Ludwig-Maximilians-Universität München

Institut für Statistik

Bachelorarbeit zum Thema:

Statistische Analysen von Assoziationsmaßen in der quantitativen Korpuslinguistik

Bearbeiterin: Ekaterina Manuilova
Ruedorffer Str. 9
83022 Rosenheim

Matr. Nr.: 4088093

Betreuer: Prof. Dr. Helmut Küchenhoff

Abgabedatum: 24. August 2009

Inhaltsverzeichnis

1 Einleitung

1.1 Quantitative Korpuslinguistik

Die Korpuslinguistik ist eine relativ alte Teildisziplin der Sprachwissenschaft, deren Anfänge sich mindestens bis ins 19.Jahrhundert zurückdatieren lassen. Sie hat aber als sprachwissenschaftliche Methode lange Zeit keine breite Beachtung gefunden. Dies hat sich in den letzten Jahren rapide geändert. Die wichtigsten Gründe dafür sind die Entwicklung der Rechenkapazitäten der Computer und die breite Verfügbarkeit sprachwissenschaftlicher Korpora in elektronischer Form.

Unter einem Korpus (Korpora) versteht man ganz allgemein eine Zusammenstellung von Texten auf der Grundlage eines oder mehrerer Kriterien. Diese Sammlung dient als Grundlage für sprachwissenschaftliche Untersuchungen; deswegen muss diese möglichst repräsentativ für die zu untersuchende Sprache sein. Üblicherweise liegt ein Korpus in elektronischer Form vor und muss unbedingt computerlesbar sein.

Korpuslinguistik untersucht sprachliche Phänomene auf Grundlage eines sprachwissenschaftlichen Korpus. Eine wichtige Rolle dabei spielt spezielle korpuslinguistische Software. Alle dazugehörigen Programme sollen in der Lage sein

- Häufigkeitslisten zu erstellen: alle Wortformen in einem Korpus erfassen und für jede Wortform errechnen, wie häufig diese im Korpus vorkommt
- Konkordanzen zu erstellen: für ein beliebiges Suchwort (oder eine Kette von Suchwörtern) alle Vorkommen in einem benutzerdefinierten Kontext auflisten
- Kollokatenlisten zu erstellen: für ein beliebiges Suchwort Häufigkeitslisten der Wörter erstellen, die an einer bestimmten Position rechts oder links von diesem Wort auftauchen.

Um Häufigkeitslisten, Kollokatelisten und Konkordanzen auf sprachwissenschaftliche Fragestellungen zu beziehen, werden bedingte Häufigkeiten berechnet, so dass erkennbar wird, wie häufig bestimmte Phänomene unter bestimmten Bedingungen auftreten.

Korpuslinguistik ist also die Operationalisierung sprachwissenschaftlicher Fragestellungen derart, dass bedingte Häufigkeiten in einem sprachwissenschaftlichen Korpus die abhängige Variable darstellen.[6]

1.2 Collostructional analysis und seine Assoziationsmaße:

Eines der sprachwissenschaftlichen Phänomene, die mit Hilfe der Korpuslinguistik in der letzten Zeit intensiv untersucht werden, ist die Assoziation zwischen grammatikalischen Strukturen und einzelnen Wörtern. Mehrere Studien zum diesen Thema wurden vor allem von A. Stefanovitsch und S.Grieß durchgeführt. Die theoretische Grundlage dafür ist die Behauptung, dass Lexika und Grammatik einer Sprache stark miteinander verbunden sind. Man betrachtet also eine Sprache als eine Menge unterschiedlicher völlig oder teilweise fester Wortverbindungen. Unter diesem Augenwinkel nähern sich die Methoden zur Untersuchung der Grammatik den Methoden zur Untersuchung der Lexika, vor allem den korpuslinguistischen Methoden. Die spezielle Methode zur Untersuchung der Assoziation zwischen Wörtern und grammatikalischen Strukturen nennen Stefanowitsch und Grieß "collostructional analysis". [8]

Ein grundlegendes Problem der "collostructional analysis" ist die Wahl des geeigneten Maßes für die Assoziation zwischen den Wörter und grammatikalischen Konstrukten. In der Literatur wurden mehr als 40 solcher Maße vorgeschlagen. Fast alle davon basieren auf 2x2 Häufigkeitstabellen folgender Gestalt [1].

	Konstruktion	¬Konstruktion	Σ
Lexem	O11	O12	R1
¬Lexem	O21	O22	R2
Σ	C1	C2	N

Tabelle 1: Basis-Kontingenztafel der Collostructional analysis

Hier:

- O11 - Häufigkeit von L in C
- O12 - Häufigkeit von L in allen anderen Konstrukten
- O21 - Häufigkeit von C zusammen mit allen Lexemen außer L
- O22 - Häufigkeit von allen Konstruktionen außer C mit allen Lexemen außer L

Die erste und einfachste Messung, die in Frage kommt, ist die absolute Häufigkeit O11. Wenn man die absoluten Häufigkeiten O11 der Lexeme vergleicht, ist zu erwarten, dass ein Lexem mit einer höheren absoluten Häufigkeit stärker mit dem Konstrukt assoziiert ist.

Dies stimmt aber nicht immer, da nicht berücksichtigt wird, wie oft das Lexem im Korpus überhaupt vorkommt (Häufigkeit R1). Bei einem seltenen Wort wird O11 klein - sogar dann, wenn die Anziehung des Wortes zur Konstruktion sehr stark ist. S.Grieß nennt drei Maße, die mehr Information über die bedingten Häufigkeiten berücksichtigen und am Öftesten benutzt werden [7]. Dies sind "pointwise Mutual Information", "t-score" und logarithmierter p-Wert des exakten Tests nach Fisher.

Viele Assotiationsmaße vergleichen die beobachtete Häufigkeit O11 mit der erwarteten Häufigkeit unter der Annahme, dass das Lexem L und das Konstrukt C stochastisch unabhängig sind. Diese erwartete Häufigkeit wird im Folgenden E11 genannt und nach folgender Formel berechnet:

$$E11 = \frac{C1 \cdot R1}{N}$$

Laut Grieß [7], liefern die Assoziationsmaße relativ ähnliche Ergebnisse, doch in manchen Fällen ist der Output sehr unterschiedlich. Deswegen muss man bei der Wahl der Messungen auf einige Eigenschaften der Maße achten [7].

Erstens ist es wichtig, zu beachten, dass die Assoziationsmaße dazu neigen, die Assoziation zu überschätzen, wenn es um niedrige Häufigkeiten geht. In diesem Fall sinkt die Zuverlässigkeit der Ergebnisse.

