Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Pfaffenwaldring 5b
D–70569 Stuttgart

Bachelorarbeit

Zusammenhang von Diskursstrukturen und Prosodie

Florin Rheinwald

Studiengang: Medieninformatik

Prüfer/in: Prof. Dr. Ngoc Thang Vu

Betreuer/in: Dr. Antje Schweitzer

Beginn am: 21. Januar 2019

Beendet am: 26. August 2019

Kurzfassung

Die Arbeit untersucht einen Zusammenhang von Diskursstrukturen von Texten und der Prosodie von SprecherInnen. Auf Basis eines Korpus bestehend aus Hörbüchern soll mit Hilfe der dazugehörigen Texte evaluiert werden, inwiefern eine Abbildung der Tonhöhe der SprecherInnen auf die unterschiedlichen inhaltlichen Diskursstrukturen möglich ist.

Diese Einteilung folgt der Extraktion der individuellen Grundfrequenz jedes Sprechers mittels der Software *Festival* und den daraus ermittelten Tonhöhen-Bereichen (*Pitch Range*) der jeweiligen Phrasen, welche Diskurssegmente enthalten. Diese Segmente werden automatisch durch zwei verschiedene Diskursparser - basierend auf der *Rhetorical Structure Theory* und der *Penn Discourse Tree Bank* - erkannt und erzeugt.

Mittels einer statistischen Analyse werden Beobachtungen zur signifikanten Unterscheidung der einzelnen Diskursstrukturen gemacht. Dabei wird auch die Klassifizierung der beiden Parser analysiert und untereinander verglichen. Aus den daraus erzielten Ergebnissen soll eine These zum allgemeinen Zusammenhang von bestimmten Diskursstrukturen und der Prosodie aufgestellt werden.

Inhaltsverzeichnis

1. Einleitung **13**

2. Realisierung **21**
2.1. Korpus 21
2.2. Alignment und Festival 23
2.3. Praat 25
2.4. Penn Discourse Tree Bank 25
2.5. Rhetorical Structure Theory 31

3. Auswertung **35**
3.1. Vorbereitung 35
3.2. Statistische Mittel 36
3.3. Durchführung 39
3.4. Ergebnisse 40
3.5. Vergleich 48
3.6. Tonhöhenregister 51

4. Zusammenfassung **53**

5. Literatur **55**

A. Anhang **59**

Abbildungsverzeichnis

1.1. Beispiel-Annotation des Tonsequenzmodells [Wennerstrom, 2001] 14
1.2. Box-Notation der Theorie der Diskursstruktur nach Asher 17
1.3. Tonhöhenregister einer Sprecherin 18

2.1. Hierarchie der semantischen Annotation in der PDTB 27
2.2. Beispiel einer RST-Hierarchie [Taboada, 2006] 31

3.1. Pitch Ranges der SprecherInnen 36
3.2. PDTB: Links wird Messung 1 gezeigt, rechts Messung 2 40
3.3. PDTB: Matrix der post-hoc t-Tests 41
3.4. PDTB: Messung 3 in Abhängigkeit zu den SprecherInnen 42
3.5. PDTB: Vergleich der Argumente von Temporal.Synchrony 43
3.6. RST: Messung 2, Standardabweichung vom Mittelwert 44
3.7. RST: Matrix der post-hoc t-Tests 45
3.8. RST: Messung 3, in Abhängigkeit zu den SprecherInnen 45
3.9. RST: Vergleich von Nukleus und Satellit 46
3.10. RST: Korrigierte und originale Pitch Range 47

Tabellenverzeichnis

2.1. Beispiel-Output von Festival (verkürzt und benannt) 24
2.2. Vergleich der Korpora 29
2.3. Relationsgruppen des RST-Parsers nach Gruppen 33

3.1. Annotation des Sprechers m5 (durchschnittlicher Pitch: 131,8 Hz) 47
3.2. Übereinstimmungen der Relationen 49
3.3. Vergleich der Pitch Ranges 50

Verzeichnis der Listings

2.1. Texte im Korpus 22
2.2. PDTB Pipe Format 28
2.3. Classifier-Ausgabe 28
2.4. Beispiel-Ausgabe des Diskursbaums 33

A.1. Angleichung der Parser-Tokenisierung an die Aligner-Tokenisierung . . 59
A.2. Extrahieren aller Diskursrelationen 59
A.3. Basic-Skript zur Identifikation aller zugeordneten Token 60
A.4. Main-Methode des Java-Codes, der die Annotation des RST-Parsers für alle Textdateien erstellt. 61
A.5. Erstellen einer Offset-Tabelle 62
A.6. Workflow zur Tokenisierung, Teil 1 62
A.7. Workflow zur Tokenisierung, Teil 2 63
A.8. Reduktion des RST-Baums 64

1. Einleitung

Die phonetische Forschung beschäftigt sich seit längerer Zeit damit, prosodische Merkmale in Diskursstrukturen zu suchen und die identifizierten Muster zu kategorisieren. Dadurch lassen sich Äußerungen, die eine bestimmte semantische Richtung vorgeben, mit einer entsprechenden Satz- bzw. Wortmelodie des Sprechers verknüpfen. Die daraus gezogenen Schlüsse sind vor allem für die Bereiche Spracherkennung und -synthese von Bedeutung, da eine synthetisierte Stimme zumeist viele prosodische Eigenschaften eines menschlichen Sprechers vermissen lässt. Besonders in längeren Texten oder Dialogen zeigt sich, dass zusammenhängende Sinneinheiten weniger über die Prosodie transportiert werden, weil die Intonation der künstlichen Stimme nur wenig variiert. Zum besseren Verständnis der hier zentralen Begrifflichkeiten werden diese im folgenden Abschnitt erläutert.

Prosodie

Prosodie spielt eine zentrale Rolle in der Sprachwissenschaft und vor allem in der Phonetik. Sie ist allgegenwärtig in der zwischenmenschlichen Kommunikation sowohl in gelesener, Spontan- als auch geschriebener Sprache. Dies lässt sich auf alle Ebenen ausweiten – von der semantischen Differenzierung durch Tonhöhen auf Silbenbasis, wie in tonalen Sprachen, über die Intonation einzelner Worte bis hin zur Akzentuierung ganzer Sätze oder Textabschnitte. Intonation ist hauptverantwortlich dafür, dass Äußerungen syntaktisch und semantisch unterschieden bzw. als Sinneinheit erkannt werden können. Dies lässt sich beispielsweise an sinkenden Tonhöhen am Ende von Äußerungen oder steigenden Tonhöhen bei Fragesätzen im alltäglichen Sprachgebrauch beobachten.

Ein gängiges Modell, welches dieses Verhalten beschreibt, ist das Tonsequenzmodell [Pierrehumbert, 1980]. Es stellt die Tonhöhenverläufe innerhalb einer Phrase auf den jeweiligen akzentuierten Silben der Worte als lokales Minimum (L-Targets für tiefe Töne) oder lokales Maximum (H-Targets für hohe Töne) dar.

In Abbildung 1.1 wird beispielhaft dargestellt, wie die Tonhöhe innerhalb einer Phrase variiert. Dabei ist anzumerken, dass es sich bei der Phrase „They have a *Western* restaurant?“ um einen Fragesatz handelt, der Überraschung ausdrücken soll. Die Annotation durch L* zeigt an, dass die Tonhöhe ein lokales Minimum beschreibt, die

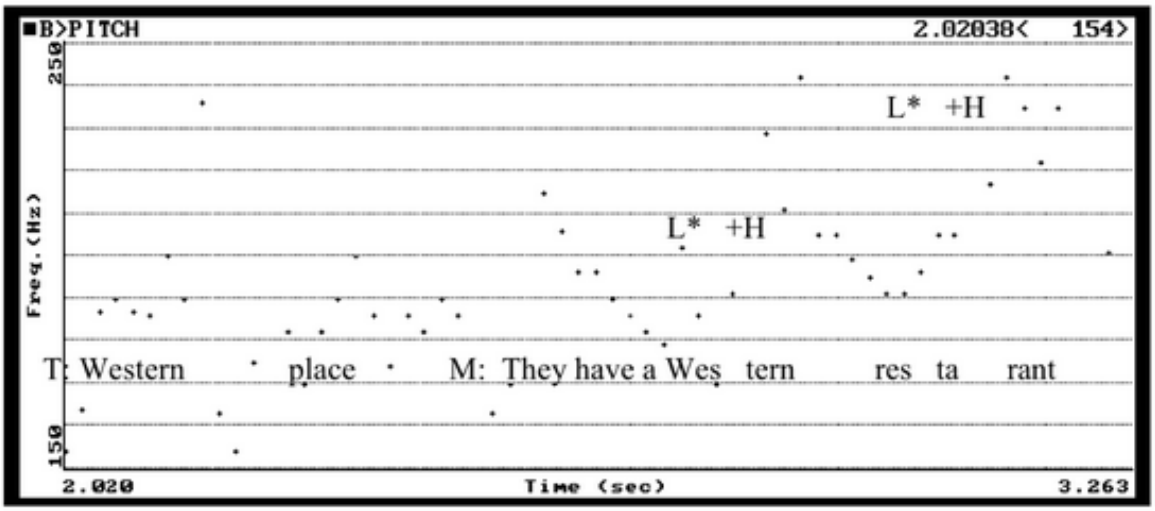

Abbildung 1.1.: Beispiel-Annotation des Tonsequenzmodells [Wennerstrom, 2001]

Annotation durch +H, dass eine Steigung mit einer gewissen Verzögerung folgt. Die Punkte im Spektrum stellen die im Intervall 2020-3263 ms gemessenen F_0-Werte dar.

Hier muss eine Unterscheidung zwischen der Verwendung des Begriffs *Pitch* und F_0 vorgenommen werden: Unter F_0 wird die Grundfrequenz verstanden, welche die Frequenz beschreibt, mit der die Stimmbänder bei der Lautproduktion vibrieren. Der Begriff Pitch dagegen bezieht sich auf die subjektive Art und Weise, wie die Grundfrequenz wahrgenommen wird [Li, 2009]. Die Grundfrequenz F_0 wird wie folgt berechnet, wobei T_0 die Periodendauer beschreibt:

$$F_0 = \left(\frac{1}{T_0}\right)$$

Die Abbildung zeigt ein Frequenzfenster von 150 Hz bis 250 Hz. In der untersuchten Phrase befindet sich das niedrigere Minimum auf einer Frequenz von ca. 170 Hz (unter dem „s“ in „Western“). Die beiden Maxima liegen jeweils auf einer Frequenz von ca. 240 Hz. Dementsprechend ist der absolute Tonhöhenbereich (die *Pitch Range*) der gesamten Phrase zwischen 170 Hz und 240 Hz. Dies lässt die Vermutung zu, dass hier eine Aufzeichnung einer weiblichen Sprecherin vorliegt. Die Grundfrequenz einer weiblichen Erwachsenen liegt in der Regel zwischen 165 Hz und 255 Hz, während die Grundfrequenz männlicher Erwachsener 85 Hz bis 180 Hz beträgt [Baken, 1987].

Das Verhältnis der beiden Begrifflichkeiten wird des Öfteren gleichgesetzt – zum Beispiel bezieht sich die Software *Praat* ebenfalls ausschließlich auf Pitch – dies wird der Einfachheit halber im Rest der Arbeit ebenfalls getan. Ist demnach von Pitch-Werten die Rede, sind damit die gemessenen F_0-Werte gemeint.

Prosodie dient außerdem dazu, Konstituenten größerer Themenfelder in der organisatorischen Struktur von Diskurs zu verbinden [Brown, 1977; Yule, 1980]. Auf einem lokalen Level zeigt Prosodie an den Grenzen zu Beginn und Ende einer Äußerung deren Abhängigkeitsverhältnis [Pierrehumbert & Hirschberg, 1990], welches relevant für das System des Sprecherwechselns (*turn-taking*) ist [Wennerstrom, 2001]. Diese Unterscheidung wird im weiteren Verlauf der Arbeit präzisiert. Speziell auf dem Bezug der Prosodie zu den organisatorischen Strukturen von Diskurs soll im Folgenden jedoch der Fokus liegen.

Diskurs

Der Begriff *Diskurs* lässt sich sehr weit fassen; er geht in vielen Teilen über die für die Arbeit relevanten linguistischen Grenzen hinaus. Unter diesem Begriff lassen sich die Beschäftigung der Öffentlichkeit mit einem bestimmten Thema, eine Diskussion unter Personen oder auch ein geschriebener Aufsatz zusammenfassen; es ist demnach nicht vorgeschrieben, wie viele Parteien an einem Diskurs beteiligt sind.

Er wird zudem sehr oft in philosophischen Kontexten verwendet, für die Definition in diesem Fachbereich zeichnen sich vor allem Michel Foucault, Jürgen Habermas oder Jean-Francois Lyotard aus. Nach Foucault bezeichnet Diskurs keinen Sprechakt, sondern vielmehr eine Verkettung von Aussagen oder eine Formation. Diskurs gebe einer solchen Verkettung eine sinngebende Struktur; er fragt: „[...] wie kommt es, daß eine bestimmte Aussage erschienen ist und keine andere an ihrer Stelle?“ [Foucault, 1973]. Auf Basis von Foucaults Diskursbegriff wird in der linguistischen Diskursanalyse untersucht, welche Zusammenhänge in Texten vorherrschen, welche Hintergründe für deren Aussagen gelten und inwiefern die genutzte Sprache dabei eine Rolle spielt [Bluhm et al., 2000].

Die Kohärenz von Texten ist relevant für deren inhaltliches Verständnis. Zum Beispiel kann ein Text oder eine Abfolge von Aussagen Köhasion beinhalten, also syntaktisch zusammenhängen, jedoch nicht kohärent sein[1]:

a. Peter steht um 8 Uhr auf. Er isst gerne Schokolade.

Auf dieser Basis steht bei der linguistischen Analyse von *Diskursstrukturen* das Einteilen solcher Texte in Diskurssegmente, welche in einer bestimmten semantischen und hierarchischen Relation zueinander stehen, im Vordergrund. Diskursstrukturen bezeichnen das allgemeine Auftreten von Diskurs in der Sprache, während Diskurssegmente einen

[1] Nach https://www.ling.uni-potsdam.de/s̃cheffler/teaching/2015pragmatics/pragmatics2015-09-discstructure.pdf

Teil der Sprache ausmachen, der nicht zwingend mit einem Textabschnitt oder einem Satz gleichgesetzt werden muss, und das Grundgerüst repräsentieren, auf dem Diskurs aufbaut [Gustafson-Čapková, 2005]. Jedes Diskurssegment verfügt dabei über eine Aussage (*Discourse Purpose (DP)*) [Grosz & Sidner, 1986].

Bei Diskurssegmenten lässt sich beobachten, dass einzelne Äußerungen (hier Segmente) immer einen unterschiedlichen Bezugsrahmen zueinander aufweisen können [nach Webber, 1991]:

a. Als seinen Anteil am gemeinsamen Projekt baute John einen zweiarmigen Roboter.

b. Er hatte am Institut die Robotik studiert.

c. Als ihren Anteil brachte Mary dem Roboter bei, Saxophon zu spielen.

Es lässt sich zu jedem einzelnen Segment a, b, c ein weiteres kohärentes Segment finden, das in einer semantischen Relation zu diesem steht. Zusätzlich lassen sich auch zu den kombinierten Segmenten a-b und a-c weitere Segmente finden:

d. Das brachte ihnen beiden die Note 1 ein.

Es lässt sich allerdings kein weiteres kohärentes Segment für die Kombination b-c finden, da diese ohne den Kontext zu a in keinem semantischen Zusammenhang stehen. Es muss dementsprechend immer ein Bezugssegment sowie ein anschließendes Segment bestehen, damit eine Diskursrelation vorliegt; dabei ist die Größe des Bezugssegments nicht von Bedeutung. Im Folgenden wird auf das Bezugssegment mit *Nukleus*, dem Kern der Relation, und auf alle anderen Segmente mit *Argument* oder *Satellit* referiert. Um welche Relationen es sich im Speziellen handeln kann, wird bei den jeweiligen Umsetzungen der Diskurs-Annotationen im zweiten Kapitel behandelt.

Verwandte Arbeiten

In der Fachliteratur gibt es einige Ansätze, den Zusammenhang von Diskursstrukturen und Prosodie zu untersuchen. Als Motivation für diese Arbeit soll derjenige von Mayer [Mayer, 1999; Möhler & Mayer, 2001] dienen. Er untersucht die Relation zwischen globalen Diskursstrukturen und prosodischen Merkmalen - spezifisch Pausendauer und Pitch Range-Variation - mit dem Ziel, „einer formal-linguistischen Definition der Prosodie-Semantik-Schnittstelle [...] näher zu kommen“ [Mayer, 1999]. Mayer bezieht sich auf die Theorie der Diskursstruktur von Asher [Asher, 1993], welche auf der Diskursrepräsentationstheorie (DRT) von Kamp [Kamp, 1981; Kamp & Reyle, 1993] fußt.

