Probleme der Graphem-Phonem-Beziehung im Deutschen in KI-gestützten Spracherkennungssystemen

مشكلات العلاقة بين الحرف والصوت في الألمانية بأنظمة التعرف الصوتي على الكلام المدعومة بالذكاء الاصطناعي

Dr. Hanaa Ahmed Sayed Abuelela
Associate Professor, Department of German Language
Faculty of Al-Alsun, Minia University

د. هناء أحمد سيد أبوالعلا
أستاذ مساعد بقسم اللغة الألمانية
كلية الألسن، جامعة المنيا

Copyright©2025Faculty of Al-Alsun | Ain Shams University | All right reserved

Problems of the Grapheme-Phoneme Relationship in German in AI-based Speech Recognition Systems

Abstract:

This study aims to evaluate the linguistic accuracy and efficiency of artificial intelligence systems in speech recognition, especially given its increasing use and integration into daily life. The focus is on identifying common transcription errors in German when using AI tools, with special attention to the grapheme-phoneme relationship, phonological deviations, morphological structures, and challenges in word segmentation.

Adopting an analytical and comparative approach, the study examines and contrasts machine-generated transcriptions with human-produced reference transcriptions. It classifies recurring error patterns within linguistic frameworks, based on the analysis of 200 audio files for which human transcriptions are available.

The findings highlight several key results, most notably that AI-generated transcriptions exhibit significant weaknesses, specifically in handling the complicated relationship between graphemes and phonemes. In addition to the language-specific phenomena in German, such as homophones, inflectional forms and present major challenges that AI systems often struggle to process accurately.

Keywords: Speech recognition, transcription, grapheme-phoneme relationship, AI, error types

مشكلات العلاقة بين الحرف والصوت في الألمانية بأنظمة التعرف الصوتي على الكلام المدعومة بالذكاء الاصطناعي

الملخص:

تحاول هذه الدراسة تقييم الدقة اللغوية وكفاءة أنظمة الذكاء الاصطناعي في التعرف على الكلام، وذلك في ظل انتشار هذه الأنظمة وتوسع استخدامها في الحياة اليومية، وتهتم الدراسة بالأخطاء الشائعة في التفريغ الصوتي للكلام في اللغة الألمانية بواسطة أدوات الذكاء الاصطناعي؛ وذلك بهدف التعرف على العلاقة بين الحروف والأصوات والانحرافات الفونولوجية، إضافة إلى البنى الصرفية، ومعوقات تقسيم الكلمات. وتعتمد الدراسة على المنهج التحليلي والمقارن من خلال تحليل ومقارنة تفريغات آلية بأخرى بشرية، وتصنيف الأخطاء الشائعة كأنماط ضمن أطر لغوية، وذلك بالتطبيق على ٢٠٠ ملف صوتي والتي يتوافر لها تفريغات مرجعية بشرية، ولقد توصلت الدراسة لعدد من النتائج أبرزها: أن التفريغات الناتجة عن أنظمة الذكاء الاصطناعي تُظهر ضعفًا ملحوظًا خاصة في التعامل مع العلاقة المعقدة بين الحرف والصوت، وأن الظواهر اللغوية الخاصة باللغة الألمانية كحروف الإمالة والكلمات المتجانسة صوتيًا، وأشكال التصريف تُعد من أبرز التحديات التي يصعُب على أنظمة الذكاء الاصطناعي معالجتها بدقة.

الكلمات المفتاحية: التعرف على الكلام، التفريغ الصوتي، العلاقة بين الحرف والصوت، الذكاء الاصطناعي، الأخطاء الشائعة

Probleme der Graphem-Phonem-Beziehung im Deutschen in KI-gestützten Spracherkennungssystemen

1. Einleitung:

1.1.Bedeutung und Ziel der Untersuchung

Die Computerlinguistik als interdisziplinäre Schnittstelle zwischen Sprachwissenschaft und Informatik spielt eine zentrale Rolle bei der Entwicklung und Optimierung von ASR-Systemen. Sie befasst sich u. a. mit der algorithmischen Modellierung sprachlicher Strukturen zur Reduktion von Transkriptionsfehlern (vgl. Andresen, 2024; Kriesch, 2023; Weigelt, 2022, S. 23). Im Zentrum steht dabei die Frage, wie Phoneme (Laute) korrekt in Grapheme (Buchstaben) überführt werden können, insbesondere im Deutschen, das trotz regelhafter Zuordnungen - im Vergleich zu andren Sprachen- durch Dialekte, Umlaute und Homophone zusätzliche Komplexität aufweist (Khosravani et al., 2021; Pellegrino et al., 2022). Diese Studie widmet sich dem Vergleich zwischen menschlicher und KI-gestützter Transkription, um Unterschiede und Fehlerarten bei der Umsetzung der Graphem-Phonem-Korrespondenz zu identifizieren.

1.2. Fragestellungen

Der vorliegende Beitrag untersucht, wie genau KI-gestützte Systeme die Graphem-Phonem-Beziehung im Deutschen und welche Fehler dabei typischerweise auftreten. Ziel ist es systematische Unterschiede zwischen maschinellen und menschlichen Referenztranskriptionen aufzuzeigen mit Fokus auf phonologische Abweichungen, morphologische Strukturen und die Wortsegmentierung. Darüber hinaus wird analysiert, in welchen Bereichen KI-Systeme an ihre Grenzen stoßen, etwa bei sprachspezifischen Phänomenen wie Homophonen und Umlautveränderungen, die besonders häufig zu Transkriptionsfehlern führen.

2. Theoretischer Hintergrund

2.1.Spracherkennung und KI-Technologien

Die automatische Spracherkennung ist eine zentrale Technologie in der Mensch-Maschine-Kommunikation. Frühere Spracherkennungssysteme basieren auf regelbasierten Ansätzen, die eine festgelegte Grammatik und ein vordefiniertes Phonemsystem verwenden,

um die Sprache in Text umzuwandeln. Diese Systeme sind jedoch sehr anfällig für Fehler, insbesondere bei variierenden Aussprachen oder komplexeren sprachlichen Strukturen. Mit dem Aufkommen neuronaler Netze und *Deep-Learning-Algorithmen* -u. a. durch Arbeiten von Hinton et al. (2012) und Graves et al. (2013)- wurde die Genauigkeit erheblich gesteigert. Moderne ASR-Systeme wie *Google Speech-to-Text*, *Siri* oder *Deep Speech* integrieren neben akustischen auch kontextuelle Informationen zur Verbesserung der Transkription (vgl. Weigelt, 2022, S. 40).