Zweitens verlangen viele Maße eine Verteilungsannahme, meistens die Normalverteilung. Doch die Korpusdaten sind in der Regel nicht normalverteilt. Deswegen muss man bei der Wahl der approximativen Assoziationsmaße aufpassen, oder solche Maße auswählen, die keine Verteilungsannahmen verlangen, wie zum Beispiel den exakten Test nach Fisher oder den Binomialtest.

Drittens ist es wichtig zu wissen, dass die Assoziationsmaße unterschiedlich auf die niedrigen Häufigkeiten der Wörter reagieren, deswegen liefern sie unterschiedliche Ranglisten in Abhängigkeit von den absoluten Häufigkeiten. Zum Beispiel ist bekannt, dass MI eine sehr hohe Assoziation bei niedrigen Häufigkeiten wiedergibt, solange diese eine kleine Variation zeigen. Der t-Score ist in diesem Fall robuster.

1.3 Vergleich der Assoziationsmaße

D. Wichman [2] betrachtet in seinem Artikel 47 Assoziationsmaße, untersucht ihre Ähnlichkeit und versucht, ihre Güte mit Hilfe psycholinguistischer Daten zu beurteilen.

Die in Betracht bezogene Maße teilt er in sieben Gruppen auf:

1. **Likelihoods** berechnen die Wahrscheinlichkeit des beobachteten gemeinsamen Auftretens von L und C O11 unter der Annahme, dass die Wahrscheinlichkeit des Auftretens von L in C sich von den anderen Lexemen nicht unterscheidet. Beispiel: Binomial-likelihood, Poisson-likelihood, Hypergeometrische-Likelihood usw.
2. **Genaue Tests** berechnen die p-Werte der Tests ohne Verteilungsannahmen, die Wahrscheinlichkeit unter H0-Hypothese den beobachteten oder einen extremeren Wert in Richtung der Alternative zu erhalten. H0-Hypothese: Assoziation zwischen L und C unterscheidet sich nicht von den anderen Lexemen. Beispiel: Binomialtest, Fisher-Test.
3. **Approximative Tests** berechnen die p-Werte unter der gleichen H0-Hypothese. Diese Tests verlangen aber bestimmte Verteilungsannahmen von L und C im Korpus. Beispiele: t-test, Chi-quadrat-Test. Dabei werden nicht nur die p-Werte, sondern auch die Teststatistiken selbst als Assoziationsmaße benutzt.[1]
4. **Ml-Schätzungen** der unterschiedlichen Koeffizienten der Assoziationsstärke. Beispiele: pointwise mutual information (MI), odds ratio, jaccard, minimum sensitivity, relative risk usw.
5. Konservative Schätzung. der Assoziationsstärke (Schätzung eines doppelseitigen Konfidenzintervalls für MI)
6. Messungen aus der Informationstheorie. Beispiele: MI, local MI usw.
7. Heuristische Messungen. Beispiele: Absolute Häufigkeit, Random, MI^2.

Auf Basis von "Britisch National Corpus World Edition" berechnet er die Assoziationsmaße zur bestimmten grammatikalischen Konstruktionen für 21 ausgewählte Verben. Nach jedem Maß erstellt er die Ranglisten für die Assoziation und vergleicht diese mit Hilfe einer Korrelationsmatrix nach der agglomerativen Clusteranalyse mit Average Linkage miteinander.

Danach teilt Wiechmann alle Koeffizienten in 7 Gruppen, die relativ homogene Ergebnisse liefern. Die Güte der Messungen aus den einzelnen Gruppen untersucht Wichman auf Basis des psycholinguistischen Experiments von Kennison 2001. Mittels eines Regressionsmodells wurde die Beziehung zwischen der Assoziationsstärke von Wörtern und Konstruktionen und der Leseschwierigkeit untersucht. Für jedes Maß wurde ein Regressionsmodell erstellt und ein adjustierters Bestimmtheitsmaß berechnet. Das Ergebnis sieht man in Tabelle 2.

[1] S.Gries. Useful statistics for corpus linguistics.

1	Minimum Sensitivity ($R^2_{adj} = 0.34$)
2	Fisher Exact Test (0.29)
3	Korrigierter Chi-Quadrat Test (0.28)
4	Binomiallikelihood (0.27)
5	Absolute Häufigkeit O11 (0.25)
6	Mutual Information (0.23)
7	odds ratio (0.22)
...	...
...	Poisson.mu10 (0.14)

Tabelle 2: Rangliste der Assoziationsmaße in der Untersuchung von D.Wiechmann

1.4 Zielsetzung

Das Ziel der vorliegenden Arbeit sind Untersuchung und Vergleich der Eigenschaften im Rahmen der Collostructional analysis folgender Assoziationsmaße:

- Jaccard-Koeffizient (jac)
- Minimum Sensitivity (MS)
- pointwise Mutual Information (MI)
- relatives Risiko (rr)
- Odds Ratio (or)
- z-Score (zscore)
- p-Wert des exakten Test nach Fischer (ft)
- p-Wert des χ^2-Unabhängigkeitstests

Da die meisten Assoziationsmaße von den Häufigkeiten C1,R1 und N abhängen ist es im Allgemeinen problematisch, die Messungen aus unterschiedlichen Korporas und die Messungen für unterschiedliche Konstrukte zu vergleichen. In den meisten Fällen erstellt man mittels eines Assoziationsmaßes eine Rangliste der Lexeme nach der Stärke ihrer Anziehung zu einem Konstrukt.

Auf Basis des gegebenen Datensatzes APPENNEU.XLS werden die entsprechenden Ranglisten für die einzelnen Maße und Konstrukte erstellt. Da wir uns nur für die Reihenfolge der Lexeme in den Ranglisten interessieren, vergleichen wir diese mittels des Spearman'schen Korrelationskoffizienten. Die entsprechenden Korrelationsmatrizen werden danach für die Clusteranalyse benutzt, welche die gegebenen Maße nach ihrer Ähnlichkeit klassifiziert.

In der Arbeit wird ein spezifisches Problem der Collostructional analysis bei der Errechnung der Basis-Kontingenztafel berücksichtigt, das H.-J. Schmidt und H. Küchenhoff als "Problem der 4.Zelle" bezeichnen [1]. Während die Bestimmung der Häufigkeiten $O11$, $O12$, $O21$ im Korpus relativ eindeutig ist, stellt die Errechnung der Häufigkeit $O22$ (Anzahl aller anderer Konstrukte außer C, die alle andere Lexeme außer C enthalten) eine besondere Herausforderung für den Spezialisten dar. Diese hängt von der Interpretation des Wortes "andere" ab und wird deswegen relativ willkürlich ausgewählt. Bei den fixen $O11$, $O12$ und $O21$ bestimmt $O22$ den Stichprobenumfang N. Die Robustheit der Maße gegen der Veränderung des Stichprobenumfanges ist also eine wichtige Eigenschaft für die Assoziationsmaße der Collostructional analysis. In der Arbeit werden die Messungen für zwei Stichprobenumfänge $N = 10^6$ und $N = 10^7$ untersucht und verglichen.