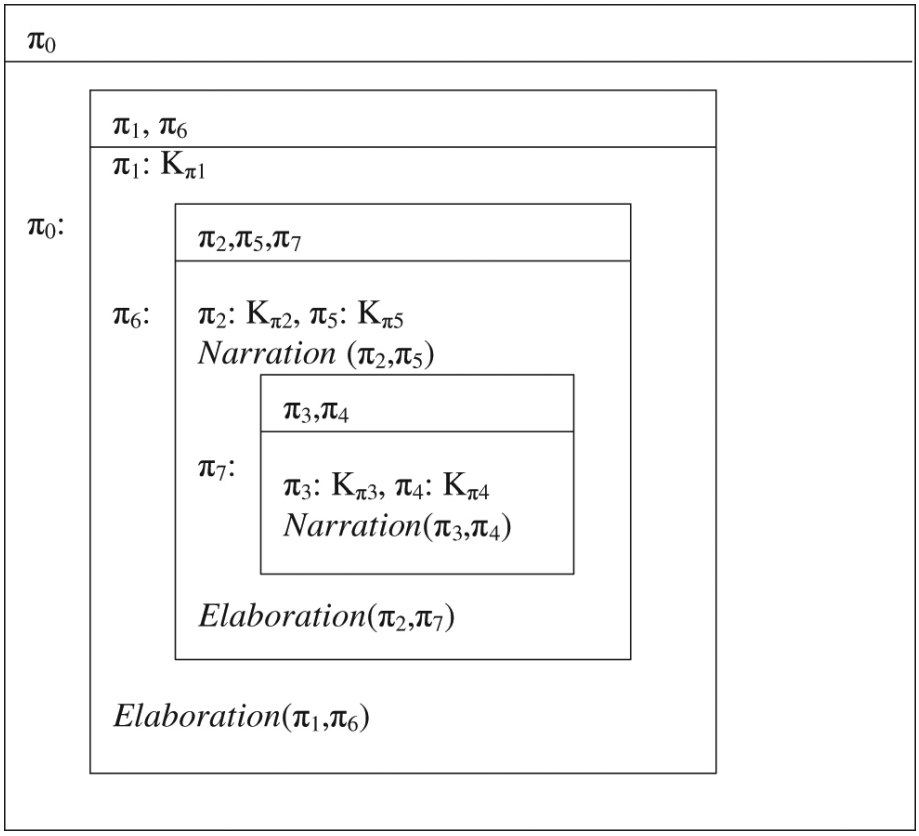

Abbildung 1.2.: Box-Notation der Theorie der Diskursstruktur nach Asher

Ashers Theorie der Diskursstruktur nutzt eine Box-Notation, welche bereits in der DRT eingeführt wurde. Diese lässt einen Überblick über alle hierarchischen Beziehungen zwischen Diskurssegmenten (bei Mayer wird von SDRSen, also segmentierten Diskursrepräsentationsstrukturen gesprochen) im Text zu. Zu einer SDRS gehört ebenfalls eine Erläuterung, welche Art von Diskursrelation zwischen den Konstituenten besteht [Mayer, 1999].

In Abbildung 1.2 ist die Box-Notation beispielhaft gezeigt, hier stellt jede Box eine SDRS mit einer dazugehörigen Menge an Bedingungen für die Diskursreferenten dar. In diesem Fall werden zwei koordinierende *Narration*-Relationen von zwei subordinierenden *Elaboration*-Relationen verbunden [Zulaica, 2018]. Die Bezeichnungen bzw. Klassifikationen dieser Relationen unterliegen verschiedenen Modellen der Diskurs-Annotation, zwei dieser Modelle werden im nächsten Kapitel vorgestellt.

Um die Zusammenhänge der vorliegenden Theorie und Prosodie zu untersuchen, führt Mayer zunächst auch die Pausendauer als prosodisches Element der Diskursstrukturen an. Da diese aber in dieser Arbeit eine untergeordnete Rolle spielt, wird darauf nicht näher eingegangen. Anschließend widmet er sich dem zum damaligen Zeitpunkt wenig untersuchten Punkt der Pitch Range des Sprechers als Tonhöhenregister.

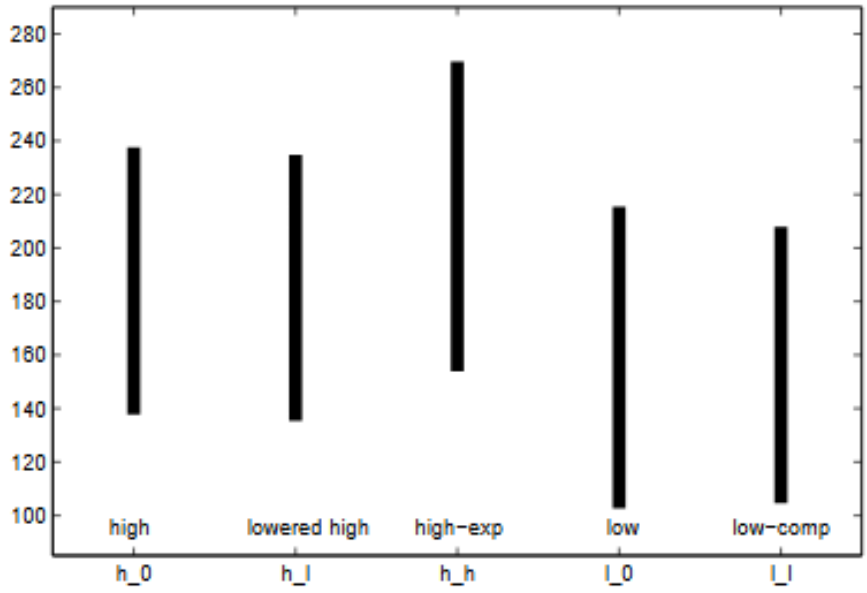

Abbildung 1.3.: Tonhöhenregister einer Sprecherin

Für deren Analyse wird für jede Phrase einer SDRS-Konstituente die Pitch Range aufgestellt. Das hierarchische, autosegmentale Modell von Clements [Clements, 1981] erweiternd, werden zunächst zwei Primärregister – {h} und {l} – aufgestellt, welche von zwei gleichnamigen Subregistern unterteilt werden. Dabei steht {h} für relativ hohe und {l} dementsprechend für relativ niedrige Tonhöhen. Als resultierende Register existieren {h,h}, {h,l}, {l,h} und {l,l}. Die Unterscheidung zum originalen Modell ist die Einordnung der oben genannten Pitch Ranges anstatt der absoluten Tonhöhen (*Pitch Levels*); gerade aus diesem Grund seien Überlappungen der Register jedoch auch nicht auszuschließen. Mayer weist zuvor eigens aufgestellte, individuelle Registerklassen diesem erweiterten Modell zu; dabei lässt er unter anderem auch für Phrasen mit einer expandierten Pitch Range die Zuweisung beider Primärregister gleichzeitig zu. Eine beispielhafte Zuweisung in Register einer weiblichen Sprecherin ist in Abbildung 1.3 zu sehen [Möhler & Mayer, 2001].

Analog zu vorangehenden Arbeiten, welche Pitch-Verläufe von Phrasen in Abhängigkeit zu der Position in Diskursstrukturen untersuchten [Hirschberg & Pierrehumbert, 1990], werden SDRS-finale Phrasen dem l-Register zugeordnet, nicht-finale dem h-Register. Mayer stellt darauf basierend folgende (hier vereinfachte) Hypothesen auf:

1. Finale Phrasen hierarchisch höherer SDRS-Konstituenten fallen in das Register {l,l}

2. Beginnt ein neues Topik oder wird ein altes wieder aufgegriffen, so wird allen Phrasen der SDRS-Konstituente das Subregister {l} zugeteilt

3. Die erste Phrase eines Sub-Topiks fällt in das Register {h,h}

Obwohl er die in seiner Untersuchung aufgestellten SDRS-Konstituenten nicht mit einer bestimmten Diskursrelation betitelt, führt Mayer die hierarchische Charakteristik des Diskursmodells von Asher als maßgeblich an, um deren Struktur zu erkennen. Die Hierarchie in 3. ist beispielsweise mit der einer ELABORATION-Relation zu vergleichen, die in 2. mit der der BACKGROUND-Relation.

Aufgabenstellung

Ziel dieser Arbeit ist es, die bisher erläuterten Überlegungen zu Zusammenhängen von Diskurssegmenten und der Prosodie - insbesondere in puncto *Pitch Range* der Sprecher pro Phrase – auf bereits existierende, unterschiedliche Diskursklassifizierungen anzuwenden. Diese werden zum einen durch die *Rhetorical Structure Theory* und zum anderen die *Penn Discourse Tree Bank* bereitgestellt, auf welche im weiteren Verlauf eingegangen wird. Da die Fachliteratur zu diesem Thema zum Großteil ein gewisses Alter aufweist, sind diese inzwischen mitunter feineren Klassifizierungen bisher kaum Gegenstand von Untersuchungen gewesen, welche einen Zusammenhang zur Prosodie des Sprechers suchen. Mayer zum Beispiel untersucht ausschließlich die Zusammenhänge zu Topik-Strukturen. Der Fokus auf diese speziellen Konzepte verfolgt die Intention, eben diese bestehenden Unterteilungen von Diskursstrukturen auf mehreren Ebenen zu beleuchten und anschließend statistisch voneinander abzugrenzen. Zudem bietet sich ein Vergleich dieser Konzepte an.

2. Realisierung

2.1. Korpus

Für die Durchführung der geplanten Untersuchungen musste demnach ein Korpus bestehend sowohl aus Texten als auch dazugehörigen Sprechern, welche diese vortragen, zusammengestellt werden. Dazu wurden Hörbücher aus der digitalen Bibliothek LibriVox[1] genutzt, welche dort lizenzfrei zur Verfügung stehen. Die dazugehörigen Texte stammen aus der Bibliothek Project Gutenberg[2]. Eine Einschränkung stellte dabei die verwendete Sprache dar; der Aligner, welcher darauf benutzt wird, verfügt zwar über die Erkennung anderssprachiger Texte, jedoch sind beide verwendeten Diskursparser ausschließlich auf die englische Sprache anwendbar.

Entscheidend für die Auswahl war dabei die Textgattung; sie musste für die Suche nach möglichst vielen diversen Diskurstypen geeignet sein. Auszuschließen waren dementsprechend Texte aus Dramatik - hier würden Regieanweisungen und fehlende Prosa zum Problem werden – und Lyrik, da diese Texte zumeist in einer nicht für die Erkennung von Parsern geeigneten Sprache geschrieben sind.

Dies stellte auch für die verwendeten Prosa-Texte eine signifikante Einschränkung dar, da die lizenzfreien Bücher alle zu Beginn des vergangenen Jahrhunderts entstanden sind. Daher mussten alle Bücher auf ihre Tauglichkeit, von einem Diskursparser mit ausreichenden Ergebnissen annotiert werden zu können, untersucht werden, bevor sie in das Korpus aufgenommen wurden.

Letztlich wurden sieben verschiedene Texte mit insgesamt zehn verschiedenen Sprecherinnen und Sprechern ausgewählt, fünf davon männlich, fünf weiblich. Von allen Sprecherinnen und Sprechern waren 15 Minuten Material in Text und vorgelesener Sprache vorhanden.

Bei der Auswahl der spezifischen Textstellen innerhalb der Texte galt es zu beachten, dass deren Inhalt nicht zu sehr aus Dialogen verschiedener Charaktere der Handlung

[1] librivox.org
[2] gutenberg.org

Listing 2.1 Texte im Korpus

```
-     Ethel Lilian Voynich: The Gadfly (1897)
-     Fergus Hume: The Crimson Cryptogram (1900)
-     Warwick Deeping: Mad Barbara (1908)
-     Booth Tarkington: The Flirt (1913)
-     Abraham Cahan: The Rise of David Levinsky (1917)
-     Charles G. D. Roberts: Jim: The Story of a Backwoods Police Dog (1919)
-     Milo Hastings: City of Endless Night (1920)
```

besteht. Das heißt, die Anzahl der durch Anführungszeichen gekennzeichneten wörtlichen Rede pro Kapitel wurde als Filterkriterium herangezogen. Entsprechend wurden Textstellen mit tendenziell weniger wörtlicher Rede bei der Auswahl priorisiert.

Bezogen auf die Definition von Diskurs muss zwischen Diskurs mit und Diskurs ohne Sprecherwechsel unterschieden werden. Diskurs mit turn-taking bezeichnet eine Konversation zwischen mehreren Gesprächsteilnehmern, bei denen der Inhalt des Textes auf die jeweiligen Teilnehmer aufgeteilt wird. Diskurs ohne Sprecherwechsel lässt sich in Bezug auf die verwendeten Texte als eine Narration jeglicher Perspektive zusammenfassen.

Ein Fokus auf letzteren lässt sich dadurch begründen, dass sich bisherige Studien ausschließlich auf Diskurs ohne Sprecherwechsel beziehen und zudem ungeklärt ist, inwiefern die Annotationen der beiden Parser dementsprechend beeinflusst werden. Dies wird bei den jeweiligen Ausführungen zu den Parsern näher betrachtet.

Ein weiterer Faktor ist die Adaption der Sprechstimme verschiedener Charaktere durch die SprecherInnen. Dies kann einen großen Einfluss auf die durchschnittlichen Pitch Ranges haben, da je mehr Sprecherwechsel im vorgelesenen Text vorkommen, desto mehr die Stimmlage verstellt bzw. den jeweiligen Charakteren angepasst wird. Es ist anzunehmen, dass männliche Sprecher für Abweichungen von der standardmäßigen Pitch Range „anfälliger“ sind, sollten sie weibliche Stimmen imitieren, ein Beispiel dafür wird im Kapitel zur Auswertung gezeigt.

Es seien außerdem einige Bemerkungen zu den Sprechern festgehalten: Sprecher m5 spricht mit einem slawischen Akzent, Sprecherin w2 mit einem indischen Akzent. Bei m1 und w4 liegt ein britisches Englisch vor, alle restlichen SprecherInnen sprechen mit amerikanischem Dialekt. Dies lässt zunächst auf keine Probleme bei der Auswertung bezüglich des Pitch schließen, jedoch waren Fehler bei der automatischen Annotation zu befürchten. Dies wird im folgenden Teil eruiert. Bei Sprecherinnen w3 und w5 fällt auf, dass beide im Vergleich zu anderen weiblichen Sprecherinnen mit einer sehr tiefen - und w3 betreffend - monotonen Stimme sprechen; ein Fakt, der bei der späteren Auswertung berücksichtigt werden muss.

2.2. Alignment und Festival

Um die vorliegenden Korpus-Daten aus Text- und Audiodateien aneinander anzugleichen (*Alignment*), wurde der Aligner des Instituts für Maschinelle Sprachverarbeitung der Universität Stuttgart verwendet, es wurde also ein *forced alignment* vorgenommen. Anschließend wurden diese Daten in die Sprachsynthese-Software *Festival* konvertiert. Damit der Aligner das automatische Alignment korrekt vornehmen kann, mussten alle Audiodateien in circa fünfminütige Teile geschnitten werden; die dazugehörigen Textstellen dem Inhalt dieser Teile entsprechend ebenfalls. Der Output des Aligners besteht aus einer Vielzahl von Dateien, deren Inhalt wird in den Relationen weiter unten aufgelistet. Diese Dateien werden in die Äußerungsstruktur von *Festival* übernommen [Schweitzer, 2018].

Festival wurde vom Center for Speech Technology Research der University of Edinburgh entwickelt und ist in C++ geschrieben. Die Software wurde primär zur Synthese von Sprache auf Basis von Textdateien entwickelt, für die Bedürfnisse dieser Arbeit wurden jedoch nur die Funktionen, Datenbanken mit Kennzeichnung (*Labeln*) zu speichern und daraus relevante Features zu extrahieren, benötigt.

Anschließend werden dann bestimmte Relationen pro Teil und Zeitpunkt gelabelt, in diesem Fall wird auf folgende Relationen zurückgegriffen:

- Phrase (insbesondere die Phrasengrenze)

- Segment (bezeichnet einzelne Phone)

- Syllable (Silben)

- Target (Pitch-Werte; hier von jeder Hundertstel-Sekunde)

- Token (Worte inklusive der Interpunktion)

- Word

Aus diesen erstellten Datensätzen werden Äußerungen (*Utterances*) kreiert, die all diese Relationen und zusätzlich Relationen als Zeiger zueinander beinhalten, sodass die jeweiligen Werte zusammengeführt werden können. Mittels dieser Äußerungen lassen sich dann individuelle Features aus dem Korpus extrahieren, welche in eine Tabelle geschrieben werden. Zu den Features gehören alle oben aufgeführten Relationen, die auf Basis der Syllable-Relation zusammengeführt wurden, sowie zusätzliche Werte, die die spätere Zusammenführung mit den Diskursrelationen und die Auswertung erleichtern sollen. Dazu zählen zum Beispiel die Anzahl der Silben pro Wort, die Nummer des Wortes im jeweiligen Teil oder der Name der Datei, aus der die Daten stammen. [3]

[3] http://www.cstr.ed.ac.uk/projects/festival/

Dabei ist die Feature-Extraktion des Pitch besonders zu erwähnen, sie erfolgt mittels des Befehls *Syllable.syl_midpitch*. Sie gibt den Pitch-Wert am mittleren Vokal der Silbe an und keinen Durchschnittswert aller Phone der Silbe. Dies könnte anderenfalls zu Problemen führen, beispielsweise in dem Fall, dass Pausen in diesen enthalten sind; deren Pitch wäre entsprechend niedriger.