Ein zentrales Problem stellt die Graphem-Phonem-Beziehung dar, die stark von Dialekten, regionalen Varianten und historischen Sprachentwicklungen beeinflusst wird. *Deep Neural Networks* und *Recurrent Neural Networks* konnten durch kontextsensitives Lernen diese Herausforderung teilweise reduzieren.

Das Verständnis gesprochener Sprache stellt ein anspruchsvolles Teilgebiet der natürlichen Sprachverarbeitung dar. Spontane Äußerungen sind oft unvollständig, enthalten Disfluenzen wie Pausen oder Füllwörter („ähm“) und weisen individuelle Merkmale wie Dialekt, Akzent oder Wortwahl auf (Weigelt, 2022, S. 47–48). Diese Faktoren führen zu typischen Fehlern bei der automatischen Transkription – besonders bei der Interpunktion, da spontane Sprache selten klar segmentiert ist. Besonders relevant ist dies für Anwendungen wie virtuelle Assistenten (z. B. *Siri*, *Google Assistant* oder *Amazon Alexa*), die auf eine präzise Spracherkennung angewiesen sind. Ein Beispiel für fehlerhafte Umsetzung ist das Missverständnis der Redewendung „mit Auge, Herz und Hirn“, die als „mit Auge, Herz und Heroin“ transkribiert wurde (Ruhaltinger, 2022, S. 58). Die hier angesprochenen Aspekte sind eng verknüpft mit den Grundlagen der Graphematik, auf die im Folgenden eingegangen wird.

2.2. Die Graphematik und Orthografie

Die Graphematik als linguistische Teildisziplin untersucht die Struktur und Funktion von Graphemen sowie deren Beziehung zur Lautstruktur. Im Gegensatz zur Orthografie, die sich mit den durch eine Norm festgelegten Schreibweisen und deren Optimierung befasst, beschäftigt sich die Graphematik mit der Struktur und Funktion von

Graphemen. Sie grenzt sich dabei auch zur Phonologie ab, die sich auf das Lautsystem fokussiert (vgl. Bußmann, 2008; Dürscheid, 2002, 2016; Maas, 2010).

Das Verhältnis zwischen gesprochener und geschriebener und Sprache wird durch zwei zentrale Positionen beschrieben: die Dependenzhypothese -vertreten etwa durch Ferdinand de Saussure- sieht die Schrift als sekundäres Zeichensystem, während die Autonomiehypothese ihr eine eigenständige sprachliche Funktion zuschreibt. (vgl. Dürscheid, 2016, S. 23; Evertz-Rittich et al., 2023, S. 199)

Typologisch lassen sich Schriftsysteme in logographische, syllabische und alphabetische Systeme unterteilen (vgl. Dürscheid, 2016, S. 65). Das Deutsche als alphabetisches System zeichnet sich durch die Repräsentation einzelner Phoneme mittels Buchstaben aus, ist jedoch durch komplexe Graphem-Phonem-Zuordnungen geprägt, die den Schriftspracherwerb sowie das automatische Transkribieren erschweren können (vgl. Treutlein, 2011, S. 56).

2.3.Die Graphem-Phonem-Beziehung im Deutschen

Das Deutsche wird als eine „lautorientierte“ [1] Alphabetschrift bezeichnet. Es weist im Vergleich zu Sprachen wie dem Englischen eine relativ regelmäßige, aber dennoch komplexe Graphem-Phonem-Beziehung auf. Auch wenn es keine eindeutige Eins-zu-eins-Zuordnung zwischen Lauten und Buchstaben gibt, orientiert sich das deutsche Schriftsystem grundsätzlich am phonologischen Prinzip (vgl. Schründer-Lenzen, 2013, S. 15-16). Die zentrale Herausforderung besteht darin, dass ein Phonem mehreren Graphemen zugeordnet sein kann, insbesondere in der automatisierten Verschriftlichung gesprochener Sprache. Die linguistische Forschung (vgl. Dürscheid, 2002; Eisenberg, 1988; Fuhrhop & Peters, 2013; Schründer-Lenzen, 2013) hat gezeigt, dass morphologische, phonologische und orthografische Faktoren die Verschriftlichung beeinflussen. Phoneme sind die kleinsten bedeutungsunterscheidenden Einheiten einer Sprache, während Grapheme die kleinsten bedeutungstragenden Einheiten des Schriftsystems darstellen. Die realisierte Form eines Phonems wird als Phon bezeichnet; Varianten eines Phonems -sogenannte Allophone- können je nach Dialekt

oder Sprachsituation variieren, ohne die Bedeutung zu verändern. Da sich die vorliegende Arbeit auf Transkriptionsfehler konzentriert, die auf Schwierigkeiten bei der Umsetzung der Graphem-Phonem-Beziehung zurückzuführen sind, wird im Folgenden nur eine Auswahl derjenigen Phoneme dargestellt, bei denen in der Analyse besonders häufig Fehler aufgetreten sind. Dabei handelt es sich um Laute, die entweder mehrdeutig verschriftet werden können oder besonders häufig zu phonetischen Verwechslungen geführt haben.

Tabelle 1: Ausgewählte problematische Phoneme des Deutschen mit ihren Basisgraphem und Orthographemen (vgl. Sapp, 2019, S. 13,30; Schründer-Lenzen, 2013, S. 22).