1.5 Datensatzbeschreibung

Der Datensatz APPENNEU enthält absoluten Häufigkeiten des gemeinsamen Auftretens (Häufigkeit O11) für 670 Lexeme und 8 grammatikalische Konstrukte. Die absoluten Häufigkeiten der Lexeme im Korpus (Häufigkeiten R1) sind im Datensatz extra angegeben. Die Verteilung der Häufigkeiten ist linkssteil.

Min	1.Quartil	Median	Mittelwert	3.Quartil	Maximum
10	1073	3472	9416	9886	342900

Tabelle 3: Verteilung der Häfigkeiten O11 in dem Datensatz APPENNEU

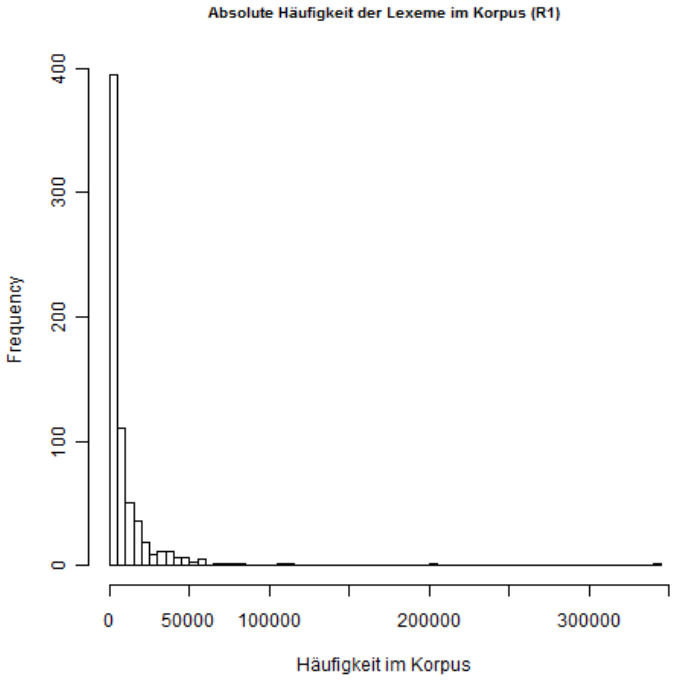

Abbildung 1: Verteilung O11 in dem Datensatz APPENNEU

Durch Summieren des gemeinsamen Auftretens aller Lexeme mit dem einzelnen Konstrukt wurden die absoluten Häufigkeiten der Konstrukte im Korpus berechnet (Häufigkeit C1). Selbstverständlich, tritt nicht jedes Lexem mit jedem Konstrukt zusammen auf. Dieser Situation entsprechen Null-Einträge im Datensatz. Die Tabelle 4 zeigt die Häufigkeit C1 für alle Konstrukte und die Anzahl der Lexeme, die gemeinsam mit dem gegebenen Konstrukt vorkommen. Die Abbildung 2 stellt diese Informationen grafisch dar.

Drei der Konstrukte (SNTH,THSN,SNTO) kommen besonders häufig vor. Der Anteil der Lexeme, die gemeinsam mit diesen Konstrukten auftreten unterscheidet sich: während mit dem Konstrukt SNTH 97% aller Lexeme vorkommen, treten nur 30% der Lexeme mit dem Konstrukt SNTO. Das Konstrukt SNBEWH ist sehr selten; die restlichen Konstrukte haben vergleichbare absolute Häufigkeiten.

2 Assoziationsmaße und ihre Eigenschaften

Die untersuchten Assoziationsmaße unterteilen wir nach Beziehung zu N und nach ihrer Herkunft in 4 Gruppen:

Konstrukt	C1	Anzahl der Lexeme mit $O11 \neq 0$	Anteil der Lexeme mit $O11 \neq 0$
SNBEWH	1712	21	3%
SNBETO	21876	162	24%
THBESN	23509	565	84%
SNWH	29492	31	5%
SNBETH	30992	366	55%
SNTH	141476	350	52%
THSN	186433	650	97%
SNTO	228165	200	30%

Tabelle 4: Häufigkeiten C1 und Anteil der Lexeme mit $O11 \neq 0$

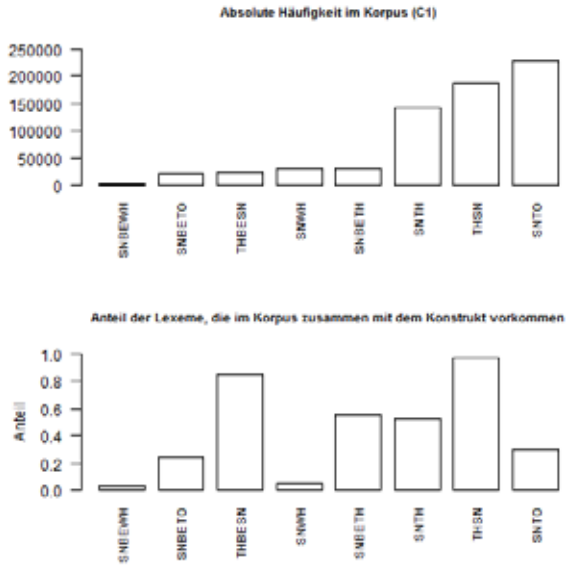

Abbildung 2: Barplots für C1 und Anteil der Lexeme mit $O11 \neq 0$

- Von der 4.Zelle unabhängige Assoziationsmaße (MS und Jaccard)
- Von der 4.Zelle abhängige Assoziationsmaße, aber keine p-Werte (Odds Ratio, relatives Risiko, MI und z-Score)
- p-Werte (p-Werte des exakten Test nach Fisher und des χ^2-Tests)
- Likelihood der Poissonverteilung

In der ersten Gruppe, zusammen mit MS und Jaccard, werden zusätzlich Reliance, Attraction und das neu eingeführte Maß L2 betrachtet.

2.1 Von der 4.Zelle unabhängige Assoziationsmaße

2.1.1 Reliance und Attraction

Um den ersten Eindruck von der Assoziation zwischen den Lexemen und Konstrukten im Datensatz APPENNEU zu bekommen, betrachten wir die geschätzten bedingten Wahrscheinlichkeiten Reliance und Attraction [2]:

$reliance = \frac{O11}{R1} = \widehat{P}\,(Konstrukt|Lexem)$

$attraction = \frac{O11}{C1} = \widehat{P}\,(Lexem|Konstrukt)$.

