Pause im Text. Zur Exploration semantisch konditionierter Sprechpausen in Hörbüchern

Stiemer, Haimo
stiemer@linglit.tu-darmstadt.de
TU Darmstadt, Deutschland
ORCID: 0000-0002-4407-2415

Hatzel, Hans Ole
hans.ole.hatzel@uni-hamburg.de
Universität Hamburg, Deutschland
ORCID: 0000-0002-4586-7260

Biemann, Chris
chris.biemann@uni-hamburg.de
Universität Hamburg, Deutschland
ORCID: 0000-0002-8449-9624

Gius, Evelyn
evelyn.gius@tu-darmstadt.de
TU Darmstadt, Deutschland
ORCID: 0000-0001-8888-8419

1. Sprechpausen als Vehikel der Textsegmentierung

Die für die Textanalyse grundlegende Segmentierung von Prosatexten, also deren Zerlegung in diskrete Einheiten, ist in den Computational Literary Studies (CLS) weiterhin ein prominentes Problem. Obgleich viele computationelle Verfahren die vorhergehende Unterteilung von Texten voraussetzen, fehlt es bislang an standardisierten Segmenten (cf. Bartsch et al. 2023). In Abhängigkeit von der jeweiligen Forschungsfrage und der zum Einsatz bestimmten computationellen Methoden finden sich sowohl Segmentierungen von Layoutelementen bzw. Einheiten der materiellen Textgestaltung (cf. Herzog 2018), die Tokenisierung (auf Wort- oder Satzebene) oder aber das *Unitizing*, also die nach inhaltsanalytischen Kategorien erfolgte Identifikation von größeren Einheiten in Texten wie z. B. Szenen (cf. Gius et al. 2019).

Ein in den CLS bislang nicht geprüfter Ansatz der Segmentierung, die Zergliederung von Erzähltexten mittels der in der Rezitation emergenten Sprechpausen, ist Gegenstand dieses Beitrags. Wir präsentieren erste Beobachtungen hinsichtlich der Möglichkeiten wie Bedingungen für die Identifikation und Analyse semantisch konditionierter Sprechpausen in Hörbüchern. Ein solcher Zugang würde die genannten, allein vom Text ausgehenden Zugänge zur Segmentierung um eine in der Textrezeption erzeugte Segmentierung komplementieren, auch um die bestehenden Segmentierungsoptionen überprüfen zu können.

Ausgangspunkt unserer Analyse ist das in der Sprechwissenschaft und Sprecherziehung entwickelte Konzept des interpretierenden Textsprechens (cf. Geißner 1981:175; Brand 2021), mit welchem jedwede Rezitation von Texten vor Publikum als originärer Interpretationsvorgang und mithin die Sprechfassung eines Textes als dessen Interpretation verstanden werden. In der Konsequenz verstehen wir die Pausensetzung und -länge, welche der prosodischen Dimension eines literarischen Werks angehören, als die von Vortragenden ad hoc oder planmäßig vorgenommene sinnhafte Segmentierung eines Textes. Unbestimmt bleibt dabei zunächst, inwiefern es sich bei den potentiell multifunktionalen Sprechpausen um syntaktische und typographisch bedingte, auf z.B. Interpunktion oder Absätze rekurrierende, oder aber um semantisch konditionierte Unterbrechungen handelt, denen eine noch näher zu bestimmende handlungsrelevante Funktion zugewiesen werden kann. In diesem Beitrag untersuchen wir demzufolge die semantisch-interpretative Qualität von Sprechpausen in Tonaufzeichnungen von eingelesenen Erzähltexten. Der Beschreibung unseres Audio-Korpus (2.) sowie der Transkriptions-Methode (3.) folgt die Exploration der aus diesem Korpus gewonnenen, transkribierten Daten, um die mögliche Motivierung der Sprechpausen zu erfassen. Analysiert wird hierfür die Verteilung der Anzahl wie Länge der in den Daten detektierten Pausen (4.). Anschließend diskutieren wir den Zusammenhang von Sprechkompetenz und Pausensetzung (5.) und ziehen aus der Datenexploration Rückschlüsse für weitergehende Forschungsaktivitäten (6.).