Der Output des automatischen Alignments enthielt mitunter einige Ungenauigkeiten; einige davon lassen sich durch die eingangs angesprochenen Charakteristika der Sprecher, andere durch spezielle Strukturen im Text erklären. Tabelle 2.1 zeigt einen Beispielausschnitt des tabellarischen Outputs von Festival, bei dem einige Fehler des Alignment enthalten sind. Die phonetische Transkription der Silben basiert auf dem ARPABET, einer für Sprachsynthesizer gängigen Transkription [Klatau, 2001].

Silbe	Wort	Part	Pitch	End	Begin	Silben/Wort	Nächste Silbe
rae	sluggish	m1-3.part45	0	1.1	0.88999999	4	dh
dher	sluggish	m1-3.part45	86	1.1900001	1.1	4	s
slah	sluggish	m1-3.part45	56	1.64	1.1900001	4	g
gihsh	sluggish	m1-3.part45	86	1.9	1.64	4	f

Tabelle 2.1.: Beispiel-Output von Festival (verkürzt und benannt)

In den ersten beiden Zeilen ist zu erkennen, dass die Silben „rae“ und „dher“ fälschlicherweise dem Wort „sluggish“ zugeordnet wurden, anstatt dem eigentlichen Wort im Text „rather“. Dementsprechend werden dem Wort „sluggish“ auch vier Silben zugeschrieben. Zusätzlich fehlt der Pitch-Wert der ersten Silbe und der Wert der dritten Silbe fällt ungewöhnlich stark ab. Gerade solche inkonsistenten Pitch-Werte innerhalb eines Wortes lassen sich in erster Linie durch die Stimme des Sprechers erklären. Die Stimme des im Beispiel gezeigten Sprechers m1 ist vergleichsweise kratzig; ihre Erkennung fällt infolgedessen schwerer.

Weitere falsche Zuordnungen zum nachfolgenden bzw. vorherigen Wort kommen demzufolge primär bei Sprecher m1 vor. Einige davon sind allerdings auch der Transkription geschuldet; ein Beispiel dafür ist „Pius IX.“, das als „Pius the ninth“ gelesen nicht korrekt zugeordnet werden kann. Die im Voraus getroffene Annahme, dass der Akzent bei den Sprechern für Probleme beim Alignment sorgen könnte, wurde jedoch nicht bestätigt. Keiner der Datensätze der Sprecher m5 oder w2 beinhaltet übermäßig viele Fehlzuordnungen.

Ein größeres Problem stellen hingegen die bei manchen Texten fehlenden Daten an deren Ende dar; meist sind entsprechend für den letzten Satz eines Textes keine Pitch-Werte vorhanden. All diese Fehler blieben bis zur Zusammenführung der Feature-Tabelle und der ersten Diskurs-Annotation unentdeckt und wurden danach manuell berichtigt.

2.3. Praat

Die ursprüngliche Herangehensweise für das Extrahieren der Pitch Range sah die Nutzung eines *Praat*-Skriptes vor. Jedoch wurde dies durch die Daten, die durch Festival in tabellarischer Form gegeben waren, mehr oder weniger obsolet. Der Fokus auf silbenbasierte Pitch-Werte, die wie oben erwähnt, auf den mittleren Vokal der Silbe zugeschnitten sind, ist für die Analyse auf größeren Ebenen ausreichend. Daher wird hier nur oberflächlich auf die Nutzung von *Praat* eingegangen.

Praat ist ein von Paul Boersma und David Weenink an der Universität Amsterdam entwickeltes Programm, mit dessen Hilfe phonetische Analysen natürlicher Sprache durchgeführt werden können. Vor allem dessen Visualisierungen von Wellenform und Spektrogramm des Sprachsignals stellten für die Arbeit eine hilfreiche Unterstützung dar, genauso wie die Frequenz-Analyse individuell angepasster Zeiträume im Sprachsignal.

Damit die Sprachsignale der Sprecher gelabelt vorliegen konnten, wurden aus dem Output des Aligners Textgrids mit den Ebenen (*Tiers*) Phone, Silben und Worte erstellt. Dabei trat allerdings die Schwierigkeit auf, dass Pausen im Silben-Tier dennoch als Teil der jeweils davor oder danach stehenden Silbe gezählt wurden. Auch aus diesem Grund wurde auf die tiefergehende Analyse der Daten mit Praat verzichtet.

2.4. Penn Discourse Tree Bank

2.4.1. Hintergrund

Das erste Konzept, Diskursrelationen zu annotieren und zu verarbeiten, welches in dieser Arbeit behandelt wird, ist das der *Penn Discourse Tree Bank*(PDTB) [Miltsakaki et al., 2004]. Es handelt sich um ein groß angelegtes Korpus, in dem Diskurs-Relationen mitsamt ihren Argumenten annotiert vorliegen. Das Projekt wurde 2004 vorgestellt. Ihren Namen sowie die syntaktische Annotation auf Satzebene verdankt die PDTB der *Penn Tree Bank* - ebenfalls ein annotiertes Korpus, der unter anderem *Part-of-speech-Tags* beinhaltet [Marcus et al., 1993]. Vom Fokus auf diese klar vordefinierte Diskursstruktur erhoffen sich Miltsakaki et al. „ein reichhaltiges Substrat für die Entwicklung [...] praxisorientierter Algorithmen“ [Miltsakaki et al., 2004]. Dies soll ein alternatives Konzept zur zum damaligen Zeitpunkt bereits existierenden *Rhetorical Structure Theory* darstellen. Auf diese wird aus chronologischen Gründen im Ablauf der Arbeit im nächsten Teilkapitel eingegangen.

Die Erklärung des Aufbaus der PDTB und deren Parser in diesem Kapitel bezieht sich ausschließlich auf die oben genannte Version und die vier Jahre später erschienene PDTB 2.0 [Prasad et al., 2008]. Version 3.0 wurde erst im Verlauf der Arbeit im März

2019 veröffentlicht und konnte dementsprechend nicht mit einem Parser angewandt werden; daher wird diese hier nicht weiter beleuchtet, stellt aber eine erwähnenswerte Alternative für weitere Untersuchungen zu diesem Thema dar.

Nach Prasad et al. beschreibt die PDTB 2.0 - als erste ihrer Art - einen lexikalisch-basierten Ansatz zur Diskursannotation; dies geschieht auf dem Wall Street Journal-Korpus, welcher rund eine Million Worte umfasst. Damit ist unter anderem der Fokus auf ein *discourse connective* (CONN) gemeint, welches eine Diskursrelation auf zwei Argumente aufspannt. Diese CONNs können in vier verschiedenen Formen mit folgenden Beispielen auftreten:

- subordinating conjunction: Verben aus den Klassen {temporal, causal, concessive, purpose, conditional}

- coordinating conjuctions: „and“, „but“, „or“

- adverbial connectives: „however“, „therefore“

- implicit connectives: zwischen zwei in Relation stehenden Argumenten fehlendes, implizites CONN, das in der Annotation um ein explizites ergänzt wird

Die Argumentstruktur der PDTB sieht vor, dass jedes CONN mit einem ersten Argument (Arg1) und einem zweiten Argument (Arg2), das zu diesem in einer Diskursrelation steht, versehen ist. Dabei kann die Reihenfolge beliebig sein; Arg1 – CONN - Arg2 und Arg2 – CONN - Arg1 sind je nach Typ des CONN zulässig. Ebenso kann ein Argument in ein anderes eingebettet sein oder die gesamte Diskursstruktur umschließen: Arg2.1 – CONN - Arg1 - Arg2.2.

Wie schon bei der obigen Auflistung zu erkennen ist, muss nicht zwingend ein CONN im Text stehen, damit eine Relation zwischen zwei Argumenten vorliegt; Token, die impliziten Relationen angehören, machen im Korpus die zweitgrößte Anzahl aus. Relationen, denen kein implizites CONN zugeordnet werden kann, sind ebenfalls im Korpus vertreten. Dazu zählen die AltLex-Relation, bei der ein CONN Redundanz hervorrufen würde, und die EntRel-Relation, die bereits eine entitätsbasierte Kohärenz aufweist. Stehen zwei Sätze in keinerlei Zusammenhang, werden sie mit der Relation NoRel vermerkt.

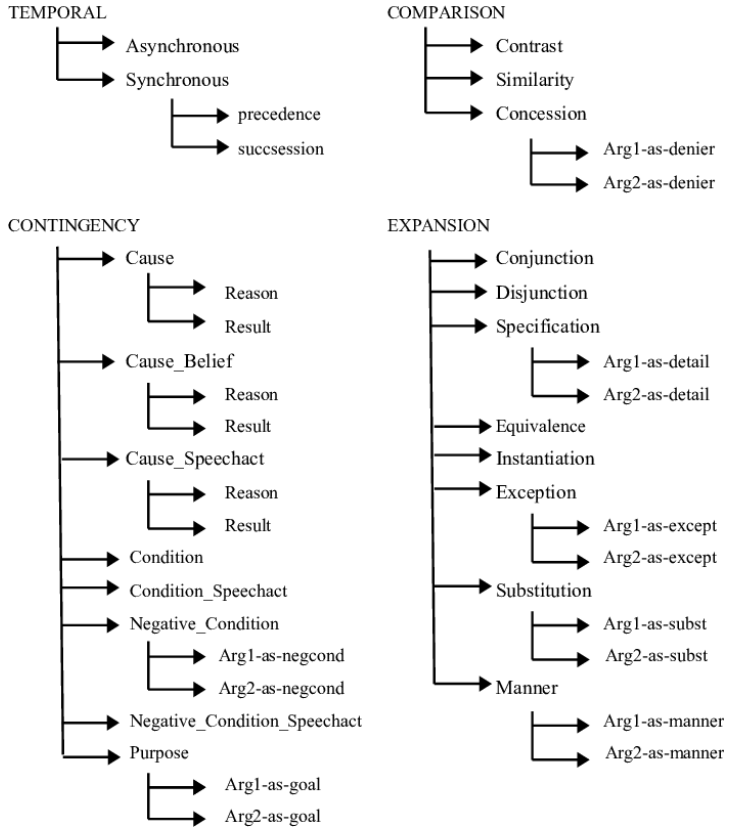

Abbildung 2.1.: Hierarchie der semantischen Annotation in der PDTB

Für die Untersuchung mit dem Parser im Vordergrund steht die semantische Annotation der CONNs und somit der gesamten Diskursrelation. Diese ist in der PDTB 2.0 in Klassen, Typen und Subtypen aufgeteilt; siehe Abbildung 2.1.

Möglich sind diese Kennzeichnungen (*Tags*) für explizite, implizite und AltLex-Relationen. Die angegebene Unterteilung enthält die vier Klassen TEMPORAL, CONTINGENCY, COMPARISON und EXPANSION. Problematisch wird diese bei bestimmten CONNs wie „since“, das mehreren Typen zugeordnet werden kann. Für solche Fälle wurde das *Inter-annotator agreement* zu Rate gezogen, das heißt die Übereinstimmung verschiedener Annotatoren bei ambigen CONNs. Die *agreement*-Rate sank proportional zum Detailgrad der Annotation, bei Subtypen war sie dementsprechend am geringsten [Prasad et al., 2008].

2.4.2. Parser

Seit der Veröffentlichung der PDTB 2.0 wurden einige Konzepte ausgearbeitet, welche die durch die vordefinierte Diskursstruktur existierende Schnittstelle für die Entwicklung eines Parsers nutzen. Die erste Veröffentlichung eines auf der PDTB basierenden

Diskursparsers wurde von Lin et al. von der National University of Singapore präsentiert [Lin et al., 2014]. Infolgedessen wurde im Rahmen der CoNNL 2015 der Shared Task zum Thema „Shallow Discourse Parsing“ ebenfalls auf Basis der PDTB organisiert. Damit wurde man dem Ansatz der PDTB gerecht, die Diskursrelationen unabhängig voneinander zu betrachten.[4]

Es wurden 16 verschiedene Parser eingereicht, die nach einem „Blind Test“, einem „Standard WSJ Test“ und der Entwicklung bewertet wurden. Als Kriterium für die Tests galt der F_1-Score für die Annotation aller Relationen [Xue et al., 2015]. Der Parser mit den besten Ergebnissen (F_1-Scores von 0.46, 0.49 und 0.57 [Xue et al., 2015]) wurde veröffentlicht und stellt eine Erweiterung des von Lin et al. entwickelten Parsers dar [Wang & Lan, 2015]. Für diese Arbeit wurde jedoch die Implementierung des Parsers von Lin et al. genutzt.[5]

Die Verwendung des Parsers stellte sich intuitiv dar; jedoch musste der Input vor dessen Parsing zunächst bearbeitet werden, da die Parserstruktur einen neue Zeile für jeden Satz im Text verlangte. Als Output liefert er eine Vielzahl an Dateien, darunter einen Syntaxbaum (.ptree), einen Dependenzbaum (.dtree), der aus dem *Stanford dependency parser*[6] gewonnen wird, und für die Annotation der Diskursrelationen relevante Ausgaben. Zu diesen zählt unter anderem die Verteilung des verwendeten Maximum Entropy Classifiers, die den Klassifikationen zugrunde liegenden Werte sind also einsehbar. Für die Darstellung der jeweiligen Annotationen nutzt die Implementierung das eigens kreierte *PDTB Pipe Format*, wie in Listing 2.2 gezeigt. Listing 2.3 zeigt die dazugehörige Klassifizierung der Diskursrelation.

Listing 2.2 PDTB Pipe Format

```
Explicit|||309..311||as||1|as|||Temporal.Synchrony||||||||||297..308|1|fitting
    him||||||||312..337|1|women's clothes fit women||||||||||||
```

Listing 2.3 Classifier-Ausgabe

```
Conjunction[0,0017] Condition[0,0165] Cause[0,3201] Concession[0,0109]
    Asynchronous[0,0135] Alternative[0,0024] Contrast[0,0141] Restatement[0,0000]
    Synchrony[0,5947] List[0,0005] Instantiation[0,0073]
    Pragmatic\_condition[0,0012] Exception[0,0003] Pragmatic\_contrast[0,0073]
    Pragmatic\_concession[0,0073] Pragmatic\_cause[0,0022] Synchrony
```

[4] https://www.cs.brandeis.edu/~clpconll15st/results.html
[5] https://github.com/WING-NUS/pdtb-parser
[6] http://nlp.stanford.edu/software/lex-parser.shtml

Der erste Eintrag zeigt an, dass es sich um eine explizite Diskursrelation handelt, der Bezug auf das CONN ist als Offset, als Volltext und als Satznummer gegeben. Mit „Temporal.Synchrony“ werden Klasse und Typ der Relation benannt. Es folgen Offset-Werte (*SpanList*) und Volltext der beiden Argumente.

In der Ausgabe des Maximum Entropy Classifiers werden alle möglichen Typen angegeben, über die der Parser verfügt, sowie die Wahrscheinlichkeiten für deren Zugehörigkeit. An diesem Beispiel lässt sich die Problematik der Ambiguität vieler CONNs erkennen. Der vorliegende (Teil-)Satz

„[. . .] fitting him as women's clothes fit women [...]“

lässt die Einordnung von „as“ als temporal offensichtlich als falsch erscheinen, obwohl die Wahrscheinlichkeit dafür mit knapp 60% beziffert wird. Hier handelt es sich um einen Vergleich, der dementsprechend der Klasse COMPARISON zugewiesen werden müsste. Jedoch kommen weder die hier angegebenen Typen Contrast oder Concession der Bedeutung nahe. Der Typ Similarity (vgl. Abbildung 2.1), der die Relation am besten repräsentiert, wird vom Parser nicht unterstützt. Dies stellt für solche Fälle auf Level 2 der Annotationsebene also eine vom Parser gegebene Einschränkung dar. Ein Training des Parsers mit dem Datensatz der PDTB 2.0 schlug aus Kompatibilitätsgründen fehl. In Tabelle 2.2. wird die Verteilung der explizit annotierten Klassen beider Korpora gezeigt; hier bestätigt sich, dass die Klasse TEMPORAL vom Parser tendenziell öfter annotiert wurde.