Der **Wertebereich** der beiden Messungen liegt zwischen 0 und 1.

Bei den meisten Konstrukten ist Attraction viel kleiner als Reliance. Die Ausnahmen sind Konstrukte SNWH (dieser ist nicht sehr selten aber tritt mit sehr wenigen Lexemen auf), SNBEWH (ein sehr seltenes Konstrukt), THSN und THBESN (diese kommen mit sehr vielen Lexemen vor). Die meisten Lexeme haben viele Ausreißer. Die Verteilungen sind also sehr stark nach links verschoben.

Konstrukt	median(attraction)	median(reliance)	C1	Anteil der Lexeme mit $O11 \neq 0$
SNTO	0.0014	0.0564	228165	30%
SNTH	6e-04	0.0287	141476	52%
THSN	3e-04	0.018	186433	87%
SNBEWH	0.007	0.0023	1712	3%
SNWH	0.0105	0.0161	29492	5%
SNBETO	0.0019	0.0061	21876	24%
SNBETH	5e-04	0.0056	30992	55%
THBESN	5e-04	0.0034	23509	84%

Tabelle 5: Vergleich der Medianen von Reliance und Attraction

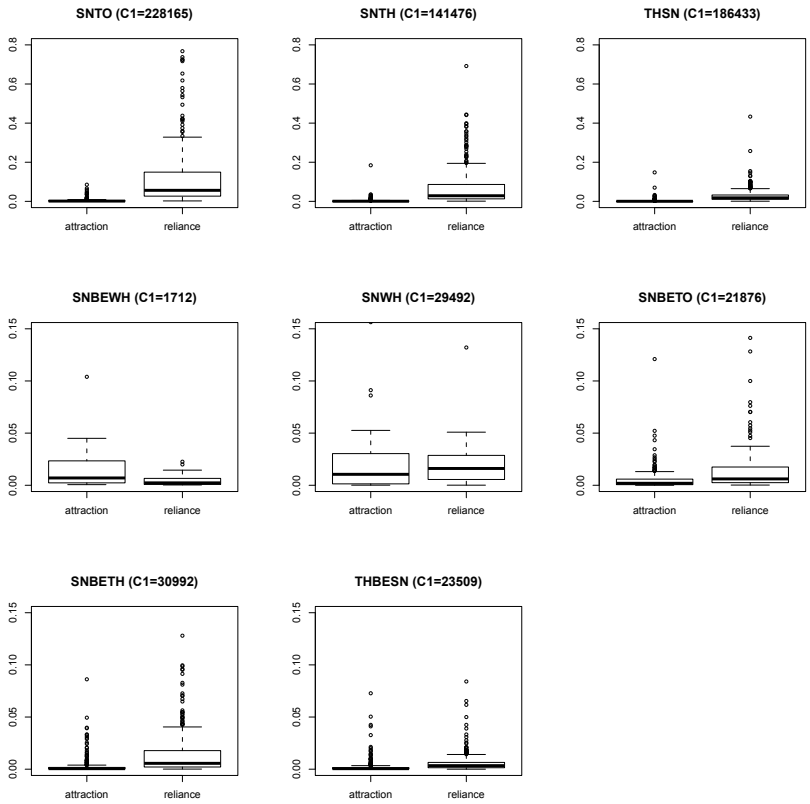

Abbildung 3: Verteilung von Reliance und Attraction in den Konstrukten

Der Wertebereich von Reliance und Attraction, ihre Verteilungen und ihre Beziehung zueinander sind für jedes Konstrukt sehr individuell: sie hängen von der Häufigkeit C1 und von der Anzahl der Lexeme mit

$O11 \neq 0$ ab. Deswegen ergibt es nicht viel Sinn die Werte von Reliance und Attraction ohne Bezug zum Konstrukt zu betrachten. Ein Wert von 0.05 für Reliance wird im Konstrukt SNBETO als sehr hoch und im Konstrukt SNTO als sehr klein eingeschätzt.

Das Verhältnis zwischen Reliance und Attraction hängt direkt von dem Verhältnis zwischen den absoluten Häufigkeiten C1 und R1 ab:

$$\frac{reliance}{attraction} = \frac{O11/R1}{O11/C1} = \frac{C1}{R1} \tag{1}$$

Daraus folgt, dass wenn $R1 \approx C1$ ist, dann ist $reliance \approx attraction$, und die Information über die Assoziationsstärke zwischen dem Konstrukt und dem Lexem ist eindeutig. Im Fall $R1 \neq C1$ ist $reliance \neq attraction$. Die Frage, welche der beiden die Assoziation richtig einschätzt bleibt offen.

H.-J. Schmidt und H.Küchenhoff [1] vor, Reliance und Attraction als ein zweidimensionales Assoziationsmaß zu betrachten. Das Paar enthält sowohl Informationen über Anziehung des Lexems zum zur Konstrukt als auch die Anziehung des Konstrukts zum Lexem. Diese Art der Beschreibung ist auch aus psychologischer Sicht sinnvoll, da das Paar (Reliance, Attraction) die kognitiven Prozesse im Bewusstsein des Sprechenden widerspiegelt.