Konsonanten		
Phoneme	Basisgraphem	Orthographeme
/f/	<f>	<v>, <ff>, <ph>
/k/	<k>	<c>, <ch>, <ck>, <g>
/ŋ/	<ng>	<n>
/p/	<p>	<pp>, <b>
/ʁ/	<r>	<rr>, <rh>
/s/	<s>	<ss>, <ß>, <c>
/ʃ/	<sch>	<sk>, <s>, <ch>, <sh>
/t/	<t>	<tt>, <th>, <dt>, <d>
/ts/	<z>	<tz>, <zz>, <c>, <t>
/ç/	<ch>	<g>
/x/	<ch>	-
/z/	<s>	<z>
Diphthonge		
Phoneme	Basisgraphem	Orthographeme
/ɔɪ/	<eu>	<äu>, <oy>
/aɪ/	<ai>	<ei>

Diese Tabelle zeigt deutlich, dass viele Phoneme mit mehreren Graphemen wiedergegeben werden können, was bei der automatisierten Transkription zu Unsicherheiten führt. Die sogenannte Phonem-Graphem-Korrespondenz (PGK) stellt daher eine der zentralen linguistischen Hindernisse für KI-basierte Systeme dar (vgl. Schründer-Lenzen, 2013, S. 20–21). Besonders deutlich wird dies bei Phonemen wie /ʃ/, das durch <sch>, <s>, <ch> oder <sh> realisiert werden kann, oder beim langen /i:/, das als <ie> („vier“), <i> („Fabrik“), <ih> („ihr“) oder <ieh> („fliehen“) schriftlich wiedergegeben wird.

Diese theoretischen Grundlagen liefern die Basis für die Fehleranalyse im empirischen Teil dieser Arbeit.

2.4.Fehlerarten der Spracherkennung im Deutschen

Sprache ist eine zentrale Kommunikationsform, die spezielle Fähigkeiten erfordert. Die Spracherkennung als eine Anwendung künstlicher Intelligenz nutzt vernetzte Modelle zur Reduzierung von Fehlern, die durch Dialekte, Hintergrundgeräusche oder falsche Aussprache entstehen können (vgl. Täschner, 2020, S. 41). Dennoch treten verschiedene Fehlertypen auf, insbesondere in Bezug auf die Graphem-Phonem-Beziehung. Dazu gehören Homophone, also Wörter, die gleich klingen, aber unterschiedliche Schreibweisen haben (z. B. Seite & Saite, Lehre & leere). Dialektale Unterschiede führen zu unterschiedlichen Phonem-realisierungen, die von den Spracherkennungssystemen oft nicht korrekt verarbeitet werden. Auch Umlautfehler sind häufig, da viele KI-Systeme Schwierigkeiten haben, Umlautregelungen korrekt umzusetzen (z. B. *a* wird *ä* in *Mäuse*). Solche Aussprachevarianten wurden bereits in der Dissertation von Schaden (2006) linguistisch klassifiziert.

Die Schwierigkeiten der Spracherkennung hängen eng mit der Graphem-Phonem-Umsetzung zusammen. Das Deutsche verfügt über etwa 42 Phoneme, deren schriftliche Repräsentation nicht immer eindeutig ist. Die Erkennung dieser Schrift erfordert phonologische Bewusstheit, was auch für KI-Systeme gilt (vgl. Naumann, 2009, S. 24). Orthographische Prinzipien wie Morphemkonstanz, Großschreibung oder die Kennzeichnung von Vokallängen erleichtern zwar das Lesen, schützen jedoch nicht vor Fehlern in der automatisierten Transkription. Besonders die regionale und dialektale Variation, semantische Ambiguitäten sowie die Integration von Fremdwörtern stellen erhebliche Herausforderungen dar. Die häufigsten Fehlertypen lassen sich wie folgt kategorisieren:

1. Phonem-Verwechselung

Das Wort *Berg* kann fälschlicherweise als *Bär* wiedergegeben werden, insbesondre wenn /b/ und /bɛr/ in bestimmten Dialekten ähnlich ausgesprochen werden.

2. Graphematische Fehler

In Wörter wie *Käse* kann ein zusätzliches /s/ eingefügt werden (*Kässe* bzw. *Käße*), was auf eine unzureichende phonemische Modellierung hindeutet.

3. Kontextuelle Fehler

Das Wort *Bank* kann je nach Kontext eine andere Bedeutung haben. KI-Systeme erkennen diese Unterschiede oft nicht korrekt und führen dadurch zu semantisch fehlerhaften Transkriptionen. Es kann richtig, als *Finanzbank* transkribiert werden, aber auch falsch, wenn *Parkbank* gemeint ist. In diesem Bezug fällt der KI schwierig, das Wort nach dem Kontext zu bestimmen, was wegen falscher Transkription zu Missverständnis führen kann.

4. Homophone

Ohne Kontext ist die Unterscheidung von Wörtern wie *Seite* und *Saite* problematisch.

5. Dialektale Unterschiede

Hochdeutsche Modelle haben Schwierigkeiten, regionale Varietäten zu erkennen, wenn sie nicht speziell darauf trainiert wurden.

Köhler et al. (2019) untersuchten die Nutzung von *Audio-Mining* für die automatische Transkription von *Oral-History-Interviews*. Trotz Vorteilen wie reduzierter Transkriptionszeit bleibt die Herausforderung bestehen, dass ältere Interviews oft eine schlechtere Aufnahmequalität und spontane Sprache aufweisen. Im Rahmen des Projekts KA³ wird an der Verbesserung der Spracherkennung gearbeitet. Die genannten Probleme zeigen, dass trotz technologischer Fortschritte die exakte Umsetzung der Graphem-Phonem-Beziehung im Deutschen eine zentrale Schwierigkeit für KI-gestützte Spracherkennung bleibt. Genau hier setzt die empirische Analyse dieser Arbeit an.

