Die Annotation von rhetorischen Strukturen im Kobalt-DaF-Korpus

Version 1.0

Shujun Wan
Humboldt-Universität zu Berlin

Zusammenfassung

Der vorliegende Beitrag bezieht sich auf den Annotationsprozess von rhetorischen Strukturen der argumentativen Texte im Kobalt-DaF-Korpus. Die folgenden Fragestellungen werden vorwiegend beantwortet: Vor welchem Forschungshintergrund wurde die Anntation vorgenommen? Welche Daten wurden annotiert? Wie wurden sie annotiert (Annotationsframework, -richtlinie und -verfahren)? Die Beantwortung dieser Fragen zielt darauf ab, (1) eine Referenz für weitere Annotationsarbeit bzgl. rhetorischer Strukturen zu bieten und (2) weitere linguistische Forschungen zur Lernersprache auf der Diskursebene mit dem Kobalt-DaF-Korpus zu ermöglichen.

Inhaltsverzeichnis

1 Hintergrund

Die Annotation der rhetorischen Strukturen erfolgte im Rahmen einer Doktorarbeit (Wan, Vorb), bei der es sich um eine korpusbasierte kontrastive Studie zu Argumentationstrategien von chinesischen Deutschlerner:innen und deutschen L1-Sprecher:innen in argumentativen Texten handelt. Einer der wichtigsten Bestandteile der Studie ist dabei, die rhetorischen Strukturen der Texte zu erforschen. Dafür sollten die Texte im Hinblick auf ihre rhetorischen Strukturen annotiert werden.

2 Das Kobalt-DaF-Korpus

Das Kobalt-DaF-Korpus[1] ist ein systematisch erhobenes Deutschlernerkorpus. Es ist frei zugänglich und kann durch den Suchwerkzeug *ANNIS* online direkt abgefragt werden. Insgesamt stehen 80 deutschsprachige argumentative Texte zur Verfügung: 60 Texte von fortgeschrittenen[2] Deutschlerner:innen aus drei Gruppen mit unterschiedlichen L1 (Chinesisch, Russisch und Schwedisch – jeweils 20 Texte) und 20 Texte von deutschen L1-Sprecher:innen. Alle Texte befassen sich mit demselben Thema „Geht es der Jugend heute besser als früher?“. Initial wurde das Kobalt-DaF-Korpus bereits mit folgenden Annotationsebenen annotiert: Wortarten und Lemmata (Schiller et al., 1999), Grammatische Funktionen (Brants et al., 2004), Topologische Felder (Telljohann et al., 2006) und Zielhypothesen und Fehlertags (Reznicek et al., 2012).

Angesichts des Forschungsziels wurden erstmal 40 der 80 argumentativen Texte (20 von den chinesischen Deutschlerner:innen und 20 von den deutschen L1-Sprecher:innen) aus dem Kobalt-DaF-Korpus herausgezogen[3]. Eine Übersicht über die grundlegenden Parameter dieser Texte findet sich in Tabelle 1.

[1] Mehr Infomationen befinden sich unter https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/kobalt-daf (Zugriff: 20.11.2021).

[2] Mindestvoraussetzung für die Teilnahme war Erreichen des B2-Niveaus im onDaF-Test

[3] All die 40 Texte wurden bereits mit Zielhypothesen annotiert. Das heißt, dass mögliche Fehler in den Texten sowohl niederiger sprachlicher Ebenen (z.B. Orthografie) als auch höherer Ebenen (z.B. Semantik) schon korrigiert wurden. Mehr Infos über das Thema Zielhypothesen findet sich in der Richtlinie der Falko-Korpusfamilie (Reznicek et al., 2012)

L1	Genre	Anzahl der Texte	Textlänge i.D.	Token
Chinesisch	argum. Texte	20	521	13949
Deutsch	argum. Texte	20	488	12984

Tabelle 1: Parameter der für die Studie adoptierten Texte aus dem Kobalt

3 Annotation

Obwohl das Kobalt-DaF-Korpus bereits tief annotiert wurde, können solche vorhandenen Annotationen natürlich nicht allen Forschungszwecken gerecht werden. Um die rhetorischen Strukturen der Texte eingehend zu erforschen, muss eine zusätzliche Annotationsebene diesbezüglich hinzufügt werden.