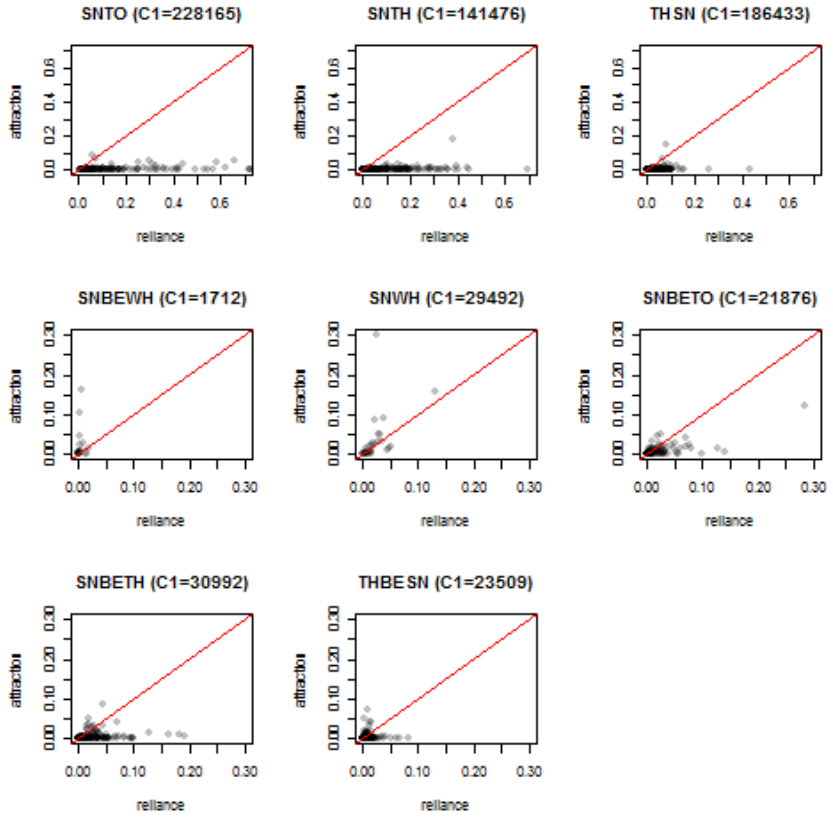

Abbildung 4: Zweidimensionale Darstellung von Reliance und Attraction

Mit Hilfe 2-dimensionaler Scatterplots (Abbildung 4) erhält man einen guten visuellen Eindruck der Beziehung von Reliance und Attraction. Die rote Hilfslinie zeigt Punkte mit $Reliance = Attraction$. Aus dem Verhältnis (1) folgt, dass für die Lexeme unter der roten Linie $R1 < C1$ gilt. Die Lexeme über der roten Linie haben $R1 > C1$.

Als Beispiel betrachten wir das häufige Konstrukt THSN (C1=186433). Hier gilt $R1 << C1$ für fast alle Lexeme, deswegen ist Attraction sehr klein und fast alle Beobachtungen liegen weit unter der roten Linie. Es gibt nur zwei Ausnahmen mit relativ hohen Werten für Attraction und Reliance. Dies sind die sehr häufigen Wörter "time" (attraction=0.148, reliance=0.080, R1=342869) und "way" (attraction=0.070, reliance=0.065, R1=201366). Man könnte vermuten, dass diese Lexeme zusammen mit dem Konstrukt THSN einen oder einige feste Bausteine im Korpus bilden und dadurch gegenseitig die Häufigkeit beeinflussen.

Die Wörter "finding" (attraction= 0.0014, reliance=0.4334, R1=586) und "brief" (attraction=0.0012, reliance=0.2575, R1=835) sind selten, deswegen haben sie keine Chance die Häufigkeit von THSN sichtbar zu beeinflussen. Andererseits, kommt das Wort "finding" in über 40% und das Wort "brief" in ca. 26% der Fälle zusammen mit dem Konstrukt THSN vor, was man als eine starke Assoziation bewerten könnte. In der Grafik liegen diese Wörter sehr nah zur x-Achse und trotzdem weit entfernt vom Zentrum $(0, 0)$.

Intuitiv könnte man sagen, dass es in beiden Fällen eine starke Assoziation zwischen dem Konstrukt und den Lexemen gibt. Je weiter also ein Punkt vom Zentrum entfernt ist, desto stärker wird die Assoziation zwischen dem gegebenem Konstrukt und dem entsprechenden Wort. Diesem Gedanken zufolge scheint es sinnvoll, als Assoziation zwischen einem Lexem und einem Konstrukt, den euklidischen Abstand des Punktes vom Zentrum zu nehmen:

$$L2 = \sqrt{attraction^2 + reliance^2}$$

Dafür spricht auch die Tatsache, dass die euklidische Distanz eine Norm für den Vektor (Reliance, Attraction) ist. Die l_1-Norm könnte man auch dem Vektor (Reliance, Attraction) zuordnen:

$$L1 = reliance + attraction$$

Es gilt: $L1 \in [0, 2]$, $L2 \in [0, \sqrt{2}]$. Für das betrachtete Beispiel erhalten wir folgende Werte (Tabelte 6):

Wort	Reliance	Attraction	L2	L1
finding	0.4334	0.0014	0.4334	0.4348
brief	0.2575	0.0012	0.2575	0.2587
time	0.080	0.148	0.1686	0.2280
way	0.065	0.070	0.09567	0.1350

Tabelle 6: L2 und L1 für die ausgewählte Wörter

Die Rangkorrelation nach Spearman für die beide Normen ist sehr hoch (siehe Tabelle 2). Im Folgenden schließen wir L2 in die Analyse ein, um ihre Messungen mit den anderen Assoziationsmaßen zu vergleichen.

Konstrukt	Korrelation
SNBETH	0.9993567
SNBETO	0.9998844
SNBEWH	0.9999978
SNTH.R	0.9997913
SNTO.R	0.9999410
SNWH.R	0.9999973
THBESN	0.9961381
THSN.R	0.9978662

Tabelle 7: Rangkorrelation für L1 und L2

2.1.2 Minimum Sensitivity

Der Koeffizient Minimum Sensitivity liefert das Minimum von Reliance und Attraction:

$MS = \min\{reliance, attraction\}$

Wertebereich: 0 bis 1 Dieser Koeffizient zeigte in der Studie von D.Wiechman [2] die besten Prognoseeigenschaften. Da die Gleichung (1) gilt, liefert Minimum Sensitivity im Fall $C1 > R1$ Attraction und im Fall $C1 < R1$ Reliance. Zum Beispiel, bei der Messung der Assoziationsstärke zwischen dem Konstrukt THSN und dem Wort "finding" (attraction= 0.0014, reliance=0.4334) wird gelten:

$MS = Attraction = 0.0014.$

Die Assoziation wird also als nicht besonders hoch eingeschätzt, obwohl das Lexem in 43% der Fälle zusammen mit diesem Konstrukt vorkommt. Überhaupt geht bei den seltenen Wörtern die Information über ihre Anziehung zu den Konstrukten verloren.

2.1.3 Jaccard

Der Koeffizient von Jaccard stammt aus der Mengentheorie und misst die Assoziation zwischen zwei Mengen. Die Basis-Kontingenztafel kann man mit Hilfe eines Diagramms von Euler folgendermaßen darstellen (Abbildung 5).

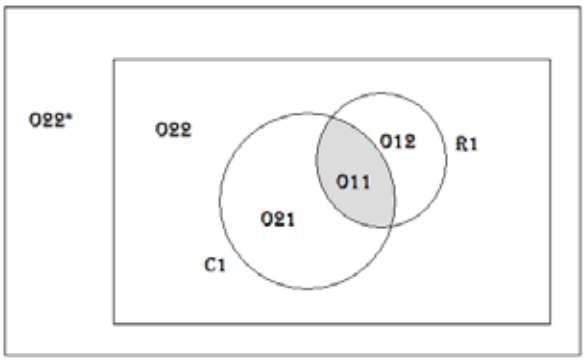

Abbildung 5: Eulerdiagramm für die Basis-Kontingenztafel

Bemerkung: aus dem Diagramm sieht man, dass (aus der Sicht der Mengentheorie), die Mächtigkeit der Menge O22 keinen Einfluss auf die gegenseitige Anziehung der Mengen $C1$ und $R1$ hat.