	PDTB-Korpus		**Eigener Korpus**	
Klassen	**Absolut**	**Prozentual**	**Absolut**	**Prozentual**
TEMPORAL	3612	19%	259	27%
CONTINGENCY	3581	19%	127	13%
COMPARISON	5516	29%	239	25%
EXPANSION	6424	34%	321	34%

Tabelle 2.2.: Vergleich der Korpora

2.4.3. Datenverarbeitung

Bei der Verarbeitung der Daten fielen einige bemerkenswerte Konstellationen im Output auf. Beispielsweise lassen sich des Öfteren CONNs finden, die gleichzeitig in mehreren Relationen als Argument stehen. Dies machte die Diskursstruktur entsprechend unübersichtlich und die Schritte zur Auswertung komplexer. Ein Beispiel, bei dem dieser Umstand gut zur Geltung kommt, ist folgender Satz:

„[...] for his solemn call. And if indeed it has been [...]“

Die ersten drei Worte des zweiten Satzes werden jeweils als CONN einer eigenen Relation annotiert, welche rückgreifend auf den ersten Satz Argumente aufspannen, die sich nach dem Schema Arg2.1 – CONN - Arg1 - Arg2.2 verhalten. Das Phänomen der Umschließung von anderen Diskursrelationen lässt sich vor allem bei Dialogen beobachten, da hier die Argumente oft auf die Teilnehmer des Dialogs aufgeteilt werden. So kann ein Token bis zu vier verschiedenen Argumenten zugeordnet werden, was die jeweilige SpanList dementsprechend vergrößert. Wiederum trat auch bei fünf Relationen der Fall auf, dass nur ein Argument gefunden wurde, was der von der PDTB vorgegebenen Struktur widerspricht.

Die Verarbeitung hatte die Erstellung einer gemeinsamen Tabelle bestehend aus der Feature-Tabelle von Festival und den extrahierten Diskursrelationen des PDTB-Parsers. Dabei sollte für jede Silbe ein Eintrag bestehen, ob sie einem CONN oder Argument zugeordnet wurde und welcher Klasse bzw. welchem Typ sie angehört. Alle dafür relevanten Daten wurden dem Pipe Format entnommen, die Offsets entstammen dem Syntaxbaum.

Die erste Herausforderung bestand darin, die Tokenisierungen des Aligners an die des PDTB-Parsers anzugleichen. Dafür wurde eine Offset-Tabelle für alle Texte erstellt und die im Anhang zu findenden Befehle ausgeführt. Es galt hier zu beachten, dass die durch Bindestriche getrennten Worte in der Parser-Tokenisierung aufgeteilt werden mussten. Als nächsten Schritt wurden die Diskursrelationen der CONNs sowie die Argumente extrahiert (Listing A.2) und mit der erstellten Tokenisierung als *RIGHT OUTER JOIN* am jeweiligen Offset-Wert verknüpft.

Anschließend wurden die Tabellen mittels eines *FULL JOIN* auf Basis der Schlüssel Text und Token zusammengeführt. Dies war nötig, da die Token in der Feature-Tabelle pro Silbe und somit mehrmals angegeben waren. Die daraus folgende Redundanz – es war nun jeder Typ eines Textes auf die gleiche Diskursstruktur bezogen – wurde mit einem *Basic*-Skript aufgehoben, das sich die individuelle WortID jedes Token zu Nutze macht und falsche Zuordnungen löscht (Listing A.3). Dabei kam es öfter zum Auslösen der Abbruchbedingung, welche falsches Alignment des Aligners offenlegte; diese wurden manuell korrigiert.

2.5. Rhetorical Structure Theory

2.5.1. Hintergrund

Das zweite Diskursmodell, das vorgestellt werden soll, ist das der *Rhetorical structure theory* (RST). Hierbei handelt es sich um eine Theorie für eine generelle Darstellung von Strukturen in Texten, die eine Basis für das *text planning* am Computer sein soll [Mann & Thompson, 1986]. Der Begriff rhetorische Relation lässt sich dabei mit dem Begriff der Diskursrelation gleichsetzen. Die RST ist – vergleichbar mit der DRT – ebenfalls hierarchisch aufgebaut, nutzt aber eine andere Notation, wie in Abbildung 2.2 zu erkennen ist. Entsprechend wird bei der RST eine gesamte Aussage bzw. der gesamte Text zusammenhängend hierarchisch strukturiert. Die Segmente, auf denen diese Hierarchie aufgebaut ist, werden als Schemata bezeichnet. Mit zunehmend vielen Segmenten – z.B. bei langen Texten – werden die Diagramme allerdings sehr groß und unübersichtlich.

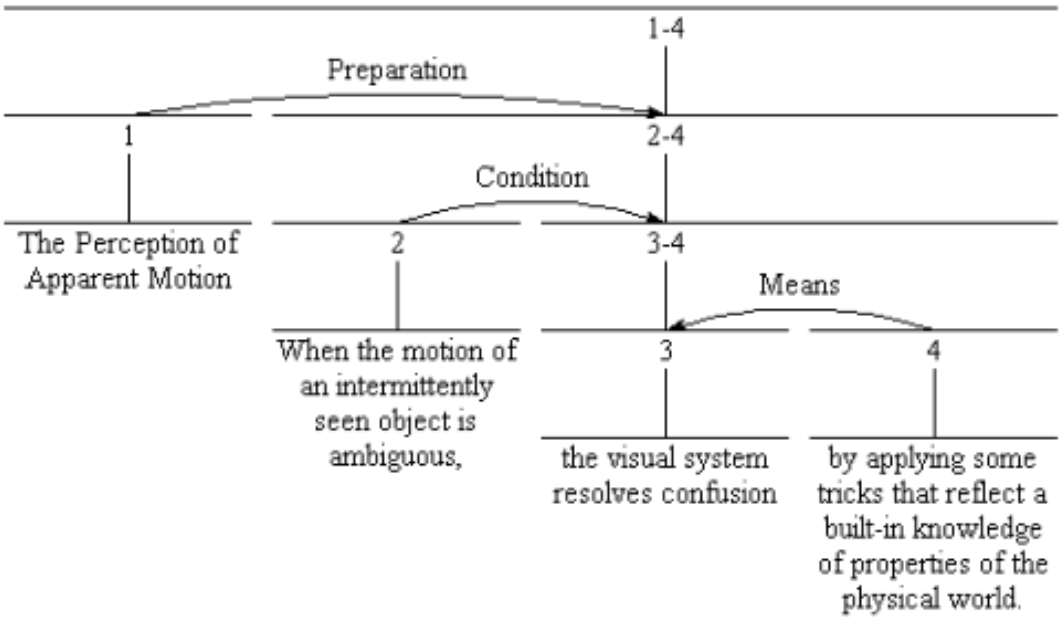

Abbildung 2.2.: Beispiel einer RST-Hierarchie [Taboada, 2006]

Das generische RST-Schema besteht aus der Aufteilung zweier in Relation stehender Textpassagen (*Spans*) bzw. Diskurssegmente in Nukleus und Satellit. Dabei ist die Heuristik gegeben, dass eine Relation möglichst in Leserichtung von rechts nach links gezogen wird. Dieses Konzept wird unter dem Begriff *nuclearity* zusammengefasst. Es besagt, dass der Satellit semantisch nicht alleine stehen kann, ein Text ohne Satelliten dennoch verständlich sein kann. Eine vollständige Annotation einer Relation beinhaltet neben deren Klasse und Richtung - sprich von links nach rechts im Text oder

umgekehrt – Auflagen (*Constraints*) für die einzelnen Diskurssegmente Nukleus und Satellit. Darunter wird verstanden, wie der Leser die einzelnen Konstituenten oder die Kombination aus ihnen wertet.

Eine erste Menge der möglichen Diskursrelationen wurde in Mann & Thompson, 1988 vorgestellt und vollständig definiert, seitdem wurde diese Menge einige Male erweitert und präzisiert. Alle inklusive ihrer Auflagen aufzulisten, wäre hier nicht sinnvoll; zumal der genutzte Parser eine eigene Menge an Relationen nutzt. Es sei jedoch erwähnt, dass sich die Relationen in die Gruppen *Presentational*, *Subject-Matter*, und *Multinuclear* auftrennen lassen.

Als Presentational Relations werden Relationen bezeichnet, mit denen eine Neigung im Leser ausgelöst werden soll, etwas zu tun oder zu denken. Subject-Matter Relations haben den Effekt, dass der Leser die Relation als solche erkennt. Eine Relation wird als Multinuclear eingestuft, wenn mehrere Nuklei an der Bildung dieser beteiligt sind.[7] Hierzu gehört auch die Relation *joint*, welche ausdrückt, dass keine semantische Relation vorliegt.

Mann und Taboada weisen zudem darauf hin, dass RST-annotierte Diskursstrukturen nicht immer auf Basis eines Signals – einer Konjunktion etwa – entstanden sind.[8] Relationen werden demnach auch für die RST sehr oft implizit dargestellt.

2.5.2. Parser

Bedingt durch ihr langes Bestehen, ist die Entwicklung von Parsern für die RST in dem Fachbereich schon seit längerem ein Thema; vor allem die Veröffentlichung der RST Discourse Treebank [Carlson et al., 2001], einem weiteren Korpus bestehend aus RST-annotierten Daten, war maßgeblich daran beteiligt. Die Auswahl an Parsern für eine RST-Annotation war folglich sehr groß; nicht zuletzt wurden sogar Korpora und Parser für andere Sprachen neben Englisch zusammengestellt [Braud et al., 2017]. Letztendlich wurde ein Parser des CLU Lab [Surdeanu et al., 2015] für diese Arbeit ausgewählt; dafür entscheidend waren dessen Programmiersprache Java und der Umfang der Funktionen.

Es handelt sich präzise um zwei verschiedene Parser, die von zwei bereits existierenden inspiriert wurden [Feng & Hirst, 2014; Hernault et al., 2010] und das Stanford CoreNLP-Toolkit nutzen. Beide Parser funktionieren auf derselben Basis bzw. demselben Framework, unterscheiden sich jedoch in ihrem Ansatz. Der erste Parser nutzt die Dependenzsyntax des Textes, der zweite zusätzlich Informationen zur Koreferenz, um

[7] Vgl. https://www.sfu.ca/rst/01intro/definitions.html
[8] Vgl. http://www.sfu.ca/rst/02analyses/index.html

Diskursrelationen zu finden. Letzterer schneidet bei den von Surdeanu et al. durchgeführten Tests um F_1-Scores von 0.1 marginal besser ab, mit der Einschränkung, dass die Annotation länger dauert. Da dieser Umstand für die Arbeit nicht von Bedeutung ist, wird dieser Parser verwendet.

Dem Algorithmus zur manuellen Annotation von RST-Strukturen folgend [Mann & Thompson, 1986], teilen die Parser den Text zunächst in Diskurseinheiten (*EDUs*) ein. Beide Parser nutzen den Perceptron-Classifier für diese Aufteilung. Für das darauffolgende Labeling der Relationen wird *Logistische Regression* (LR) verwendet.

Surdeanu et al. bieten ebenfalls ein Tool zum direkten Vergleich der beiden Parser an. Es lassen sich bei der Betrachtung beider Outputs einige Unterschiede in der Annotation feststellen, jedoch wenige Regelmäßigkeiten finden. Der letztendlich für die Auswertung verwendete Parser scheint auf höheren Ebenen öfter die Annotation *joint* zu vergeben, also weniger semantische Relationen zwischen großen Konstituenten zuzuweisen. Es folgt eine Auflistung aller 16 Relationsgruppen in Carlson et al., 2001. Relationen, die keiner der zuvor vorgestellten Gruppen angehören, wurden erst in der RST Discourse Treebank eingeführt.

Presentational	Subject-Matter	Multinuclear	RST-TB
Background	Cause	Contrast	Attribution
Enablement	Condition	Joint	Comparison
Summary	Elaboration		Explanation
	Evaluation		Same-unit
	Manner-means		Temporal
			Topic-comment
			Topic-change

Tabelle 2.3.: Relationsgruppen des RST-Parsers nach Gruppen

Der Output der Parsers repräsentiert den Diskursbaum, dessen Blätter die jeweiligen EDUs darstellen. Diese werden mit dem Tag *TEXT:* markiert. Die Elternknoten geben den Relationsnamen und die Relationsrichtung vom Nukleus aus an (Left-To-Right oder umgekehrt, siehe Listing 2.4).

Listing 2.4 Beispiel-Ausgabe des Diskursbaums

```
contrast (RightToLeft)
 elaboration (LeftToRight)
  TEXT:This is an example .
  condition (RightToLeft)
   TEXT:If it gets processed ,
   TEXT:it will show the RST discourse tree .
 TEXT:But not every discourse relation possible
```

2.5.3. Datenverarbeitung

Da der Parser in seiner Quell-Version nur einzelne Strings einlesen konnte, wurde ein Java-Skript geschrieben, das das Korpus einlesen konnte und annotiert zurückgab; siehe Anhang A.4. Dabei trat einmalig das Problem auf, dass die Zeichenkombination [‘“] am Ende eines Dialogs den Parsing-Vorgang unterbrach. Anschließend sollte eine Offset-Tabelle aus den Ausgaben des CoreNLP-Teils erstellt werden; hier war entscheidend, ob die Offsets denen des PDTB-Parsers entsprachen (vgl. Listing A.5). Da sich diese Vermutung bestätigte, erleichterte dies das Zusammenfügen mit der zuvor erstellten Tabelle aus PDTB und Festival zu einem späteren Zeitpunkt.

Da der vom Parser ausgegebene RST-Baum in keinem gängigen Hierarchie-Format (z.B. XML oder JSON) war, wurde für die Zuordnung der Relationen aller Ebenen zu den EDUs ein Workaround benutzt. Dieser sah vor, die Tags aller Relationen nach unten auszufüllen, sodass jede Zeile jeden Knoten der Hierarchie anzeigte. Anschließend erfolgte die Tokenisierung der EDUs nach dem Workflow in Listing A.6, hierfür wurden reguläre Ausdrücke und Makros in dem Textbearbeitungsprogramm Notepad++[9] genutzt. Es folgte - wie zuvor erwähnt – das Zusammenführen der tokenisierten Tabelle mit der bereits zusammengeführten Tabelle aus Kapitel 2.4.3., nun anhand der Schlüssel Text und Offset mit einem *RIGHT OUTER JOIN*.

Aufgrund der Vorgehensweise, die dem Workaround folgte, zeigte sich nun die Größe der von der RST erzeugten Bäume; die Hierarchie umfasst teilweise 21 Ebenen von Kopf- zu Blattknoten. Es ist naheliegend, dass Annotationen für lange Prosatexte auf höheren Ebenen nicht sonderlich aussagekräftig sind. Die Semantik eines nicht dafür ausgelegten Textes, der in den meisten Fällen noch Dialoge enthält, auf eine Relation zusammenfassen zu können, scheint unrealistisch. Es wurde dementsprechend auch bei vielen Texten die Relation *joint* auf oberster Ebene ausgegeben.

Daher wurden für die Auswertung die Bäume bzw. die pro Wort bestehenden Relationen auf die letzten drei Ebenen reduziert, um nur Relationen zu zählen, deren Schemata nicht zu groß und Klassifizierung nahe an der Semantik der Inhalte sind. Dafür wurde der Workflow in Listing A.8 verwendet und aus den extrahierten Blattknoten eine finale Tabelle für die Auswertung erstellt.

[9] https://notepad-plus-plus.org/

3. Auswertung

3.1. Vorbereitung

Vor Beginn der eigentlichen Auswertung muss geklärt werden, welche Pitch-Werte dafür in Frage kommen und auf welcher Basis sie zusammengefasst werden sollen. Sie liegen – mit Ausnahme einiger Annotationslücken – pro Silbe vor; das bedeutet, es stehen somit alle Auswertungsräume, wie Wort, Phrase, Satz oder Text zur Verfügung. Um dem Prinzip der Diskursstrukturen zu folgen, ist die Auswertung auf Phrasenebene naheliegend. Der Festival-Datensatz liefert eine Angabe zu den jeweiligen Phrasengrenzen und zeigt diese unter *Phrase Breaks* an. Diese werden durch ein probabilistisches Modell, das auf die Part-of-speech-Tags der umliegenden Token angewendet wird, statistisch annotiert und mit *NB* für „no break“ oder *BB* für „big break“ angegeben[1].

Die Betrachtung des kombinierten Datensatzes zeigt, dass die Aufteilung der Argumente beider Parser mit der Phrase Break-Annotation zu sehr großen Teilen übereinstimmt. Da die Argumente in der Analyse ohnehin voneinander getrennt untersucht werden, erfolgt die Zusammenfassung der Pitch-Werte also auf Argument-Basis. Dies ermöglicht zudem eine unabhängige Untersuchung beider verschiedener Argumentstrukturen und die Vermeidung von Fehlern bei der Phrase Break-Annotation. Vor allem für die Auswertung der RST-Argumente ist dies wichtig, da sie auf höheren Hierarchieebenen mehrere Phrasen umfassen können.

Problematisch für die korrekte Darstellung der Pitchverläufe wurden die Silben, die vom Aligner fehlerhaft mit einem Pitch von 0 angegeben wurden. Erste Messungen zeigten, dass dadurch die Durchschnittswerte signifikant sanken. Deshalb wurde davon abgesehen, diese Silben in die Auswertung aufzunehmen und alle Werte = 0 wurden gelöscht.