2. Professionelle und Laien-Lesungen

Für unsere Untersuchung haben wir die aufgezeichneten Lesungen von drei, aus dem EvENT-Projektkorpus (Vauth et al. 2021) entnommenen Erzähltexten analysiert. Bei der Auswahl der Texte wurde darauf geachtet, dass sich diese durch verschiedene narrative Profile bzw. Charakteristika auszeichnen. Den analysierten Vortragsaufzeichnungen zugrunde lagen somit Franz Kafkas hypotaktisch geprägte Erzählung *Die Verwandlung*, Annette von Droste-Hülshoffs dialogreiche, viele Passagen mit direkter Rede aufweisende Erzählung *Die Judenbuche* sowie Heinrich von Kleists *Das Erdbeben in Chili*, in welchem kontrastiv zeitraffende und szenisch erzählte Passagen vorhanden sind. Für jeden dieser Texte haben wir eine professionelle, von einer/m Schauspieler:in eingesprochene Lesung sowie eine über das frei zugängliche Online-Portal *LibriVox* erhältliche Laien-Le-

sung analysiert. Durch die Auswahl von professionellen sowie laienhaften Vorträgen enthält unser Audio-Korpus mutmaßlich Aufzeichnungen von Vortragenden mit unterschiedlicher Text- und Vortragskompetenz, was zumindest eine tentative Auswertung der Auswirkungen dieser Kompetenzen auf die Pausensetzungen in den verschiedenen Sprechfassungen ermöglicht. Die Annahme ist dabei, dass die professionell Vortragenden die Pausen bewusster einsetzen und damit auch mehr semantisch konditionierte bzw. relevante Sprechpausen vornehmen, die entsprechend stärker an Bedeutungselementen der Texte ausgerichtet sind. Solche Pausen sind an ihrer verhältnismäßig längeren Dauer erkennbar. Überdies muss berücksichtigt werden, dass die professionellen Lesungen redaktionell betreut sind und die Tonaufzeichnungen zudem eine Postproduktion, also Nachbearbeitung, durchlaufen haben.

3. Pausenerkennung mit WhisperX

Für die automatische Pausenerkennung verwendeten wir Whisper (Radford et al. 2023), ein neuronales Modell zur Transkription von gesprochener Sprache, welches annähernd die Fehlerraten von professionellen Transkriptor:innen erreicht (Radford et al. 2023, Abbildung 7). Prinzipiell wäre es wünschenswert, einen Alignierungsansatz zu nutzen, die Audiodaten also mit einem – in unserem Fall verfügbaren – Originaltext abzugleichen. Schiel et al. (2017) beschreiben ein derartiges System. In der Praxis war jedoch keine Implementation einfach auf unseren Daten anwendbar, sodass die automatische Transkription für diese explorative Arbeit passend war. Konkret setzen wir die Implementation WhisperX (Bain et al. 2023) ein, welche zahlreiche zusätzliche Funktionen im Vergleich zur ursprünglichen Whisper-Implementation bietet. Die Anwendung von WhisperX auf eine beliebige Audiodatei erzeugt ein Transkript eben dieser, in dem einzelne Segmente mit Zeitcodes versehen sind. Die Segmente entsprechen dabei linguistischen Sätzen (da wir WhisperX mit den Standardoptionen aufrufen und somit *Sentence* als Wert für die *Segment-Resolution* wählen), wobei diese Separierung unter anderem auf Grundlage des transkribierten Textes erfolgt, dessen Satzzeichen naturgemäß nicht immer dem Originaltext entsprechen. Für unseren Anwendungsfall sind die ebenfalls verfügbaren Zeitinformationen auf der Wortebene – im Hintergrund durch wav2vec2 (Baevski et al. 2020) umgesetzt – essentiell, denn Whisper bietet in der Ursprungsimplementierung nur Anfangs-Zeitcodes für Segmente, sodass unklar ist, wieviel Zeit zwischen zwei Sätzen oder Segmenten verbleibt. Mit WhisperX erstellen wir also eine Transkription, in der Start- und Endzeitstempel auf der Segment- sowie auf der Tokenebene enthalten sind. Unsere Transkriptionen wurden mit der large-v2 Variante von Whisper angefertigt. Die sechs so bearbeiteten Audio-Dateien werden von dem Modell in insgesamt 4’542 Segmente aufgeteilt, die wir im Folgenden weiter untersuchen.