3.1 Annotationsframework: Rhetorical Structure Theory

Zur Beschreibung der rhetorischen Struktur eines Texts wurde hier das funktionale Framework *Rhetorical Structure Theory* (Abk. RST) (Mann und Thompson, 1988; Taboada und Mann, 2006) eingesetzt. Die RST geht davon aus, dass die Organisation eines Textes durch eine Baumstruktur dargestellt werden kann. Drei Bestandteile sind für die Zusammenstellung des Baumes besonders wichtig: (1) die Blätter - Minimale Diskurseinheiten (auf Englisch *Elementary Discourse Units*, Abk. EDUs), (2) die internen Knoten - Rhetorische Relationen, und (3) die Struktur - Schemata.

Abb. 1 zeigt ein RST-Analysebeispiel eines Textauszugs aus dem Kobalt-DaF-Korpus, visualisiert mithilfe des Werkzeugs *RSTWeb* (Zeldes, 2016).

3.2 Annotationsrichtlinie

Grundsätzlich orientiert sich die Annotationsrichtlinie an der Richtlinie des *Potsdamer Commentary Corpus* (Abk. PCC) (Stede, 2016), die speziell für deutschsprachige Meinungstexte entwickelt wurde und somit sowohl der Textsprache als auch der Textsorte der Korpusdaten entspricht. Der PCC-Richtlinie zufolge wird die gesamte Annotationsarbeit in *Segmentierung der EDUs* und *Annotation der rhetorischen Strukturen* eingeteilt.

3.2.1 Segmentierung der EDUs

Das vollständiges Segmentierungsverfahren befindet sich in der PCC-Richtlinie (Stede, 2016). Zusammengefasst verläuft die Segmentierung der Texte in EDUs gemäß

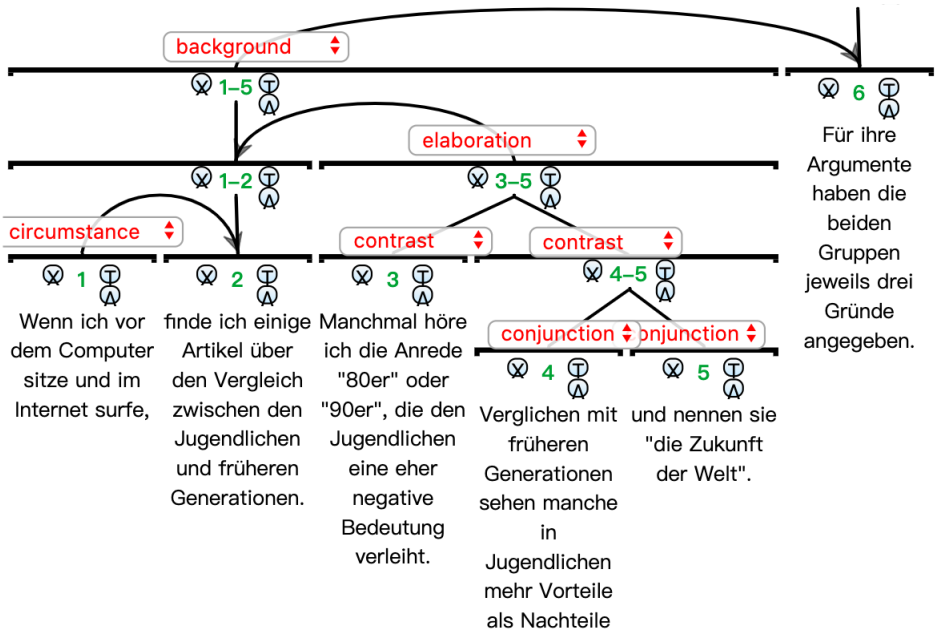

Abb. 1: Ein RST-Analysebeispiel (Textauszug aus KobaltCMN017)

der PCC-Richtlinie in den folgenden vier Schritten: (1) Teilung in Sinneinheiten, (2) Unterteilung der Sinneinheiten in Segmente, (3) Unterteilung zusammengesetzter Sätze, und (4) Partitionierung in EDU. Am Ende des gesamten Segmentierungsprozesses sollte ein Text vollständig in eine Sequenz von EDUs aufgeteilt sein.