3. Methodik

3.1. Forschungsdesign und Korpus

Diese Studie verfolgt das Ziel, die Leistungsfähigkeit KI-gestützter Spracherkennungssysteme im Vergleich zu menschlichen Transkriptionen zu evaluieren. Insbesondere wird untersucht, inwieweit automatisierte Transkriptionssysteme die Graphem-Phonem-Beziehungen im Deutschen adäquat abbilden und welche Fehlerarten dabei auftreten. Die Analyse der Transkriptionsgenauigkeit basiert auf einer

vergleichenden Analyse verschiedener linguistischer Faktoren wie phonologische Besonderheiten und Textkomplexität. Dafür wurde das *Spoken Wikipedia Corpus* als primäre Datenquelle gewählt, da es eine umfangreiche Sammlung deutscher Audioaufnahmen und menschlicher Transkriptionen in Form von Alignment-Dateien (XML) enthält. Diese Datenbasis eignet sich hervorragend, um die Herausforderungen der Graphem-Phonem-Beziehung im Deutschen zu untersuchen. Die *Spoken Wikipedia* ist eine wertvolle und frei zugängliche Quelle gesprochener Sprache, die bisher wenig in der Forschung genutzt wurde. Während viele existierende Sprachkorpora kostenpflichtig sind oder begrenzte Mengen an Daten enthalten, stellt die *Spoken Wikipedia* eine alternative Quelle gesprochener Sprache dar, die kontinuierlich durch freiwillige Beiträge wächst. Obwohl *Wikipedia* als eine der am häufigsten verwendeten Internetquellen für linguistische und computerlinguistische Forschung gilt, wurde die gesprochene Variante kaum wissenschaftlich verwendet. Die zentrale Herausforderung für die Forschung ist die fehlende Verknüpfung zwischen gesprochenem und geschriebenem Text. Zur Lösung diese Problem wurde eine zeitlich alignierte gesprochene *Wikipedia* entwickelt, die die Navigation und den Zugriff auf Informationen deutlich verbessert. Die Nutzer können auf Inhalte effizienter zugreifen als bei herkömmlichen, nicht-interaktiven Sprachaufnahmen (vgl. Baumann et al., 2019, S. 304-305).

Eine der wichtigsten Arbeiten, die die *Spoken Wikipedia* behandeln, ist „The Spoken Wikipedia Corpus Collection“. In dieser Arbeit stellen Baumann, Köhn, und Hennig (2019) die wichtigsten Vorteile der *Spoken Wikipedia* dar, und zwar:

- Ein neues Annotationsschema und eine Software zur Zeit-Alignierung gesprochener Wikipedia-Artikel.
- Die Erstellung und Analyse großer, frei verfügbarer Sprachkorpora für Deutsch, Englisch und Niederländisch.
- Eine Untersuchung des Interaktionsverhaltens („Hyperlistening“) mit gesprochenem Hypertext, die zeigt, dass Nutzer durch Zeit-Alignierung schneller auf Informationen zugreifen können.

Die *Spoken Wikipedia* ist somit eine nachhaltige und erweiterbare Datenquelle für die Sprach- und Computerlinguistik sowie für die

Weiterentwicklung von Spracherkennungssystemen (vgl. Baumann, Köhn, & Hennig, 2019, S. 304-305).

Das Untersuchungsmaterial umfasst 200 Audiodateien aus dem *Spoken Wikipedia Corpus*, die zufällig ausgewählt wurden.

3.2.Menschliche Referenztranskriptionen

Als Referenz wurden menschliche Transkriptionen aus dem gleichen Datensatz herangezogen. Diese stammen aus dem *Spoken Wikipedia Corpus*, wo sie zusammen mit den zugehörigen Audiodateien bereits von professionellen Transkribierenden manuell erstellt und veröffentlicht wurden. Sie dienen als Vergleichsgrundlage zur Bewertung der KI-Transkriptionen. Um die in den Alignment-Dateien enthaltenen manuellen Transkriptionen weiterzuverarbeiten, wurde ein Python-basiertes Skript entwickelt. Dabei werden alle Wörter zu einem zusammenhängenden Fließtext verbunden und anschließend als Text- bzw. Word-Dateien gespeichert, um eine einfache Weiterverarbeitung und Analyse zu ermöglichen.

3.3.KI-gestützte Transkription

Parallel zu den manuellen Transkriptionen wurden die entsprechenden Audioaufnahmen mittels modernes KI-basiertes Spracherkennungssystems automatisch transkribiert, und zwar *OpenAI Whisper*. Für die automatische Transkription wurde das Modell *Whisper*-Small verwendet. Die Transkriptionen erfolgten unter standardisierten Bedingungen mit einer Sampling-Rate von 16 kHz, in deutscher Spracheinstellung, und wurden im TXT-Format ausgegeben. Die automatisch erstellten KI-Transkripte wurden in einem separaten Ordner gespeichert, um für die anschließende Fehleranalyse mit den manuellen menschlichen Transkriptionen vergleichen zu können.

3.4.Vergleichsmethodik

Die Vergleiche zwischen menschlichen und KI-gestützten Transkriptionen basieren auf einer Kombination quantitativer und qualitativer Methoden. Die quantitative Fehleranalyse umfasst die Messwerte Word Error Rate (WER), Character Error Rate (CER) und Levenshtein-Distanz. Die WER misst die Fehler auf der Wortebene und bezeichnet das Verhältnis der falsch transkribierten Wörter zur Gesamtanzahl der Wörter. Eine hohe WER bedeutet, dass viele Wörter

falsch erkannt oder weggelassen wurden, z. B. wenn WER zwischen 10 und 20 % liegt, bedeutet dies, dass die Transkription sehr gut ist. Die Transkription bei WER zwischen 30 und 50 % ist mittelmäßig. Wenn die Fehlerrate mehr als 50 % liegt, wird die Transkription als schlecht bezeichnet.

Die Character Error Rate (CER) misst die Fehler auf Buchstabenebene und die Fehlerquote auf Zeichenebene zur Bewertung orthographischer Abweichungen. Hohe CER bedeutet, dass viele einzelne Buchstaben falsch erkannt wurden.

Die Levenshtein-Distanz zeigt die Anzahl der Einfügungen, Löschungen und Ersetzungen, die nötig sind, um die KI-Transkription in die menschliche Transkription umzuwandeln.

Die Berechnung der Metriken wurde mit Hilfe der *Jiwer-Bibliothek* in *Python* durchgeführt und die Ergebnisse wurden in einer Excel-Datei gespeichert und systematisch analysiert.

Neben den quantitativen Metriken wurden qualitative Analysen der häufigsten Fehlerarten durchgeführt, darunter:

- Substitutionsfehler: Verwechslung ähnlich klingender Wörter
- Tilgungen: Auslassungen bestimmter Laute oder Silben
- Einfügungen: Hinzufügung von nicht vorhandenen Elementen

Die qualitative Analyse wurde durch eine manuelle Beobachtung der Transkripte durchgeführt, um systematische Muster und wiederkehrende Fehlerkategorien zu identifizieren.

