikator der Äusserungslänge (MLU, segmentiert nach Sprecherwechseln) und einen Indikator der semantisch-lexikalischen Ebene (MATTR). Leider konnte für die Analyse nur ein Teil der Gesamtstichprobe der im DigiSpon 1 gesammelten Daten verwendet werden, da noch nicht alle Sprachproben transkribiert werden konnten. Obwohl die Stichprobe mit n = 49 Sprachproben vergleichsweise bereits nicht klein ist, wäre es aufschlussreich die Analysen nochmals mit der gesamten Stichprobe zu wiederholen. Je grösser die Stichprobe ist, desto kleiner werden die Konfidenzintervalle der instrumentellen Reliabilität (der Testhalbierungsreliabilitäten). Da das Konfidenzintervall bei der MLU aktuell die Schwelle von 0.7 noch knapp miteinfasst, könnte eine Überprüfung dieser Analyse mit einer grösseren Stichprobe beispielsweise hier ein klareres Bild zeichnen.

Eine Schwäche der Analysen der MLU in dieser Arbeit ist, dass diese anhand der Sprecherwechsel segmentiert wurden. Die allermeisten Studien wie beispielsweise Guo und Eisenberg (2015) und Wilder und Redmond (2022) und Pavelko et al. (2020) handhaben dies anders. Für die weitere Entwicklung des Tools und der Vergleichbarkeit der Resultate sollte daher in Betracht gezogen werden, die Segmentation der Äusserungen anzupassen. Dies würde auch die Interpretation der B-A Plots erleichtern. Bezüglich des MATTR müssten zur erleichterten Analyse der B-A Plots genauere Kenntnisse zur Entwicklung bei Kindern erfasst werden. Dazu würde eine standardisierte Fenstergrösse enorm helfen, da die Studien so vergleichbar würden. Je mehr Wissen bezüglich der Entwicklung der MATTR bei Kindern besteht, desto klarer kann auch bestimmt werden, wie gross die Differenz zwischen zwei Werten sein darf ohne klinisch relevant zu sein.

Auch wurden in dieser Arbeit keine Untersuchungen nur anhand von Daten von Kindern mit Spracherwerbsstörungen durchgeführt. Dazu wäre neben der Sprachprobe-

nerhebung die Durchführung eines standardisierten Verfahrens sinnvoll, um die Kinder zweifelsfrei klassifizieren zu können. Da aber genau diese Kinder das logopädische Klientel darstellen, wären Aussagen spezifisch zu dieser Population von grosser Bedeutung.

Um die kürzeren Sprachproben mit den Gesamtsprachproben zu vergleichen, wurden die Äusserungen zu Beginn einer Sprachprobe analysiert. Dieses Vorgehen findet sich so auch in Guo und Eisenberg (2015) und Wilder und Redmond (2022). Es ist aber zu beachten, dass dieses Vorgehen dazu führt, dass sich die analysierten Daten der kurzen und der gesamten Sprachprobe teilweise überschneiden, was die Reliabilitätsschätzungen möglicherweise beeinflusst haben könnte.

Diese Arbeit beschäftigte sich ausschliesslich mit der Reliabilität von kurzen Sprachproben. Weder die Validität noch die Objektivität wurden untersucht. Aber auch auf der Seite der Reliabilität bleiben Fragen offen, die in zukünftigen Studien untersucht werden könnten. In Bezug auf die erarbeitete Elizitationsmethode stellt sich beispielsweise die Frage nach der Testwiederholungsreliabilität. Da ein Teilziel des DigiSpon 1-Projektes auch die (semi-) automatisierte Transkription und Analyse der Sprachproben ist, könnte des weiteren die Interrater-Reliabilität untersucht werden. Um auf die Validität zurückzukommen, bleibt die Frage ungeklärt, ob die gemessenen Werte auch den wahren Fähigkeiten der Kinder entsprechen. Um diese Frage zu beantworten, könnte beispielsweise ein Vergleich mit einem standardisierten Test zur Rate gezogen werden. Auch sollte die zukünftige Forschung das Augenmerk auf die Sensitivität und Spezifität von kurzen Spontansprachproben legen. Denn auch wenn aus kürzeren Sprachproben reliable Werte gewonnen werden können, bleibt unklar, ob die einzelnen Kinder aufgrund der Ergebnisse korrekt in die Gruppe der Kinder mit und ohne Spracherwerbsstörungen klassifiziert würden. Es sollte daher überprüft werden, ob ein Zusammenhang der Sprachprobenlänge und der diagnostischen Güte besteht. Diese Untersuchungen würden eine klarere Entscheidungsgrundlage bieten, wie lange eine Sprachprobe sein muss.

Eine klare Limitation dieser Arbeit besteht darin, dass die gefundenen Resultate nicht zwingend auf andere Indikatoren übertragen werden können. Um generelle Empfehlungen abgeben zu können, müsste auch die Reliabilität dieser anderen Indikatoren untersucht werden. Aufgrund dessen, dass auch die MATTR keinen perfekten Indikator für die semantisch-lexikalische Ebene darstellt, wäre die Entwicklung anderer Indikatoren auf dieser Ebene sinnvoll.

Erstrebenswert als Fernziel wäre der Aufbau einer Vergleichsdatenbank für Spontansprachproben in der schweizerdeutschen Sprache. Dadurch könnten Normen generiert werden, die im diagnostischen Prozess mittels Spontansprachanalyse sehr hilfreich sein könnten.

In Bezug auf die angesprochene Option, die LSA als Tool zur Verlaufsdiagnostik einzusetzen, bleiben weiterhin Fragen ungeklärt: Die Leistungen von Kindern können von

ihrer Tagesverfassung (Motivation, Aufmerksamkeit und Stimmung) abhängen und sich so möglicherweise von Tag zu Tag unterscheiden und so unabhängig von der instrumentellen Messgenauigkeit eines diagnostischen Verfahrens zu unterschiedlichen Resultaten führen. Für die regelmässige Fortschrittsdiagnostik im Sinne des Response-to-Intervention-Ansatzes der inklusiven Schule sind weitere Untersuchungen bezüglich dieser Schwierigkeit wünschenswert, um sicherzustellen, dass diejenigen Kinder Unterstützung erhalten, die tatsächlich in einem Lerninhalt abgehängt wurden und nicht diejenigen, die einen schlechten Tag hatten.

Literatur

Bawayan, R., & Brown, J. A. (2022). Language Sample Analysis Consideration and Use: A Survey of School-based Speech Language Pathologists. *Clinical Archives of Communication Disorders*, *7*(1), 15–28.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical methods in medical research*, *8*(2), 135–160.

Brotz, J., & Döring, N. (2006). *Forschungsmethoden und evaluation für human- und sozialwissenschaftler* (4. Aufl.). Springer Medizin Verlag.

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it and how is it measured? *Physiotherapy, 86*(2), 94–99.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition*, *1*(1).

Bundesrat der Schweizerischen Eidgenossenschaft. (2004). Behindertengleichstellungsgesetz (BehiG) [SR 151.3].

Casby, M. W. (2011). An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. *Child Language Teaching and Therapy*, *27*(3), 286–293.

Channell, M. M., Loveall, S. J., Conners, F. A., Harvey, D. J., & Abbeduto, L. (2018). Narrative language sampling in typical development: Implications for clinical trials. *American Journal of Speech-Language Pathology*, *27*(1), 123–135.

Charest, M., & Skoczylas, M. J. (2019). Lexical Diversity Versus Lexical Error in the Language Transcripts of Children With Developmental Language Disorder: Different Conclusions About Lexical Ability. *American Journal of Speech-Language Pathology*, *28*(3), 1275–1282.

Charest, M., Skoczylas, M. J., & Schneider, P. (2020). Properties of Lexical Diversity in the Narratives of Children With Typical Language Development and Developmental Language Disorder. *American journal of speech-language pathology*, *29*(4), 1866–1882.

Clahsen, H. (1986). *Die Profilanalyse: ein linguistisches Verfahren für die Sprachdiagnose im Vorschulalter*. Marhold.

Cole, K. N., Mills, P. E., & Dale, P. S. (1989). Examination of test-retest and split-half reliability for measures derived from language samples of young handicapped children. *Language, Speech, and Hearing Services in Schools*, *20*(3), 259–268.

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspectives on Language Learning and Education*, *17*(1), 9–15.