3.2.2 Annotation der rhetorischen Strukturen

Die Annotation der rhetorischen Strukturen setzt der Segmentierung voraus. Sie liegt ebenfalls der PCC-Richtlinie (Stede, 2016) zugrunde, jedoch mit Modifizierungen. Nach einigen Problemen, die in der Pilotannotation (s. Kapitel 3.3) entdeckt wurden, wurde die PCC-Richtlinie hauptsächlich bei den folgenden zwei Punkten angepasst.

1. **Die Ergänzung der Relationen *Question* und *Solutionhood-S***

Germäß der PCC-Richtlinie werden insgesamt 31 rhetorische Relationen definiert. Nach der inhaltlichen und strukturellen Unterscheidung gliedern sich die Relationen in vier Gruppen: Primär pragmatische Relationen beschreiben die Argumentation des Autors. Primär semantische Relationen kommen zum Einsatz, wenn im Text ein (komplexer) Sachverhalt der Welt lediglich beschrieben wird. Textuelle Relationen erfüllen einen organisierenden Zweck. Während die ersten drei Grup-

pen als mononukleare Relationen einzustufen sind, werden Relationen mit zwei oder mehr Nuklei von der vierten Gruppe multinukleare Relationen umfasst.

In der Pilotannotation wurde allerdings entdeckt, dass noch eine Relation fehlt, die die rhetorische Relation zwischen einer Fragestellung und der entsprechenden Antwort/den Antworten beschreibt. Diese Relation wird als *Question* bezeichnet und ist zur Kategorie der primär pragmatischen Relationen zuzuordnen, erfüllt aber nebenbei auch eine strukturell organisierende Funktion.

Question

- N: Eine subjektive These/Aussage/Einschätzung.
- S: Eine Fragestellung.
- N/S: Der Inhalt von N kann als „Antwort“ auf die in S gestellte Frage aufgefasst werden. S geht N normalerweise im Text voraus.
- Effekt: Durch N erhält der Leser eine Antwort auf S.
- Typische Konnektoren: selten durch Konnektoren angezeigt; gelegentlich durch das Interpunktionszeichen Fragezeichen.
- Beispiel: [Wie sieht die Situation bei den heutigen Jugendlichen aus?]$_{\mathrm{S}}$ [Diese Fragestellung kann man aus verschiedenen Perspektiven beantworten:]$_{\mathrm{N}}$

Darüber hinaus tauchte ein weiteres Problem auf: Die originale Definition der Relation „Solutionhood“ umfasst nur die Situation, in der lediglich der Satellit als das „Problem“ betrachtet werden darf, während der Nukleus die „Lösung“ des Problems darstellt.

Bsp. [Mit der Verabschiedung des Nichtraucherschutzgesetzes sitzen viele Kneipen in der Falle.]$_{\mathrm{S}}$ [Es empfiehlt sich, früh genug auf die Einrichtung abtrennbarer Räume zu achten.]$_{\mathrm{N}}$ (PCC-Richtlinie)

Das folgende Beispiel zeigt jedoch, dass es noch eine Möglichkeit gibt, bei der das „Problem“ eine wichtigere Rolle als die „Lösung“ übernimmt:

Bsp. [Trotz einigen guten Charaktereigenschaften bei der heutigen Jugend ist sie aber schlechter als frühere Generationen.]$_{\mathrm{N}}$ [Die heutige Jugend sollte vieles von früheren Generationen lernen.]$_{\mathrm{S}}$ (KobaltCMN010)

In diesem Fall repräsentiert der „Problemteil“ - „die heutige Jugend ist schlechter als frühere Generationen“ - die zentrale Meinung der Autors. Insofern sollte sie gemäß der *centrality for the author's purposes* sowie der These *strong nuclearity principle* als Nukleus betrachtet werden. Demzufolge haben wir die Definition der Relation „Solutionhood“ modifiziert, und sie je nach dem Status des Nukleus in zwei Relationen gliedert: *Solutionhood-N* und *Solutionhood-S*. Während die Relation *Solutionhood-N* mit der originalen Relation *Solutionhood* identisch ist, notiert *Solutionhood-S* die Relation, in der der Nukleus das „Problem“ darstellt und der Satellit die „Lösung“ erläutert.

Solutionhood-S

- N: Der Inhalt von N kann als „Problem“ aufgefasst werden.
- N/S: S kann als Lösung des in N dargestellten Problems aufgefasst werden.
- Effekt: Der Leser erkennt S als Lösung des Problems in N.
- Typische Konnektoren: selten durch Konnektoren angezeigt.
- Beispiel: [Meiner Ansicht nach geht es der Jugend heute aber nicht besser als früheren Generationen.]$_{\mathrm{N}}$ [Wenn die Jugend ein besseres führen soll, muss man zuerst diese Probleme lösen.]$_{\mathrm{S}}$ (KobaltCMN009)

Zusammenfassend lässt sich sagen, dass die modifizierte Richtlinie schließlich insgesamt 33 rhetorische Relationen umfasst. Für einen Überblick aller dieser Relationen siehe Tabelle 2. Die Definitionen der übrigen Relationen befinden sich in der PCC-Richtlinie.

2. **Genauere Differenzierung verwandter Relationen**

Bereits in der PCC-Richtlinie wird auf die Differenzierung einiger bedeutungsähnlicher Relationen hingewiesen. Allerdings wurde darüber hinaus bei der Annotation der Probetexte entdeckt, dass es noch einige Relationen gibt, wo in der Praxis tatsächlich ebenfalls eine Entscheidung schwierig fallen mag, weil diese Relationen für bestimmte Kontexte zugleich gelten können. Diese sind:

- *Evidence* & *Elaboration*
- *Evidence* & *Reason*

Primär pragmatische Relationen	**Primär semantische Relationen**	**Textuelle Relationen**	**Multinukleare Relationen**
Background	Circumstance	Preparation	Contrast
Antithesis	Condition	Restatement	Sequence
Concession	Otherwise	Summary	List
Evidence	Unless		Conjunction
Reason	Elaboration		Joint
Reason-N	E-Elaboration		
Justify	Interpretation		
Evaluation-S	Means		
Evaluation-N	Cause		
Motivation	Result		
Enablement	Purpose		
Question	Solutionhood-N		
	Solutionhood-S		

Tabelle 2: Überblick der rhetorischen Relationen der in dieser Forschung verwendeten Annotationsrichtlinie

- *List & Cause*

Eine Entscheidung zwischen diesen rhetorischen Relationen sollte dann mit besonderer Sorgfalt, im Kontext und unter Berücksichtigung der Richtlinie getroffen werden. Ein Beispielfall dafür findet sich unten:

Bsp. [Auch der Begriff der Familie rückt für die meisten immer mehr in den Hintergrund und die Karriere gewinnt immer mehr an Bedeutung.]$_{\text{N}}$ [Generell lässt sich beobachten, dass Geld und Macht über die Zeit an Bedeutung gewonnen haben.]$_{\text{S}}$ (DEU001[4])

Laut den Definitionen beschreibt *Evidence* eine subjektive Aussage, die „objektiv“ dargestellt wird, während *Elaboration* genauere Informationen zu dem Nukleus liefert. In diesem Beispielfall kann man den Satellit - Geld und Macht haben an Bedeutung gewonnen - als eine Erklärung und Ergänzung zu dem Nukleus - die Karriere gewinnt immer mehr an Bedeutung - verstehen. Allerdings kann man den Satellit auch als einen Beweis für den Fall im Nukleus - Familie rückt immer mehr in den

[4]Der Beispielsatz hier wurde ursprünglich in verschiedenen EDUs segmentiert, und mit unterschiedlichen Relationen hierarchisch verbunden. Allerdings werden lediglich die zwei EDUs, die mit den betroffenen Relation verbunden sind, gezeigt.

Hintergrund - sehen. Laut der PCC-Richtlinie haben pragmatische Relationen in der Regel Vorrang, das heißt, wenn beide *Evidence* und *Elaboration* bestehen, liegt die Priorität auf *Evidence*.

Abgesehen von den beiden oben genannten Modifizierungen folgt die Annotation der rhetorischen Strukturen streng der PCC-Richtlinie.

3.3 Annotationsverfahren

Der RST-Analyse ist eine gewisse Subjektivität inhärent (Stede, 2016). Damit die Daten aber trotzdem nachvollziehbar und zuverlässig bleiben, wurde die ganze Annotationsarbeit anhand eines ausführlichen Annotatationsplans von zwei fachlich ausgebildeten Linguistinnen manuell durchgeführt. Der Arbeitsprozess als Ganzes ist in Abb. 2 illustriert.

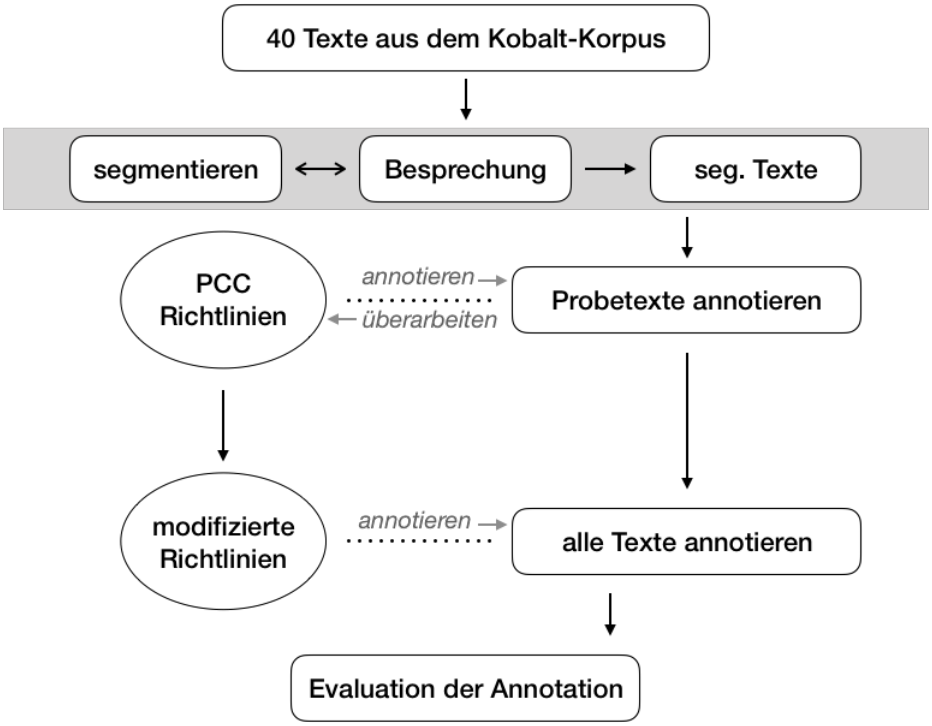

Abb. 2: Arbeitsprozess der Segmentierung und Annotation

Zunächst haben die beiden Annotatorinnen alle 40 Texte aus dem Kobalt-DaF-Korpus anhand der Annotationsrichtlinie getrennt segmentiert. Der Werkzeug *RST-Tool*[5] (O'Donnell, 1997), der vor allem für die RST-Analyse entwickelt wurde, wurde für die Segmentierung eingesetzt. Anschließend wurden die ersten Entwürfe der

[5] http://www.wagsoft.com/RSTTool/ (Zugriff: 20.11.2021)

Segmentierung von den beiden Annotatoren Wort für Wort verglichen und besprochen, weil es im praktischen Vorgehen mit der Richtlinie oft Zweifelssfälle gibt, und ein Austausch zwischen den Annotatorinnen zu einem besseren Verständnis jedes Einzelfalls und zu einem zuverlässigeren Segmentierungsergebnis beitragen kann. Letztendlich entstand die endgültige Version der segmentierten Texte, die dann die Grundlage für die weitergehende Annotation bildeten.

Als nächstes wurde eine Pilotannotation durchgeführt. Die Annotatorinnen haben auf Basis der PCC-Richtlinie zuerst fünf Texte aus dem Kobalt als Probe annotiert. Mit dem Annotationstool *RSTWeb*[6] (Zeldes, 2016), das stetig für die RST-Analyse weiter entwickelt und aktualisiert wird, wurde die Annotation der Probetexte von den beiden Annotatorinnen getrennt und ohne gegenseitige Absprache vorgenommen. Anschließend wurde das Inter-Annotator-Agreement mithilfe des Tools *RST-Tace*[7] (Wan et al., 2019) berechnet[8]. Dies dient dem Ziel, Ungenauigkeiten und Unklarheiten der PCC-Richtlinie zu aufzudecken, um darauf basierend eine besser geeignete Richtlinie festzulegen. Mit einem Wert von Cohens Kappa von $\kappa = 0,5429$ liegt das Inter-Annotator-Agreement der Probetexte im Bereich von 0,41 bis 0,60, was darauf hinweist, dass die Annotationsergebnisse beider Annotatorinnen miteinandern mittelmäßig übereinstimmen (Landis und Koch, 1977)[9]. Das heißt: die originale PCC-Richtlinie ist bereits für die Korpusdaten annehmbar, jedoch gibt es noch Räume für eine Verbessung des Annotationsergebnisses. Wie bereits im Kapitel 3 erwähnt, wurden anschließend zwei Modifizierungen zu der PCC-Richtlinie hinzufügt.

Nach der Pilotannotation und der Festlegung der endgültigen Annotationsrichtlinie haben die Annotatorinnen angefangen, die rhetorischen Strukturen aller Texte zu annotieren. Das Annotationsvorgehen richtet streng nach der Annotationsrichtlinie: Text Lesen -> Thematische Grenzen -> Starke Nuklei wählen -> Lokale EDUs verbinden -> Größere Einheiten verknüpfen -> Hierarchisierung -> Vergleichen und bearbeiten.

---

[6] `https://github.com/amir-zeldes/rstWeb` (Zugriff: 20.11.2021)

[7] `https://github.com/tkutschbach/RST-Tace` (Zugriff: 20.11.2021)

[8] Eine RST-Analyse von *RST-Tace* wird in vier Faktoren evaluiert: (1) Constituent, (2) Attachment Point (A), (3) Nuklearity (N), und (4) Relation (R). Mehr Info darüber s. Wan et al. (2019)

[9] Die Interpretation vom Kappawert ist tatsächlich ziemlich umstritten (Banerjee et al., 1999). Je nach Fachgebiet gibt es auch viele andere Interpretationsmöglichkeiten, wie die von McHugh (2012). Die hier verwendete Interpretation von Landis und Koch (1977) ist eine der am häufigsten eingesetzte Versionen.

Literatur

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeney und Debajyoti Sinha (1999). Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics 27*(1), 3–23.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith und Hans Uszkoreit (2004). TIGER: Linguistic interpretation of a German corpus. *Research on language and computation 2*(4), 597–620.

Landis, J Richard und Gary G Koch (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Mann, William C und Sandra A Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse 8*(3), 243–281.

McHugh, Mary L (2012). Interrater reliability: the kappa statistic. *Biochemia medica 22*(3), 276–282.

O'Donnell, Michael (1997). RST-Tool: An RST analysis tool.

Reznicek, Marc, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann und Torsten Andreas (2012). Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01.

Schiller, Anne, Simone Teufel, Christine Thielen und Christine Stöckert (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Technical Report. Institut für maschinelle Sprachverarbeitung, Stuttgart.

Stede, Manfred (2016). *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*, Bd. 8. Universitätsverlag Potsdam.

Taboada, Maite und William C Mann (2006). Rhetorical structure theory: Looking back and moving ahead. *Discourse studies 8*(3), 423–459.

Telljohann, Heike, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister und Kathrin Beck (2006). Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In *Seminar fur Sprachwissenschaft, Universitat Tubingen, Tubingen, Germany*. Citeseer.

Wan, Shujun (i. Vorb.). *Argumentationsstrategien von chinesischen Deutschlerner/-innen – Eine korpusbasierte Studie im Vergleich zu deutschen Muttersprachler/-innen.* Unveröff.diss, Humboldt-Universitä zu Berlin.

Wan, Shujun, Tino Kutschbach, Anke Lüdeling und Manfred Stede (2019). RST-Tace A tool for automatic comparison and evaluation of RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, S. 88–96.

Zeldes, Amir (2016). rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, S. 1–5.