4. Pausen im Korpus

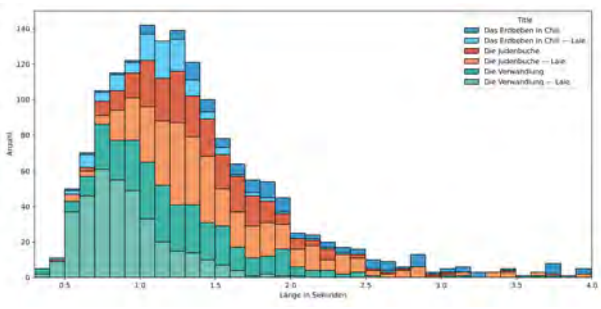

Abbildung 1: Verteilung der Pausen nach Länge und Anzahl in den Sprechfassungen

Die Verteilung der insgesamt 4.542 im Audio-Korpus mit WhisperX detektierten Pausen (an Satzenden) nach den Sprechfassungen der Texte auf die Pausenlängenwerte (x-Achse) sowie die Pausenanzahl (y-Achse) findet sich in Abbildung 1. Mit der Auswertung der Daten und in Anlehnung an die vom Grammatischen Informationssystem grammis (cf. Institut für Deutsche Sprache 2013) vorgeschlagene Differenzierung von Sprechpausen ergeben sich zunächst folgende Beobachtungen: In allen Sprechfassungen dominieren erwartungsgemäß die sehr kurze Pausen, die lediglich auf Atemeinschnitte bei der Satzbeendigung verweisen ($x<0{,}3$ Sek.), während Verzögerungspausen ($0{,}3<x<1$) deutlich seltener auftreten. Wird ab dem Längenwert 1 Sekunde den Pausen eine dann noch zu bestimmende semantische Relevanz zugemessen (Relevanzpause; cf. ebd.), weisen beide Sprechfassungen der *Verwandlung* in diesem Sektor die wenigsten Relevanzpausen auf (professionelle Lesung 24 % und Laienlesung 12 % aller Pausen der jeweiligen Sprechfassung), was im Zusammenhang mit der hypotaktischen Struktur des Textes oder aber dem deutlich geringeren Orts- und Zeitenwechsel im Text stehen könnte. Die Verwandlung Gregor Samsas in ein Ungeziefer und die sich anschließenden Interaktionen mit seiner Familie finden ausschließlich in der Wohnung der Samsas statt, die erzählte Zeit im ersten Kapitel beträgt ca. eineinhalb Stunden.

Die dialogreiche Gestaltung der *Judenbuche* wiederum scheint, obschon der Text kürzer als *Die Verwandlung* ist, zu der höchsten Anzahl an Pausen im Audio-Korpus beizutragen (1.108 bei der Laienlesung, 1.064 bei der professionellen Lesung): Allein bei den Relevanzpausen beträgt der Anteil jener Pausen, die nach oder vor der direkten Redewiedergabe einer Figur gesetzt werden, 30 % (n = 66). Bei der *Verwandlung* hingegen beläuft sich der Anteil der anscheinend durch direkte Rede motivierten Pausen im gleichen Pausenspektrum auf gerade einmal 3 % (n = 7).

Die höchsten Anteile an Relevanzpausen weisen die beiden Sprechfassungen des Kleist-Textes auf (professionelle Lesung 44 % und Laienlesung 41 % aller Pausen der jeweiligen Sprechfassung). Es handelt sich im Korpusvergleich um den Text mit den meisten Figuren, mit einem sehr handlungsintensiven, also auch diverse Ortswechsel enthalten-

den, Plot. Nachdem das Verhältnis der Segmentlänge zur durchschnittlichen Satzlänge der Texte unterschiedlich ist, nehmen wir an, dass eher bestimmte narrative Eigenschaften für Pausen relevant sind. Zumindest scheint die höhere Heterogenität, die durch vermehrte narrative Elemente wie Dialoge, Figuren oder auch weitere, handlungsbezogene Elemente entsteht, auch die Anzahl an Relevanzpausen zu steigern.