Daran anschließend stellte sich die Frage nach der Grundfrequenz der individuellen SprecherInnen, die für die Einordnung der argumentspezifischen Pitch-Werte zu Rate gezogen werden sollte. Diese ist nach Hess, 1983 algorithmisch nicht zu bestimmen. Unterstrichen wird dies durch viele nicht kontrollierbare Faktoren bei der aktuellen Untersuchung, wie das Verstellen der Stimme sowie Qualität der Aufnahme und

[1] http://www.festvox.org/docs/manual-2.4.0/festival_17.html#Phrase-breaks

Messung. Folglich wurde die Grundfrequenz der SprecherInnen angenähert, indem der Durchschnittswert der absoluten Pitch-Werte – exklusive der Messfehler des Aligners - berechnet wurde. Abbildung 3.1 zeigt die Pitch Ranges aller SprecherInnen.

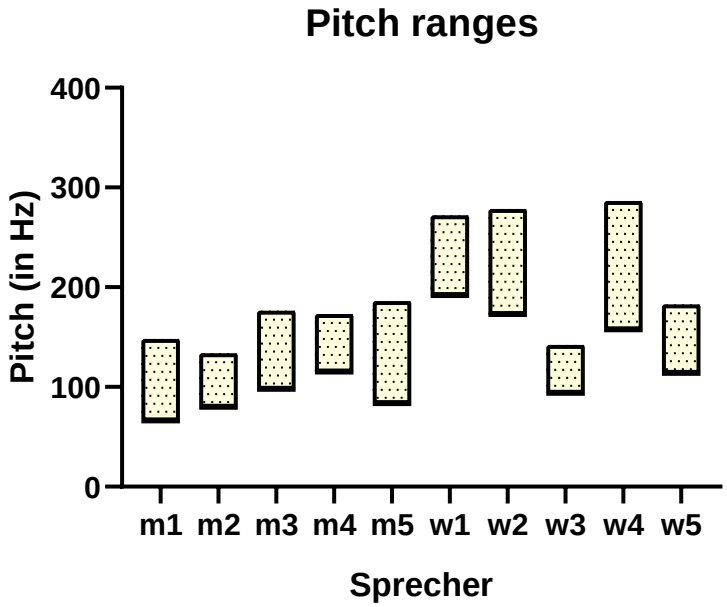

Abbildung 3.1.: Pitch Ranges der SprecherInnen

Bei Betrachtung des Diagramms bestätigen sich zuvor getroffene Annahmen. Zum Beispiel, dass Sprecherin w3 mit einem für weibliche Sprecherinnen verhältnismäßig niedrigen Pitch spricht und zusätzlich sehr monoton, was sich an der geringen Range erkennen lässt. Auch w5 spricht durchschnittlich nur unwesentlich höher als die meisten männlichen Sprecher. Aus diesen Beobachtungen lässt sich der offensichtliche Schluss ziehen, dass eine durchschnittliche Auswertung nicht auf den absoluten Pitch-Werten durchgeführt werden kann, da sie zu sehr vom Sprecher abhängen.

3.2. Statistische Mittel

Median und Standardabweichung

Für die Analyse der Verteilung einzelner Datenpunkte bestehend aus Pitch-Werten besteht die Möglichkeit, Streuungsmaße entweder um den Mittelwert oder den Median der Datenpunkte zu betrachten. Dies ist notwendig, um Ausreißer oder falsche Messwerte in der Auswertung nicht zu stark zu gewichten und demnach ein realistisches Maß für die durchschnittliche Pitch Range zu erhalten.

Der Median bezeichnet den expliziten mittleren Wert einer nach Größe sortierten Auflistung aller Datenpunkte. Die Streuung wird anhand der expliziten Werte, die den Bereich der oberen bzw. unteren 75% der Auflistung abgrenzen, angezeigt. Diese Grenzen werden auch Quartile (Quartil 1 für 25%, Quartil 2 für den Median, Quartil 3 für 75%) genannt, den Bereich dazwischen Interquartilsabstand (*interquartile range*).

Die *Standardabweichung* stellt eine Berechnung der Streuung um den Mittelwert dar. Das Ergebnis der Standardabweichung ist abhängig von der Differenz jedes einzelnen Datenpunktes vom Mittelwert. Sie wird folgendermaßen berechnet,

$$\sigma = \sqrt{\frac{1}{n}\sum_{1}^{n}(X_i - \bar{X})^2}$$

wobei $\bar{X}$ den Mittelwert angibt, und n die Anzahl der Datenpunkte. Diese Formel beschreibt die für die Auswertung verwendete Standardabweichung, welche nicht für eine Stichprobe berechnet wird. Eine Untersuchung auf eine Stichprobe ist hier nicht vonnöten, da nicht in allen Fällen auf andere Sprecher geschlossen werden kann. Abbildung 3.1 wurde auf Basis der Standardabweichung vom Pitch-Mittelwert erstellt.

Während die Nutzung des Median und des Interquartilsabstands bei der Analyse von Pitch Ranges ein oft verwendetes Mittel ist [Carlson, 2004], wird bei dieser Arbeit ausschließlich die Standardabweichung genutzt. Der Bezug auf den Median ist bei großen Datensätzen zwar durchaus sinnvoll, allerdings sind die Daten unter den einzelnen Typen bzw. Klassen, die untersucht werden, nicht sonderlich ausgeglichen verteilt. Es existieren einige Klassen, deren geringe Anzahl die Auswertung mit dem Median dementsprechend stark beeinflussen kann und eine potenziell zu große Range zur Folge haben. Ignoriert man diese kleineren Klassen, lassen sich jedoch zueinander proportionale Ergebnisse zwischen der Auswertung von Median und Standardabweichung beobachten.

Z-Scores

Damit die Pitch-Werte aller SprecherInnen für eine allgemeine Auswertung zusammengefasst werden können, mussten deren absoluten Pitch-Werte vereinheitlicht werden. Dafür stehen in der Regel mit der Skalierung und der Normalisierung zwei gängige Methoden zur Verfügung.

Skalierung (hier S_i) bezeichnet das Verteilen der Datenpunkte in einem neuen numerischen Bereich, beispielsweise zwischen 0 und 1. Dieses Vorgehen wird oft bei Classifier- oder Clustering-Algorithmen wie SVMs oder kNN benutzt.

Bei der Normalisierung (hier Z_i) werden die Werte auf den Mittelwert und die Standardabweichung angepasst, sodass sich eine Verteilung um 0 ergibt. Diese neu entstehenden

Werte werden auch *Z-Scores* genannt, sie verändern Form und Bezugspunkt der Verteilung. Z-Scores zeigen an, wie viele Standardabweichungen der Datenpunkt vom Mittelwert entfernt ist.

$$S_i = \frac{(X_i - min(X_{Sprecher}))}{max(X_{Sprecher}) - min(X_{Sprecher})}, \qquad Z_i = \frac{(X_i - \bar{X}_{Sprecher})}{\sigma_{Sprecher}}$$

Wie bereits beim Median ist das Nutzen absoluter Werte, in diesem Fall der jeweiligen Minima und Maxima, nicht sonderlich hilfreich für die Auswertung, da vor allem dabei Ausreißer einer größeren Klasse einen Einfluss auf das Endergebnis haben. Folglich werden Z-Scores genutzt, um die Daten statistisch auswerten zu können. Vor allem dafür ist entscheidend, dass als 0 erkannte Datenpunkte im Pitch eliminiert werden, da der Z-Score ansonsten für jeden sehr niedrig wäre.

Analysis of variance und t-Test

Um eine statistische Unterscheidung zwischen den Diskursrelationen machen zu können, wird zunächst eine *Analysis of variance* (ANOVA) durchgeführt. Der ANOVA-Test untersucht, ob sich in der Werteverteilung von mehr als zwei Gruppen signifikante Unterschiede finden lassen.

Wird auf nur eine Variable überprüft, wird ein *One-way ANOVA*, anderenfalls ein *Two-way ANOVA* durchgeführt. Letzteres gilt beispielsweise für die Suche nach Entwicklungen unter einer bestimmten Variable. Für die Überprüfung auf die Pitch Range ist ein One-way ANOVA ausreichend. Da die Daten jeweils einem Sprecher oder einer Sprecherin entstammen, wird der ANOVA als *repeated measures* bezeichnet.

Da das Ergebnis des ANOVA zwar Aufschluss darüber gibt, ob es einen signifikanten Unterschied zwischen den untersuchten Gruppen gibt, aber nicht zwischen welchen, wird ein *post-hoc t-Test* angeschlossen. Die Ergebnisse eines T-Tests machen eine Aussage über den signifikanten Unterschied zweier spezifischer Gruppen.

Dabei muss zwischen einem gepaarten und einem ungepaarten t-Test unterschieden werden. Gepaarte t-Tests zeigen meist einen „vorher-nachher“-Vergleich an, während der ungepaarte t-Test zwei Stichproben aus den Datenmengen erhebt. Der gepaarte t-Test ist für die Analyse der Diskursrelationen ungeeignet, da er zwei gleich große Mengen benötigt und die Klassen der Diskursrelationen unterschiedlich oft in den Texten vorkamen.

Signifikanz bedeutet für beide Tests, dass der Ergebniswert p kleiner als der vorgegebene Signifikanzwert sein muss, in der Regel gilt $\alpha = 0.05$. Das drückt den Kehrwert der Wahrscheinlichkeit aus, dass man die zuvor aufgestellte Hypothese, dass ein signifikanter Unterschied besteht, annehmen kann. Da der t-Test jedoch mehrmalig ausgeführt werden muss, wird die Bonferroni-Korrektur angewandt. Sie besagt, dass für eine n-fache Durchführung eines t-Tests der Signifikanzwert um n geteilt werden muss.

$$\alpha' = \frac{\alpha}{n}$$

Dies soll dem durch mehrfaches Durchführen steigenden Fehlerpotenzial des Signifikanzwertes entgegenwirken. Die Ergebnisse der einzelnen t-Tests werden also mit dem neu berechneten Bonferroni-Signifikanzwert auf Signifikanz verglichen.

3.3. Durchführung

Die Auswertung sah vor, die Pitch Ranges für alle Argumente der jeweiligen Diskursrelationen auszulesen und statistisch zusammenzufassen. Die Bestimmung der Pitch Range sollte dabei sowohl H- als auch L-Targets abdecken, das Bestimmen ausschließlich anhand der H-Targets ist eine öfter praktizierte [Ayers, 1994; Carlson, 2004], aber hierfür nicht unbedingt geeignete Herangehensweise. Für jedes Argument wurden anhand der dazugehörigen Pitch-Werte Z-Scores berechnet. Aus diesen wurden die H- und L-Targets, der Mittelwert und die Standardabweichung aller einzelnen Argumente gespeichert. Pro Parser und Diskursrelation wurden folgende Messungen durchgeführt:

1. Pitch Range anhand der Mittelwerte aller H- und L-Targets

2. Pitch Range anhand der Silben-Standardabweichung vom Mittelwert (sprecherunabhängig)

3. Pitch Range anhand der Silben-Standardabweichung vom Mittelwert (sprecherabhängig)

3a. Vergleich der Pitch Ranges von Nukleus und Satellit

Diese verschiedenen Messungen wurden mit der Intention ausgewählt, möglichst viele Faktoren der Datenverteilungen untersuchen zu können. Messung 1 legt den Fokus auf die oberen und unteren Begrenzungen der Pitch Range, um deren Differenz zu vergrößern und so aussagekräftigere Daten zu erhalten. Messung 2 stellt eine vergleichsweise neutrale Untersuchung dar, da hier jeder einzelne Datenpunkt ohne eine Gewichtung in den Ergebnissen repräsentiert wird. Aus diesem Grund werden hauptsächlich hierfür

der ANOVA inklusive der t-Tests ausgewertet, um signifikante Unterschiede zwischen den Diskursrelationen zu finden.

In Messung 3 werden die individuellen Pitch Ranges aller SprecherInnen pro Relation ausgewertet. Diese Messung reduziert die Daten drastisch auf maximal zehn Werte – es kam nicht jede Relation pro SprecherIn vor – und zeigt stärkere Tendenzen auf; damit ist sie aber auch anfälliger für Ausreißer. Außerdem soll eine zusätzlich durchgeführte Messung Aufschluss darüber geben, inwiefern sich die Pitch Range der SprecherInnen zwischen den beiden Argumenten einer Relation verändert; sie ist daher zwangsläufig sprecherabhängig.

3.4. Ergebnisse

3.4.1. PDTB

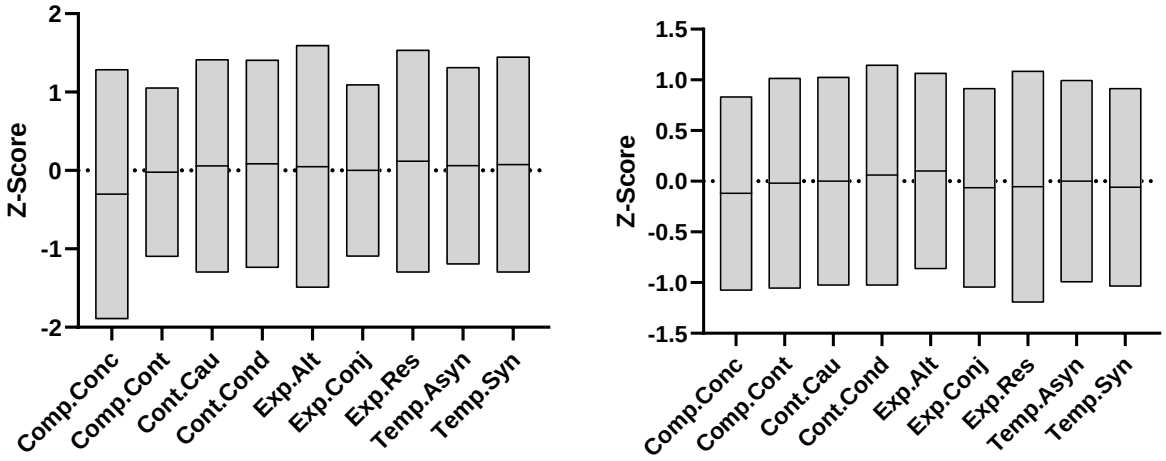

Abbildung 3.2.: PDTB: Links wird Messung 1 gezeigt, rechts Messung 2

Die Ergebnisse der ersten beiden Messungen weisen eine ähnliche Form auf, erwartungsgemäß ist die Skalierung der ersten Messung jedoch größer. Die Pitch Range für die Betrachtung der H- und L-Targets ist jeweils mindestens eine Standardabweichung von dem durchschnittlichen Pitch der SprecherInnen entfernt, im Falle des Typs *Concession* sogar nahezu zwei. Die Verteilung von Messung 2 ist sehr ausgeglichen und liegt angenähert jeweils genau eine Standardabweichung vom Mittelwert entfernt. Auffällig ist beim Großteil der Relationen die sehr symmetrische Verteilung und dass der jeweilige Mittelwert (als Trennlinie) auch bei Messung 1 hauptsächlich im Intervall

von [-0,1…0,1] Standardabweichungen liegt. Der grafische Vergleich dieser beider Messungen ist entsprechend nicht sonderlich aussagekräftig, auch eine Unterscheidung zwischen den Relationen fällt so zunächst schwer.

Aus diesem Grund werden in einem nächsten Schritt die zuvor erklärten Signifikanztests durchgeführt. Der ANOVA liefert p=0,0000056 und sagt damit signifikante Unterschiede zwischen den Klassen voraus. Um diese genauer zu untersuchen, wird für jede Kombination der Diskursrelationen ein post-hoc t-Test durchgeführt und dessen p-Wert in die Matrix in Abbildung 3.3 eingetragen.

Comp.Conc	Comp.Cont	Cont.Cond	Exp.Alt	Exp.Res	Temp.Asyn	Temp.Syn	Exp.Conj	Cont.Cau	
1,000	0,135	0,018	0,005	0,509	0,069	0,325	0,374	0,101	**Comp.Conc**
	1,000	0,011	0,014	0,519	0,327	0,068	0,004	0,586	**Comp.Cont**
		1,000	0,476	0,102	0,081	0,000	0,000	0,180	**Cont.Cond**
			1,000	0,042	0,043	0,001	0,000	0,074	**Exp.Alt**
				1,000	0,330	0,964	0,943	0,403	**Exp.Res**
					1,000	0,014	0,001	0,959	**Temp.Asyn**
						1,000	0,718	0,106	**Temp.Syn**
							1,000	0,049	**Exp.Conj**
								1,000	**Cont.Cau**

Abbildung 3.3.: PDTB: Matrix der post-hoc t-Tests

Die Matrix zeigt für jedes Feld die p-Werte der t-Tests beider verglichenen Relationen. Sind diese gleich, ist der Wert entsprechend mit p=1 auf dem Maximum, da kein Unterschied besteht. Rote Felder zeigen einen p-Wert, der über dem im Voraus festgelegten Signifikanzwert von 0,05 liegt; die jeweiligen Relationen sind also nicht signifikant zu unterscheiden. Gelbe Felder stellen einen Grenzfall dar; sie zeigen zunächst, dass ein signifikanter Unterschied besteht, da sie unter dem Signifikanzwert von 0,05 liegen. Wird für diesen Signifikanzwert jedoch die Bonferroni-Korrektur angewandt, ergibt sich der unten berechnete korrigierte Signifikanzwert.

$$\alpha' = \frac{0.05}{9} = 0.005$$

Zieht man diesen Wert in Betracht, sollten die gelb markierten Relationspaare nur unter Vorsicht als signifikant heranziehen. Grüne Werte hingegen zeigen einen klar signifikanten Unterschied. Aus diesen Werten lassen sich mit *Expansion.Conjunction* und *Expansion.Alternative* zwei Relationen finden, deren t-Tests ein größtenteils signifikantes Ergebnis liefern. Restliche Relationen scheinen sich weniger voneinander abgrenzen zu lassen. Unter Berücksichtigung dieser Daten lassen sich in Hinblick auf Messung 3 einige weitere Aussagen treffen.