Zur Berechnung des Koeffizienten von Jaccard wird folgende Formel verwendet:

jaccard= $\frac{|R1 \cap C1|}{|R1 \cup C1|} = \frac{|O11|}{|O11|+|O12|+|O21|}$.

Die **Werte** liegen zwischen 0 und 1.

Im Fall $R1 << C1$ verliert Jaccard genauso wie MS die Information über die starke Anziehung eines Lexems zur Konstruktion (und umgekehrt). Die entsprechende Situation stellt das nächste Eulerdiagramm dar (Abbildung 6): obwohl die Menge R1 fast vollständig in der Menge C1 liegt, liefert Jaccard kein außergewöhnlich großes Ergebnis, da die Menge $C1$ viel größer als $R1$ ist.

MS und Jaccard liefern oft ähnliche Ergebnisse:

$C1 >> R1 \Rightarrow$ jaccard=$\frac{O11}{C1+R1-O11} \approx \frac{O11}{C1} = MS$

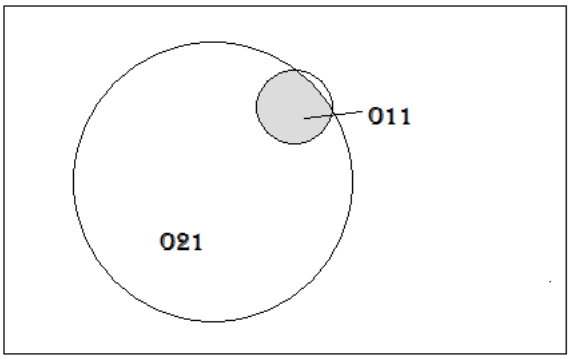

Abbildung 6: Beispiel

$C1 << R1 \Rightarrow$ jaccard=$\frac{O11}{C1+R1-O11} \approx \frac{O11}{R1} = MS.$

Tabelle 8 stellt die Rangkorrelation für Jaccard, MS und L2 für alle Konstrukte dar.

Konstrukt	C1	Jaccard-MS	Jaccard-L2
SNBEWH	1712	0.9660	0.7509
SNBETO	21876	0.9930	0.9701
THBESN	23509	0.9840	0.9835
SNWH	29492	0.9888	0.9773
SNBETH	30992	0.9882	0.9905
SNTH	141476	0.9975	0.9994
THSN	186433	0.9874	0.9997
SNTO	228165	0.9764	0.9989

Tabelle 8: Rangkorrelation für Jaccard, MS und L2

Aus der Tabelle sieht man, dass alle Koeffizienten miteinander sehr stark korrelieren. Nur im Fall SNBEWH unterscheiden sich L2 wesentlich von Jaccard und MS. SNBEWH ist das seltenste Konstrukt des Datensatzes mit wenigen Lexemen. Aus der Grafik (Abbildung 7) sieht man, dass die schwächere Korrelation an den Ausreißern liegt, die aufgrund der wenigen Beobachtungen(nur 21 Lexeme kommen mit dem Konstrukt zusammen vor) stark die Korrelationskoeffizienten beeinflussen.

Die Tabelle 9 vergleicht die Wörter-Ausreißer. Man merkt, dass bei dem starken Unterschied zwischen Reliance und Attraction Jaccard das Minimum und L2 das Maximum der beiden liefert.

noun	O11	R1	C1	attraction	reliance	MS	jaccard	L2
issue	77	32881	1712	0.0450	0.0023	0.0023	0.0022	0.0450
problem	277	59600	1712	0.1618	0.0046	0.0046	0.0045	0.1619
question	959	42406	1712	0.5602	0.0226	0.0226	0.0222	0.5606
thing	178	80013	1712	0.1040	0.0022	0.0022	0.0022	0.1040
dilemma	19	2862	1712	0.0111	0.0066	0.0066	0.0042	0.0129
mystery	46	5949	1712	0.0269	0.0077	0.0077	0.0060	0.0280
problem	277	59600	1712	0.1618	0.0046	0.0046	0.0045	0.1619
puzzle	27	1349	1712	0.0158	0.0200	0.0158	0.0089	0.0255
question	959	42406	1712	0.5602	0.0226	0.0226	0.0222	0.5606

Tabelle 9: Lexeme-Ausreißer für SNBEWH

Alle Koeffizienten aus der 1.Gruppe haben folgende angenehme Eigenschaften: sie sind leicht zu rechnen,

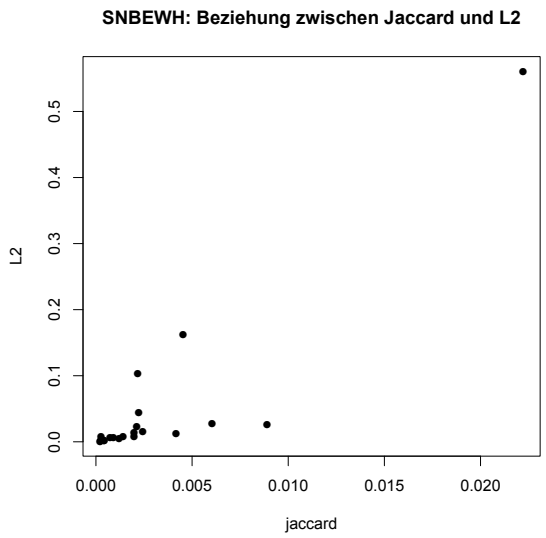

Abbildung 7: Jaccard und L2 für den Konstrukt SNBEWH

gut interpretierbar, frei von Verteilungsannahmen und hängen nicht von N ab. MS und Jaccard neigen dazu, im Fall $C1 >> R1$ und $C1 << R1$ die Assoziation zwischen Lexem und Konstrukt zu unterschätzen. L2 verhält sich in dieser Situation besser. Zu den Maßen, die versuchen, alle Informationen der Basis-Kontingenztafel zu berücksichtigen, zählen die Koeffizienten aus der 2.Gruppe.

2.2 Von der 4.Zelle abhängige Assoziationsmaße, aber keine p-Werte

In dieser Gruppe betrachten wir die Koeffizienten, die vom Stichprobenumfang abhängen aber keine p-Werte sind.