Sprechfassung	Token Textvorlage	Gesamtanzahl Pausen	Atemeinschnitte und Verzögerungspausen (x<1) in Sekunden und Prozent	Relevanzpausen (x>1) in Sek. und Prozent
Die Judenbuche professionelle Lesung	16'205	1'064	844 (79,3 %)	220 (20,7 %)
Die Judenbuche Laienlesung	16'205	1'108	744 (67,2 %)	364 (32,8 %)
Die Verwandlung professionelle Lesung	19'153	937	708 (75,6 %)	229 (24,4 %)
Die Verwandlung Laienlesung	19'153	962	851 (88,5 %)	111 (11,5 %)
Das Erdbeben in Chili professionelle Lesung	5'390	270	150 (56,0 %)	120 (44,0 %)
Das Erdbeben in Chili Laienlesung	5'390	201	119 (59,2 %)	82 (40,8 %)

Tabelle 1: Verteilung der Pausen nach Länge in den Sprechfassungen

5. Kompetenz und Pausensetzung

Abgesehen von der Laienfassung des *Erdbebens* ist die Anzahl der Gesamtpausen bei den beiden anderen, von Laien eingesprochenen Texten leicht höher als bei jenen der professionell Vortragenden. Eine Tendenz bezüglich der möglichen Auswirkung der Sprechkompetenz zeichnet sich jedoch ab, wenn der Fokus auf die Relevanzpausen gesetzt wird. Hier ist bei der *Verwandlung* und, wenn auch nur leicht, beim *Erdbeben* der prozentuale Anteil bei den professionellen im Vergleich zu den Laien-Sprecher:innen höher. Die Tendenz verstärkt sich, wenn im Spektrum der Relevanzpausen zwischen jenen Pausen unterschieden wird, die mutmaßlich stärker typographisch motiviert sind, weil sie nach Absätzen eingelegt werden, und jenen, die innerhalb von Absätzen gemacht wurden. Die Auswertung der von uns dementsprechend manuell annotierten Whisper-Transkripte ergab, dass die professionellen Leser:innen in allen Texten mehr Relevanzpausen in den Absätzen setzen als die Laien-Leser:innen.

Zu bemerken ist, dass sich in Absätzen auch weitere, möglicherweise durch typographische Trigger erzeugte Pausen befinden. Neben den Anführungszeichen der direkten Rede zählt dazu z.B. ungewöhnliche Interpunktion am Satzende.

Sprechfassung	Relevanzpausen im Absatz	Relevanzpausen vor dem Absatz
Die Judenbuche professionelle Lesung	82 %	18 %
Die Judenbuche Laienlesung	78 %	22 %
Die Verwandlung professionelle Lesung	79 %	21 %
Die Verwandlung Laienlesung	40 %	60 %
Das Erdbeben in Chili professionelle Lesung	64 %	36 %
Das Erdbeben in Chili Laienlesung	43 %	57 %

Tabelle 2: Verteilung der Relevanzpausen nach Textposition

Wir gehen zugleich davon aus, dass die Textkenntnis bei den professionellen Sprecher:innen in der Regel höher ist, nicht zuletzt aufgrund der redaktionellen Betreuung, ihrer Ausbildung und der im Vorfeld der Aufnahme mutmaßlich erfolgten, eingehenden Beschäftigung mit dem Text.

6. Fazit und Ausblick

Unsere tentative Annäherung an die Segmentierungsfunktion von Sprechpausen hat ergeben, dass es lohnend erscheint, diesen Ansatz weiter zu verfolgen und durch ein größeres Audio-Korpus zu validieren. Die Exploration unserer Daten bislang deutet an, dass mittels der sprechpausenbezogenen Segmentierung Textprofile erstellt werden können und eine größere Sprecher:innenkompetenz vermutlich zu mehr nicht primär typographisch getriggerten Sprechpausen führt.