Messung 3, welche die Standardabweichung ausschließlich aus den individuellen Pitch Ranges der SprecherInnen anzeigt, lässt einige Tendenzen aus den vorangehenden

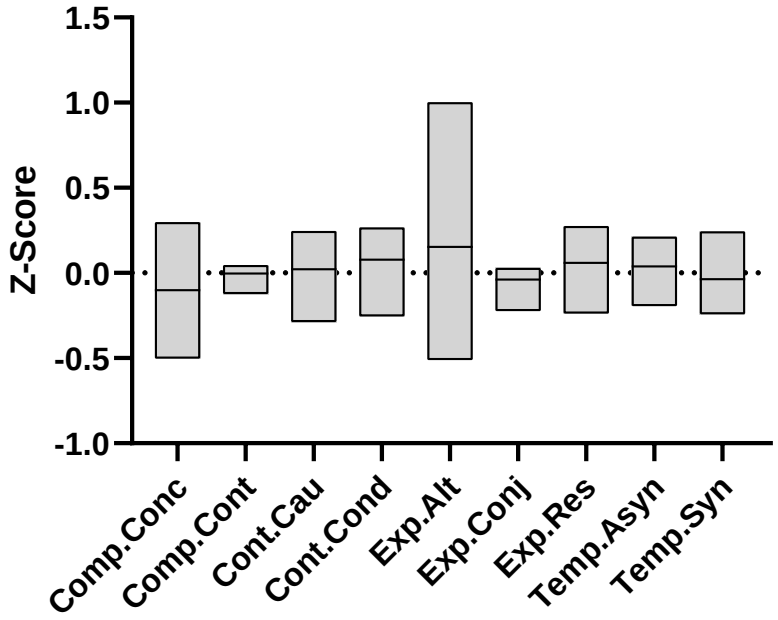

Abbildung 3.4.: PDTB: Messung 3 in Abhängigkeit zu den SprecherInnen

Diagrammen erkennen. Die beiden Relationen, welche am ehesten signifikante Unterschiede beinhalten, zeigen eine jeweils individuelle Pitch Range. *Expansion.Alternative* hat eine sehr großflächige Verteilung, während die Verteilung von *Expansion.Conjunction* relativ konzentriert auf negativen Z-Scores ist. Die Relation *Comparison.Contrast* bestätigt vorherige Beobachtungen und hat ebenfalls eine sehr kleine Verteilung, ist jedoch laut t-Tests nicht signifikant unterschiedlich. Insgesamt zeigt diese Messung besser erkenntliche Unterscheidungen; zu erwähnen ist jedoch, dass der ANOVA speziell für Messung 3 mit p = 0,58 keine Signifikanz zwischen den Klassen aussagt. Dies ist unter anderem auf die geringe Zahl der Datenpunkte pro Relation zurückzuführen.

Es wurde außerdem untersucht, ob signifikante Unterschiede zwischen den einzelnen Argumenten der Relationen bestehen. Im Falle der PDTB existiert die Aufteilung in Arg1, Arg2 und CONN. Letzteres Element wurde für die Untersuchung nicht berücksichtigt, um die beiden Argumente komplett voneinander abzutrennen. Eine einzelne Betrachtung der CONNs ist nicht aussagekräftig, da sie zumeist nur aus einem Wort bestehen; daraus lässt sich folgern, dass deren durchschnittlicher Pitch-Wert deutlich höher ist als der beider Argumente.

Es wurden t-Tests zwischen allen Argumentpaaren in jeder Relation durchgeführt; da in diesem Fall ein direkter Effekt bei allen SprecherInnen untersucht wird, wurde der gepaarte t-Test verwendet. Es zeigte sich nur bei einer Klasse ein signifikanter Effekt

der Argument-Richtung von Nukleus zu Satellit, sprich es ergab nur ein t-Test einen Wert von $p < 0{,}05$. Dies traf auf die Relation *Temporal.Synchrony* zu; der Nukleus (Arg1) wird hier bei allen SprecherInnen im Schnitt mit einem deutlich höheren Pitch-Wert artikuliert als der Satellit (Arg2). Es sei dabei erneut zu erwähnen, dass die Begriffe nicht mit der Nummer des Arguments gleichzusetzen sind, die ausgewerteten Daten es in Bezug auf diese Relation aber zulassen. Dieses Verhalten anhand der Mittelwerte wird in Abbildung 3.5 gezeigt, die daraus resultierenden Pitch Ranges sind erneut quasi-gleichverteilt.

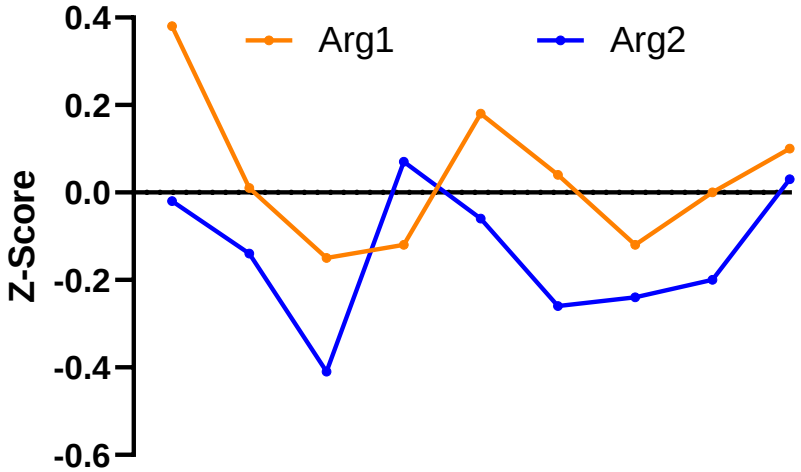

Abbildung 3.5.: PDTB: Vergleich der Argumente von Temporal.Synchrony

3.4.2. RST

Da die Untersuchung der RST-Daten im Workflow chronologisch nach der Fertigstellung der Auswertung der PDTB-Daten stand, wurde dieselbe Vorgehensweise wie zuvor angewendet. Es zeigte sich das gleiche Verhalten der ersten beiden Messungen in Bezug auf grafische Darstellung und Skalierung der Verteilungen, wie es bereits bei der PDTB beobachtet wurde; deshalb wird in Abbildung 3.6 nur Messung 2 gezeigt.

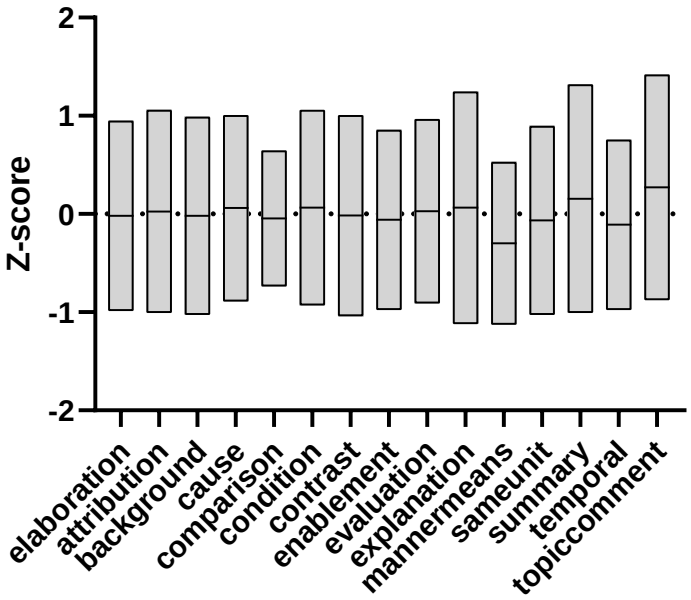

Abbildung 3.6.: RST: Messung 2, Standardabweichung vom Mittelwert

Es wurden insgesamt alle 16 Diskursrelationen vom RST-Parser erkannt und annotiert, 15 davon auf den untersten drei Hierarchie-Ebenen, diese sind in dem Diagramm zu sehen. Es sind erneut nicht viele Beobachtungen möglich, da die Pitch Ranges erneut durchschnittlich rund eine Standardabweichung vom Mittelwert entfernt sind. An den Mittelwerten ist allerdings teilweise eine Tendenz in entweder negativer oder positiver Richtung vom neutralen Z-Score zu erkennen, Beispiele sind *manner-means*, *temporal* oder *topic-comment*. Die Relation *comparison* hat als einzige eine proportional kleine Verteilung.

Es wurde auf Messung 2 ein weiterer ANOVA mit post-hoc t-Tests für alle Relationen durchgeführt (siehe Abbildung 3.7). Die Bonferroni-Korrektur ergab folgenden neuen Signifikanzwert:

$$\alpha' = \frac{0.05}{15} = 0.003$$

Die in Ansätzen erkennbaren Tendenzen aus Messung 2 werden durch das Ergebnis der t-Tests unterstützt, da die Relationen *manner-means* und *topic-comment* mit (nahezu) allen Ergebnissen unter dem korrigierten Signifikanzwert liegen. Auch für die Relation *temporal* liegen mehr signifikante Ergebnisse vor als nicht-signifikante. Es lässt sich also eine klare Abgrenzung dieser Relationen voneinander vornehmen. Weitere potenziell differenzierbare Relationen sind *comparison* und *same-unit*. Der Großteil der anderen Relationen wird jedoch als nicht signifikant unterschiedlich eingestuft.

elab	attr	backg	cause	comp	cond	contrast	enab	eval	expl	mMean	unit	summ	temp	tComm	
1,000	0,003	0,957	0,122	0,795	0,007	0,670	0,325	0,530	0,028	0,000	0,041	0,069	0,021	0,000	**elab**
	1,000	0,086	0,531	0,449	0,239	0,169	0,044	0,926	0,343	0,000	0,000	0,208	0,001	0,000	**attr**
		1,000	0,167	0,795	0,026	0,788	0,379	0,565	0,070	0,000	0,122	0,091	0,040	0,000	**backg**
			1,000	0,248	0,907	0,222	0,054	0,781	0,931	0,000	0,024	0,397	0,005	0,007	**cause**
				1,000	0,209	0,727	0,857	0,456	0,296	0,001	0,777	0,101	0,397	0,002	**comp**
					1,000	0,047	0,009	0,712	0,989	0,000	0,001	0,403	0,000	0,001	**cond**
						1,000	0,290	0,632	0,106	0,000	0,073	0,113	0,027	0,000	**contrast**
							1,000	0,310	0,032	0,000	0,831	0,037	0,284	0,000	**enab**
								1,000	0,763	0,000	0,261	0,377	0,096	0,031	**eval**
									1,000	0,000	0,004	0,480	0,002	0,007	**expl**
										1,000	0,000	0,000	0,001	0,000	**mMean**
											1,000	0,027	0,307	0,000	**unit**
												1,000	0,007	0,360	**summ**
													1,000	0,000	**temp**
														1,000	**tComm**

Abbildung 3.7.: RST: Matrix der post-hoc t-Tests

Das Diagramm zu Messung 3 (Abbildung 3.8) konkretisiert die Beobachtungen aus den t-Tests auf der grafischen Ebene. Besonders die Pitch Range der als signifikant unterschiedlich erkannten Relationen *manner-means* und *topic-comment* lässt sich klar von anderen unterscheiden. Interessant zu beobachten sind auch die sehr dicht zusammenliegenden Grenzen, beispielsweise *elaboration*, *attribution* oder *contrast*. Es scheinen hier alle SprecherInnen eine sehr gleichmäßige Intonation zu haben, da keine Ausreißer für diese Relationen auftreten.

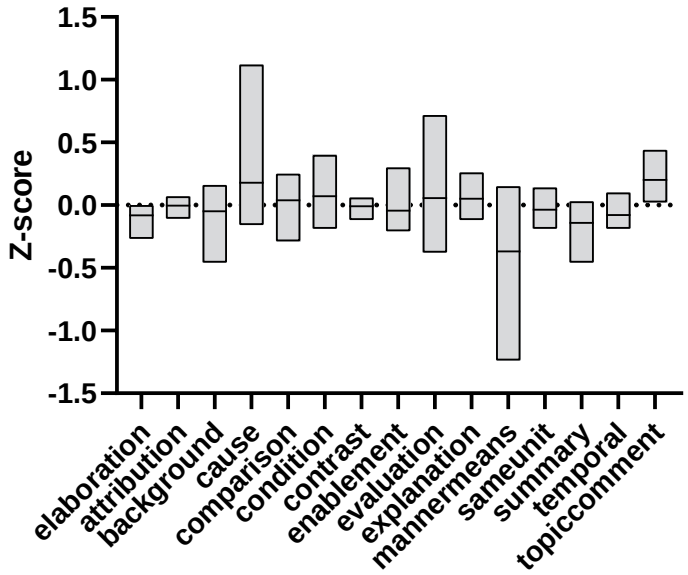

Abbildung 3.8.: RST: Messung 3, in Abhängigkeit zu den SprecherInnen

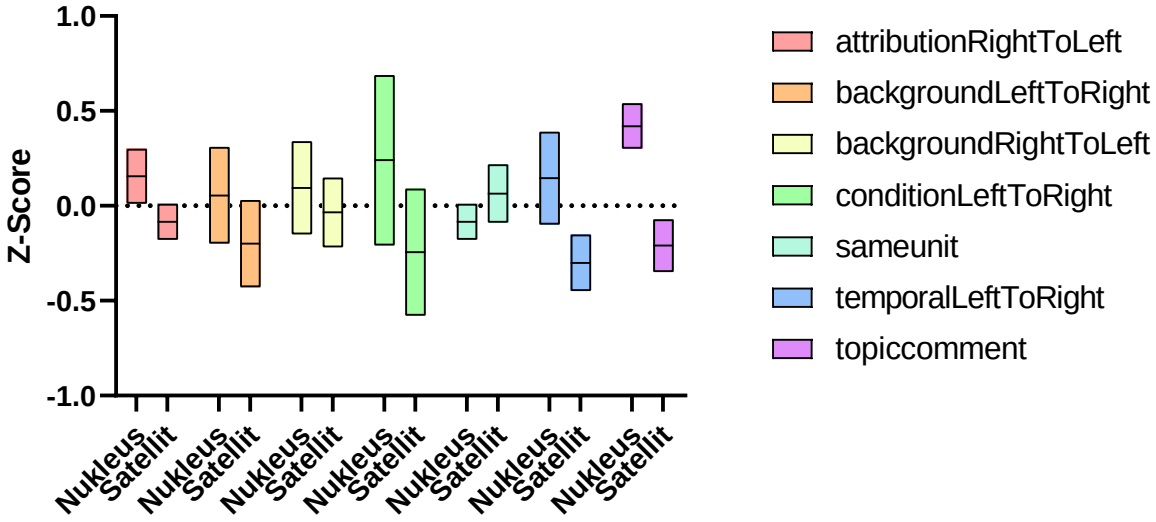

Abbildung 3.9.: RST: Vergleich von Nukleus und Satellit

Die Untersuchung auf die Unterschiede einzelner Argumente erwies sich bei der RST in Bezug auf die Signifikanz ebenfalls als erfolgreich. Wie bereits bei der PDTB festgestellt, besteht ein Zusammenhang zwischen der Relationsrichtung und der Pitch Range beider Argumente. Abbildung 3.9 zeigt alle Argumentpaare, welche sich signifikant unterscheiden. Es gilt dabei zu beachten, dass der RST-Parser die Relationsrichtung für einige Klassen unterschiedlich definiert hat.

So kamen bei einigen Relationen keine Informationen zu Nukleus oder Satellit vor, das lässt sich daran erkennen, dass die Relationsbezeichnung kein Suffix enthält. Dieser Umstand sollte demnach bei Betrachtung des Diagramms berücksichtigt werden, die Beschriftungen „Nukleus“ und „Satellit“ bezeichnen hier ausschließlich die Reihenfolge – Nukleus steht hier also für das erste Argument. Es existieren zusätzlich ein paar Relationen, für die nur manchmal keine Richtung angegeben wurde, ansonsten aber dennoch zwei gerichtete Angaben vorherrschen; insgesamt finden sich für diese also drei unterschiedliche Datensätze.

Es fällt auf, dass der Nukleus – insofern ein Argument als solcher angegeben ist – bei einem signifikanten Unterschied immer die höhere Pitch Range innehat. Bei der *background*-Relation ist dies sogar wechselseitig zu erkennen.