2.2.1 Odds Ratio

Odds Ratio berechnet man nach der Formel: $or = \frac{O11 \cdot O22}{O12 \cdot O21}$.

Wertebereich: $[0, \infty)$

Dieses Maß vergleicht die Chance für das Lexem L, zusammen mit dem Konstrukt C vorzukommen, mit der gleichen Chance für alle anderen Lexeme. Allgemein interpretiert man die Werte von Odds Ratio folgendermaßen: Wenn Odds Ratio eins ist, sind L und C voneinander unabhängig. Bei den Werten aus $[0, 1)$ stoßen das Lexem und das Konstrukt einander ab, im Fall, wenn Odds Ratio größer als 1 ist besteht eine Assoziation. Je größer Odds Ratio, desto größer die Assoziation. Später wird gezeigt, dass im Rahmen der Collostruction analysis die Interpretation der Werte problematisch ist.

Berechnen wir die Odds Ratios für den Datensatz mit $N = 10^7$. Abbildung 8 zeigt die Verteilung der Werte dieses Maßes. Aus der Grafik sieht man, dass die Wertebereiche des Odds Ratios für die einzelnen Konstrukte sich unterscheiden. Die größten Werte hat das seltenere Konstrukt SNBEWH, das nur mit wenigen Lexemen auftritt. Die kleinste Odds Ratios haben Konstrukte, die zusammen mit vielen Lexemen vorkommen: THBESN (kommt mit 84% der Lexeme vor) und THSN (kommt mit 97% der Lexeme vor).

Alle Verteilungen sind linkssteil und haben viele Ausreißer nach oben. Bei diesen Lexemen vermutet man

die starke Assoziation mit dem Konstrukt. Der auffälligste Ausreißer mit dem $or = 305$ ist das häufige Wort "question", das in 56% der Fälle mit dem seltenen Konstrukt SNBEWH vorkommt.

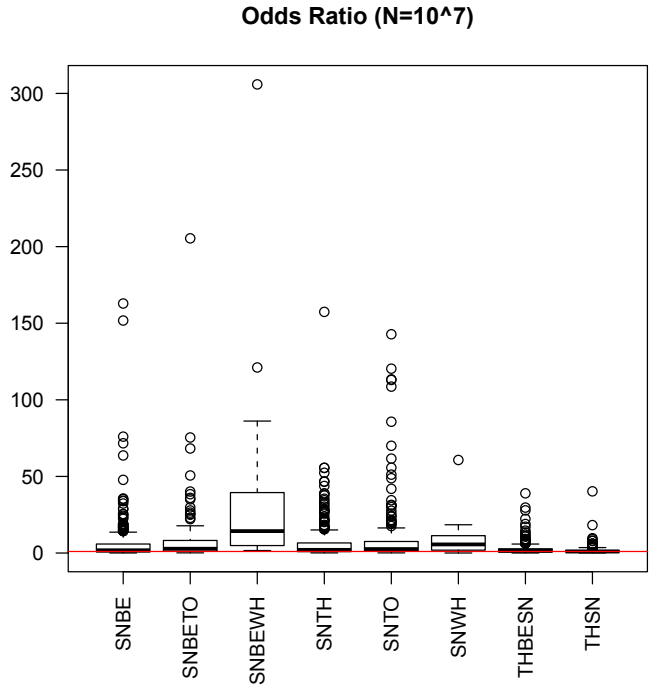

Abbildung 8: Verteilung von Odds Ratio

Die Tabelle 10 stellt die Medianen der Verteilungen für $N = 10^7$ und $N = 10^6$ dar.

Konstrukt	C1	median(odds ratio) bei $N = 10^7$	median(odds ratio) bei $N = 10^6$
SNBEWH	1712	14.30	1.37
SNBETO	21876	2.80	0.27
THBESN	23509	1.45	0.14
SNWH	29492	5.62	0.52
SNBE	30992	1.79	0.17
SNTH	141476	2.06	0.17
THSN	186433	0.96	0.07
SNTO	228165	2.57	0.19

Tabelle 10: Medianen von Odds Ratio

Aus der Tabelle sieht man, dass die Medianen bei $N = 10^6$ ca. 10 Mal kleiner als bei $N = 10^7$ sind. Dies verwundert nicht, da:

$$\frac{oddsratio(10^7)}{oddsratio(10^6)} = \frac{\frac{O11}{O12} \cdot \frac{O21}{O22}}{\frac{O11}{O12} \cdot \frac{O21}{O22*}} = \frac{O22}{O22^*} \approx \frac{N}{N^*} = \frac{10^7}{10^6} = 10$$

Während die meisten Werte für Odds Ratio bei $N = 10^7$ über Eins liegen (Abbildung 9), hat die überwiegende Mehrheit der Lexeme bei $N = 10^6$ den Odds Ratio kleiner eins. Das gleiche Lexem kann also bei unterschiedlichen N über oder unter Eins liegen. Dies macht eine direkte Interpretation des Wertes von Odds Ratio problematisch.

Es bleibt die Frage, ob die Reihenfolge der Lexeme in den Ranglisten, die mittels Odds Ratios erstellt