De Anda, S., Cycyk, L. M., Durán, L., Biancarosa, G., & McIntyre, L. L. (2023). Sentence Diversity in Spanish-English Bilingual Toddlers. *American journal of speech-language pathology*, *32*(2), 576–591.

Downing, S. M., & Yudkowsky, R. (Hrsg.). (2009). *Assessment in health professions education*. Routledge.

Ebert, K. D., & Pham, G. (2017). Synthesizing information from language samples and standardized tests in school-age bilingual assessment. *Language, Speech, and Hearing Services in Schools*, *48*(1), 42–55.

Ebert, K. D., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools*, *45*(4), 337–350.

Eisenberg, S. L., Fersko, T. M., & Lundgren, C. (2001). The Use of MLU for Identifying Language Impairment in Preschool Children: A Review. *American journal of speech-language pathology*, (10), 323–342.

Eisenberg, S. L., & Guo, L.-Y. (2015). Sample size for measuring grammaticality in preschool children from picture-elicited language samples. *Language, Speech, and Hearing Services in Schools*, *46*(2), 81–93.

Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, cronbach, or spearman-brown? *International Journal of Public Health*, *58*(4), 637–642.

Escobedo, A. G., Gallagher, J. F., Potapova, I., Pham, G., & Pruitt-Lord, S. (2023). Understanding (un)grammaticality in context: Evidence from young Spanish-English bilinguals over time. *Journal of Communication Disorders*, *101*, 106281.

Evans, J. L., & Miller, J. (1999). Language sample analysis in the 21st century. *Seminars in speech and language*, *20*(2), 101–15, quiz 116.

Gallagher, J. F., & Hoover, J. R. (2020). Measure What You Treat: Using Language Sample Analysis for Grammatical Outcome Measures in Children With Developmental Language Disorder. *Perspectives of the ASHA Special Interest Groups*, *5*(2), 350–363.

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research*, *39*(6), 1258–1262.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151.

Guo, L.-Y., & Eisenberg, S. (2015). Sample length affects the reliability of language sample measures in 3-year-olds: Evidence from parent-elicited conversational samples. *Language, Speech, and Hearing Services in Schools*, *46*(2), 141–153.

Guo, L.-Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent grammatical utterances between 4 and 9 years of age for the edmonton narrative norms instrument: Reference data and psychometric properties. *American Journal of Speech-Language Pathology*, *28*(4), 1448–1462.

Guo, L.-Y., Schneider, P., & Harrison, W. (2021). Clausal density between ages 4 and 9 years for the edmonton narrative norms instrument: Reference data and psy-

chometric properties. *Language, Speech, and Hearing Services in Schools*, *52*(1), 354–368.

Heilmann, J., DeBrock, L., & Riley-Tillman, T. C. (2013). Stability of measures from children's interviews: the effects of time, sample length, and topic. *American Journal of Speech-Language Pathology*, *22*(3), 463–475.

Heilmann, J., Miller, J. F., Iglesias, A., Fabiano-Smith, L., Nockerts, A., & Andriacchi, K. D. (2008). Narrative Transcription Accuracy and Reliability in Two Languages. *Topics in language disorders*, *28*(2), 178–188.

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: does the length of the transcript matter? *Language, Speech, and Hearing Services in Schools*, *41*(4), 393–404.

Heilmann, J., & Westerveld, M. F. (2013). Bilingual language sample analysis: Considerations and technological advances. *Journal of Clinical Practice in Speech-Language Pathology*, *15*(2), 87–93.

Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders*, *38*(3), 197–213.

Justice, L. M. (2006). Evidence-based practice, response to intervention, and the prevention of reading difficulties. *Language, Speech, and Hearing Services in Schools*, *37*(4), 284–297.

Kapantzoglou, M., Fergadiotis, G., & Auza Buenavides, A. (2019). Psychometric Evaluation of Lexical Diversity Indices in Spanish Narrative Samples From Children With and Without Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *62*(1), 70–83.

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the united states. *Child Language Teaching and Therapy*, *13*(2), 161–176.

Kempe Preti, S. (2023). *Digital unterstützte Spontansprachanalyse - DigiSpon 1* [HfH]. Verfügbar 7. August 2023 unter https://www.hfh.ch/projekt/digital-unterstuetzte-spontansprachanalyse-digispon-1

Klatte, I. S., Van Heugten, V., Zwitserlood, R., & Gerrits, E. (2022). Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs. *Language, Speech, and Hearing Services in Schools*, *53*(1), 1–16.

Leonard, L. B., Haebig, E., Deevy, P., & Brown, B. (2017). Tracking the growth of tense and agreement in children with specific language impairment: Differences between measures of accuracy, diversity, and productivity. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3590–3600.

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*, *66*(7), 2421–2433.

Lüdtke, U., Bornman, J., De Wet, F., Heid, U., Ostermann, J., Rumberg, L., Van Der Linde, J., & Ehlert, H. (2023). Multidisciplinary perspectives on automatic analysis of children's language samples: Where do we go from here? *Folia Phoniatrica et Logopaedica*, *75*(1), 1–12.

Lüdtke, U., Ehlert, H., Gaigulo, D., & Bornman, J. (2023). Research on the methodology of LSA with preschool children: A scoping review. *Clinical Archives of Communication Disorders*, *8*(2), 29–46.

Lundine, J. P. (2020). Assessing Expository Discourse Abilities Across Elementary, Middle, and High School. *Topics in language disorders*, *40*(2), 149–165.

MacFarlane, H., Salem, A. C., Bedrick, S., Dolata, J. K., Wiedrick, J., Lawley, G. O., Finestack, L. H., Kover, S. T., Thurman, A. J., Abbeduto, L., & Fombonne, E. (2023). Consistency and reliability of automated language measures across expressive language samples in autism. *Autism Research*, *16*(4), 802–816.

MacWhinney, B., & Fromm, D. (2022). Language sample analysis with TalkBank: An update and review. *Frontiers in Communication*, *7*, 865498.

McDaniel, J., & Brady, N. C. (2022). The Influence of Communication Sample Length on Reliability and Convergent Validity of Vocal Measures Derived From the Communication Complexity Scale. *Journal of Speech, Language, and Hearing Research*, *65*(10), 3881–3889.

McKee, G. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*(3), 323–338.

Miller, J. F. (1991). *Research in Child Language Disorder: A Decade of Progress*. Pro-Ed.

Moosbrugger, H., & Kelava, A. (Hrsg.). (2020). *Testtheorie und Fragebogenkonstruktion*. Springer Berlin Heidelberg.

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: examining the use of complex syntax. *Journal of Speech, Language, and Hearing Research*, *57*(3), 876–886.

Overton, C., Baron, T., Pearson, B. Z., & Ratner, N. B. (2021). Using free computer-assisted language sample analysis to evaluate and set treatment goals for children who speak african american english. *Language, Speech, and Hearing Services in Schools*, *52*(1), 31–50.

Owen, A. J., & Leonard, L. B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: application of D. *Journal of Speech, Language, and Hearing Research*, *45*(5), 927–937.

Owens, R. E., & Pavelko, S. L. (2020). Sampling Utterances and Grammatical Analysis Revised (SUGAR): Quantitative Values for Language Sample Analysis Measures in 7- to 11-Year-Old Children. *Language, Speech, and Hearing Services in Schools*, *51*(3), 734–744.

Park, E., Cho, M., Ki, C.-S., et al. (2009). Correct use of repeated measures analysis of variance. *Korean J Lab Med*, *29*(1), 1–9.

Paul, R. (Hrsg.). (2007). *Language disorders from a developmental perspective: Essays in honor of Robin S. Chapman*. Erlbaum.

Pavelko, S. L., & Owens, R. E. (2019a). Diagnostic Accuracy of the Sampling Utterances and Grammatical Analysis Revised (SUGAR) Measures for Identifying Children With Language Impairment. *Language, Speech, and Hearing Services in Schools*, *50*(2), 211–223.

Pavelko, S. L., & Owens, R. E. (2019b). SUGAR (Sampling Utterances and Grammatical Analysis Revised): Breaking Tradition. *Language, Speech, and Hearing Services in Schools*, *50*(3), 452–456.

Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, *47*(3), 246–258.

Pavelko, S. L., Price, L. R., & Owens, R. E. (2020). Revisiting reliability: Using sampling utterances and grammatical analysis revised (SUGAR) to compare 25- and 50-utterance language samples. *Language, Speech, and Hearing Services in Schools*, *51*(3), 778–794.

Price, L. R. (2016). *Psychometric methods: Theory into practice*. Guilford Publications.

Ramos, M. N., Collins, P., & Peña, E. D. (2022). Sharpening Our Tools: A Systematic Review to Identify Diagnostically Accurate Language Sample Measures. *Journal of Speech, Language, and Hearing Research*, *65*(10), 3890–3907.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological assessment*, *31*(12), 1395.

Rice, M. L., Redmond, S. M., & Hoffman, L. (2006). Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, and Hearing Research*, *49*(4), 793–808.

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, *53*(2), 333–349.

Roberts, J. A., Altenberg, E. P., Ferrugio, H. R., & Rosenberg, J. E. (2022). How to Use the Index of Productive Syntax to Select Goals and Monitor Progress in Preschool Children. *Language, Speech, and Hearing Services in Schools*, *53*(3), 803–824.

Ryser, A. (2023). Development of a swiss german language sample analysis tool - legal and ethical aspects of data collection.

Schmidt-Atzert, L., & Amelang, M. (2012). *Psychologische Diagnostik* (5., vollständig überarbeitete und erweiterte Aufl). Springer.

Scott, A., Gillon, G., McNeill, B., & Kopach, A. (2022). The Evolution of an Innovative Online Task to Monitor Children's Oral Narrative Development. *Frontiers in Psychology*, *13*, 903124.

Soleymani, Z., Mahmoodabadi, N., & Nouri, M. M. (2016). Language skills and phonological awareness in children with cochlear implants and normal hearing. *International Journal of Pediatric Otorhinolaryngology*, *83*, 16–21.

Southwood, F., & Russell, A. F. (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research*, *47*(2), 366–376.

Spencer, E., Bryant, L., & Colyvas, K. (2020). Minimizing Variability in Language Sampling Analysis a Practical Way to Calculate Text Length and Time Variability and Measure Reliable Change When Assessing Clients. *Topics in language disorders*, *40*(2), 166–181.

Stockman, I. J. (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools*, *27*(4), 355–366.

Thordardottir, E., & Weismer, S. E. (2001). High-frequency verbs an d verb diversity inthe spontaneous speech of school-agechildren wit h speci c language impairment. *International Journal of Language & Communication Disorders*, *36*(2), 221–244.

Thurman, A. J., Edgin, J. O., Sherman, S. L., Sterling, A., McDuffie, A., Berry-Kravis, E., Hamilton, D., & Abbeduto, L. (2021). Spoken language outcome measures for treatment studies in Down syndrome: feasibility, practice effects, test-retest reliability, and construct validity of variables generated from expressive language sampling. *Journal of Neurodevelopmental Disorders*, *13*(1), 13.

Tomas, E., & Dorofeeva, S. (2019). Mean Length of Utterance and Other Quantitative Measures of Spontaneous Speech in Russian-Speaking Children. *Journal of Speech, Language, and Hearing Research*, *62*(12), 4483–4496.

Tommerdahl, J., & Kilpatrick, C. (2013). Analysing frequency and temporal reliability of children's morphosyntactic production in spontaneous language samples of varying lengths. *Child Language Teaching and Therapy*, *29*(2), 171–183.

Ukrainetz, T. A. (2006). The implications of RTI and EBP for SLPs: Commentary on l. m. justice. *Language, Speech, and Hearing Services in Schools*, *37*(4), 298–303.

Van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. *Modelling and assessing vocabulary knowledge*, *93*, 115.

van Severen, L., van den Berg, R., Molemans, I., & Gillis, S. (2012). Consonant inventories in the spontaneous speech of young children: A new procedure.

Voniati, L., Tafiadis, D., Armostis, S., Kosma, E. I., & Chronopoulos, S. K. (2021). Lexical Diversity in Cypriot-Greek-Speaking Toddlers: A Preliminary Longitudinal Study. *Folia Phoniatrica et Logopaedica*, *73*(4), 277–288.

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring Children's Lexical Diversity: Differentiating Typical and Impaired Language Learners. *Jounal of Speech, Language an Hearing Research*, *38*, 1349–1355.

Westerveld, M. F. (2011). Sampling and analysis of children's spontaneous language: From research to practice. *13*(2), 63–67.

Westerveld, M. F. (2019). Language Sampling. In J. S. Damico & M. J. Ball (Hrsg.), *The SAGE Encyclopedia of Human Communication Sciences and Disorders*. SAGE Publications, Inc.

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in australia. *International Journal of Speech-Language Pathology*, *16*(3), 242–249.

Westerveld, M. F., & Gillon, G. (2002). Language sampling protocol. *Christchurch: University of Canterbury*.

Westerveld, M. F., & Moran, C. A. (2011). Expository language skills of young school-age children. *Language, Speech, and Hearing Services in Schools*, *42*(2), 182–193.

Westerveld, M. F., & Vidler, K. (2016). Spoken language samples of Australian children in conversation, narration and exposition. *International journal of speech-language pathology*, *18*(3), 288–298.

Wilder, A., & Redmond, S. M. (2022). The reliability of short conversational language sample measures in children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *65*(5), 1939–1955.

Wofford, M. C., Cano, J., Goodrich, J. M., & Fitton, L. (2022). Tell or Retell? The Role of Task and Language in Spanish-English Narrative Microstructure Performance. *Language, speech, and hearing services in schools*, *53*(2), 511–531.

Wu, S.-Y., Huang, R.-J., & Tsai, I.-F. (2019). The applicability of D, MTLD, and MATTR in Mandarin–speaking children. *Journal of Communication Disorders*, *77*, 71–79.

Yang, J. S., Rosvold, C., & Bernstein Ratner, N. (2022). Measurement of Lexical Diversity in Children's Spoken Language: Computational and Conceptual Considerations. *Frontiers in psychology*, *13*, 905789.

8 Anhang

Sprachprobenlänge	Mdn	Q1	Q3
10 Äusserungen	5.50	3.70	7.50
20 Äusserungen	5.25	4.05	7.00
30 Äusserungen	5.67	4.43	7.23
40 Äusserungen	5.97	4.67	7.50
50 Äusserungen	5.96	4.40	7.96
Gesamtsprachprobe	6.63	4.67	7.68

Tabelle 13: Übersicht über die MLU-Werte bei den verschiedenen Sprachprobenlängen

Sprachprobenlänge	Mdn	Q1	Q3
35 Wörter	88.76	84.38	93.52
70 Wörter	88.00	85.24	90.95
105 Wörter	88.76	84.63	91.24
140 Wörter	88.38	85.10	90.48
175 Wörter	88.04	85.18	90.48
Gesamtsprachprobe	87.78	85.97	90.20

Tabelle 14: Übersicht über die MATTR-Werte bei den verschiedenen Sprachprobenlängen

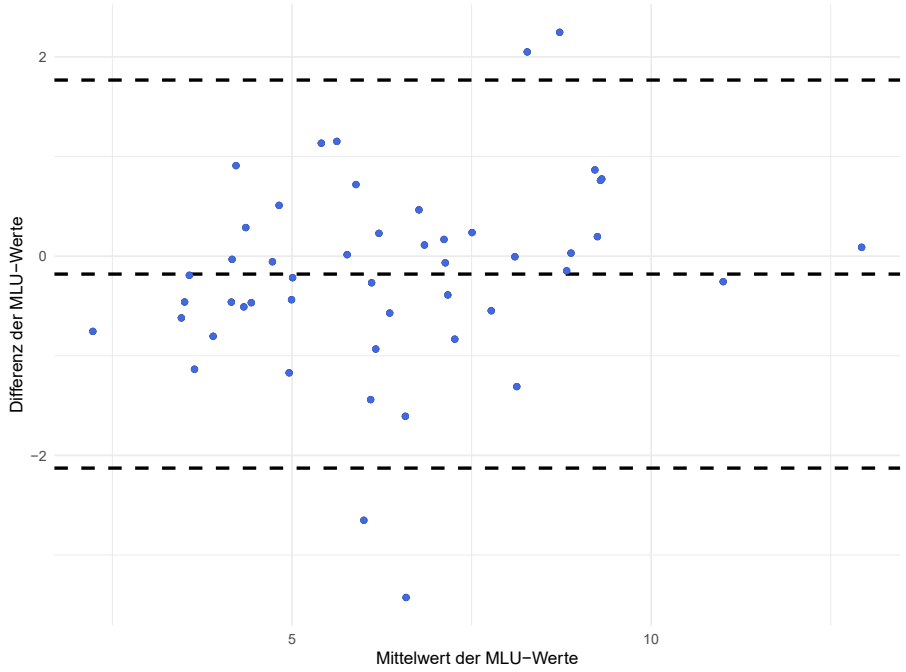

Abbildung 16: Bland-Altman Plot für den Vergleich der MLU-Werte von 40 Äusserungen und der gesamten Sprachprobe