Korrektur und Diskussion zu Messung 3

Obwohl die t-Tests bzw. ANOVAs einige positive Resultate erbrachten, so gab es doch mangelnde Differenzierungsmöglichkeiten der verschiedenen Diskursrelationen durch die ersten beiden Messungen. Aus diesem Grund und um zusätzlich die Argumente einzeln betrachten zu können, wurde Messung 3 durchgeführt. Das Problem potenzieller Ausreißer, welche die durchschnittlichen Pitch Ranges beeinflussen können, kam allerdings zum Tragen. Dies kann an folgendem Beispiel ausgemacht werden:

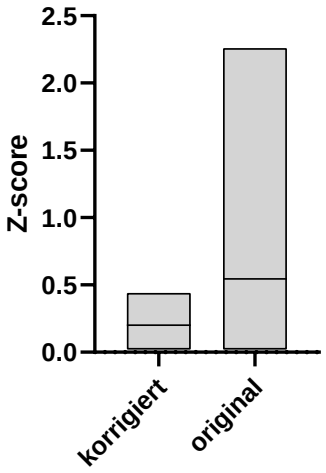

Abbildung 3.10.: RST: Korrigierte und originale Pitch Range

Links wird die Pitch Range der Relation *topic-comment* angezeigt, wie sie im Diagramm zu Messung 3 angegeben wurde, rechts die ursprüngliche Version. Diese Diskrepanz kam durch die Pitch Range-Werte eines einzigen Sprechers zustande, der nur einmalig ein Argument der Relation artikulierte. Dieser Umstand alleine war allerdings nicht für den Anstieg der durchschnittlichen Pitch Range entscheidend; ausschlaggebend war, dass der (männliche) Sprecher dabei eine weibliche Stimme imitiert hatte, was aus dem Inhalt hervorgeht.

Wort	Silbe	Part	Pitch
Hold	hhowld	m5-2.part073	326
on	aan	m5-2.part073	108
What	waht	m5-2.part074	331
are	aar	m5-2.part074	304
you	yuw	m5-2.part074	325
hiding	hhay	m5-2.part074	327
hiding	dihng	m5-2.part074	108
there	dhehr	m5-2.part074	111

Tabelle 3.1.: Annotation des Sprechers m5 (durchschnittlicher Pitch: 131,8 Hz)

Der Inhalt dieser Stelle - ein erstaunter Fragesatz - ist ein Fakt, der dafür spricht, dass die Relation *topic-comment* richtig annotiert wurde; zu der Relation gehört die Unterkategorie *question-answer* [Carlson et al., 2001]. Die Sprünge, welche im Pitch zu beobachten sind, stellen ein Indiz dar, dass der Sprecher seine Stimme verstellt hat.

Das hatte zur Folge, dass für diesen Sprecher mit Z-Scores von 3,65 für das H-Target in Silbe drei und -0,43 für das L-Target zum Ende der Intonationsphrase ein entsprechend hoher Mittelwert bzw. eine hohe Standardabweichung für die gesamte Relation gespeichert wurde. Da dies mit den Werten der anderen Sprecher verrechnet wurde, kam diese große Abweichung zustande und wurde aus der Auswertung (auch bei der Untersuchung zwischen Nukleus und Satellit) gelöscht. Es war jedoch der einzige Fall, bei dem dieses Vorgehen nötig war.

Die angesprochenen Verhältnismäßigkeiten lassen allerdings die Vermutung zu, dass andere SprecherInnen und insbesondere eine größere Anzahl derer ein verändertes Ergebnisbild bewirkt hätten. Auf Messung 3 durchgeführte t-Tests ergeben wiederum für beide Parser sehr ähnliche Signifikanzwerte, wenn auch weniger als bei den großen Datensätzen.

Generell sollten auch Ergebnisse von „kleineren“, sprich spezifischeren Relationen kritisch begutachtet werden. Die Verteilung zwischen den Relationen ist bei beiden Parsern sehr unausgeglichen, z.B. *elaboration* oder *Expansion.Conjunction* nehmen einen großen Anteil an der Gesamtzahl der Annotationen ein. Daraus ist zu folgern, dass diese Relationen umfangreiche Bedeutungsdefinitionen haben und die Durchschnittswerte daher keine spezifische Aussagekraft haben. Eine genauere Untersuchung der Bedeutung einzelner Relationen folgt im nächsten Teil. Messung 3 gleicht diese Diskrepanzen aus, indem nur die individuellen Sprecherwerte betrachtet werden.

3.5. Vergleich

Nach der Beschäftigung mit beiden Konzepten sowie der Nutzung und Auswertung beider Parser lassen sich ein Vergleich ziehen und mögliche Gemeinsamkeiten und Unterschiede festhalten. Es existieren bereits Arbeiten, die den Versuch unternommen haben, ein Mapping zwischen PDTB- und RST-Annotationen in einem doppelt annotierten Korpus vorzunehmen. Scheffler & Stede, 2016 gehen dabei auf die CONNs ein, Poláková et al., 2017 spezifisch auf die impliziten Relationen. Ersterer Ansatz lässt sich mit dem dieser Arbeit vergleichen, jedoch gibt es einige Unterschiede in den Herangehensweisen. Scheffler & Stede, 2016 führen ihre Untersuchung auf einem Korpus, welches nicht von Parsern annotiert wurde und zusätzlich in deutscher Sprache vorliegt. Die Sprache spielt für die RST-Annotation keine primäre Rolle, da jedes Token annotiert vorliegt; es ist jedoch nicht geklärt, wie der lexikalische Ansatz der PDTB-Annotation die Anzahl der erkannten Relationen zwischen den Sprachen beeinflusst.

Es ist dennoch anzunehmen, dass die Unterschiede speziell zwischen diesen Sprachen nicht entscheidend sind.

In Scheffler & Stede, 2016 wird das Mapping auf Basis jedes einzelnen CONNs vorgenommen und die vorliegende RST-Relation aus der RST-TreeBank zugeordnet. Es wird allerdings keine Angabe zu Klassen oder Typen der CONNs aus der PDTB-Annotation gemacht, nur das CONN selbst wird angegeben.

Im für diese Arbeit verwendeten Korpus konnten 74% aller CONNs einem Tag der RST-Annotation auf den untersten drei Ebenen zugeordnet werden, der nicht der Relation *joint* angehört. 51% aller RST-Tags wurden ebenfalls mit einem PDTB-Argument oder -CONN annotiert. Obwohl die beiden Annotationen verschiedenen Ansätzen entstammen, lassen sich deren Relationen mehr oder weniger semantisch vergleichen. Einige dieser Schlüsse konnten aus den Unterteilungen der 16 RST-Relationsgruppen gezogen werden, in denen beispielsweise der PDTB-Typ *Concession* zur Relationsgruppe *Contrast* gezählt wird. Bei der Zuordnung wurde die RST-Relation *joint* ignoriert. Daraufhin wurden für jede PDTB-Relation sowohl die Übereinstimmungen mit der semantisch nächsten RST-Relationsgruppe als auch die Relationsgruppe mit den meisten Übereinstimmungen gesucht – insofern diese nicht übereinstimmen.

Dabei fiel auf, dass die Relationsgruppe *elaboration* – wie bereits auch im gesamten Korpus – 54% aller RST-Relationsgruppen ausmacht. Das lässt sich dadurch erklären, dass die Relationsgruppe per Definition sehr viele Unterteilungen besitzt (z.B. *elaboration-additional*, *elaboration-general-specific* oder *elaboration-object-attribute*) und diese vom Parser entsprechend oft annotiert wurden. In Tabelle 3.2 werden daher zunächst alle Übereinstimmungen mit der Gruppe *elaboration*, dann Übereinstimmungen mit der semantisch nächsten Gruppe und schließlich die Gruppe mit den meisten Übereinstimmungen exklusive *elaboration* angezeigt. Die semantisch nächste Gruppe zu *Expansion.Conjunction* ist nach Carlson et al., 2001 *joint* und hat deshalb keinen Wert.

PDTB	elaboration	semantisch	Anteil	beste	Anteil
Comparison.Concession	63%	contrast	0%	background	43%
Comparison.Contrast	53%	**contrast**	48%	**contrast**	48%
Contingency.Condition	52%	**condition**	43%	**condition**	43%
Expansion.Alternative	37%	condition	0%	attribution	21%
Expansion.Restatement	63%	summary	0%	topic-comment	25%
Temporal.Asynchronous	65%	**temporal**	30%	**temporal**	30%
Temporal.Synchrony	54%	temporal	2%	background	44%
Expansion.Conjunction	49%	joint	-	attribution	25%
Contingency.Cause	56%	cause	9%	explanation	23%

Tabelle 3.2.: Übereinstimmungen der Relationen

Es zeigt sich, dass nur bei drei Relationen die erwartete Gruppe auch die meisten Übereinstimmungen erzielt. Wie bei Scheffler & Stede, 2016 beschrieben, lassen sich viele

der neu entstandenen Übereinstimmungen mit den CONNs erklären. Zum Beispiel ist für *Temporal.Synchrony* die erwartete Gruppe *temporal* sehr schlecht repräsentiert, die Gruppe *background* trifft jedoch eindeutig am häufigsten ein. Die primär zur Relation zählenden CONNs sind „when“ und „as“, die darauffolgenden Argumente stellen offensichtlich einen Indikator für die Gruppe *background* dar. *Expansion.Restatement* wird mit den CONNs „rather“ oder „indeed“ assoziiert, diese werden ebenfalls in der Gruppe *topic-comment* genannt, die zu großen Teilen aus wörtlicher Rede besteht. *Expansion.Alternative* wurde wiederum explizit wegen seiner CONNs „unless“ und „otherwise“, welche in der RST eine Sonderstellung einnehmen[2], der Gruppe *condition* zugeordnet; dieser Zusammenhang konnte allerdings nicht bestätigt werden. Gruppen, die keine PDTB-Relationen zugeordnet werden konnten, da sie nur < 2% derer Anteile stellten, waren *enablement*, *evaluation*, *manner-means* und *summary*.

Es gelingt nicht, bei diesem Vergleich zwei zusammengehörige Relationen zu finden, welche ähnlich positive Signifikanzwerte haben. Dies ist allerdings auch dem Umstand geschuldet, dass bei der PDTB nur wenige dieser Werte vorliegen. Das beste Ergebnis liefern die neu zugeordneten Relationen *Expansion.Conjunction* und *attribution* mit vier bzw. fünf signifikanten p-Werten, die Bonferroni-korrigiert wurden.

Obwohl keine Signifikanz festgestellt werden kann, ist es möglich, die Pitch Ranges relativ zueinander zu betrachten. Für die drei Relationen, die in Tabelle 3.2 hervorgehoben sind, kommen sich zusätzlich deren Pitch Ranges aus Messung 3 am nächsten, wie in Tabelle 3.3 gezeigt wird.

Relation	**Pitch Range**	
Comparison.Contrast	-0,05	0,05
contrast	-0,06	0,04
Contingency.Condition	-0,09	0,25
condition	-0,11	0,25
Temporal.Synchrony	-0,17	0,10
temporal	-0,24	0,15

Tabelle 3.3.: Vergleich der Pitch Ranges

Die zusätzliche Beobachtung der Differenz zweier voneinander abhängiger Argumente bzw. Nukleus und Satellit brachte bereits einige positive Ergebnisse. Vor allem zwischen den Relationen *Temporal.Synchrony* und *background*, welche nach Tabelle 3.2 eng miteinander verknüpft sind, besteht ein Zusammenhang dieser Beobachtung. Für beide konnte gleichermaßen gezeigt werden, dass die Position des Arguments bzw. seine Funktion die Pitch Range und den durchschnittlichen Pitch beeinflusst.

[2] https://www.sfu.ca/rst/01intro/definitions.html

3.6. Tonhöhenregister

Es stellt sich abschließend zur Auswertung die Frage, ob die zu Beginn erläuterte Registeranalyse auf den erhobenen Daten möglich ist. Grundsätzlich unterscheiden sich die Voraussetzungen in einigen Punkten; es bestehen hier beispielsweise deutlich umfangreichere Datensätze, die zudem über Z-Scores normalisiert sind. Die Normalisierung hilft zwar dabei, eine bessere Aufteilung in hohe und tiefe Register vornehmen zu können, verändert das Erscheinungsbild dennoch merklich.

Ein weiterer Unterschied besteht in der Art der untersuchten Diskursrelationen; Mayer,1999 bezieht sich auf eine Topikstruktur, welche ihren Ursprung ihren der DRT findet und nicht komplett mit den genutzten Diskursmodellen darstellbar ist. Dies trifft vor allem für die PDTB-Darstellung zu, in der RST sind einer Topikstruktur verwandte Relationsgruppen wie *topic-comment*, *topic-change* oder auch *background* und *elaboration*. In Möhler & Mayer, 2001 wird unter anderem auf die RST zurückgegriffen, um ähnliche Untersuchungen durchzuführen.

Als Kriterium für die Einteilung auf einem Primärregister kann für jede Relation der Mittelwert aller Z-Scores auf das Vorzeichen überprüft werden; ist es negativ, findet eine Zuordnung zum {l}-Register statt und umgekehrt zum {h}-Register. Für die Subregister ist wiederum ein Schwellenwert (z.B. ein Z-Score von 0,7) möglich, der anzeigt, wie groß der Anteil des positiven oder negativen Pitch Range sein muss. Liegt dieser Anteil über dem Schwellenwert, wird das jeweilige Subregister {h} oder {l} vergeben. Relationen mit ausgedehnten Pitch Ranges können analog zu Mayer,1999 beide Primärregister zugewiesen werden; hier kann als Kriterium zum Beispiel ein Z-Score von mindestens 1 und -1 auf beiden Seiten gelten.

Unter den aufgestellten Bedingungen könnten beispielsweise folgende Zuordnungen erfolgen:

manner-means = {l,l}

background = {l};

topic-comment = {h};

cause = {h,h};

Zuordnungen unter diesen Bedingungen gelten allerdings nur für Messung 3, für andere Messungen müssen folglich andere Kriterien gelten. Zieht man die Hypothesen von Mayer zur Position von Intonationsphrasen in Betracht, lassen sich aus den gewonnenen Erkenntnissen zum Verhalten von Nukleus und Satellit auch messungsunabhängige Register vergeben. Relationen, bei denen eine sinkende Pitch Range im Satelliten zu erwarten ist, könnten demnach für diesen das Subregister {l} und für den Nukleus das Subregister {h} vergeben.

4. Zusammenfassung

In dieser Arbeit wurde eine detaillierte Analyse von Diskursstrukturen zweier verschiedener Diskursmodelle und ihrem Zusammenhang mit der Pitch Range der SprecherInnen auf einem selbst zusammengestellten Korpus durchgeführt. Das Ziel war es dabei, anhand der Pitch Range Differenzierungen zwischen den einzelnen Diskursrelationen dieser beiden Modelle zu finden und daraus Aufschluss über deren Wirkung auf die Prosodie der SprecherInnen zu erhalten.

Das aus Hörbüchern aufgebaute Korpus hat eine ausgeglichene Anzahl an männlichen und weiblichen Sprechern, die Texte entstammten dem vergangenen Jahrhundert, wurden aber so ausgewählt, dass sie mit einem Parser annotiert werden konnten. Damit eine Verbindung zwischen Text und Sprachsignal hergestellt werden konnte, wurden die Hörbücher mittels eines Aligners und der Software Festival in das passende Format gebracht; es kamen allerdings einige Annotationsfehler vor.

Die Texte wurden von jeweils einem Parser eingelesen, die Ausgaben wurden anschließend in ein auswertbares Format konvertiert und zusammengefügt. Es wurden Parser der Penn Discourse Tree Bank und der Rhetorical Structure Theory untersucht.

Die Penn Discourse Tree Bank stellt einen lexikalischen Ansatz auf Basis von Verbindungswörtern dar, Diskursstrukturen unabhängig voneinander zu annotieren. Konzept und Parser waren intuitiv verständlich, jedoch war die Datenverarbeitung des Parsers komplizierter. Die Rhetorical Structure Theory ist ein bewährtes Konzept zur Darstellung von Strukturen in Texten. Sie werden in einer Hierarchie angegeben, welche den gesamten Text in Abhängigkeit voneinander beschreibt. Hier stellte sich das Konzept als komplexer dar, die Verwendung des Parsers und seines Outputs hingegen als einfach.

Die Auswertung sollte die einzelnen Diskursrelationen als Pitch Range auf ihren Strukturen in den Texten darstellen. Gleichzeitig sollte eine statistische Studie zu signifikanten Unterschieden zwischen ihnen sowie ihren Argumenten eine Abgrenzung ermöglichen. Dies gelang im Falle der PDTB nur bedingt, bei der RST konnten ein paar signifikante Unterscheidungen gefunden werden. Die Analyse der Argumente zeigte einen signifikanten Zusammenhang zwischen Pitch Range und Art des Diskursarguments.

Diese Beobachtungen wurden anschließend in einem direkten Vergleich beider Parser-Annotationen evaluiert und zusammengefasst. Außerdem wurde ein Versuch unternommen, die Ergebnisse in das vorgestellte Prinzip der Tonhöhenregister einzuordnen.