4. Die Ergebnisse

In diesem Kapitel werden die im Rahmen der Analyse gewonnenen Ergebnisse systematisch dargestellt. Es wird zwischen qualitativen und quantitativen Verfahren unterschieden, um sowohl einzelne Fehlertypen detailliert zu erfassen als auch deren statistische Häufigkeiten zu quantifizieren. Ziel ist es, Unterschiede zwischen menschlichen und KI-gestützten Transkriptionen aufzuzeigen und diese im Hinblick auf die Graphem-Phonem-Korrespondenz sowie relevante phonologische und morphologische Strukturen linguistisch einzuordnen. Aufgrund fehlender Unterschiede in Bezug auf Dialektvariation und Sprechtempo konzentriert sich die vorliegende Darstellung auf

übergreifende Transkriptionsfehler, deren sprecherabhängige Ausprägungen Gegenstand zukünftiger Untersuchungen sein können.

4.1.Die quantitative Analyse

Tabelle 2 zeigt eine Übersicht über die Häufigkeit der erfassten Fehlertypen in den KI-Transkriptionen.

Tabelle 2: Statistische Übersicht der allgemeinen Transkriptionsfehler.

Fehlerrate	Mittelwert	Standardabweichung
WER (%)	58,71 %	89,41
CER (%)	31,15 %	22,53
Levenshtein-Distanz	926924	244638

Die Analyse der Transkriptionsqualität ergab eine durchschnittliche Wortfehlerrate (WER) von 58,71 %, was darauf hindeutet, dass mehr als die Hälfte der transkribierten Wörter fehlerhaft wiedergegeben wurden. Die Zeichenfehlerrate (CER) liegt bei 31,19 %, was darauf schließen lässt, dass Fehler auf Wortebene deutlich stärker ins Gewicht fallen als auf Zeichenebene. Die hohe Standardabweichung des WER (89,41) verweist auf eine große Qualitätsunterschiede zwischen den Transkripten, was unter anderem auf phonetische Reduktionen oder eine komplexe Lexik zurückzuführen sein könnte. Demgegenüber deutet die vergleichsweise geringere Standardabweichung des CER (22,65) darauf hin, dass die Anzahl fehlerhafter Zeichen in den Transkriptionen relativ stabil bleibt, auch wenn ganze Wörter häufig vollständig falsch transkribiert werden. Zur Ergänzung der WER- und CER-Messung wurde die Levenshtein-Distanz als weiteres Maß herangezogen. Der durchschnittliche Wert von 10.133,40 Zeichenoperationen (Einfügungen, Löschungen oder Ersetzungen) zwischen KI- und Referenztranskriptionen verdeutlicht das Ausmaß der Abweichungen. Eine Standardabweichung von 26.787,40 zeigt zudem, dass einige Transkripte erheblich von den menschlichen Referenzen abweichen, während andere deutlich näher an der Zielversion liegen.

4.2.Die qualitative Analyse

Die qualitative Analyse konzentriert sich auf typische Fehlerarten in KI-gestützten Transkriptionen im Vergleich zu menschlichen Transkriptionen. Dabei geht es nicht nur um die Anzahl der Fehler, sondern vor allem um deren sprachliche Beschaffenheit,

Regelmäßigkeiten und mögliche Ursachen. Die Auswertung zeigt, dass die KI-Transkriptionen häufig kleinere Rechtschreibfehler sowie Probleme bei der Zeichensetzung aufweisen. Besonders auffällig sind fehlerhafte Wiedergaben von Eigennamen und Fachbegriffen, die teilweise durch ähnlich klingende Wörter ersetzt oder ausgelassen wurden. Auch grammatikalische Veränderungen traten auf, die zu inhaltlichen Abweichungen führten. Zusätzlich zeigten sich strukturelle Schwächen bei der Satz- und Absatzbildung. Tabelle 3 gibt einen Überblick über die häufigsten Fehlerkategorien und stellt ausgewählte Beispiele aus dem Korpus dar.

Tabelle 3: Typische Fehlerkategorien in KI-gestützten Transkriptionen mit Beispielen.

Fehlerkategorie	Beschreibung	Beispiel: Menschlich → KI
Substitutionsfehler	Ein Wort wird durch ein anderes Wort ersetzt.	Städten → Stitten
Auslassungen	Ein Wort oder ein Laut fehlt.	Herbsttagen → Herbstagen
Einfügungen	Ein zusätzliches Wort oder Segment werden eingefügt.	Wertach → Werthach
Phonetische Fehler	Ähnliche Laute werden verwechselt. Lautähnliche Wörter werden verwechselt, insbesondere bei Eigennamen oder seltenen Begriffen.	Singold → Zingold
Morphosyntaktische Fehler	Endungen oder Flexionen sind falsch, d. h. Wortformen werden falsch erkannt oder die Satzstruktur ist fehlerhaft.	Gebietsreformen → Gebietreform
Falsche Segmentierung	Wörter werden falsch getrennt oder verbunden, wodurch neue lexikalische Einheiten entstehen.	… bilden dabei im Osten beginnend und dem Uhrzeigersinn → bilden dabei im ostenbeginnend und dem Uhrzeige sind …

Die identifizierten Fehlertypen lassen sich in folgende Kategorien gliedern:

a. Orthographische Fehler und falsche Worttrennung

Menschliche Transkription	KI-Transkription	Fehlerart
„Wertach“	„Werthach“	Falsche Schreibweise
„Singold“	„Zingold“	Verwechslung ähnlich klingender Wörter
„Aichach-Friedberg“	„Eichach-Friedberg“	Buchstabenvertauschung
„UPM Kymmene“	„UPM Kümene“	Falsche Schreibweise des Eigennamens

Viele Fehler resultieren aus einer falschen phonemischen Zuordnung von Wörtern. Besonders betroffen sind ungewöhnliche oder regionale Namen. Die KI scheint Schwierigkeiten mit Doppelkonsonanten und Umlauten zu haben.