Die Erstellung eines größeren Audio-Korpus für weitere Untersuchungen erscheint dabei nicht zuletzt aufgrund der spezifischen Rhetorizität von Vortragenden geboten, ihrer idiosynkratischen Realisierung der Sprechfassung eines Textes. Um die damit verbundenen Parameter der Pausensetzung und -längen weitestmöglich zu neutralisieren, wäre die Untersuchung von deutlich mehr Sprechfassungen nur eines Textes notwendig. In einem nächsten Schritt werden wir daher acht zusätzliche Audioaufnahmen von Laien-Lesungen des *Erdbeben*-Textes der gleichen Analyse unterziehen, auch um die bisherigen Ergebnisse einer Prüfung zu unterziehen. Auf diese Weise wollen wir ebenso eruieren, welche Textcharakteristika für die spezifische Pausenverteilung in den Sprechfassungen maßgeblich sind. Sollten sich in der Analyse eines größeren Audio-Korpus Übereinstimmungen diesbezüglich zeigen, wäre eine weitere Auswertung dieser Segmentgrenzen interessant. Zudem könnten die Daten Grundlagen für die Entwicklung eines neuen Ansatzes zur automatischen Textsegmentierung sein. Wenn die Segmente aus erzähltheoretischer Sicht stimmige bzw. konsistente Einheiten ergeben, wäre ihre automatische Erkennung eine geeignete Grundlage für darauf aufbauende narrative Analysen wie etwa von Figuren oder Handlung.

Bibliographie

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, und Michael Auli. 2020. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

Bain, Max, Jaesung Huh, Tengda Han und Andrew Zisserman. 2023. "Whisperx: Time-accurate speech transcription of long-form audio." In *Proc. INTERSPEECH 2023*. 4489-4493.

Bartsch, Sabine, Evelyn Gius, Marcus Müller, Andrea Rapp und Thomas Weitin. 2023. "Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft." *Zeitschrift für digitale Geisteswissenschaften* 10.17175/2023_003.

Brand, Svenja. 2021. " Textsprechen als Interpretation. Gedanken zur ästhetischen Kommunikation in der literaturwissenschaftlichen Lehre." In PhiN. Philologie im Netz, Beiheft 27/2021, 66-76.

Geißner, Hellmut. 1981. "Sprechwissenschaft. Theorie der mündlichen Kommunikation." Königstein: Scriptor.

Gius, Evelyn, Carla Sökefeld, Lea Dümpelmann, Lucas Kaufmann, Annekea Schreiber, Svenja Guhr, Nathalie Wiedmer und Fotis Jannidis. 2021. "Guidelines for Detection of Scenes (1.0)." Zenodo. https://doi.org/10.5281/zenodo.4457177 (zugegriffen: 19.Juli 2024) .

Herzog, Rainer. 2018. "Ein generischer Ansatz zur digitalen Layoutanalyse von Manuskripten". https://ediss.sub.uni-hamburg.de/handle/ediss/6036 (zugegriffen: 19.Juli 2024).

Kisler, Thomas, Uwe Reichel, Florian Schiel. 2017. "Multilingual processing of speech via web services." Computer Speech & Language. 47:326-347

Klos, Thomas und Heiner Ellgring. 1987. " Manuelle versus elektronische Analyse von Sprechpausen" In Zeitschrift für experimentelle und angewandte Psychologie XXXIV, 64-71.

Leibniz-Institut für Deutsche Sprache. 2013. "Propädeutische Grammatik". Grammatisches Informationssystem grammis. https://doi.org/10.14618/programm (zugegriffen: 19.Juli 2024).

Meyer-Kalkus, Reinhart. 2020. "Geschichte der literarischen Vortragskunst." Berlin: J. B. Metzler. https://doi.org/10.1007/978-3-476-04802-8 (zugegriffen: 19.Juli 2024).

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey und Ilya Sutskever. 2023. "Robust speech recognition via large-scale weak supervision." In *International conference on machine learning* , 28492-28518.

Trouvain, Jürgen und Bernd Möbius. 2018. "Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache". Proc. 29th Conf. Elektronische Sprachsignalverarbeitung (ESSV '18) . 334-341.

Vauth, Michael, Hans Ole Hatzel, Evelyn Gius und Chris Biemann. 2021. "Automated Event Annotation in Literary Texts". In: CHR 2021: Computational Humanities Research Conference, 333-345. Amsterdam. http://ceur-ws.org/Vol-2989/short_paper18.pdf