Ausblick

Die Untersuchungen haben einige Einschränkungen und Möglichkeiten aufgezeigt, welche in zukünftigen vergleichbaren Studien verbessert oder erweitert werden können. Das Korpus hat keine optimale Datenbank dargestellt, da es nicht aus kontrollierten Aufnahmen und Texten besteht und so einige Unausgeglichenheiten und Fehler enthalten sein können. Dies resultierte in teilweise fehlerhaft annotierten Daten, ein Umstand, der mit anderen Korpora und SprecherInnen verbessert werden kann. Zudem fehlten korrigierende Elemente in der Auswertung, wie zum Beispiel, dass derselbe Text von zwei verschiedenen SprecherInnen vorgelesen wird.

Es wäre zudem möglich gewesen, weitere gemittelte Faktoren zur Differenzierung der Relationen heranzuziehen, wie zum Beispiel Pausendauer, Sprechgeschwindigkeit oder Lautstärke. Eine zusätzliche Erweiterung, welche es vom Umfang her nicht in diese Arbeit geschafft hatte, stellt die Analyse von wörtlicher Rede dar; es wurde gezeigt, dass das Verstellen der Stimme bei unterschiedlichen Charakteren zu einer großen Varianz in den Ergebnissen führen kann.

Obwohl diese Beziehungen selbst für die englische Sprache nach wie vor sehr vage sind, wären zusätzliche Untersuchungen zu anderen Sprachen interessant, die durch eine abweichende Prosodie womöglich ausgeprägtere Ergebnisse erzielen. Das Finden von weiteren Zusammenhängen würde dazu beitragen, die Prosodie von Sprachsynthesen in Erzählungen und Dialogen deutlich zu verbessern.

5. Literatur

Asher, Nicholas.: Reference to Abstract Objects in Discourse, Dordrecht: Kluwer Academic Publishers 1993.

Ayers, G.M.: Discourse functions of pitch range in spontaneous and read speech, Ohio State University Working Papers in Linguistics 44, 1994, S.1-49.

Baken, R. J.: Clinical Measurement of Speech and Voice. London: Taylor and Francis 1987

Bluhm, Claudia; Deissler, Dirk; Scharloth, Joachim; Stukenbrock, Anja: Linguistische Diskursanalyse: Überblick, Probleme, Perspektiven. In: Sprache und Literatur in Wissenschaft und Unterricht 86 (2000). Paderborn, München: Schöningh, S. 3-19

Boersma, Paul: Praat, a system for doing phonetics by computer. Glot international 5.9/10 (2002), S. 341-345.

Braud, Chloe; Coavoux, Maximin; Søgaard, Anders: Cross-lingual RST Discourse Parsing. Univ. of Copenhagen 2017

Carlson, Lynn; Marcu, Daniel; Okurowski, Mary Ellen: Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In: Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, Vol 16 (2001), S. 1-10

Carlson, R., Elenius K. and Swerts M.: Perceptual Judgments of Pitch Range. Proc Speech Prosody 2004

Clements, G.N.: The hierarchical representation of tone features. In: I.R. Dihoff (Hrsg.): Current approaches to African linguistics vol 1. Foris: Dordrecht. 1981, S. 145-176

Feng, Vanessa Wie; Ziheng Lin; Hirst, Graeme: The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014), Dublin, Ireland 2014.

Foucault, Michel: Archäologie des Wissens. Frankfurt am Main: Suhrkamp 1973

Grosz, Barbara J., Sidner, Candace L.: Attention, intentions, and the structure of discourse. In: Computational Linguistics 12/3 1986. Cambridge, MA: MIT Press S. 175-204

Gustafson-Čapková, Sofia: Integrating Prosody into an Account of Discourse Structure. Stockholm: Stockholm Univ. 2005

Hernault, Hugo; Prendinger, Helmut; du Verle, David A.; Ishizuka, Mitsuru: HILDA: A discourse parser using support vector machine classification. In: Dialogue and Discourse, 1/2010, S. 1–33.

Hess, Wolfgang: Pitch Determination of Speech Signals. Berlin: Springer, 1983

Kamp, Hans, Uwe Reyle: From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Boston, MA: Springer 1993

Kamp, Hans: A theory of Truth and Semantic Representation. In: P. Portner & B. H. Partee (Hrsg.): Formal Semantics - the Essential Readings. Blackwell, 1981 S. 189-222

Klautau, Aldebaro (2001). ARPABET and the TIMIT alphabet, [http://speech.ucsd.edu/aldebaro/papers]

Li S.Z., Jain A. (Hrsg.): Fundamental Frequency, Pitch, F0. In: Encyclopedia of Biometrics., Boston, MA: Springer 2009

Mann, William C.; Thompson, Sandra A: Rhetorical Structure Theory: Toward a functional theory of text organization. In: Text 8(3) 1988, S. 243-281.

Mann, William C.; Thompson, Sandra A.: Rhetorical Structure Theory: description and construction of text structures. Information Sciences Institute, Nijmegen, The Netherlands, 1986, S. 1-15.

Marcus, Mitchell P.; Santorini, Beatrice; Marcinkiewicz, Mary Ann: Building a Large Annotated Corpus of English: The Penn Treebank. In: Computational Linguistics - Special issue on using large corpora: II. Volume 19 Issue 2, 1993. MIT Press Cambridge, MA, S. 313-330

Mayer, Jörg: Prosodische Merkmale von Diskursrelationen. In: Linguistische Berichte, Vol. 177 (1999). Opladen: Westdeutscher Verlag, S. 65–86

Miltsakaki, Eleni; Prasad, Rashmi; Joshi, Aravind; Webber, Bonnie: The Penn Discourse Treebank. In Proceedings of the Language Resources and Evaluation Conference. Lisbon, Portugal 2004.

Möhler, Gregor, Mayer, Jörg: A Discourse Model for Pitch-Range Control. SSW, 2001

Pierrehumbert, Janet B.: The Phonology and Phonetics of English Intonation. Bloomington: Indiana Univ. 1987

Pierrehumbert, Janet und Hirschberg, Julia: The Meaning of Intonational Contours in the Interpretation of Discourse. In P. Cohen, J. Morgan, and M. Pollack, (Hrsg.) Intentions in Communication. Cambridge MA., Bradford Books, MIT Press, 1990, S. 271-311.

Poláková, Lucie; Jirí Mírovský, Pavlína Synková: Signalling Implicit Relations: A PDTB – RST Comparison, in Dialogue Discourse 8(2), 2017, S. 225-248

Prasad, Rashmi; Dinesh, Nikhil; Lee, Alan; Miltsakaki, Eleni; Robaldo, Livio; Joshi, Aravind; Webber, Bonnie: The Penn Discourse Treebank 2.0. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco, 2008.

Scheffler, Tatjana & Manfred Stede: Mapping PDTB-style connective annotationto RST-style discourse annotation, in: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), 2016

Schweitzer, Antje: Building a Unit Selection Synthesis Voice, 2018

Surdeanu, Mihai; Hicks, Thomas; Valenzuela-Escarcega, Marco A: Two Practical Rhetorical Structure Theory Parsers. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: Software Demonstrations (NAACL HLT), 2015.

Taboada, Maite; Mann, William C.: Rhetorical structure theory: looking back and moving ahead. In: Discourse Studies. 8 (3) 2006, S. 423–459.

Wang, Jianxiang and Man Lan: A Refined End-to-End Discourse Parser, 2015

Webber, Bonnie: Structure and Ostension in the Interpretation of Discourse Deixis. In: Language and Cognitive Processes, 6:2 (1991). London: Taylor and Francis, S. 107-135.

Wennerstrom, Ann: The Music of Everyday Speech: Prosody and Discourse Analysis. Oxford: Oxford University Press 2001

Xue, Nianwen; Hwee Tou Ng; Sameer Pradhan; Rashmi Prasad; Christopher Bryant; Attapol T. Rutherford: The CoNLL-2015 Shared Task on Shallow Discourse Parsing. (2015)

Ziheng Lin; Hwee Tou Ng; Min-Yen Kan: A PDTB-Styled End-to-End Discourse Parser. Natural Language Engineering, Cambridge University Press, 20 (2014), S. 151-184.

Zulaica, Iker: Resolving abstract anaphors in Spanish discourse: Underspecification and mereological structures.

Alle URLs wurden zuletzt am 24. 08. 2019 geprüft.

A. Anhang

Listing A.1 Angleichung der Parser-Tokenisierung an die Aligner-Tokenisierung

```
 cat v1 | sed s/'\.\.'/'\t'/g | column -t > v2
 cat v2 | sed s/'\-'/'\t'/g > v3
 cat v3 | sed 's/'.txt'//g' > v4
 cat v4 | sed '/'[[:punct:]]'/d' > v5
//Dos2Unix
 sed -e "s/\r//g" v5 > tokenized
```

Listing A.2 Extrahieren aller Diskursrelationen

```
// Diskursrelationen (ohne Argumente) extrahieren
awk '{print FILENAME "\t" $0}' *.exp.res | sed 's/'.txt.exp.res'//g' | sed
    's/'\|'/\t /g' | sed 's/\.\./\t/g' > allDiscs

//Offsets der Argumente extrahieren
grep "" *.txt.pipe | tr -s [A-Za-z] ' ' | sed s/'| [[:punct:]]* |'//g | sed s/'
    [[:punct:]] '//g | sed s/'|[[:punct:]]*'/'|'/g | sed s/'| [[:punct:]]*'/'|'/g
    | tr '[\, \" \-]' ' ' | tr ';' '^' | tr '| |' '\t'| sed s/'.:'//g > allPipes
```

Listing A.3 Basic-Skript zur Identifikation aller zugeordneten Token

```
Attribute VB_Name = "Modul1"
Sub deleterows()

Dim lastrow
lastrow = ThisWorkbook.Sheets(1).Range("A194005").End(xlUp).Row
Dim temp As Long

'iteriert die Tabelle von unten nach oben
For i = lastrow To 2 Step -1

      'Ueberprueft, ob die WordNo absteigend ist oder gleich
   If (ThisWorkbook.Sheets(1).Cells(i, 2).Value - ThisWorkbook.Sheets(1).Cells(i
       - 1, 2).Value > 1) Or (ThisWorkbook.Sheets(1).Cells(i, 2).Value -
       ThisWorkbook.Sheets(1).Cells(i - 1, 2).Value < 0) Then

            'loesche ueberzaehlige Zeilen
            ThisWorkbook.Sheets(1).Rows(i - 1).EntireRow.Delete
      temp = temp + 1
   Else
       temp = 0
   End If

      'Abbruchbedingung bei falschem Alignment
   If temp = 40 Then
      Exit Sub
   End If

Next i


End Sub
```

Listing A.4 Main-Methode des Java-Codes, der die Annotation des RST-Parsers für alle Textdateien erstellt.

```
public static void main(String[] args) throws IOException {
      //set input directory
      File folder = new File("PathToDirectory");
      File[] list = folder.listFiles();

      for (int i = 0; i < listOfFiles.length; i++) {
            if (list[i].isFile()) {
                  String name = list[i].getName();
                  PrintStream out = new PrintStream(new
                      FileOutputStream(name));
                  System.setOut(out);

                  // read file to string
                  String content = new String(
                              Files.readAllBytes(Paths.get("PathToDirectory"
                                  + name)),
                              Charset.forName("UTF-8"));

            /*
             * CLU Lab Processors Discourse Parse
             * Surdeanu et al.
             * 2015
             * https://github.com/clulab/processors
             */

                  // create the processor
                  Processor proc = new CoreNLPProcessor(true, false, 1,
                      100);

                  // the actual work is done here
                  Document doc = proc.annotate(content, true);

                  [...]

                  out.close();
            }
      }
}
```

Listing A.5 Erstellen einer Offset-Tabelle

```
//Konsolenausgabe am Anfang ausschneiden
for x in *.txt; do tail -n +6 <"$x" >"$x.tmp"; mv "$x.tmp" "$x"; done

//Tokens und Offsets extrahieren...
for x in *.txt; do sed -e '/Tokens:/,/Start character offsets:/!d' <"$x"
    >"$x.tmp"; mv "$x.tmp" "$x.token"; done

//...und in neues File schreiben
for x in *.txt; do sed 's/Tokens: //g' | sed 's/Start character offsets: //g' |
    tr " " "\t" <"$x.token" >"$x.tmp"; mv "$x.tmp" "$x.token.offsets"; done

//konkatenierte Offsets in eine transponierte Tabelle umwandeln
 less all | sed -e "s/\r//g" | sed -e 's/[0-9] /&\n/g' | less
```

Listing A.6 Workflow zur Tokenisierung, Teil 1

```
//1. Loeschen aller Sonderzeichen
[!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~.]

//2. behaelt die Einrueckung der Hierarchie bei
<Macro name="Indent" Ctrl="no" Alt="no" Shift="no" Key="0">
         <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
         <Action type="3" message="1601" wParam="0" lParam="0" sParam=" " />
         <Action type="3" message="1625" wParam="0" lParam="2" sParam="" />
         <Action type="3" message="1702" wParam="0" lParam="1794" sParam="" />
         <Action type="3" message="1701" wParam="0" lParam="1" sParam="" />
         <Action type="1" message="2003" wParam="119" lParam="0"
             sParam="&#x0009;" />
         <Action type="1" message="2003" wParam="110" lParam="0" sParam="" />
         <Action type="1" message="2170" wParam="0" lParam="0" sParam="&#x000A;"
             />
      </Macro>
```

Listing A.7 Workflow zur Tokenisierung, Teil 2

```
//3. Erstellt aus den EDUs eine Wortliste inklusive der Relationen und loescht
    das erste Wort der EDU
<Macro name="CopyLine" Ctrl="no" Alt="no" Shift="no" Key="0">
         <Action type="0" message="2452" wParam="0" lParam="0" sParam="" />
         <Action type="0" message="2452" wParam="0" lParam="0" sParam="" />
         <Action type="0" message="2178" wParam="0" lParam="0" sParam="" />
         <Action type="0" message="2306" wParam="0" lParam="0" sParam="" />
         <Action type="1" message="2170" wParam="0" lParam="0" sParam="&#x000D;"
            />
         <Action type="1" message="2003" wParam="191" lParam="0" sParam="" />
         <Action type="1" message="2003" wParam="0" lParam="0" sParam="" />
         <Action type="1" message="2170" wParam="0" lParam="0" sParam="&#x000A;"
            />
         <Action type="0" message="2179" wParam="0" lParam="0" sParam="" />
         <Action type="0" message="2453" wParam="0" lParam="0" sParam="" />
         <Action type="0" message="2453" wParam="0" lParam="0" sParam="" />
         <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
         <Action type="3" message="1601" wParam="0" lParam="0"
            sParam='[A-Za-z:&quot;,;.0-9]* ' />
         <Action type="3" message="1625" wParam="0" lParam="2" sParam="" />
         <Action type="3" message="1602" wParam="0" lParam="0" sParam="" />
         <Action type="3" message="1702" wParam="0" lParam="1792" sParam="" />
         <Action type="3" message="1701" wParam="0" lParam="1608" sParam="" />
         <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
         <Action type="3" message="1601" wParam="0" lParam="0"
            sParam='[A-Za-z:&quot;,;.0-9]* ' />
         <Action type="3" message="1625" wParam="0" lParam="2" sParam="" />
         <Action type="3" message="1602" wParam="0" lParam="0" sParam="" />
         <Action type="3" message="1702" wParam="0" lParam="1792" sParam="" />
         <Action type="3" message="1701" wParam="0" lParam="1608" sParam="" />
         <Action type="0" message="2453" wParam="0" lParam="0" sParam="" />
         <Action type="0" message="2453" wParam="0" lParam="0" sParam="" />
      </Macro>

//4. Loescht jeweils die letzten (n-1) Worte
( ([A-Za-z0-9:]*)+)+\t*$

//5. Extrahiert alle Zeilen, deren EDU nicht-leer ist
<Macro name="RegExtract" Ctrl="no" Alt="no" Shift="no" Key="0">
         <Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
         <Action type="3" message="1601" wParam="0" lParam="0"
            sParam="([A-Za-z\(\)-]*\t)*TEXT[A-Za-z -_!?.,]*(\t)*" />
         <Action type="3" message="1625" wParam="0" lParam="2" sParam="" />
         <Action type="3" message="1702" wParam="0" lParam="1810" sParam="" />
         <Action type="3" message="1701" wParam="0" lParam="1615" sParam="" />
         <Action type="2" message="0" wParam="43051" lParam="0" sParam="" />
      </Macro>

//6. Loescht ueberzaehlige Leerzeichen
([A-Za-z]*\t)+\t*$\r\n
```

Listing A.8 Reduktion des RST-Baums

```
//1. Entfernen der Blattknoten der Token
\t[A-Za-z0-9]*\t*\r\n

//Fuehre jeweils 3x durch:
//2. RST-Blattknoten abfragen und kopieren
^([A-Za-z]*\t)*

//3. RST-Blattknoten entfernen
\t*[A-Za-z]+$
```

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift