Department for Linguistics
Universität Potsdam

Disambiguierung von Diskurskonnektoren im Deutschen

Bachelorarbeit

Angela Schneider
Matrikel-Nummer 735923

Betreuer	Prof. Dr. Manfred Stede
Erstprüfer	Prof. Dr. Manfred Stede
Zweitprüfer	Florian Kuhn

Inhaltsverzeichnis

Tabellenverzeichnis

Zusammenfassung

In dieser Arbeit widme ich mich der Disambiguierung von ambigen Diskurskonnektoren im Deutschen mit Hilfe von Part-of-Speech-Tagging und der Betrachtung des unmittelbaren Kontextes. Durch das Anwenden von Part-of-Speech-Regeln und Kontextregeln werden die ambigen Diskurskonnektoren in die Konnektor-Lesart und die Nicht-Konnektor-Lesart klassifiziert.

1 Einleitung

Diskurskonnektoren sind der Klebstoff zwischen den Sätzen eines Textes. Sie verknüpfen Sätze miteinander und setzen die von diesen Sätzen dargestellten Sachverhalte zueinander in Beziehung. Sie sorgen dafür, dass aus einer Aneinanderreihung von Sätzen ein Text entsteht.

Die korrekte automatische Identifizierung von Konnektoren ist für viele computerlinguistische Anwendungen von entscheidender Bedeutung. Zu wissen in welcher Beziehung die durch Sätze repräsentierten Sachverhalte zueinander stehen, helfen beispielsweise bei der Sentiment Analyse (siehe zum Beispiel [TBT$^+$][1]), bei der automatischen Textzusammenfassung (siehe [SL02]) oder bei der automatischen Question Generierung ([ASM11]).

Allerdings ist die eindeutige Identifizierung von Konnektoren nicht einfach. Renate Pasch sagt folgendes über die Identifizierung von Konnektoren[2]:

> Als „**Konnektoren**“ sehen wir im Deutschen diejenigen Ausdrücke x an, die folgende Merkmale (M) aufweisen:
>
> (M1) x ist **nicht flektierbar**.
>
> (M2) x vergibt **keine Kasusmerkmale** an seine syntaktische Umgebung.
>
> (M3) Die Bedeutung von x ist eine **zweistellige Relation**.
>
> (M4) Die **Relate** der Bedeutung von x sind **Sachverhalte**.
>
> (M5) Die **Relate** der Bedeutung von x müssen **durch Sätze bezeichnet** werden können.

1 In dieser Arbeit beziehen sich Buchstaben- und Zahlenkürzel in eckigen Klammern auf Literaturangaben, die im Literaturverzeichnis nachzuschlagen sind.

2 siehe [Pas03] S. 1

Um als Konnektor identifiziert zu werden muss ein Ausdruck also fünf sehr unterschiedliche Merkmale (ein morphologisches, ein syntaktisches, zwei semantische und ein Merkmal der Laut-Bedeutungs-Zuordnung) erfüllen.

Obwohl Präpositionen gegen das Merkmal (M2) verstoßen und Kasusmerkmale vergeben, werden sie manchmal doch als Konnektoren anerkannt. Auch das Merkmal (M5) wird oftmals aufgeweicht, sodass die Sachverhalte nicht unbedingt durch Ausdrücke, die ein finites Verb enthalten, bezeichnet werden müssen, sondern sie müssen ohne den Inhalt zu verfälschen in Sätze umformuliert werden können. Somit wären die Wörter *während* und *trotz* in den Sätzen (1) und (3) Konnektoren, obwohl sie gegen die Merkmale (M2) und (M5) verstoßen.

(1) ***Während*** *ihrer Beerdigung gingen ihm noch mal seine letzten Worte zu ihr durch den Kopf.*

(2) ***Während*** *sie beerdigt wurde, gingen ihm noch mal seine letzten Worte zu ihr durch den Kopf.*

(3) ***Trotz*** *seiner Warnung ging sie allein zu dem Treffen.*

(4) ***Obwohl*** *er sie gewarnt hatte, ging sie allein zu dem Treffen.*

Da sich der Satz (1) zum Satz (2) umformulieren lässt, ohne dass sich etwas an der Bedeutung verändert, kann man das Wort *während* auch im Satz (1) als Konnektor interpretieren, obwohl es Kasusmerkmale vergibt und sich auf eine Nicht-Satz-Phrase bezieht. Satz (3) lässt sich als Satz (4) formulieren. Das Wort *obwohl* in Satz (4) ist ein Konnektor. Dadurch, dass die Phrase *trotz seiner Warnung* die gleiche Funktion erfüllt wie der Nebensatz *obwohl er sie gewarnt hatte* kann man das Wort *trotz* in Satz (3) auch als Konnektor interpretieren, obwohl es gegen die Merkmale (M2) und (M5) verstößt.

1.1 Das Problem der Ambiguität

Diskurskonnektoren sind keine feste Wortklasse. Es können sowohl Konjunktionen, Adverbien als auch Partikel oder Präpositionen sein. Diskurskonnektoren sind zwar eine geschlossene Klasse[3], allerdings sind einige Diskurskonnektoren ambig. Dabei gibt es zwei Arten von Ambiguität. Die erste Art ist die Ambiguität zwischen einer Konnektor-Lesart

3 Wenn man nur Ein-Wort-Konnektoren betrachtet.

und einer Nicht-Konnektor-Lesart. Das Wort *als* im Satz (5) ist Teil eines Komparativs und im Satz (6) ein Konnektor, der die zwei Sachverhalte in eine temporale Beziehung zueinander setzt.

(5) *Er wünschte sich oft größer zu sein **als** ihm durch die Natur vergönnt war.*

(6) *Ihm wurde flau im Magen, **als** ihm das volle Ausmaß seines Handelns bewusst wurde.*

Die zweite Art der Ambiguität ist die der Art des Konnektors. Manche Konnektoren können zum Beispiel einmal temporal und ein anderes Mal kausal interpretiert werden. In Satz (7) beschreibt *dann* die Reihenfolge der zwei Sachverhalte, während im Satz (8) *dann* eine kausale Folge beschreibt.

(7) *Der Antrag wird von der Kommission geprüft und kann **dann** vom Vorsitzenden unterzeichnet werden.*

(8) *Wenn die Staatsanwaltschaft den Antrag rechtzeitig einreicht, **dann** können wir noch vor März in Berufung gehen.*

Mit der zweiten Art von Ambiguität werde ich mich in dieser Arbeit nicht befassen. Hierzu lese man [Hut05] für das Englische oder [Bay04] für das Deutsche.

1.1.1 Relevante Arbeiten

Da sich die meisten Arbeiten mit der zweiten Art von Ambiguität befassen, habe ich zur ersten Art von Ambiguität nur eine englische Arbeit von Emily Pitler und Ani Nenkova gefunden, die sich damit befasst, ob Syntaxinformationen dazu beitragen können, diese Art von Ambiguität bei englischen Konnektorenkandidaten aufzulösen (siehe [PN09]). Die einzige Arbeit zur Disambiguierung von deutschen Diskurskonnektoren liefern Stefanie Dipper und Manfred Stede [DS06], welche die Grundlage dieser Arbeit liefern.

1.2 Ausgangspunkte

Als Ausgangspunkt für diese Arbeit gilt das Paper *Disambiguating potential connectives* von Stefanie Dipper und Manfred Stede (2006) [DS06]. In einer Fußnote erwähnen sie 42

Wörter die sowohl eine Konnektor-Lesart besitzen als auch eine Nicht-Konnektor-Lesart. Diese Wörter lauten: *aber, allein, allenfalls, allerdings, als, also, auch, aufgrund, außer, da, dabei, dafür, dagegen, daher, danach, dann, darauf, darum, denn, doch, entgegen, ferner, nebenher, nur, seit, seitdem, so, sonst, soweit, statt, trotz, und, während, wegen, weshalb, weswegen, wie, wogegen, womit, wonach, worauf, zugleich*. In [DS06] werden für 9 dieser Wörter Disambiguierungsstrategien entwickelt. In dieser Arbeit werde ich die Ansätze auffassen und mich dabei allen 42 Konnektorenkandidaten widmen.

2 Daten

Um zu erfahren, ob die Disambiguierung von Diskurskonnektoren von gängigen Part-of-Speech-Taggern bereits gut gelöst wird oder noch nicht habe ich zu jedem der 41 ambigen Konnektoren[1] ein Korpus von ca. 200-250 Sätzen aus dem Kernkorpus des Digitalen Wörterbuchs der Deutschen Sprache[2] (DWDS) extrahiert und manuell auf Konnektor-Lesart und Nicht-Konnektor-Lesart annotiert.

2.1 Tagger

Danach habe ich die Korpora mit einem Part-of-Speech-Tagger annotiert. Als Tagger habe ich den TreeTagger der Universität Stuttgart benutzt ([Sch95]; [Sch94]). Dieser Tagger benutzt das STTS Tagset[3]. In der Tabelle 2.1 sind die Verteilungen der Part-of-Speech-Tags für die Konnektorenkandidaten in der Konnektor- und Nicht-Konnektor-Lesart zu sehen. Um nun abzulesen, ob man allein mit der Hilfe des Part-of-Speech-Tags Konnektor- von Nicht-Konnektor unterscheiden kann, muss man die Part-of-Speech-Tag-Verteilungen zu der Verteilung Konnektor-/Nicht-Konnektor in Beziehung setzen. Beispielsweise kann man bei dem Wort *allein* davon ausgehen, dass, wenn es mit dem POS-Tag PTKVZ annotiert ist, es sich nicht um einen Konnektor handelt. Dies ist aber nur bei 1.9 % der Vorkommen von *allein* der Fall. Für den Großteil der Vorkommen, kann man nicht sagen, ob es sich um einen Konnektor handelt oder nicht. Bei dem Wort *als* wiederum lässt sich eine gewisse Tendenz ausmachen, dass es sich bei einem mit dem POS-Tag KOKOM versehenen Wort nicht um einen Konnektor handelt und bei einem mit dem POS-Tag KOUS wohl. Benutzt

1 Bei dem Wort *aufgrund* stellte sich heraus, dass es immer eine Präposition ist und somit nicht ambig. Die Ambiguität tritt erst bei der Rechtschreibvariante *auf Grund* auf. Da ich mich aber bei dieser Arbeit auf Ein-Wort-Konnektoren beschränke, gehe ich von nun an nicht mehr auf *aufgrund* ein.

2 zu erreichen unter `http:\www.dwds.de` Digitales Wörterbuch der deutschen Sprache, DWDS-Projekt, Berlin-Brandenburgische Akademie der Wissenschaften, 2008-2011.

3 siehe `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html`

man nun diese Tendenz, um zu entscheiden, ob es sich bei einem Vorkommen des Wortes *als* um einen Konnektor handelt, wird die Vorhersage *POS-Tag KOKOM bedeutet Nicht-Konnektor* zu 98.5 % zutreffen, die Vorhersage *POS-Tag KOUS bedeutet Konnektor* allerdings nur zu 35.4 %.

	Konnektor		Nicht Konnektor	
	Anzahl	% POS-Tags	Anzahl	% POS-Tags
aber	170	45.9 % ADV; 54.1 % KON	67	58.2 % ADV; 41.8 % KON
allein	18	100 % ADV	196	98.0 % ADV; 2.0 % PTKVZ
allenfalls	23	100 % ADV	184	100 % ADV
allerdings	169	100 % ADV	33	100 % ADV
als	21	19.0 % KOKOM; 81.0 % KOUS	295	89.5 % KOKOM; 10.5 % KOUS
also	63	100 % ADV	162	100 % ADV
auch	64	100 % ADV	286	100 % ADV
außer	15	100 % APPR	187	100 % APPR
da	99	31.3 % ADV; 68.7 % KOUS	118	84.7 % ADV; 7.6 % KOUS; 7.6 % PTKVZ
dabei	19	100 % PAV	188	98.9 % PAV; 1.1 % PTKVZ
dafür	24	100 % PAV	184	100 % PAV
dagegen	152	100 % PAV	54	100 % PAV
daher	196	100 % PAV	16	100 % PAV
danach	119	100 % PAV	91	100 % PAV
dann	161	100 % ADV	62	100 % ADV
darauf	14	100 % PAV	194	100 % PAV
darum	82	100 % PAV	126	100 % PAV
denn	120	5.8 % ADV; 94.2 % KON	91	79.1 % ADV; 20.9 % KON
doch	83	16.9 % ADV; 83.1 % KON	140	92.1 % ADV; 7.9 % KON
entgegen	70	1.4 % APPO; 64.3 % APPR; 34.3 % PTKVZ	145	0.7 % APPR; 99.3 % PTKVZ
ferner	187	5.3 % ADJA; 94.7 % ADJD	18	55.6 % ADJA; 44.4 % ADJD
nebenher	108	100 % ADV	93	98.9 % ADV; 1.1 % NN
nur	62	100 % ADV	234	100 % ADV
seit	25	16.0 % APPR; 84.0 % KOUS	207	97.1 % APPR; 2.9 % KOUS

seitdem	61	77.0 % KOUS; 23.0 % PAV	140	7.9 % KOUS; 92.1 % PAV
so	45	100 % ADV	205	100 % ADV
sonst	58	100 % ADV	152	100 % ADV
soweit	196	38.8 % ADV; 61.2 % KOUS	31	100 % ADV
statt	41	48.8 % APPR; 42.5 % KOUI; 2.4 % NN; 7.3 % PTKVZ	170	22.9 % APPR; 5.3 % KOUI; 1.2 % NN; 70.6 % PTKVZ
trotz	188	100 % APPR	20	5.0 % APPR; 95.0 % NN
und	219	100 % KON	329	100 % KON
während	103	3.9 % APPR; 96.1 % KOUS	103	78.6 % APPR; 21.4 % KOUS
wegen	235	3.4 % APPO; 96.6 % APPR	10	10.0 % APPO; 30.0 % APPR; 60 % NN
weshalb	76	100 % PWAV	133	1.5 % NN; 98.5 % PWAV
weswegen	89	100 % PWAV	114	6.1 % NN; 93.9 % PWAV
wie	16	12.5 % KOKOM; 87.5 % KOUS	268	63.1 % KOKOM; 3.4 % KON; 33.2 % KOUS; 0.4 % PWAV
wogegen	148	100 % PWAV	55	1.8 % NN; 98.2 % PWAV
womit	92	100 % PWAV	112	0.9 % NE; 4.5 % NN; 94.6 % PWAV
wonach	15	100 % PWAV	189	100 % PWAV
worauf	99	100 % PWAV	103	1.0 % NN; 99.0 % PWAV
zugleich	83	100 % ADV	124	100 % ADV

Tabelle 2.1: Anzahl der Vorkommen der Konnektorenkandidaten in den Korpora und Verteilung der POS-Tags für die Konnektor- und Nicht-Konnektor-Lesart

2.2 Auswertung

Die Auswertung der Tabelle 2.1 ergibt, dass es für einige der Konnektorenkandidaten durchaus ausreicht, allein Part-of-Speech-Tagging zur Disambiguierung heranzuziehen. Wenn man einen Grenzwert von 80 %-iger Genauigkeit anlegt, kann man für die folgenden 7 Konnektorenkandidaten anhand ihres Part-of-Speech-Tags zwischen Konnektor- und

Nicht-Konnektor-Lesart disambiguieren: *denn*, *doch*, *entgegen*, *seitdem*, *trotz*, *während*, *wegen*. In Tabelle 2.2 sind die POS-Tags dargestellt, die sich dafür eignen, POS-Regeln zu erstellen, um die Konnektor-/Nicht-Konnektor-Frage zu klären.

	POS = Konnektor	Precision	Recall[4]	POS = Nicht-Konnektor	Precision	Recall
denn	KON	85.6 %	94.2 %	ADV	91.9 %	79.1 %
doch	KON	86.3 %	83.1 %	ADV	90.2 %	92.1 %
entgegen	APPO APPR	100 % 97.8 %	65.7 %	PTKVZ	85.7 %	99.3 %
seitdem	KOUS	81.0 %	77.0 %	PAV	90.2 %	92.1 %
trotz	APPR	99.5 %	100 %	NN	100 %	95.0 %
während	KOUS	81.8 %	96.1 %	APPR	95.3 %	78.6 %
wegen	APPO APPR	88.9 % 98.7 %	100 %	NN	100 %	60.0 %

Tabelle 2.2: POS-Regeln und deren Precision und Recall für die 7 Konnektorenkandidaten, bei denen die Precision über 80 % liegt.

Bei anderen Konnektorenkandidaten liefern die POS-Tags zwar einen Hinweis auf die Konnektor-/Nicht-Konnektor-Eigenschaft eines Wortvorkommens, allerdings reicht diese Information nicht aus, um sichere Entscheidungen (mit einer Genauigkeit von über 80 %) treffen zu können. In Tabelle 2.3 sind diese Konnektorkandidaten aufgezählt.

Die Regel, dass es sich bei Vorkommen des Wortes *aber*, die mit dem POS-Tag KON annotiert sind, um Konnektoren handelt, hat zwar eine knapp unter dem Grenzwert von 80 % liegende Genauigkeit von 76.6 %; allerdings hat die Umkehrregel, dass es sich bei Vorkommen, die mit dem POS-Tag ADV annotiert sind, um Nicht-Konnektoren handelt, nur eine Genauigkeit von 33.3 % und ist somit eher unbrauchbar.

Bei dem Wort *als* bietet sich das umgekehrte Bild. Die Regel, dass Wortvorkommen, die mit dem POS-Tag KOKOM annotiert sind, Nicht-Konnektoren sind, hat eine sehr hohe Genauigkeit von 98.5 %. Die Umkehrregel, dass es sich bei Vorkommen des Wortes *als*, die mit dem POS-Tag KOUS annotiert sind, um Konnektoren handelt, hat wiederum nur eine sehr niedrigere Genauigkeit von 35.4 %.

Die POS-Regeln, die den Konnektorkandidaten *da* betreffen, sind an sich gesehen brauchbar mit dem kleinen Manko, dass die Regel, dass es sich bei Vorkommen von *da*, die mit

4 Der Recall wird, wenn es mehr als eine POS-Regel gibt, für alle zusammen berechnet.

	POS = Konnektor	Precision	Recall	POS = Nicht-Konnektor	Precision	Recall
aber	KON	76.6 %	54.1 %	ADV	33.3 %	58.2 %
als	KOUS	35.4 %	81.0 %	KOKOM	98.5 %	89.5 %
da	KOUS	88.3 %	68.7 %	PTKVZ ADV	100 % 76.3 %	92.4 %
ferner	ADJD	95.7 %	94.7 %	ADJA	50.0 %	55.6 %
seit	KOUS	77.8 %	84.0 %	APPR	98.0 %	96.6 %
soweit	KOUS	100 %	61.2 %	ADV	29.0 %	100 %
statt	APPR KOUI	33.9 % 65.4 %	90.2 %	NN PTKVZ	66.7 % 97.6%	71.8 %
wie	KOUS	13.6 %	87.5 %	KOKOM KON PWAV	98.8 % 100 % 100 %	66.8 %

Tabelle 2.3: POS-Regeln und deren Precision und Recall für die 8 Konnektorenkandidaten, bei denen die Precision unter 80 % liegt.

dem POS-Tag ADV annotiert sind, um Nicht-Konnektoren handelt, eine Genauigkeit von nur 76.3 % hat und damit leicht unter dem Grenzwert liegt.

Bei dem Wort *ferner* gibt es eine sehr genaue POS-Regel, um die Konnektoren zu identifizieren, allerdings ist die POS-Regel, um Nicht-Konnektoren zu identifizieren mit einer Genauigkeit von 50 % unbrauchbar.

Die POS-Regeln, die das Wort *seit* betreffen sind recht gut brauchbar, wobei die Regel, dass Wortvorkommen, die mit dem POS-Tag KOUS annotiert sind, Konnektoren sind, eine leicht unter dem Grenzwert liegende Genauigkeit von 77.8 % hat.

Bei dem Wort *soweit* hat die POS-Regel, dass es sich bei jedem Vorkommen von *soweit*, wenn es mit dem POS-Tag KOUS annotiert ist, um einen Konnektor handelt, sogar eine Genauigkeit von 100 %. Allerdings hat die Umkehrregel mit dem POS-Tag ADV nur eine Genauigkeit von 29 %.

Bei dem Konnektorkandidaten *statt* ist es schwierig POS-Regeln mit einer hohen Genauigkeit zu finden. Zwar ist die POS-Regel, die besagt, dass alle Vorkommen, die mit dem POS-Tag PTKVZ annotiert sind, Nicht-Konnektoren sind zu 97.6 % genau, aber es lässt sich nicht so einfach eine genaue POS-Regel zur Konnektoridentifikation finden.

Die POS-Regeln, die bestimmen, wann das Wort *wie* kein Konnektor ist, sind sehr genau. Allerdings beträgt die Genauigkeit der POS-Regel, die besagt, dass ein Vorkommen von *wie* dann ein Konnektor ist, wenn es mit dem POS-Tag KOUS annotiert wird, gerade einmal 13.6 %.

Bei über der Hälfte der Konnektorkandidaten gibt das POS-Tag keinerlei Aufschluss darüber, ob es sich um einen Konnektor handelt oder nicht. Sieht man von den wenigen Vorkommen ab, die mit den POS-Tags NE, NN oder PTKVZ (Eigenname, normales Nomen, abgetrennter Verbzusatz) annotiert und damit immer ein Nicht-Konnektor sind, kann man für die folgenden 26 Konnektorkandidaten allein durch Part-of-Speech-Taggen nicht zwischen Konnektor- und Nicht-Konnektor-Lesart disambiguieren: *allein*, *allenfalls*, *allerdings*, *also*, *auch*, *außer*, *dabei*, *dafür*, *dagegen*, *daher*, *danach*, *dann*, *darauf*, *darum*, *nebenher*, *nur*, *so*, *sonst*, *und*, *weshalb*, *weswegen*, *wogegen*, *womit*, *wonach*, *worauf*, *zugleich*.

2.3 Ergebnisse

Zusammenfassend lässt sich über die Frage, ob Part-of-Speech-Tagging dazu geeignet ist, das Ambiguitätsproblem von Diskurskonnektoren im Deutschen zu lösen, sagen, dass es für einige der 41 ambigen Diskurskonnektoren durchaus ausreicht, allein die POS-Tags zurate zu ziehen. Wenn man das Mittel des f-score (Formel 2.1) benutzt, um zu bestimmen, ob eine POS-Regel genau genug ist, und man einen Grenzwert von 0.75 anlegt, kann man für die folgenden Konnektorenkandidaten sagen, dass die POS-Regeln, die in den Tabellen 2.2 und 2.3 angegeben sind, ausreichen, um sie erfolgreich zu disambiguieren: *denn*, *doch*, *entgegen*, *seit*, *seitdem*, *trotz*, *während* und *wegen*.

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.1}$$

Für die Konnektorenkandidaten *als*, *da*, *ferner*, *soweit*, *statt* und *wie* überschreitet jeweils der f-score nur eines Regelsets, entweder dessen für die Konnektor- oder dessen für die Nicht-Konnektor-Lesart, den Grenzwert von 0.75.

Zusätzlich lassen sich allgemeine POS-Regeln aufstellen, die für alle Konnektorenkandidaten gelten. Diese wären, dass es sich niemals um einen Konnektor handelt, wenn ein Konnektorenkandidat mit einem der POS-Tags NE, NN oder PTKVZ annotiert ist. Es

gibt zwar einige wenige Vorkommen in den Korpora, wo ein als Konnektor eingestuftes Wort mit einem dieser POS-Tags annotiert wurde. Diese sind allerdings unter der üblichen Fehlerquote von Part-of-Speech-Taggern vernachlässigbar.

Für die restlichen 33 Konnektorenkandidaten, die nicht eindeutig durch ihre Part-of-Speech-Tags disambiguiert werden können, müssen weitere Informationen herangezogen werden, um das Ambiguitätsproblem zu lösen. Dies sind diese Konnektorenkandidaten: *aber, allein, allenfalls, allerdings, als, also, auch, außer, da, dabei, dafür, dagegen, daher, danach, dann, darauf, darum, ferner, nebenher, nur, so, sonst, soweit, statt, und, weshalb, weswegen, wie, wogegen, womit, wonach, worauf* und *zugleich*. Im nächsten Kapitel werde ich mich diesen Wörtern zuwenden.

3 Betrachten des unmittelbaren Kontextes

Der nächstliegende Schritt, um die Konnektorenkandidaten zu disambiguieren, ist es, sich ihren Kontext im Satz anzugucken. Dazu reicht erst einmal ein Kontext von zwei Wörtern davor und zwei Wörtern danach und von diesen Wörtern auch jeweils nur ihre POS-Tags.

In Tabelle 3.1 sind die 11 Wörter aufgelistet, bei denen man mit Hilfe des unmittelbaren Kontextes so disambiguieren kann, dass der Grenzwert von 0.75 für den f-score erreicht wird. Die Kontextregeln in dieser Tabelle sind so zu lesen, dass das unterstrichene POS-Tag dem Vorkommen des Konnektorenkandidatens zugeordnet wird und die davor (danach) stehenden POS-Tags den unmittelbar davor (danach) stehenden Kontext beschreiben. Wenn sich in einem POS-Tag ein Punkt befindet, steht dieser für einen beliebigen Buchstaben, beispielsweise steht das POS-Tag V.FIN für die POS-Tags VVFIN, VAFIN oder VMFIN, also für alle finiten Verben. Da im STTS das Tag für satzbeendende Interpunktion bereits einen Punkt enthält, benutze ich das Zeichen $ für jegliche Form der Interpunktion, also für $, $. oder $(. Das $.^+$ steht für mindestens einen beliebigen Buchstaben. $V.^+$ steht also für alle Verben unabhängig von Zeitform oder Modus.

Regeln mit einem detaillierteren Kontext werden nicht von einer Regel mit gröberen Kontext eingeschlossen. Ein Beispiel: Die Regel $. ADV VVFIN für das Wort *nebenher* besagt, dass es sich um einen Konnektor handelt, wenn *nebenher* mit dem POS-Tag ADV versehen ist, unmittelbar hinter einer satzbeendenden Interpunktion steht und direkt dahinter ein finites Vollverb steht. Die Regel ADV VVFIN auf der Nicht-Konnektor-Seite der Tabelle besagt nun, dass es sich um einen Nicht-Konnektor handelt, wenn *nebenher* das POS-Tag ADV erhält, direkt darauf ein finites Vollverb folgt und davor **alles andere** außer einer satzbeendenden Interpunktion steht.

	Kontext-regeln Konnektor	Precision	Recall	Kontext-regeln Nicht-Konnektor	Precision	Recall
also	$, ADV V.FIN $. ADV V.FIN V.FIN ADV	83.3 %	87.3 %	alle anderen	95.0 %	93.2 %
auch	ADV VVFIN	98.0 %	78.1 %	alle anderen	95.3 %	99.7 %
außer	$ APPR $, $ APPR KOUS	100 %	86.7 %	alle anderen	98.9 %	100 %
da	$, ADV $ KON ADV KOUS	76.9 %	86.9 %	ADV PTKVZ	88.1 %	81.4 %
darum	alle anderen	77.2 %	95.1 %	PAV $($ PAV $, PAV $. PAV VV	96.3 %	77.4 %
nebenher	$. ADV VVFIN alle anderen	88.2 %	83.3 %	ADV $ ADV VVFIN ADV VVPP ADV KON ADV VVINF	81.8 %	87.1 %
nur	$. ADV VVFIN	100 %	74.2 %	alle anderen	93.6 %	100 %
so	$, ADV KOUS $, ADV V.FIN KON ADV V.FIN	77.8 %	77.8 %	alle anderen	95.1 %	95.1 %
sonst	$ ADV V.FIN	87.2 %	70.7 %	alle anderen	89.6 %	96.1 %
soweit	$ ADV KOUS	98.9 %	92.3 %	alle anderen	65.9 %	93.5 %
zugleich	$ ADV V.FIN $ KON ADV V.$^+$ KON ADV VVFIN ADV	86.6 %	69.9 %	alle anderen	82.1 %	92.7 %

Tabelle 3.1: Kontextregeln und deren Precision und Recall für die 11 Konnektorenkandidaten, bei denen der f-score über 0.75 liegt.

Bei dem Wort *da* lag der f-score für die Nicht-Konnektor-Lesart bei den POS-Regeln in Tabelle 2.3 noch knapp unter dem Grenzwert von 0.75. Nimmt man die zwei Kontextregeln aus der Tabelle 3.1 dazu, erreicht man jetzt f-scores von 0.82 für die Konnektor-Lesart und 0.85 für die Nicht-Konnektor-Lesart, die somit über dem Grenzwert liegen.

Bei dem Wort *soweit* lag der f-score für die Nicht-Konnektor-Lesart noch deutlich unter dem Grenzwert. Durch das Hinzufügen einer einzigen Kontextregel stiegen die f-scores auf 0.95 für die Konnektor-Lesart und auf 0.77 für die Nicht-Konnektor-Lesart.

Bei den weiteren 8 Konnektorenkandidaten (*also, auch, außer, darum, nebenher, nur, so* und *sonst*), für die es genaue Kontextregeln gibt, gab es ohne den Kontext keinerlei Möglichkeit sie zu disambiguieren.

Viele der Kontextregeln orientieren sich an in der Nähe des Konnektorenkandidatens stehenden Interpunktionszeichen oder Verben. Interpunktionszeichen markieren in der Regel das Ende eines Satzes oder Nebensatzes und den Anfang eines Neuen. Da viele Konnektoren auch gerade zwischen zwei Sätzen stehen, das heißt am Anfang des zweiten Satzes oder am Ende des Ersten, wobei es sich auch hier um Nebensätze handeln kann, erscheint es plausibel, dass sich die Nähe eines Wortvorkommens zu einem Interpunktionszeichen dazu eignet, zu entscheiden, ob es sich um einen Konnektor handelt oder nicht. Bei Verben kommt es zu einem ähnlichen Phänomen, da es im Deutschen[1] so ist, dass die Position von Verben auch dazu geeignet ist, um die Grenze zwischen Sätzen bestimmen zu können.

Für die restlichen 22 Konnektorenkandidaten lassen sich auch teilweise Kontextregeln finden, die aber nicht mit ihrem f-score an den Grenzwert von 0.75 heranreichen. Oder aber es müssen sehr viele detaillierte Regeln konstruiert werden, um die Konnektor-Lesart von der Nicht-Konnektor-Lesart ausreichend unterscheiden zu können. Da die Korpora, auf deren Grundlage diese Regeln konstruiert werden mit 200-250 Sätzen sehr klein sind, beruhen manche dieser Regeln nur auf einem oder zwei Sätzen und sind damit nicht ausreichend belegt. Bei manchen Wörtern sind die Kontexte auch so vielfältig, dass es unmöglich erscheint, überhaupt manuell Kontextregeln aufstellen zu können. Hierzu zählen auch die sehr häufigen Wörter *aber*, *dann* und *und*.

In der Tabelle 3.2 sind alle Kontextregeln für die verbleibenden 19 Konnektorenkandidaten mit Precision und Recall dargestellt. Einige dieser Kontextregeln liegen nur knapp unter dem f-score Grenzwert von 0.75. Dazu zählen die Kontextregeln für die Wörter *allein, allenfalls, allerdings, daher, danach, ferner* und *statt*. Wenn man also einen niedrigeren Grenzwert anlegt, sind diese Kontextregeln durchaus noch brauchbar.

1 vor allem in verschachtelten Sätzen aus Zeitungstexten, die den größten Teil des DWDS-Kernkorpus stellen und somit auch meiner Korpora

	Kontext-regeln Konnektor	Precision	Recall	Kontext-regeln Nicht-Konnektor	Precision	Recall
allein	$(ADV $. ADV	44.7 %	94.4 %	alle anderen	99.4 %	89.2 %
allenfalls	$(ADV $. ADV	53.2 %	73.9 %	alle anderen	96.6 %	91.9 %
allerdings	alle anderen	94.5 %	92.3 %	$(ADV $, ADV	64.9 %	72.7 %
als	alle anderen	42.1 %	76.2 %	KOKOM KOUS V.FIN	98.2 %	92.5 %
dabei	$, PAV $. PAV	24.2 %	84.2 %	alle anderen	97.9 %	73.4 %
dafür	$, PAV $. PAV	44.7 %	87.5 %	$. PAV $, alle anderen	98.1 %	85.9 %
dagegen	alle anderen	82.3 %	94.7 %	PAV $. PAV $,	74.2 %	42.6 %
daher	alle anderen	97.4 %	95.9 %	PAV $. PAV $, PAV KON	57.9 %	68.8 %
danach	alle anderen	75.9 %	89.9 %	PAV $ PAV KOKOM PAV VV.$^{+}$	82.6 %	62.6 %
darauf	$, PAV $. PAV	19.7 %	92.9 %	$, PAV $ alle anderen	99.3 %	72.7 %
ferner	alle anderen	97.3 %	95.7 %	ADJA NN ADV ADJD	61.9 %	72.2 %
statt	$, KOUI PPER alle anderen	58.2 %	78.0 %	$ KOUI NN PTKVZ N. APPR	94.2 %	86.5 %
weshalb	alle anderen	46.9 %	100 %	NN PWAV KON	100 %	35.3 %

				PWAV $ PWAV V.+ KON PWAV		
weswegen	alle anderen	55.8 %	96.6 %	NN PWAV V.+ PWAV $ PWAV N. KON PWAV	93.9 %	40.4 %
wie	$, KOUS	15.5%	81.3 %	alle anderen	98.5 %	73.5 %
wogegen	alle anderen	77.4 %	99.3 %	NN $, $(PWAV PWAV $ PWAV N. KON PWAV	92.3 %	21.8 %
womit	alle anderen	52.0 %	100 %	N. $. $, PWAV $, PWAV APPR PAV V.+ KON PAV PWS PAV $,	100 %	24.1 %
wonach	$, PWAV $. PWAV	10.3 %	100 %	$ PWAV $ $ PWAV V.+ $ PWAV APPRART $ PWAV N. $ PWAV CARD $ PWAV PDAT alle anderen	100 %	30.7 %
worauf	alle anderen	52.1 %	99.0 %	NN $. KON PWAV $(PAV V.+	92.9 %	12.6 %

Tabelle 3.2: Kontextregeln und deren Precision und Recall für die 19 Konnektorenkandidaten, bei denen der f-score unter 0.75 liegt.

Wo man den Vergleich mit den Ergebnissen in der Tabelle 2.3 ziehen kann, zeigt sich, dass das Hinzufügen von Kontextregeln den f-score steigern kann, wenn auch nicht über den Grenzwert von 0.75. Bei dem Wort *als* führt das Hinzufügen der einen Kontextregel zu einer Steigerung der Precision bei der Konnektor-Lesart von 35.4 % auf 42.1 %. Und auch der f-score steigt von 0.49 auf 0.54. Bei *ferner* steigt die Precision bei der Nicht-Konnektor-Lesart durch das Ersetzen der POS-Regeln durch zwei Kontextregeln von 50 % auf 61.9 %. Und bei *statt* führt das Hinzufügen von drei Kontextregeln dazu, dass die Precision bei der Konnektor-Lesart von 43.5 % auf 58.2 % ansteigt. Die Precision für die Konnektor-Lesart von *wie* verzeichnet nach dem Hinzufügen einer Kontextregel immerhin eine leichte Steigerung von 13.6 % auf 15.5 %.

Insgesamt lässt sich sagen, dass durch das zusätzliche Betrachten von unmittelbarem Kontext aus 8 allein durch ihr POS-Tag disambiguierbaren Konnektorenkandidaten nun 19 disambiguierbare Konnektorenkandidaten geworden sind. Für weitere 7 Konnektorenkandidaten existieren Kontextregeln die ein etwas weniger genaues Disambiguieren ermöglichen. Für 12 habe ich Kontextregeln gefunden, die nur geringe f-scores besitzen. Für die letzten 3 habe ich keine Kontextregeln finden können.

4 Zusammenfassung und Ausblick

Für gut die Hälfte der von mir betrachteten 41 ambigen Konnektorenkandidaten lässt sich das Ambiguitätsproblem mit Hilfe von POS- und Kontextregeln mit einer Genauigkeit von f-score > 0.75 lösen. Die disambiguierbaren Konnektorenkandidaten lauten: *also*, *auch*, *außer*, *da*, *darum*, *denn*, *doch*, *entgegen*, *nebenher*, *nur*, *seit*, *seitdem*, *so*, *sonst*, *soweit*, *trotz*, *während*, *wegen* und *zugleich*.

Um die restlichen 22 Konnektorenkandidaten (*aber*, *allein*, *allenfalls*, *allerdings*, *als*, *dabei*, *dafür*, *dagegen*, *daher*, *danach*, *dann*, *darauf*, *ferner*, *statt*, *und*, *weshalb*, *weswegen*, *wie*, *wogegen*, *womit*, *wonach*, *worauf*) zu disambiguieren reichen die gefundenen POS- und Kontextregeln nicht aus.

Für mindestens die 7 Konnektorenkandidaten, für die es Kontextregeln gibt, deren f-score leicht unterhalb den Grenzwerts von 0.75 liegt (*allein*, *allenfalls*, *allerdings*, *daher*, *danach*, *ferner*, *statt*) denke ich, dass ein maschinelles Lernverfahren basierend auf einem größeren Korpus zu genaueren Kontextregeln führen kann, die manuell nicht leicht zu erfassen sind. Ein maschinelles Lernverfahren könnte auch bei den Wörtern *aber*, *dann* und *und* nützlich sein, für die die Kontexte zu unterschiedlich und vielfältig waren, sodass ich keinerlei zufriedenstellende Kontextregeln finden konnte. Bei der manuellen Durchsicht der Kontexte fallen einem auch nicht viele Kontextregeln ins Auge, die über mehr als ein Wort zu einer Seite des Konnektorenkandidaten hinausgehen. Hier sehe ich das größte Potenzial für maschinelle Lernverfahren. Allerdings müssten die Korpora dafür deutlich größer gewählt werden, damit sich die Kontextregeln nicht auf lediglich ein oder zwei Sätze stützen müssen.

Generell lässt sich sagen, dass die Größe der Korpora viel zu klein ist. Bei vielen Konnektorenkandidaten ist die Verteilung von Konnektor-Lesart und Nicht-Konnektor-Lesart nicht gleichmäßig, da entweder die eine oder die andere Lesart im deutschen Sprachgebrauch dominiert. So kann es sein, dass das Korpus für eine der beiden Lesarten mitunter

nur 9 Sätze enthält. Größere Korpora können nicht nur die Genauigkeit der gefundenen Kontextregeln präzisieren, sondern wahrscheinlich auch Grundlage liefern für weitere Kontextregeln, die die Genauigkeit erhöhen. Allerdings muss auch gesagt werden, dass bei den seltensten Konnektorenkandidaten (*wogegen, nebenher, weswegen*) die Vorkommen im DWDS-Kernkorpus beinahe vollständig in meine Korpora geflossen sind. Will man für diese Konnektorenkandidaten größere Korpora anlegen, muss man noch andere Quellen hinzuziehen.

Da ich diese Arbeit vielen Konnektorenkandidaten gewidmet habe und die Korpora für die einzelnen Wörter jeweils nur sehr klein sind, hätte das Abtrennen eines Teils dieser Korpora zu Testzwecken dazu geführt, dass weniger Kontextregeln gefunden worden wären, was zu einer geringeren Genauigkeit geführt hätte, oder dass es keine ausreichende Grundlage für die jetzt gefundenen Kontextregeln gegeben hätte. Die Angaben von Precision und Recall beziehen sich also allein auf das Trainingskorpus und müssten auf einem davon unabhängigen Korpus getestet werden, um sagen zu können, ob sich die Kontextregeln dafür eignen auch unbesehene Vorkommen der Konnektorenkandidaten zu disambiguieren.

Leider gehören die Wörter *und, als, wie* und *aber* zu den häufigsten Konnektorenkandidaten und können durch meine Kontextregeln nicht eindeutig disambiguiert werden. Andererseits sind andere Wörter, die nicht disambiguiert werden können wie *wonach, worauf, weshalb, womit, wogegen* und *weswegen*, weniger frequent. Es ist also zu erwarten, dass auch eine nicht erfolgreiche Disambiguierung in diesen Fällen nicht zu großen Folgefehlern in den weiteren computerlinguistischen Anwendungen führen wird.

Sollten auch die Mittel von maschinellen Lernverfahren auf den POS-Tags eines Kontexts von jeweils 2 Wörtern vor und nach einem Konnektorenkandidat erschöpft sein, ist der nächste Schritt zur Disambiguierung syntaktische Informationen zurate zu ziehen. Um den Aufwand vom Parsing des ganzen Satzes zu vermeiden, wäre es interessant zu sehen, ob vielleicht Chunking ausreicht, um die Merkmale zur Disambiguierung der restlichen Konnektorenkandidaten zu liefern. Ich könnte mir vorstellen, dass ebenfalls die physische Nähe eines Konnektorenkandidaten zu Nominalphrasen, Verben und Verbphrasen ein verwendbares Merkmal für die Disambiguierung sein kann.

Da alle POS- und Kontextregeln auf den POS-Tags basieren, die der TreeTagger produziert, kann es natürlich sein, dass es auf den POS-Tags von anderen Part-of-Speech-Taggern nicht greifen. Die Frage der Verwendbarkeit der POS- und Kontextregeln in Unabhängigkeit

vom Part-of-Speech-Tagger muss also noch geklärt werden. Wobei natürlich festgelegt ist, dass die verwendbaren Part-of-Speech-Tagger mit dem STTS als Tagset arbeiten müssen.

Insgesamt bin ich positiv davon überrascht, dass man nur mit Hilfe eines Part-of-Speech-Taggers beinahe die Hälfte der ambigen Konnektorenkandidaten mit einer zufriedenstellenden Genauigkeit (f-score > 0.75) disambiguieren kann, ohne syntaktische Informationen einholen zu müssen.

A Anhang

A.1 Kriterien zur Annotation von Konnektor- bzw. Nicht-Konnektor-Lesart

Es folgen die Kriterien, die bei der Annotation der Korpora für die Konnektor- bzw. Nicht-Konnektor-Lesart der 42 ambigen Diskurskonnektoren zum Einsatz gekommen sind. In Klammern werden Beispiele für die verschiedenen Lesarten aufgeführt.

aber

Aber ist kein Konnektor, wenn es nicht Sätze, sondern Präpositionalphrasen oder Adjektive miteinander verbindet („informativ, ehrlich, aber unterhaltsam“). **Aber** ist ein Konnektor, wenn es zwei Sätze oder Haupt- und Nebensatz miteinander verknüpft („Polens Staatsgebiet erstreckte sich zwar über die endlosen Ebenen zwischen Ostsee und Schwarzem Meer, aber wie in Deutschland verhinderte der Adel, daß eine starke Erbmonarchie entstand. “). Dabei muss **aber** nicht am Anfang des Satzes stehen („Er lächelte huldvoll in alle Richtungen, rutschte aber nervös auf seinem Sitz hin und her“).

allein

Allein ist kein Konnektor, wenn es das Adjektiv **allein** ist („er wollte allein sein“). **Allein** ist kein Konnektor, wenn es anstelle von **nur** benutzt wird und sich auf eine Nicht-Satz-Phrase bezieht („nicht allein deshalb“, „allein durch Willenskraft“). **Allein** ist ein Konnektor, wenn er zwei Sätze miteinander verknüpft und einen Kontrast darstellt („Hätte Gott mich doch so geschaffen, daß ich alles so leicht nehmen möchte wie andere! Allein , ich kann es nicht.“).

allenfalls

Allenfalls ist kein Konnektor, wenn es sich auf eine NP oder PP bezieht, und man es durch **höchstens** ersetzen kann („Erdbeeren faßt man einzeln an den Stielen und taucht sie allenfalls in Staubzucker.“). **Allenfalls** ist ein Konnektor, wenn es einen Nebensatz einleitet („ Zu feiern gibt nichts - allenfalls gilt es zu trauern über die verpaßten Chancen.“).

allerdings

Allerdings ist kein Konnektor, wenn es sich auf eine NP o. Ä. bezieht und innerhalb von NP oder PP zu finden ist („Das Stück behandelt mit allerdings dürftigen dramaturgischen Mitteln das Thema »Sucht«.“). **Allerdings** ist ein Konnektor, wenn er zwei Sätze miteinander verbindet („Mit ihnen kann man einen Brief in den Computer sprechen und der Empfänger kann ihn sich anhören. Allerdings kostet so ein Hörbrief enormen Speicherplatz.“).

als

Als ist kein Konnektor, wenn es vergleichend verwendet wird („breiter als“, „sowohl Wünsche als auch reale Möglichkeiten“, „was sich anhörte, als würde man...“). **Als** ist ein Konnektor, wenn es temporal verwendet wird, um zwei Sätze miteinander zu verknüpfen („Mann, war ich froh, als die Zeit bei dem alten Uhu endlich vorbei war.“).

also

Also ist kein Konnektor, wenn es in ergänzender, bzw. präzisierender Weise gebraucht wird, um NPs, PPs oder andere Satzteile aufzulisten („in Süddeutschland, also in Bayern, Österreich und Baden, nicht aber in Württemberg“, „den Lohn von 2,30 auf 5 Dollar zu erhöhen, also auf mehr als das doppelte“). **Also** ist ein Konnektor, wenn er Sätze miteinander verknüpft und man es durch **folglich** ersetzen kann („Der Hersteller weiß aber nicht, wie die Menschen die Technik nutzen werden. Also baut man eine Plattform und läßt den Kunden entscheiden.“ $\widehat{=}$ „Folglich baut man eine Plattform...“).

auch

Auch ist kein Konnektor in der Phrase „sowohl ... als auch ...“ „wenn auch ...“, oder wenn es sich auf die dahinter stehende NP oder PP bezieht. **Auch** kann nur ein Konnektor sein, wenn es sich auf einen Satz bzw. ein finites Verb bezieht („Auch fehlte es wieder nicht an ...“). **Auch** am Satzanfang gefolgt von einem finiten Verb ist meist ein Konnektor. Wenn auf **auch** ein finites Verb folgt, kann man es umstellen und prüfen, ob sich die Semantik geändert hat. Wenn das nicht zutrifft, handelt es sich um einen Konnektor. („Josef sah, dass Johannes auch lachte.“ $\neq$ „Josef sah, dass auch Johannes lachte.“)

aufgrund

Aufgrund ist nicht ambig!

außer

Außer ist kein Konnektor, wenn es sich auf eine NP im Dativ bezieht („alle, außer dem Bundespräsidenten, waren eingeladen“), oder in bestimmten Wendungen auf Nomen („außer Betrieb“, „außer acht lassen“). **Außer** ist ein Konnektor, wenn er sich auf einen Satz oder Nebensatz bezieht („Er stellt keine besonderen Ansprüche, außer , daß ...“).

da

Da ist kein Konnektor, wenn es einen Ort beschreibt („hier und da“). **Da** ist ein Konnektor, wenn es einen kausalen Zusammenhang zwischen zwei Sätzen beschreibt („Da noch niemand jemals von der Insel zurückgekehrt ist...“). **Da** ist ein Konnektor, wenn es zwei Sätze temporal miteinander verknüpft („Fast wäre ich eingenickt, da tauchten plötzlich Leute vor meinem Auto auf.“).

dabei

Dabei ist kein Konnektor, wenn es im Sinne von „bei etwas“ verwendet wird („4711 ist immer dabei.“). **Dabei** ist ein Konnektor, wenn es am Anfang eines (Neben-)Satzes steht und zwei Sätze miteinander verknüpft und man ihn nicht in eine andere Position des Satzes verschieben kann („Sie wußte nicht recht, was sie von ihrem Chef halten sollte, dabei gefiel er ihr als Mann.“).

dafür

Dafür ist kein Konnektor, wenn es statt „für etwas“ verwendet wird („Der italienische Ausdruck dafür ist...“, „als Beweis dafür, dass...“, „die dafür erforderlichen Voraussetzungen“). **Dafür** ist ein Konnektor, wenn es einen Kontrast zwischen zwei Sätzen zum Ausdruck bringt („Er reimte sich zwar nicht, traf aber dafür den Nerv vieler Deutscher.“) **Dafür** ist ein Konnektor, wenn man es durch **stattdessen** ersetzen kann („weshalb der barocke Durchhof in Leipzig keine Galerien, dafür aber 4 mitunter kunstvolle Hoffassaden kennt“).

dagegen

Dagegen ist kein Konnektor, wenn es anstelle von „gegen etwas“ gebraucht wird („ich habe etwas dagegen“). **Dagegen** ist ein Konnektor, wenn es den Kontrast zwischen zwei Sätzen zum Ausdruck bringt („Die Beschränkungen der diplomatischen Kontakte blieben bestehen. Dagegen würden die Ausfuhrbeschränkungen aufgehoben.“). **Dagegen** ist ein Konnektor, wenn man ihn durch **hingegen** ersetzen kann („Mein linker Nebenmann dagegen schien auch noch jetzt...“).

daher

Daher ist kein Konnektor, wenn er ein Verbzusatz ist („das kommt daher, dass...“) oder eine Richtung beschreibt. **Daher** ist ein Konnektor, wenn man es durch **deswegen** ersetzen kann („Lassen Sie uns daher Laos den Laoten überlassen.“).

danach

Danach ist kein Konnektor, wenn es anstelle von „nach etwas“ benutzt wird („danach zu fragen“). **Danach** ist ein Konnektor, wenn es die zeitliche Abfolge zwischen zwei Sätzen ausdrückt („Der Staatspräsident eröffnete die Feier. Danach forderte er die Ehrengäste auf, sich mit ihm vor der ewigen Flamme zu verneigen.“).

dann

Dann ist kein Konnektor, wenn es sich auf einen bestimmten Zeitpunkt bezieht („Wie er dich dann angeschaut hat, da habe ich Angst gekriegt um dich.“). **Dann** ist ein Konnektor, wenn es sich auf einen ganzen Sachverhalt bezieht und eine Folge beschreibt („Ich würde die Grosvenor Street verlassen und dann links in die Park Lane einbiegen.“), in der Verbindung mit wenn („Wenn Nationalstaaten erodieren, dann werden neue Religionskriege ausbrechen.“) oder wenn man es durch **infolgedessen** oder **danach** ersetzen kann („Die Staaten hoben hervor, was sie alles für die Flüchtlinge getan hätten, um dann über die prekäre wirtschaftliche Lage zu jammern.“).

darauf

Darauf ist kein Konnektor, wenn es ein Verbpartikel ist („kommt darauf an“, „weißt darauf hin“), eine Zeitangabe („kurz darauf“) oder eine Ortsangabe. **Darauf** ist ein Konnektor, wenn er einen zweiten Satz einleitet und damit eine zeitliche oder kausale Folge ausdrückt und man es durch **daraufhin** ersetzen könnte („Seit 1947 saß er im Repräsentantenhaus; darauf übernahm er das Landwirtschaftsministerium.“).

darum

Darum ist kein Konnektor, wenn es anstelle von „um etwas“ gebraucht wird („es ging darum, dass...“). **Darum** ist ein Konnektor, wenn es kausal zwei Sätze miteinander verknüpft und anstelle von „deshalb“ oder „deswegen“ benutzt wird („O ja, ihr seid’ ne starke Macht im Staat! Darum macht der ja auch so allerlei Buhei um und mit euch.“).

denn

Denn ist kein Konnektor, in Phrasen wie „es sei denn, ...“ oder „wer macht denn so etwas?“ oder anstelle von **als** in einem Vergleich („Dies gilt heute mehr denn je.“). **Denn** ist ein Konnektor, wenn er zwei Sätze kausal miteinander verknüpft („So schlecht ist alles auch wieder nicht. Denn er hält noch einen Trumpf in petto.“).

doch

Doch ist kein Konnektor, wenn es innerhalb einer NP steht („in einem doch sehr rigidem System“). **Doch** ist in der Kombination mit aber („Aber einmal hat er doch gemerkt...“) oder als Interjektion in gesprochener Sprache kein Konnektor. **Doch** ist kein Konnektor wenn es bedeutet, dass ein Sachverhalt vorher wahr (falsch) war, und nun „doch“ falsch (wahr) ist („Inzwischen war das Ganze doch etwas langweilig geworden.“). **Doch** ist ein Konnektor, wenn es zwei Sätze miteinander verbindet („Er hat sich den Ruf eines guten Managers erworben. Doch die Mehrheit des Personals ist ihm feindlich gesonnen“).

entgegen

Entgegen ist ein Konnektor, wenn es eine Präposition ist und sich auf eine NP im Dativ bezieht („entgegen einer vielverbreiteten Ansicht“). **Entgegen** ist kein Konnektor, wenn es ein Verbpartikel oder eine Richtungsangabe ist („etwas entgegen nehmen“).

ferner

Ferner ist kein Konnektor, wenn es sich um den Komparativ oder eine andere Form von fern handelt („in ferner Zukunft“). **Ferner** ist ein Konnektor, wenn er zwei Sätze miteinander verknüpft und eine „zusätzlich dazu“-Beziehung ausdrückt und durch **des Weiteren** ersetzt werden kann („Sie sollten ferner von allen Feindseligkeiten gegen die CSSR in ihren Reden Abstand nehmen.“).

nebenher

Nebenher ist kein Konnektor, wenn es eine Richtungsangabe oder ein Adjektiv ist („Ich ging gemütlich nebenher , schwatzte lustig und machte kleine Scherze.“). **Nebenher** ist ein Konnektor, wenn es ausdrückt, dass zwei Sachverhalte gleichzeitig wahr bzw. eingetreten sind („Es liefert uns Schinken, Wurst, Koch- und Bratfleisch und dazu große Fettmengen für unsere Küche. Nebenher spendet es uns durch seine Haut vielerlei Leder.“).

nur

Nur ist kein Konnektor, wenn es sich auf die dahinter stehende NP oder PP bezieht. **Nur** kann nur ein Konnektor sein, wenn es sich auf einen Satz bzw. ein finites Verb bezieht („Nur reicht die Problematik weit darüber hinaus“). **Nur** am Satzanfang gefolgt von einem finiten Verb ist meist ein Konnektor. Wenn auf **nur** ein finites Verb folgt, kann man es umstellen und prüfen, ob sich die Semantik geändert hat. Wenn das nicht zutrifft, handelt es sich um einen Konnektor. („Nur erfolgt die Zurückweisung in einer vornehmen Form.“ $\neq$ „Nur die Zurückweisung erfolgt in einer vornehmen Form.“)

seit

Seit ist kein Konnektor, wenn es einen Zeitpunkt beschreibt („seit jener Zeit“) und sich auf eine NP bezieht. **Seit** ist ein Konnektor, wenn es Haupt- und Nebensatz miteinander verknüpft („Erst 200 Generationen sind vergangen, seit ein Mensch namens Abraham aufstand, um sein Land und seine Heimat zu verlassen.“).

seitdem

Seitdem ist kein Konnektor, wenn es wie eine einfache Zeitangabe verwendet wird („... und seitdem fährt er nicht mehr mit dem Auto.“). **Seitdem** ist ein Konnektor, wenn er einen neuen (Neben-)Satz einleitet und ihn somit mit einem anderen Satz in Beziehung bringt („Die Preise für Grund und Boden sind um 50 Prozent gestiegen, seitdem Wilhelm II. Schloß Doorn gekauft hat.“).

so

So ist kein Konnektor, wenn man es durch „auf diese Weise“ ersetzen kann oder wenn darauf ein Adjektiv oder ein Adverb folgt („so verrückt“) oder wenn man es vergleichend benutzt („so wie er da saß“). **So** ist ein Konnektor, wenn es zwei Sätze miteinander verbindet oder in der Verbindung mit **daß**, **so dass** („Und wenn es nachher hieß Hans-Joachim zur Linden habe den Herzinfarkt » im Laufstall seines alten Kinderzimmers erlitten«, so ist das genauso falsch.“).

sonst

Sonst ist kein Konnektor, wenn er sich auf eine NP o. Ä. bezieht („ob sich in der Nähe eine Omnibushaltestelle oder sonst eine Möglichkeit zum Weiterkommen befindet“). **Sonst** ist kein Konnektor, wenn es „zu allen anderen Zeitpunkten“ meint. **Sonst** ist ein Konnektor, wenn er zwei Sätze miteinander verbindet („Gehorchen Sie Ihrem Geliebten und fliehen Sie! Sonst sind Sie verloren.“).

soweit

Soweit ist kein Konnektor, wenn es anstelle von **wie weit**, **so weit** benutzt wird („Soweit sind wir noch nicht.“). **Soweit** ist ein Konnektor, wenn es einen Nebensatz anstelle von **wenn** anschließt („Der Präsident der Bundesregierung gibt Gelegenheit zur Stellungnahme, soweit diese nicht bereits vorliegt“).

statt

Statt ist kein Konnektor, wenn es ein Verbpartikel ist („findet statt“). **Statt** ist kein Konnektor, wenn es zwei NPs miteinander verbindet („mehr Wohnungen statt Kasernen“). **Statt** ist ein Konnektor, wenn er zwei VPs miteinander verbindet („Die afrikanischen Führer neigten dazu, ausländische Gepflogenheiten mit geringen Änderungen zu übernehmen, statt nach einer Regierungsform zu suchen, die für ihr Land am besten geeignet sei“).

trotz

Trotz ist ein Konnektor, wenn es als Präposition benutzt wird („Trotz intensiver Suche fand man keinerlei sichere Ursache.“). Es kann auch das Nomen **Trotz** gemeint sein, und ist damit kein Konnektor („Aus Trotz habe ich nicht geheult.“).

und

Wenn **und** innerhalb einer Phrase (NP, PP, ...) auftaucht, dann ist es kein Konnektor („die durch Krieg und Nachkriegszeit ins Wanken geratenen Rollenmuster“). **Und** muss zwei Sätze miteinander verbinden, um ein Konnektor zu sein („Die Herren konnten sich das gar nicht vorstellen, und so entwickelte Siesina einen Pilotfilm.“).

während

Während ist kein Konnektor, wenn es eine Zeitspanne beschreibt („während einer Konferenz“) und sich auf eine NP bezieht. **Während** ist ein Konnektor, wenn es Haupt- und Nebensatz miteinander verknüpft („Während Focus die fröhlichen Aufsteiger zu seinen Lesern zählt, ist der Spiegel-Leser kein bequemer Mensch.“).

wegen

Wegen ist ein Konnektor, wenn es wie eine Präposition verwandt wird und sich auf eine NP bezieht („wegen des Erfolgs des Buches“, „wegen des Lärms“). **Wegen** kann auch eine Form von **Weg** sein und ist dann natürlich kein Konnektor („auf neuen Wegen“).

weshalb

Weshalb ist kein Konnektor, wenn es ein Fragewort ist und durch **warum** ersetzt werden kann („Weshalb hatte es der Präsident nötig, mit der Möglichkeit eines ’Atomkrieges’ zu drohen?“). **Weshalb** ist ein Konnektor, wenn es kausal einen Nebensatz einfügt, und man es mit Verbumstellung durch **deshalb** ersetzen kann („Immerhin sind die Werke in

Halbzeug noch gut besetzt -, weshalb der Verband aus dem Auslandsmarkte auch nur wenige Geschäfte hereinnehmen konnte“).

weswegen

Weswegen ist kein Konnektor, wenn es ein Fragewort ist und durch **warum** ersetzt werden kann („die Ursache weswegen man dieser dringlichen Frage noch immer nicht näher getreten ist“). **Weswegen** ist ein Konnektor, wenn es kausal einen Nebensatz einfügt, und man es mit Verbumstellung durch **deswegen** ersetzen kann („Dagegen sind die Augen der Katze von der Beute in Anspruch genommen, weswegen das lauernde Tier dem Jäger leicht zum Opfer fällt“).

wie

Wie ist kein Konnektor, wenn es ein Interrogativadverb ist („wie ... ?“), ein vergleichendes **Wie** („sieht aus wie ...“, „so, wie er sich bewegt, ...“, „wie lang“, „wie gut“). **Wie** ist kein Konnektor, wenn man es durch „auf welche Art und Weise“ ersetzen kann („Beide lachten, fragten sich, wie er diesmal auftauchen würde.“). **Wie** ist ein Konnektor, wenn es einen Nebensatz einleitet („..., wie der zuständige Mann im Ministerium bestätigte.“).

wogegen

Wogegen ist kein Konnektor, wenn es anstelle von „gegen etwas“ benutzt wird, oder als Fragewort („wogegen er nichts einzuwenden hatte“). **Wogegen** ist ein Konnektor, wenn er zwei Sätze miteinander verbindet und einen Kontrast herstellt und wenn man es durch **wohingegen** ersetzen kann („Die Selbstverwaltungskörperschaften haben das Recht der Mitbestimmung, wogegen die Betriebsräte nur ein beratendes Organ sein sollen.“).

womit

Womit ist kein Konnektor, wenn es anstelle von „mit etwas“ benutzt wird („womit ich nicht gerechnet habe“). **Womit** ist ein Konnektor, wenn es einen Nebensatz einläutet und eine Folge beschreibt und man es mit Verbumstellung durch **damit** ersetzen kann(„Der Schubal wird mir sowieso mit der Zeit viel zu selbständig, womit ich aber nichts zu Ihren Gunsten gesagt haben will.“).

wonach

Wonach ist kein Konnektor, wenn es sich auf eine NP bezieht, und „nach dem“ bedeutet, oder sich als Relativpronomen verhält („das, wonach ich suchte“). **Wonach** ist ein Konnektor, wenn es anstelle von **danach** temporal verwendet wird („Das Treffen lief zu seinen Gunsten, wonach er beschwingt nach hause ging.“).

worauf

Worauf ist kein Konnektor, wenn es anstelle von „auf etwas“ gebraucht wird („Worauf es ankommt ist, dass...“). **Worauf** ist ein Konnektor, wenn man es durch **woraufhin** ersetzen kann („Er brach ab, Renate sah ihn im Zimmer stehn und sich mit der Hand an die Stirn schlagen, worauf er lachte und sagte:...“).

zugleich

Zugleich ist kein Konnektor, wenn es sich auf eine NP o. Ä. bezieht („von mehreren Seiten zugleich“). **Zugleich** ist ein Konnektor, wenn er zwei Sätze miteinander verbindet („Von der Romantik ausgehend, wandte er sich der Dodekaphonie zu. Zugleich griff er Liszts Art der Thementransformation auf.“).

Literaturverzeichnis

[ASM11] Agarwal, Manish; Shah, Rakshit; Mannem, Prashanth: Automatic question generation using discourse cues. In: *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2011 (IUNLPBEA '11), 1–9

[Bay04] Bayerl, Petra S.: Disambiguierung deutschsprachiger Diskursmarker: Eine Pilot-Studie. In: *Linguistik Online* 18 (2004)

[DS06] Dipper, Stefanie; Stede, Manfred: Disambiguating potential connectives. In: Butt, Miriam (Hrsg.): *Proceedings of the Konvens-2006 Workshop on the Lexicon-Discourse Interface*, 2006, S. 167–173

[Hut05] Hutchinson, Ben: Modelling the substitutability of discourse connectives. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2005 (ACL '05), 149–156

[Pas03] Pasch, R.: *Handbuch der deutschen Konnektoren: linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. De Gruyter, 2003 (Schriften des Instituts für Deutsche Sprache)

[PN09] Pitler, Emily; Nenkova, Ani: Using syntax to disambiguate explicit discourse connectives in text. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2009 (ACLShort '09), 13–16

[Sch94] Schmid, Hemut: Probablistic part-of-speech tagging using decision trees. In: *Prodeedings of International Conference on New Methods in Language Processing*, 1994, S. 44–49

[Sch95] Schmid, Hemut: Improvements in Part-of-Speech Tagging with an Application to German. In: *Prodeedings of the ACL SIGDAT-Workshop*, 1995, S. 47–50

[SL02] Saggion, Horacio; Lapalme, Guy: Generating indicative-informative summaries with sumUM. In: *Comput. Linguist.* 28 (2002), Dezember, Nr. 4, 497–526. `http://dx.doi.org/10.1162/089120102762671963`. – ISSN 0891–2017

[TBT+] Taboada, Maite; Brooke, Julian; Tofiloski, Milan; Voll, Kimberly; Stede, Manfred: Lexicon-based methods for sentiment analysis. In: *Comput. Linguist.* 37, Nr. 2, 267–307. `http://dx.doi.org/10.1162/COLI_a_00049`. – ISSN 0891–2017

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Ort, Datum Unterschrift