b. Satzzeichen- & Strukturfehler

Menschliche Transkription	KI-Transkription	Fehlerart
„Augsburg ist durch seine Lage im gewitterintensivsten Bundesland Bayern des Öfteren von heftigen Unwettern betroffen, welche zu enormen Lech- und Wertachhochwassern führen.“	„Augsburg ist durch seine Lage im Gewitter intensivsten Bundesland Bayern des äfteren von heftigen Unwetter betroffen, welche zu enormen Läch und Werthach hochwasser führen …“	Fehlende Satztrennung
„Wasser mit Härtegrad 13,5° dH (mittelhart) versorgt …“	„Wasser mit Herdegrad 13,5° deutsche Härte, das ist Mittelhard, versorgt …“	Wortersetzungsfehler („Härtegrad“ → „Herdegrad“)

Viele Sätze sind entweder zu lang oder falsch segmentiert, was zu Verständnisschwierigkeiten führt. Die Interpunktion ist uneinheitlich, insbesondere bei der Setzung von Kommas und Punkten. Manche Sätze wurden von der KI umgestellt oder ergänzt, was zu inhaltlichen Veränderungen führt und den Text schwer verständlich macht. Die KI scheint jedoch eine solide Grundstruktur beizubehalten, sodass eine manuelle Nachbearbeitung möglich ist.

c. Lexikalische Fehler & Bedeutungsveränderungen

Menschliche Transkription	KI-Transkription	Fehlerart
„Agglomeration“	„Akklimeration“	Falsches Wort gewählt
„Venedig“	„venedig“	Kleinschreibung von Eigennamen

Die KI hat Schwierigkeiten mit Bedeutungserhaltung (Semantik), seltenere Wörter werden durch falsche oder unpassende Alternativen ersetzt. Auch Groß- und Kleinschreibung von Eigennamen wird nicht immer korrekt übernommen. Diese Beobachtungen verdeutlichen die Grenzen aktueller KI-Transkriptionssysteme, insbesondere im Umgang mit Eigennamen, Fachtermini und komplexer Satzstruktur. Abschließend wurden die Ergebnisse in Bezug auf die Graphem-Phonem-Beziehung im Deutschen diskutiert. Die Analyse ermöglicht eine Einschätzung der Leistungsfähigkeit aktueller KI-Modelle in der automatischen Spracherkennung und liefert wertvolle Erkenntnisse über die spezifischen Herausforderungen bei der Transkription gesprochener Sprache.
Im Folgenden werden die wichtigsten Ergebnisse im Hinblick auf die Fragestellungen kritisch diskutiert.

5. Diskussion

In Bezug auf die erste Fragestellung „Welche Unterschiede zeigen sich zwischen KI-gestützten und menschlichen Transkriptionen im Deutschen, insbesondere in Bezug auf die Graphem-Phonem-Beziehung, phonologische Abweichungen, morphologische Strukturen und die Wortsegmentierung?“ zeigt die Analyse der Transkriptionen deutliche Unterschiede zwischen menschlicher und KI-gestützter Transkription gesprochener Sprache. Menschliche Transkriptionen weisen ein deutlich höheres Maß an linguistischer Sensitivität auf. Insbesondere bei morphologisch komplexen Strukturen –etwa Flexionsendungen oder Wortzusammensetzungen– erkennen menschliche Transkribierende durch ihr grammatisches Wissen bedeutungstragende Einheiten zuverlässig und setzen diese korrekt um. Die KI hingegen greift primär auf akustische Wahrscheinlichkeiten zurück und zeigt Schwächen bei der morphologischen Analyse. Auch im Bereich der Graphem-Phonem-Zuordnung zeigen sich Differenzen; menschliche Transkriptionen berücksichtigen standardsprachliche orthografische Normen, während KI-Systeme oft lautliche Realisierungen direkt in Schrift umsetzen, ohne morphologische oder syntaktische Regularitäten zu beachten. Dies betrifft vor allem Fälle von Reduktion, Elision oder Assimilation in der gesprochenen Sprache, z. B. („Kusee“ statt „Kuhsee“) oder („Herbstagen“ statt „Herbsttagen“). Diese Prozesse werden von KI-Systemen häufig

entweder übersehen oder fälschlich überinterpretiert, was die Genauigkeit zusätzlich mindert. Auch beim Umgang mit komplexen Wortstrukturen zeigen sich erste Unterschiede. Während Menschen auch bei schneller oder verschliffener Sprache die Wortgrenzen rekonstruieren können („Wald Besitzer“ statt „Waldbesitzer“), transkribiert die KI häufig lautgetreu, was zu syntaktisch oder semantisch fehlerhaften Strukturen führen kann. Dies verdeutlicht den Mangel an tiefgehendem sprachsystematischem Wissen bei automatisierten Systemen. Die Auswertung offenbart mehrere charakteristische Fehlertypen, die auf die technologischen Grenzen aktueller Spracherkennungssysteme hinweisen. Häufige Fehlerkategorien umfassen phonematische Substitutionen (z. B. Verwechslung von /b/ und /p/, /ɛː/ und /eː/), die auf akustische Ähnlichkeiten zurückzuführen sind und von der KI nicht durch sprachliches Vorwissen berücksichtigt werden können, Orthografische Fehlschreibungen (etwa bei langen Vokalen, Umlauten oder bei der Groß-/Kleinschreibung, wobei letztere von KI-Systemen weitgehend unbeachtet bleibt, obwohl sie semantisch und syntaktisch bedeutsam ist) und morphologische Fehler, z. B. die Auslassung oder falsche Bildung von Flexionsmorphemen (Plural, Tempus), die sich negativ auf die semantische Präzision auswirken.

Diese Fehler zeigen, dass die KI-gestützte Umsetzung der Graphem-Phonem-Beziehung im Deutschen scheitert. Der rein akustische Fokus der KI verhindert eine regelhafte und kontextangemessene Abbildung der orthographischen Form, wodurch es zu einer Vielzahl nicht standardisierter oder sinnverzerrender Transkriptionsformen kommt. Zusammenfassend lassen sich bei KI-gestützten Transkriptionen im Vergleich zu menschlichen Transkriptionen folgende typische Fehlmuster feststellen:

- Phonetische Ambiguität und Konsonantenverwechslung

Die KI weist Schwierigkeiten in der Unterscheidung zwischen stimmhaften und stimmlosen Konsonanten auf, insbesondere im Falle von <s> und <z> (z. B. „Singold“ → „Zingold“). Dies deutet auf eine unzureichende Modellierung der positionsabhängigen Phonemrealisierung hin, die für das Deutsche charakteristisch ist.

- Orthografische Abweichungen und historische Schreibweisen

Die erstellten Transkriptionen enthalten gelegentlich veraltete Schreibweisen, etwa durch die falsche Verwendung von <th> in „Werthach“. Dies könnte darauf hinweisen, dass die Trainingsdaten der KI nicht einheitlich modernisiert wurden oder dass mehrere Orthografievarianten parallel verarbeitet werden, ohne eine klare Standardisierung vorzunehmen.

- Probleme mit Wortsegmentierung und Komposita

Eine auffällige Fehlerquelle stellt die fehlerhafte Trennung von Zusammensetzungen dar (z. B. „Autobahnsee“ → „Auto Bahnsee“). Da das Deutsche eine hohe Anzahl an Komposita aufweist, sind leistungsfähige Sprachmodelle darauf angewiesen, Wortgrenzen zuverlässig zu identifizieren. Die fehlerhafte Segmentierung legt nahe, dass die KI-Transkription Probleme hat, phonologische Hinweise auf Wortgrenzen korrekt zu verarbeiten.

- Morphosyntaktische Fehler und Flexionsabweichungen

Die KI-Transkriptionen enthalten vereinzelt syntaktische und morphologische Fehler, beispielsweise durch das irrtümliche Hinzufügen eines Genitiv-Suffixes („Augsburg“ → „Augsburgs“). Dies deutet darauf hin, dass die KI-systeme gelegentlich grammatische Strukturen falsch transkribieren, anstatt sich ausschließlich auf das phonetische Signal zu stützen.

In Bezug auf die zweite Fragestellung „Welche Auswirkungen haben aktuelle Technologien der KI-gestützten Spracherkennung auf korrekte Umsetzung der Graphem-Phonem-Beziehung im Deutschen?“ machen die Ergebnisse deutlich, dass aktuelle KI-Technologien erhebliche Einschränkungen in der korrekten Umsetzung der Graphem-Phonem-Beziehung aufweisen. Diese Beziehung wird insbesondere dort fehlerhaft wiedergegeben, wo die lautliche Realisierung von der orthografischen Standardform abweicht. So werden etwa lange Vokale, Dehnungen oder stumme Buchstaben (wie das Dehnungs-„h“ oder das „ie“ für langes /iː/) häufig nicht korrekt erfasst. Ein zentrales Problem besteht darin, dass KI-Systeme typischerweise keine phonologisch-orthografische Normalisierung durchführen. Sie verschriftlichen das akustische Signal weitgehend direkt ohne die dafür notwendige konzeptuelle Rekonstruktion auf Wort- und Satzebene. Dadurch entstehen

fehlerhafte Graphem-Phonem-Umsetzungen, insbesondere bei Flexionsformen, Wortbildungen oder standardsprachlich kodifizierten Besonderheiten. Zusammenfassend lässt sich feststellen, dass die Technologie noch nicht in der Lage ist, die Komplexität der deutschen Graphem-Phonem-Korrespondenzen regelhaft abzubilden. Die Abhängigkeit von der Oberflächenform führt zu einer Vielzahl orthografischer Abweichungen von der Norm, was insbesondere in linguistisch relevanten Analysen problematisch ist.

In Bezug auf die dritte Fragestellung „Welche sprachspezifischen Herausforderungen (z. B. Homophone, Umlautveränderungen) führen besonders häufig zu Fehlern in der KI-gestützten Transkription gesprochener Sprache?“ zeigt die Analyse, dass bestimmte sprachspezifische Phänomene besonders anfällig für Fehler sind. Dazu gehören vor allem:

- Homophone, die sich lautlich gleichen, aber orthografisch und semantisch unterscheiden (z. B. „Seite“ vs. „Saite“). KI-Systeme erkennen solche Unterschiede nicht, da sie keine semantische Kontextanalyse durchführen können, wodurch es zu zufälligen oder inkorrekten Schreibungen kommt.
- Umlautveränderungen im morphologischen Paradigma (z. B. „Haus“ → „Häuser“, „Wald“ → „Wälder“) stellen eine große Herausforderung dar. Während menschliche Transkribierende diese auf Basis ihres sprachlichen Wissens korrekt umsetzen, werden sie von der KI häufig als phonetisch unabhängige Einheiten interpretiert.
- Auch trennbare Verben („abfahren“, „aufstehen“) und Komposita („Hochschulstudium“, „Sprachverarbeitung“) bereiten der KI-Schwierigkeiten, da sie häufig nicht als zusammengehörige Einheiten erkannt oder falsch segmentiert werden.

Ohne Einbindung in einen semantisch-syntaktischen Kontext bleibt die Transkription auf der lautlichen Oberfläche stehen – mit entsprechend hoher Fehleranfälligkeit.

6. Fazit

Die vorliegende Analyse zeigt, dass die fehlerhafte Umsetzung der Graphem-Phonem-Beziehung in KI-Transkriptionen maßgeblich durch phonologische, orthografische und morphologische Faktoren bedingt ist. Besonders auffällig ist die eingeschränkte Fähigkeit der KI-Systeme, stimmhafte und stimmlose Konsonanten korrekt zu differenzieren. Auch komplexe Komposita und zusammengesetzte Wörter werden häufig unzutreffend segmentiert. Darüber hinaus treten orthografische Unregelmäßigkeiten auf, die teils auf veraltete oder inkonsistente Schreibweisen zurückzuführen sind. Morphologische Fehler deuten auf eine unzureichende Modellierung der deutschen Flexionsmorphologie hin.

Diese Ergebnisse unterstreichen die Notwendigkeit einer gezielten Anpassung KI-gestützter Spracherkennungssysteme an die spezifischen Eigenschaften der deutschen Graphem-Phonem-Korrespondenz. Eine stärkere Einbindung linguistischer Regeln zu Wortbildung, Morphologie und Segmentierung erscheint besonders vielversprechend, um die Transkriptionsqualität nachhaltig zu verbessern.

Die systematische Fehlerklassifikation ermöglicht eine differenzierte Bewertung der Schwächen aktueller Systeme und bildet zugleich eine Grundlage für zukünftige Optimierungsansätze. Für die Qualitätsbeurteilung zentral sind dabei sprecherabhängige Variablen wie Sprechgeschwindigkeit, Dialektvariation und Hintergrundgeräusche. Künftige Studien könnten verstärkt untersuchen, inwieweit sich bestehende Sprachmodelle an solche variierenden linguistischen und akustischen Bedingungen anpassen lassen. Darüber hinaus leistet die linguistische Analyse einen entscheidenden Beitrag zur Weiterentwicklung KI-gestützter Systeme: Durch die systematische Beschreibung der Fehler und Schwachstellen liefert sie wertvolle Anhaltspunkte für die gezielte Optimierung durch Informatiker/innen.

Literaturverzeichnis

Andresen, M. (2024). *Computerlinguistische Methoden für die Digital Humanities: Eine Einführung für Geisteswissenschaftler: innen*: Narr Francke Attempto Verlag.

Baumann, T., Köhn, A., & Hennig, F. (2019). The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. *Language Resources and Evaluation, 53*(2), 303-329. doi:10.1007/s10579-017-9410-y

Bußmann, H. (2008). *Lexikon der Sprachwissenschaft* (4, durchgesehene und ergänzte Aufl.). Stuttgart: Kröner.

Dürscheid, C. (2002). Graphematik. In C. Dürscheid (Hrsg.), *Einführung in die Schriftlinguistik* (S. 139-179). Wiesbaden: VS Verlag für Sozialwissenschaften.

Dürscheid, C. (2016). *Einführung in die Schriftlinguistik. Mit einem Kapitel zur Typographie von Jürgen Spitzmüller* (5. aktualisierte und korrigierte Aufl. Vol. 3740). Göttingen: Vandenhoeck & Ruprecht.

Eisenberg, P. (1988). Die Grapheme des Deutschen und ihre Beziehungen zu den Phonemen. *Sprachwissenschaft. Ein Reader*, 346-360.

Evertz-Rittich, M., Berg, K., & Meisenburg, T. (2023). Graphematik – die Beziehung zwischen Sprache und Schrift. In R. Klabunde & W. Mihatsch (Eds.), *Linguistik: Eine Einführung (nicht nur) für Germanisten, Romanisten und Anglisten* (S. 197-217). Berlin, Heidelberg: Springer.

Fuhrhop, N., & Peters, J. (2013). Graphematik. In N. Fuhrhop & J. Peters (Eds.), *Einführung in die Phonologie und Graphematik* (S. 179-285). Stuttgart: J.B. Metzler.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). *Speech recognition with deep recurrent neural networks.* Paper presented at the 2013 IEEE international conference on acoustics, speech and signal processing.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine, 29*(6), 82-97.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3. Aufl.): Pearson.

Khosravani, A., Garner, P. N., & Lazaridis, A. (2021). *Modeling Dialectal Variation for Swiss German Automatic Speech Recognition.* Paper presented at the Interspeech. Doi: 10.21437/Interspeech.2021-1735.

Köhler, J., Gref, M., & Leh, A. (2019). KA³. Weiterentwicklung von Sprachtechnologien im Kontext der Oral History. *BIOS–Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen, 30*(1+ 2), 44-59.

Kriesch, L. (2023). *Web Mining und Natural Language Processing als methodisches Komplement in der Wirtschaftsgeographie.* (PhD Diss.). Justus-Liebig-Universität Gießen, Gießen.

Maas, U. (2010). *Grundzüge der deutschen Orthographie* (Vol. 120): Walter de Gruyter.

Naumann, C. L. (2009). Schrift verstehen und lernen–wie hilft die Graphematik bei Unterricht und Lerntherapie. *Individuelle Förderung: Lernschwierigkeiten als schulische Herausforderung: Lese-Rechtschreibschwierigkeiten-Rechenschwierigkeiten, 3*, 23.

Pellegrino, E., Dellwo, V., & Kathiresan, T. (2022). Vowel convergence does not affect auditory speaker discriminability in humans and machine in a case study on

Swiss German dialects. *International Journal of Speech, Language and the Law, 29*(1), 60–84. doi:https://doi.org/10.1558/ijsll.19954.

Ruhaltinger, J. (2022). Reden wie gedruckt. *Das österreichische Gesundheitswesen ÖKZ, 63*(5), 58-59. doi:10.1007/s43830-022-0114-2

Sapp, C. D. (2019). Einführung in die deutsche Linguistik/Introduction to German Linguistics (Textbooks and Open Educational Resources.1) (Publication no. https://egrove.olemiss.edu/open/1). Retrieved 26.09.2024

Schaden, S. (2006). *Regelbasierte Modellierung fremdsprachlich akzentbehafteter Aussprachevarianten.* (Dr. phil.). Duisburg Universität, Essen.

Schiel, F. (1999). *Automatic phonetic transcription of non-prompted speech*. Paper presented at the ICPhS99 San Francisco.

Schründer-Lenzen, A. (2013). *Schriftspracherwerb* (4. Aufl.). Wiesbaden: Springer-Verlag.

Täschner, D. (2020). Spracherkennung - ein Grundbaustein von Digitalisierungsstrategien in der Pflege. In V. Kubek, S. Velten, F. Eierdanz, & A. Blaudszun-Lahm (Eds.), *Digitalisierung in der Pflege: Zur Unterstützung einer besseren Arbeitsorganisation* (S. 41-48). Berlin, Heidelberg: Springer Berlin Heidelberg.

Treutlein, A. (2011). *Rekodieren im Deutschen und Englischen: Wie rekodieren Englischlerner/-innen mit deutscher Muttersprache englische Wörter?* (Dissertation). Universität Tübingen,

Weigelt, S. (2022). *Eine agentenbasierte Architektur für Programmierung mit gesprochener Sprache*. Karlsruhe: KIT Scientific Publishing.

[1] Hervorhebung im Original (Schründer-Lenzen, 2013, S. 16).