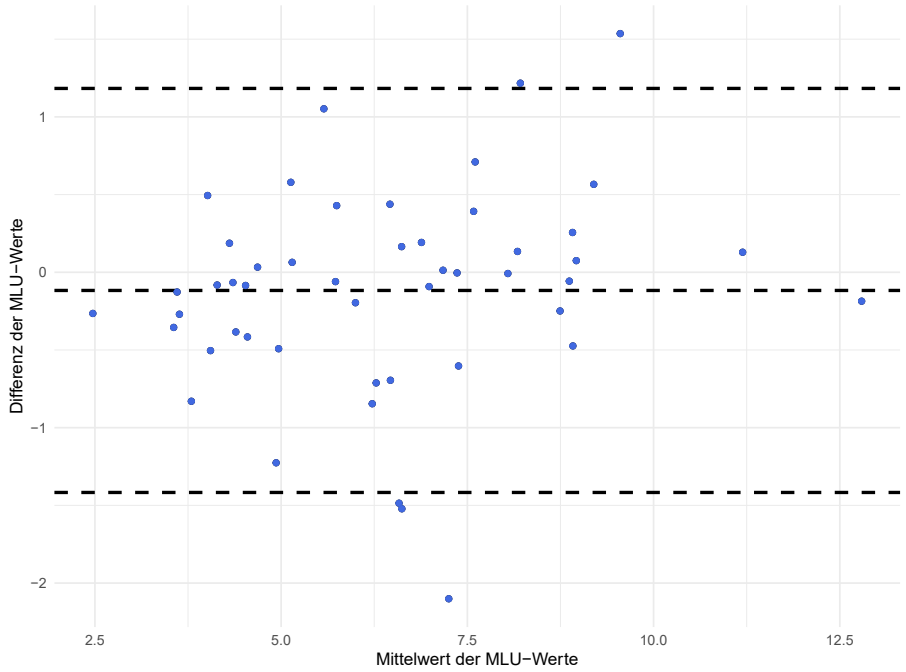

Abbildung 17: Bland-Altman Plot für den Vergleich der MLU-Werte von 50 Äusserungen und der gesamten Sprachprobe

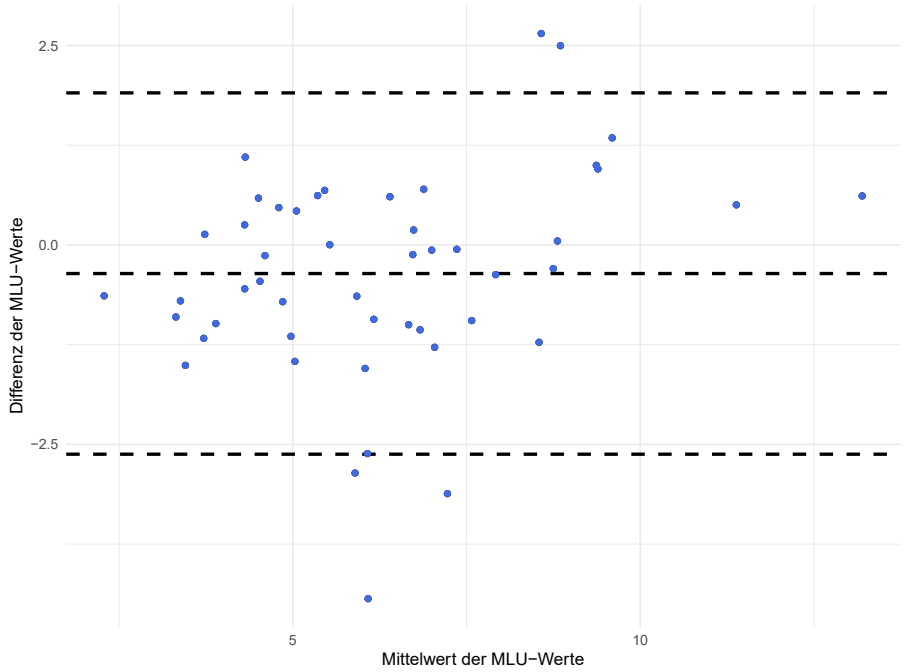

Abbildung 18: Bland-Altman Plot für den Vergleich der MLU-Werte von 30 Äusserungen und der gesamten Sprachprobe mit ± 1.75 Standardabweichungen

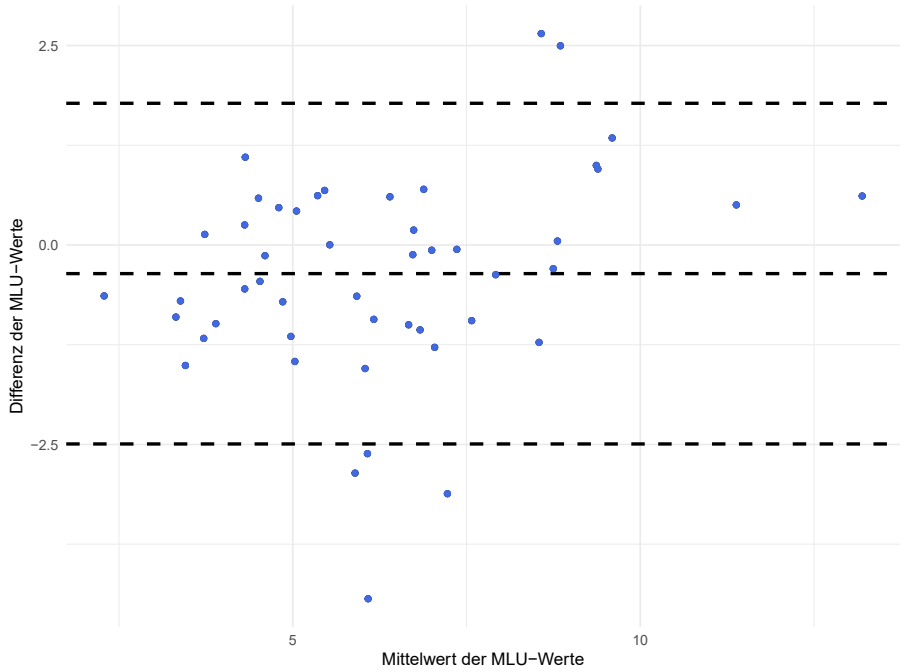

Abbildung 19: Bland-Altman Plot für den Vergleich der MLU-Werte von 30 Äusserungen und der gesamten Sprachprobe mit ± 1.65 Standardabweichungen

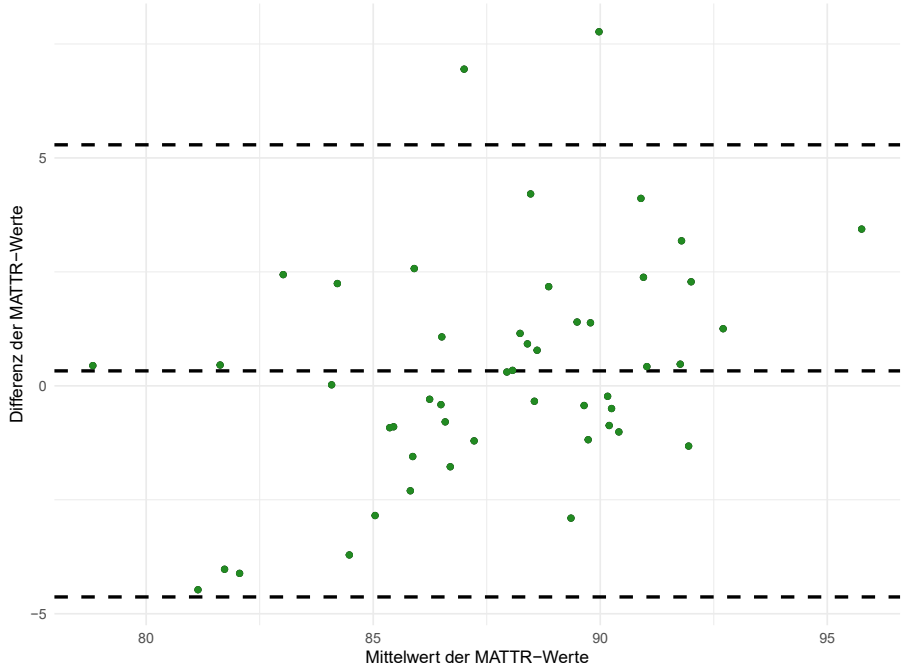

Abbildung 20: Bland-Altman Plot für den Vergleich der MATTR-Werte von 140 Wörtern und der gesamten Sprachprobe

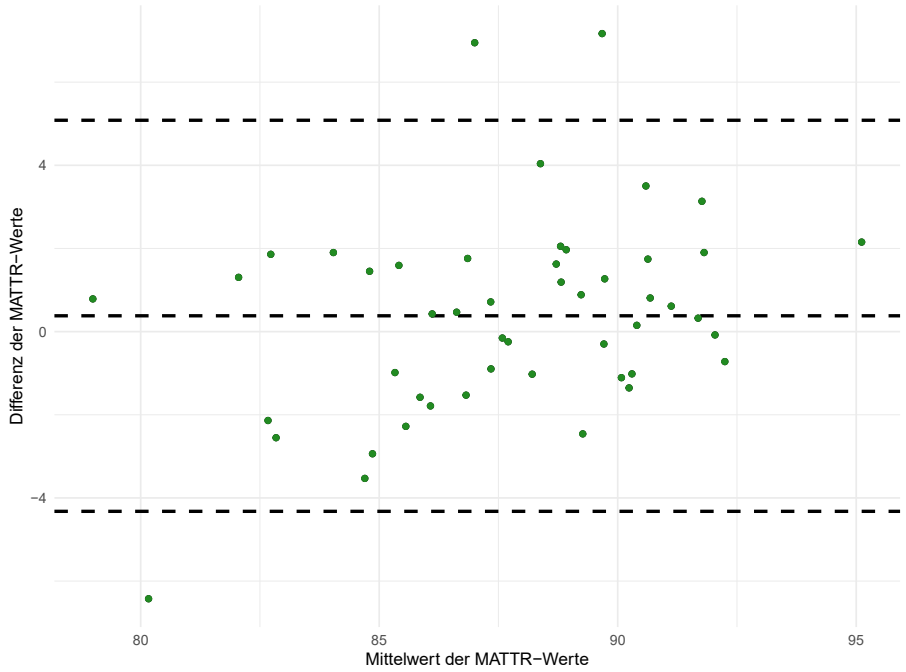

Abbildung 21: Bland-Altman Plot für den Vergleich der MATTR-Werte von 175 Wörtern und der gesamten Sprachprobe

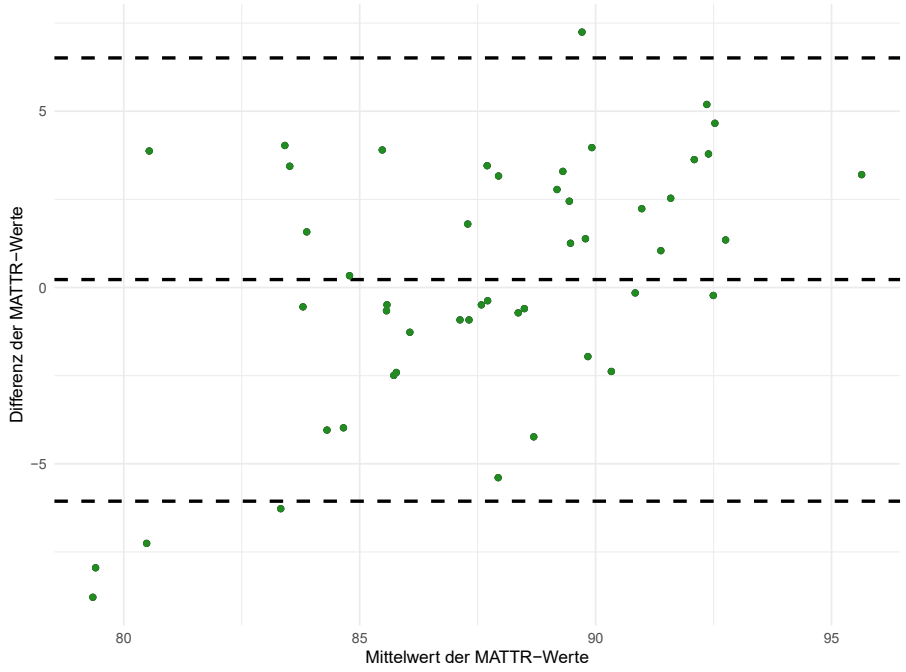

Abbildung 22: Bland-Altman Plot für den Vergleich der MATTR-Werte von 70 Wörtern und der gesamten Sprachprobe mit ± 1.75 Standardabweichungen

R-Code

```
# Load packages

library(tidyverse)
library(readxl)
library(psych)
library(ltm)
library(splithalfr)
library(BlandAltmanLeh)
library(ggplot2)
library(ggpubr)
library(rstatix)
library(xtable)
library(stringr)


#read excel-file from working directory
df_raw <- readxl::read_excel('aeusserungsanalyse.xlsx')

df <- df_raw %>%
  pivot_wider(names_from = "Indikator",
              values_from = matches("T[0-9]+")) %>%
  filter(!Datei == "FR_44_24.01.24_korr_MZ.txt")

#Crohnbachs Alpha MLU
df_mlu <- subset(df, select = c("Datei",
                                grep("MLU",
                                     colnames(df),
                                     value = TRUE)))

#Reduktion des Datensatzes auf die kürzeste Sprachprobe
df_mlu_crohn <- subset(df, select = grep("MLU", colnames(df))) %>%
  dplyr::select(!T53_MLU:T961_MLU)

#splithalf und crohnbachs alpha MLU
cor(df_mlu_crohn)

sh_mlu <- psych::splitHalf(df_mlu_crohn,
                           raw=T,
                           brute=FALSE,
```

```
                          n.sample=10000,
                          ci = .05)

pdf("plots/sh_hist_mlu.pdf",
    width = 8,
    height = 6)
hist(sh_mlu$raw, breaks = 101,
     xlab = "Testhalbierungsreliabilitäten",
     ylab = "Häufigkeit",
     main = "Testhalbierungsreliabilitäten der MLU",
     col = "lightskyblue")
dev.off()

Indikator <- c ("MLU", "MATTR")

sh_mlu1 <- data.frame(
  Indikator = "MLU",
  maxrb = sh_mlu$maxrb,
  minrb = sh_mlu$minrb,
  meanr = sh_mlu$meanr,
  alpha = sh_mlu$alpha
)

#Crohnbachs Alpha MATTR
df_mattr <- subset(df, select = c("Datei",
                                  grep("MATTR",
                                       colnames(df),
                                       value = TRUE)))

#kürzung auf kürzeste Sprachprobe
df_mattr_crohn <- subset(df, select = grep("MATTR", colnames(df))) %>%
  dplyr::select(!T184_MATTR:T961_MATTR)

df_mattr_crohn1 <- df %>%
  subset(select = grep("MATTR", colnames(df))) %>%
  filter(df$Datei != "FR_45_31.01.24_AD.txt") %>%
  dplyr::select(!T321_MATTR:T961_MATTR)


#splithalf und crohnbachs alpha MATTR
sh_mattr <- psych::splitHalf(df_mattr_crohn,
```

```
                                raw=T,
                                brute=FALSE,
                                n.sample=10000,
                                ci = 0.05)

pdf("plots/sh_hist_mattr1.pdf", width = 8, height = 6)
hist(sh_mattr$raw,
     breaks = 101,
     xlab = "Testhalbierungsreliabilitäten",
     ylab = "Häufigkeit",
     main = "Testhalbierungsreliabilitäten der MATTR",
     col = "darkseagreen1")
dev.off()

sh_mattr1 <- data.frame(
  Indikator = "MATTR",
  maxrb = sh_mattr$maxrb,
  minrb = sh_mattr$minrb,
  meanr = sh_mattr$meanr,
  alpha = sh_mattr$alpha
)

sh <- dplyr::bind_rows(sh_mlu1, sh_mattr1) %>%
  rename("Maximale Reliabilität" = maxrb,
         "Minimale Reliabilität" = minrb,
         "Durchschnittliche Reliabilität" = meanr,
         "Crohnbach's Alpha" = alpha)

print.xtable(
  xtable(sh),
  file = "tables/sh_comb.tex",
  floating = F,
  include.rownames = F
)

#nur Fenster ohne Überschneidung
df_mattr_crohn2 <- df %>%
  subset(select = grep("MATTR", colnames(df))) %>%
  filter(df$Datei != "FR_45_31.01.24_AD.txt") %>%
  dplyr::select(T1_MATTR, T15_MATTR, T30_MATTR, T45_MATTR,T60_MATTR, T75_MATTR,
                T90_MATTR, T105_MATTR, T120_MATTR, T135_MATTR, T150_MATTR,
```

```
                T165_MATTR, T180_MATTR, T195_MATTR, T210_MATTR, T225_MATTR,
                T240_MATTR, T255_MATTR, T270_MATTR, T285_MATTR, T300_MATTR,
                T225_MATTR)

sh_mattr2 <- psych::splitHalf(df_mattr_crohn2,
                              raw=T,
                              brute=FALSE,
                              n.sample=10000,
                              ci = .05)

pdf("plots/sh_hist_mattr2.pdf",
    width = 8,
    height = 6)
hist(sh_mattr2$raw,
     breaks = 101,
     xlab = "Testhalbierungsreliabilitäten",
     ylab = "Häufigkeit",
     main = "Testhalbierungsreliabilitäten der MATTR ohne Überschneidungen",
     col = "darkseagreen1")
dev.off()

sh_mattr_einzel <- data.frame(
  Indikator = "MATTR",
  maxrb = sh_mattr2$maxrb,
  minrb = sh_mattr2$minrb,
  meanr = sh_mattr2$meanr,
  alpha = sh_mattr2$alpha
)

print.xtable(
  xtable(sh_mattr_einzel),
  file = "tables/sh_mattr_einzel.tex",
  floating = F,
  include.rownames = F
)

#Die Reliabilität von kurzen Spontansprachproben

#Zusammenführen der beiden Datensätze
df2_raw <- readxl::read_excel('resultate.xlsx')
```

```
df_all <- df %>%
  left_join(df2_raw, by = "Datei") %>%
  rename(mattr_total = "Moving-average Type-token ratio",
         ttr = "Type-token ratio",
         mlu_total = "MLU",
         anz_worte = "Anzahl Worte",
         anz_auss = "Anzahl Äusserungen") %>%
  dplyr::select(!Sprecher)

haven::write_sav(df_all, "df_all.sav")


sum_mlu <- summary(df_all$mlu_total)
sum_mattr <- summary(df_all$mattr_total)

#berechnen der MLU für kürzere Sprachproben
df_all$mlu_first10 <- rowMeans(df_mlu[, c(paste0("T", 1:10, "_MLU"))])
df_all$mlu_first20 <- rowMeans(df_mlu[, c(paste0("T", 1:20, "_MLU"))])
df_all$mlu_first30 <- rowMeans(df_mlu[, c(paste0("T", 1:30, "_MLU"))])
df_all$mlu_first40 <- rowMeans(df_mlu[, c(paste0("T", 1:40, "_MLU"))])
df_all$mlu_first50 <- rowMeans(df_mlu[, c(paste0("T", 1:50, "_MLU"))])

#Spearman-Brown Korrelationen für MLU
sb_mlu_10 <- spearman_brown(df_all$mlu_total, df_all$mlu_first10)
sb_mlu_20 <- spearman_brown(df_all$mlu_total, df_all$mlu_first20)
sb_mlu_30 <- spearman_brown(df_all$mlu_total, df_all$mlu_first30)
sb_mlu_40 <- spearman_brown(df_all$mlu_total, df_all$mlu_first40)
sb_mlu_50 <- spearman_brown(df_all$mlu_total, df_all$mlu_first50)

#Tabelle erstellen
sb_coeff_mlu <- c(sb_mlu_10, sb_mlu_20, sb_mlu_30, sb_mlu_40, sb_mlu_50)
anz_äuss <- c(10, 20, 30, 40, 50)

sb_mlu_table <- data.frame(anz_äuss, sb_coeff_mlu)
colnames(sb_mlu_table) <- c("Anzahl Äusserungen", "Korrelationskoeffizient")

print.xtable(
  xtable(sb_mlu_table),
  file = "tables/sb_mlu_table.tex",
  floating = F,
  include.rownames = F
```

```
)

pdf("plots/sb_mlu.pdf",
    width = 8,
    height = 6)
plot(x = anz_äuss,
     y = sb_coeff_mlu,
     ylim = c(0.8, 1),
     xlab = "Spachprobenlänge",
     ylab = "Spearman-Brown Koeffizienten",
     main = "Spearman-Brown Reliabilität für MLU"
)
dev.off()

#berechnen der MATTR für kürzere Sprachproben
df_all$mattr_first35 <- 100*(rowMeans(df_mattr[, c(paste0("T",
                                                          1:35, "_MATTR"))]))
df_all$mattr_first70 <- 100*(rowMeans(df_mattr[, c(paste0("T",
                                                          1:70, "_MATTR"))]))
df_all$mattr_first105 <- 100*(rowMeans(df_mattr[, c(paste0("T",
                                                           1:105, "_MATTR"))]))
df_all$mattr_first140 <- 100*(rowMeans(df_mattr[, c(paste0("T",
                                                           1:140, "_MATTR"))]))
df_all$mattr_first175 <- 100*(rowMeans(df_mattr[, c(paste0("T",
                                                           1:175, "_MATTR"))]))

#Spearman-Brown Korrelationen für MATTR
sb_mattr_35 <- spearman_brown(df_all$mattr_total, df_all$mattr_first35)
sb_mattr_70 <- spearman_brown(df_all$mattr_total, df_all$mattr_first70)
sb_mattr_105 <- spearman_brown(df_all$mattr_total, df_all$mattr_first105)
sb_mattr_140 <- spearman_brown(df_all$mattr_total, df_all$mattr_first140)
sb_mattr_175 <- spearman_brown(df_all$mattr_total, df_all$mattr_first175)

#Tabelle erstellen
sb_coeff_mattr <- c(sb_mattr_35, sb_mattr_70,
                    sb_mattr_105, sb_mattr_140, sb_mattr_175)
anz_wörter <- c(35, 70, 105, 140, 175)
sb_mattr_table <- data.frame(anz_wörter, sb_coeff_mattr)
colnames(sb_mattr_table) <- c("Anzahl Wörter", "Korrelationskoeffizient")

print.xtable(
```

```
  xtable(sb_mattr_table),
  file = "tables/sb_mattr_table.tex",
  floating = F,
  include.rownames = F
)

pdf("plots/sb_mattr.pdf",
    width = 8,
    height = 6)
plot(x = anz_wörter,
     y = sb_coeff_mattr,
     ylim = c(0.7, 1),
     xlab = "Spachprobenlänge",
     ylab = "Spearman-Brown Koeffizienten",
     main = "Spearman-Brown Reliabilität für MATTR")
dev.off()

summary <- lapply(df_all[c("mlu_first10",
                           "mlu_first20",
                           "mlu_first30",
                           "mlu_first40",
                           "mlu_first50",
                           "mlu_total",
                           "mattr_first35",
                           "mattr_first70",
                           "mattr_first105",
                           "mattr_first140",
                           "mattr_first175",
                           "mattr_total")],
                  summary)
summary_matrix <- do.call(rbind, summary)
kurzSp <- as.data.frame(summary_matrix)

print.xtable(
  xtable(kurzSp),
  file = "tables/kurzSp.tex",
  floating = F
)

#Bland-Altman-Plots MLU
bland.altman.stats(df_all$mlu_first10,
```

```
                    df_all$mlu_total,
                    mode = 1)
ba_10_mlu <- bland.altman.plot(df_all$mlu_first10,
                               df_all$mlu_total,
                               mode = 1,
                               graph.sys = "ggplot2",
                               geom_count = FALSE)
pdf("plots/ba_mlu_10.pdf",
    width = 8,
    height = 6)
print(ba_10_mlu
      + xlab("Mittelwert der MLU-Werte")
      + ylab("Differenz der MLU-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal()
)
dev.off()

bland.altman.stats(df_all$mlu_first20,
                    df_all$mlu_total,
                    mode = 1)
ba_20_mlu <- bland.altman.plot(df_all$mlu_first20,
                               df_all$mlu_total,
                               graph.sys = "ggplot2",
                               geom_count = FALSE,
                               mode = 1)
pdf("plots/ba_mlu_20.pdf",
    width = 8,
    height = 6)
print(ba_20_mlu
      + xlab("Mittelwert der MLU-Werte")
      + ylab("Differenz der MLU-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

bland.altman.stats(df_all$mlu_first30,
                    df_all$mlu_total,
                    mode = 1)
ba_30_mlu <- bland.altman.plot(df_all$mlu_first30,
                               df_all$mlu_total,
```

```
                                graph.sys = "ggplot2",
                                geom_count = FALSE,
                                mode = 1)
pdf("plots/ba_mlu_30.pdf",
    width = 8,
    height = 6)
print(ba_30_mlu
      + xlab("Mittelwert der MLU-Werte")
      + ylab("Differenz der MLU-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

bland.altman.stats(df_all$mlu_first40,
                   df_all$mlu_total,
                   mode = 1)
ba_40_mlu <- bland.altman.plot(df_all$mlu_first40,
                                df_all$mlu_total,
                                graph.sys = "ggplot2",
                                geom_count = FALSE,
                                mode = 1)
pdf("plots/ba_mlu_40.pdf",
    width = 8,
    height = 6)
print(ba_40_mlu
      + xlab("Mittelwert der MLU-Werte")
      + ylab("Differenz der MLU-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

bland.altman.stats(df_all$mlu_first50,
                   df_all$mlu_total,
                   mode = 1)
ba_50_mlu <- bland.altman.plot(df_all$mlu_first50,
                                df_all$mlu_total,
                                graph.sys = "ggplot2",
                                geom_count = FALSE,
                                mode = 1)
pdf("plots/ba_mlu_50.pdf",
    width = 8,
```

```
    height = 6)
print(ba_50_mlu
      + xlab("Mittelwert der MLU-Werte")
      + ylab("Differenz der MLU-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

#MLU 30 Äusserungen mit SD = 1.75
bland.altman.stats(df_all$mlu_first30,
                   df_all$mlu_total,
                   mode = 1,
                   two = 1.75)
ba_30_mlu_75 <- bland.altman.plot(df_all$mlu_first30,
                                  df_all$mlu_total,
                                  graph.sys = "ggplot2",
                                  geom_count = FALSE,
                                  mode = 1,
                                  two = 1.75)
pdf("plots/ba_mlu_30_75.pdf",
    width = 8,
    height = 6)
print(ba_30_mlu_75
      + xlab("Mittelwert der MLU-Werte")
      + ylab("Differenz der MLU-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

#MLU 30 Äusserungen mit SD = 1.65
bland.altman.stats(df_all$mlu_first30,
                   df_all$mlu_total,
                   mode = 1,
                   two = 1.65)
ba_30_mlu_65 <- bland.altman.plot(df_all$mlu_first30,
                                  df_all$mlu_total,
                                  graph.sys = "ggplot2",
                                  geom_count = FALSE,
                                  mode = 1,
                                  two = 1.65)
pdf("plots/ba_mlu_30_65.pdf",
```

```
    width = 8,
    height = 6)
print(ba_30_mlu_65
      + xlab("Mittelwert der MLU-Werte")
      + ylab("Differenz der MLU-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

#Bland-Altman Plot MATTR
bland.altman.stats(df_all$mattr_first35,
                   df_all$mattr_total,
                   mode = 1)
ba_35_mattr <- bland.altman.plot(df_all$mattr_first35,
                                 df_all$mattr_total,
                                 mode = 1,
                                 graph.sys = "ggplot2",
                                 geom_count = FALSE)
pdf("plots/ba_mattr_35.pdf",
    width = 8,
    height = 6)
print(ba_35_mattr
      + xlab("Mittelwert der MATTR-Werte")
      + ylab("Differenz der MATTR-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

bland.altman.stats(df_all$mattr_first70,
                   df_all$mattr_total,
                   mode = 1)
ba_70_mattr <- bland.altman.plot(df_all$mattr_first70,
                                 df_all$mattr_total,
                                 mode = 1,
                                 graph.sys = "ggplot2",
                                 geom_count = FALSE)
pdf("plots/ba_mattr_70.pdf",
    width = 8,
    height = 6)
print(ba_70_mattr
      + xlab("Mittelwert der MATTR-Werte")
```

```
      + ylab("Differenz der MATTR-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

bland.altman.stats(df_all$mattr_first105,
                   df_all$mattr_total,
                   mode = 1)
ba_105_mattr <- bland.altman.plot(df_all$mattr_first105,
                                  df_all$mattr_total,
                                  mode = 1,
                                  graph.sys = "ggplot2",
                                  geom_count = FALSE)
pdf("plots/ba_mattr_105.pdf",
    width = 8,
    height = 6)
print(ba_105_mattr
      + xlab("Mittelwert der MATTR-Werte")
      + ylab("Differenz der MATTR-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

bland.altman.stats(df_all$mattr_first140,
                   df_all$mattr_total,
                   mode = 1)
ba_140_mattr <- bland.altman.plot(df_all$mattr_first140,
                                  df_all$mattr_total,
                                  mode = 1,
                                  graph.sys = "ggplot2",
                                  geom_count = FALSE)
pdf("plots/ba_mattr_140.pdf",
    width = 8,
    height = 6)
print(ba_140_mattr
      + xlab("Mittelwert der MATTR-Werte")
      + ylab("Differenz der MATTR-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()
```

```
bland.altman.stats(df_all$mattr_first175,
                   df_all$mattr_total,
                   mode = 1)
ba_175_mattr <- bland.altman.plot(df_all$mattr_first175,
                                  df_all$mattr_total,
                                  mode = 1,
                                  graph.sys = "ggplot2",
                                  geom_count = FALSE)
pdf("plots/ba_mattr_175.pdf",
    width = 8,
    height = 6)
print(ba_175_mattr
      + xlab("Mittelwert der MATTR-Werte")
      + ylab("Differenz der MATTR-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

bland.altman.stats(df_all$mattr_first70,
                   df_all$mattr_total,
                   mode = 1,
                   two = 1.75)
ba_70_mattr_75 <- bland.altman.plot(df_all$mattr_first70,
                                    df_all$mattr_total,
                                    mode = 1,
                                    graph.sys = "ggplot2",
                                    geom_count = FALSE,
                                    two = 1.75)
pdf("plots/ba_mattr_70_75.pdf",
    width = 8,
    height = 6)
print(ba_70_mattr_75
      + xlab("Mittelwert der MATTR-Werte")
      + ylab("Differenz der MATTR-Werte")
      +theme(plot.title = element_text(hjust = 0.5))
      +theme_minimal())
dev.off()

#Inferenzstatistik
#Datensatz im longformat nach MLU
df_anova_mlu <- df_all %>%
```

```
  dplyr::select(!T1_MLU:T961_MATTR) %>%
  rename(MLU_10_Äusserungen = "mlu_first10", MLU_20_Äusserungen = "mlu_first20",
         MLU_30_Äusserungen = "mlu_first30", MLU_40_Äusserungen = "mlu_first40",
         MLU_50_Äusserungen = "mlu_first50", MLU_Gesamtsample = "mlu_total") %>%
  pivot_longer(cols = c(MLU_Gesamtsample, MLU_10_Äusserungen,
                        MLU_20_Äusserungen, MLU_30_Äusserungen,
                        MLU_40_Äusserungen, MLU_50_Äusserungen),
               names_to = "length_mlu",
               values_to = "mlu") %>%
  rename(Sprachprobenlänge = "length_mlu")

#Voraussetzungen: Prüfung Normalverteilung
pdf("plots/qq_MLU_Normalverteilung.pdf",
    width = 8,
    height = 6)
ggqqplot (df_anova_mlu,
          "mlu",
          facet.by = "Sprachprobenlänge",
          color = "lightskyblue")
dev.off()

anova_mlu_overview <- df_anova_mlu %>%
  group_by(Sprachprobenlänge) %>%
  summarize(Mdn = median(mlu),
            Q1 = quantile(mlu, probs = .25),
            Q3 = quantile(mlu, probs = .75)) %>%
  as.data.frame()

print.xtable(
  xtable(anova_mlu_overview),
  file = "tables/anova_mlu_overview.tex",
  floating = F,
)

rep_anova_mlu <- anova_test(data = df_anova_mlu,
                            dv = mlu,
                            wid = Datei,
                            within = Sprachprobenlänge,
                            effect.size = "ges")
anova_mlu_spheri <- rep_anova_mlu$`Mauchly's Test for Sphericity`
anova_mlu_table <- get_anova_table(rep_anova_mlu)
```

```
print.xtable(
  xtable(anova_mlu_table),
  file = "tables/anova_mlu_table.tex",
  floating = F,
  include.rownames = F
)

print.xtable(
  xtable(anova_mlu_spheri),
  file = "tables/anova_mlu_spheri.tex",
  floating = F,
  include.rownames = F,

)

anova_mlu <- df_anova_mlu %>%
  pairwise_t_test(mlu~Sprachprobenlänge,
                  paired = TRUE,
                  p.adjust.method = "bonferroni") %>%
  as.data.frame()

mlu_eff <- df_anova_mlu %>%
  cohens_d(mlu~Sprachprobenlänge,
           paired = TRUE) %>%
  as.data.frame()

anova_mlu1 <- anova_mlu %>%
  select(group1, group2, statistic, df, p, p.adj) %>%
  mutate(effectsize = mlu_eff$effsize)

print.xtable(
  xtable(anova_mlu1),
  file = "tables/anova_mlu1.tex",
  floating = F,
  include.rownames = F,
)

#Datensatz im longformat nach MATTR
df_anova_mattr <- df_all %>%
  dplyr::select(!T1_MLU:T961_MATTR) %>%
  rename(MATTR_35_Wörter = "mattr_first35",
```

```
          MATTR_70_Wörter = "mattr_first70",
          MATTR_105_Wörter = "mattr_first105",
          MATTR_140_Wörter = "mattr_first140",
          MATTR_175_Wörter = "mattr_first175",
          MATTR_Gesamtsprachprobe = "mattr_total") %>%
  pivot_longer(cols = c(MATTR_Gesamtsprachprobe,
                        MATTR_35_Wörter,
                        MATTR_70_Wörter,
                        MATTR_105_Wörter,
                        MATTR_140_Wörter,
                        MATTR_175_Wörter),
               names_to = "length_mattr",
               values_to = "mattr") %>%
  mutate(length_mattr = factor(length_mattr,
                               levels = c("MATTR_35_Wörter",
                                          "MATTR_70_Wörter",
                                          "MATTR_105_Wörter",
                                          "MATTR_140_Wörter",
                                          "MATTR_175_Wörter",
                                          "MATTR_Gesamtsprachprobe"))) %>%
  rename(Sprachprobenlänge = "length_mattr")

pdf("plots/qq_MATTR_Normalverteilung.pdf",
    width = 8,
    height = 6)
ggqqplot(df_anova_mattr,
         "mattr",
         facet.by = "Sprachprobenlänge",
         color = "darkseagreen")
dev.off()

anova_mattr_overview <-df_anova_mattr %>%
  group_by(Sprachprobenlänge) %>%
  summarize(Mdn = median(mattr),
            Q1 = quantile(mattr, probs = .25),
            Q3 = quantile(mattr, probs = .75)) %>%
  as.data.frame()

print.xtable(
  xtable(anova_mattr_overview),
  file = "tables/anova_mattr_overview.tex",
```

```
  floating = F,
)

rep_anova_mattr <- anova_test(data = df_anova_mattr,
                              dv = mattr,
                              wid = Datei,
                              within = Sprachprobenlänge)
anova_mattr_spheri <- rep_anova_mattr$`Mauchly's Test for Sphericity`
anova_mattr_table <- get_anova_table(rep_anova_mattr)

print.xtable(
  xtable(anova_mattr_table),
  file = "tables/anova_mattr_table.tex",
  floating = F,
  include.rownames = F
)

print.xtable(
  xtable(anova_mattr_spheri),
  file = "tables/anova_mattr_spheri.tex",
  floating = F,
  include.rownames = F
)

anova_mattr <- df_anova_mattr %>%
  pairwise_t_test(mattr~Sprachprobenlänge,
                  paired = TRUE,
                  p.adjust.method = "bonferroni") %>%
  as.data.frame()

mattr_eff <- df_anova_mattr %>%
  cohens_d(mattr~Sprachprobenlänge,
           paired = TRUE) %>%
  as.data.frame()

anova_mattr1 <- anova_mattr %>%
  select(group1, group2, statistic, df, p, p.adj) %>%
  mutate(effectsize = mattr_eff$effsize)

print.xtable(
  xtable(anova_mattr1),
```

```
  file = "tables/anova_mattr1.tex",
  floating = F,
  include.rownames = F,
)

#Boxplots
ggplot(df_anova_mlu, aes(y = mlu,
                         x = Sprachprobenlänge,
                         group = Sprachprobenlänge,
                         fill = Sprachprobenlänge)) +
  geom_boxplot(notch = T) +
  theme_minimal() +
  theme(panel.grid.major.x = element_blank()) +
  labs(x = "Sprachprobenlänge",
       y = "mlu",
       title = "Boxplot MLU")
ggsave("plots/boxplot_mlu.pdf",
       width = 12,
       height = 6)

ggplot(df_anova_mattr, aes(y = mattr,
                           x = Sprachprobenlänge,
                           group = Sprachprobenlänge,
                           fill = Sprachprobenlänge)) +
  geom_boxplot(notch = T) +
  theme_minimal() +
  theme(panel.grid.major.x = element_blank()) +
  labs(x = "Sprachprobenlänge",
       y = "MATTR",
       title = "Boxplot MATTR")
ggsave("plots/boxplot_mattr.pdf",
       width = 12,
       height = 6)

#Deskriptive Statistiken Stichprobe

df3_raw <- readxl::read_excel('Metadaten.xlsx')
df_meta <- df3_raw[c(1:44,79:133),] %>%
  select("Datei":"Version")%>%
  filter(!Datei %in% c("FR_01", "FR_12", "FR_14", "FR_14", "FR_15", "FR_24",
                       "FR_26", "FR_28", "FR_30", "FR_31", "FR_33", "FR_40",
```

```
                        "FR_44", "FR_46", "FR_47", "FR_49", "FR_50", "FR_51",
                        "FR_52", "FR_53", "FR_55", "FR_56", "FR_57", "FR_58",
                        "FR_59", "FR_60", "FR_61", "FR_62", "FR_63", "FR_64",
                        "FR_65", "FR_68", "FR_69", "FR_70", "FR_71", "FR_72",
                        "FR_73", "ZH_05", "ZH_06", "ZH_08", "ZH_09", "ZH_10",
                        "ZH_14", "ZH_16", "ZH_17", "ZH_18", "ZH_19", "ZH_20",
                        "ZH_24", "ZH_25", "ZH_26")) %>%
  mutate(birth_date = as.Date(Geburtsdatum),
         rec_date = as.Date(`Datum Aufnahme`),
         age_days = as.numeric(rec_date - birth_date))

table(df_meta$Geschlecht)
table(df_meta$Mehrsprachig)
table(df_meta$Version)
table(df_meta$Logopädie)
table(df_meta$`Lern- oder Sprachstörungen`)
table(df_meta$Logopädin)
dauer <- summary(df_meta$Dauer)
auss <- summary(df_all$anz_auss)
wort <- summary(df_all$anz_worte)
sum_day <- summary(df_meta$age_days)
age_year <- floor(sum_day / 365.25)
age_month <- floor((sum_day - (age_year * 365.25)) / 30.4)

dauer_lsa <- bind_rows(dauer, auss, wort)%>%
  dauer_lsa$Einheit <- c("Minuten", "Äusserungen", "Wörter")

dauer_lsa1 <- dauer_lsa %>%
  as.data.frame() %>%
  select("Einheit",
         everything()) %>%
  select(!`NA's`)

print.xtable(
  xtable(dauer_lsa1),
  file = "tables/dauer_lsa1.tex",
  floating = F,
  include.rownames = F,
)
```

Ehrenwörtliche Erklärung

Ich, Lena Graf (geb. 12.01.1997), bestätige mit meiner Unterschrift, dass ich die Arbeit persönlich erstellt und dabei nur die aufgeführten Quellen und Hilfsmittel verwendet sowie wörtliche Zitate und Paraphrasen als solche gekennzeichnet habe.

Für die keinen Teil meiner Arbeit habe ich auf die Unterstützung einer künstlichen Intelligenz zurückgegriffen.

Brugg, 24.4.24 L. Graf

Ort, Datum Unterschrift