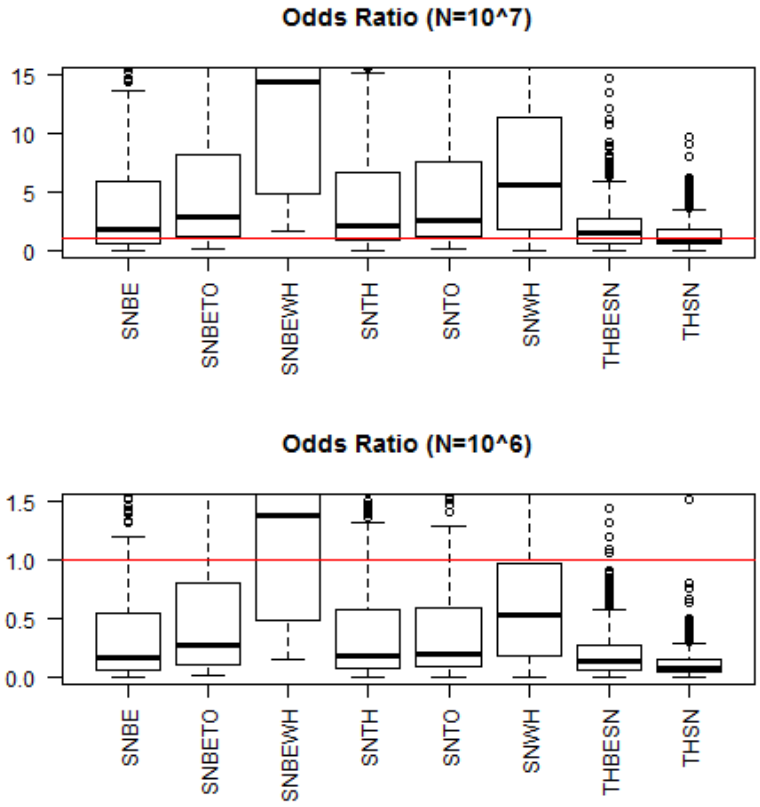

Abbildung 9: Odds Ratio bei $N = 10^6$ und $N = 10^7$

wurden, robust gegen die Veränderung des Stichprobenumfanges ist. Dafür berechnen wir die Rangkorrelation der Messungen bei $N = 10^6$ und $N = 10^7$ (Tabelle 11). Diese ist bei allen Konstrukten sehr hoch, fast eins. Die Rangfolge der Lexeme hängt also nur ganz wenig vom Stichprobenumfang ab.

Konstrukt	Rangkorrelation für Odds Ratio $N = 10^6$ vs. $N = 10^7$
SNBETH	0.99994
SNBETO	0.99975
SNBEWH	0.9948
SNTH	0.9999
SNTO	0.9991
SNWH	0.9964
THBESN	0.9997
THSN	0.9997

Tabelle 11: Rangkorrelation für Odds Ratio bei $N = 10^6$ und $N = 10^7$

Mann kann zeigen, dass Reliance, Attraction und Odds Ratio in folgender Beziehung miteinander stehen:

$$odds\,ratio = \left[\frac{N}{C_1} - 1\right] \cdot \frac{\alpha}{(1-\alpha)\,(1-\beta)} - \frac{\beta}{1-\beta}, \qquad \alpha = \frac{O11}{R1}, \quad \beta = \frac{O11}{C1} \tag{2}$$

Aus der Gleichung (2) sieht man, dass Reliance einen positiven Einfluss auf Odds Ratio hat. Der Einfluss von Attraction ist nicht so eindeutig, da diese sowohl im positiven als auch im negativen Term vorkommt.

Die Tabelle 12 zeigt die Korrelation nach Spearman zwischen Odds Ratio, Jaccard und L2. Aus der Tabelle sieht man, dass Odds Ratio bei allen Konstrukten außer SNBEWH eine sehr hohe Korrelation mit L2 hat. Bei der SNBEWH ist Attraction deutlich höher als Reliance, so dass der negativer Effekt aus dem 2.Term der Gleichung (2) besonders starken Einfluss auf Odds Ratio hat. Das ist vermutlich der Grund für die schwächere Korrelation mit L2 in diesem Konstrukt. Die Übereinstimmung mit Jaccard ist deutlich schwächer.

NAM	odds ratio vs. jaccard	odds ratio vs. L2
SNBETH	0.5905	0.9952
SNBETO	0.7017	0.9673
SNBEWH	0.9090	0.7168
SNTH	0.5106	0.9999
SNTO	0.3270	0.9991
SNWH	0.8834	0.9407
THBESN	0.5210	0.9755
THSN	0.4168	0.9998

Tabelle 12: Rangkorrelation von Odds Ratio vs. Jaccard und L2

2.2.2 Relatives Risiko

Das relative Risiko drückt aus um welchen Faktor sich ein Risiko in zwei Gruppen unterscheidet. Im Fall der Collostruction Analysis wird relatives Risiko folgendermaßen interpretiert und errechnet:

$RR = \frac{P(Lexem|Konstrukt)}{P(Lexem|\neg Konstrukt)} = \frac{O11/C1}{O12/C2}$.

Die **Werte** von RR liegen zwischen 0 und ∞. Bei $RR = 1$ ist das Risiko in den beiden Gruppen gleich. Bei $RR > 1$ wird das Risiko in der ersten Gruppe größer (in unserem Fall bedeutet es Assoziation zwischen

dem Lexem und dem Wort). Da die Werte von RR für die Lexeme oft sehr klein sind und sehr nah nebeneinander liegen, verwenden wir eine logarithmische Transformation von RR:

$rr = \log \frac{O11/C1}{O12/C2}$, $rr \in R$, $RR = 1 \Leftrightarrow rr = 0$.

Wertebereich ändert sich entsprechend nach $(-\infty, +\infty)$.

Wie im Fall mit Odds Ratio hängt rr von N ab, deswegen ist eine direkte Interpretation seiner Werte problematisch. Berechnen wir den Unterschied zwischen den Messungen bei $N = 10^6$ und $N = 10^7$.

$\log \frac{O11 \cdot (10^7 - C1)}{O12 \cdot C1} - \log \frac{O11 \cdot (10^6 - C1)}{O12 \cdot C1} = \log \frac{10^7 - C1}{10^6 - C1}$.

Im Fall $C1 << 10^6$ beträgt Differenz ca. $\log 10$.

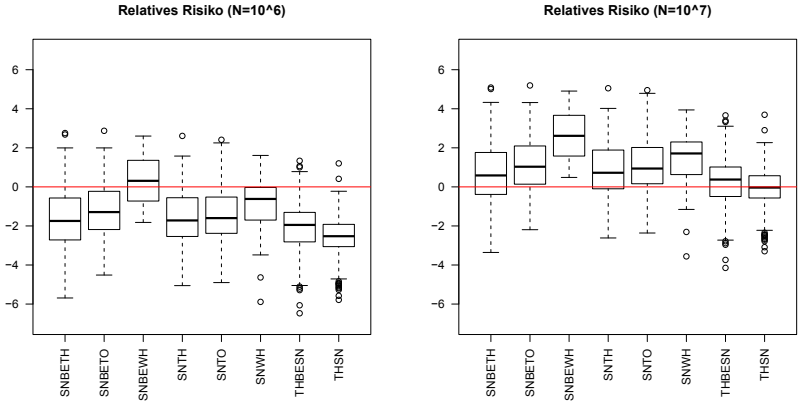

Abbildung 10: Relatives Risiko bei $N = 10^6$ und $N = 10^7$

Die Reihenfolgen der Messungen bei $N = 10^6$ und $N = 10^7$ bleiben unverändert: für alle Lexeme beträgt der Korrelationskoeffizient nach Spearman 1. Fast immer liefern relatives Risiko und Odds Ratio ähnliche Ergebnisse (Tabelle 14). Seltene Abweichungen in der Reihenfolge treten dem im Fall auf, wenn Relance oder Attraction sehr groß sind.

Konstrukt	Rangkorrelation zwischen OR und RR für $N = 10^6$
SNBETH	0.9999
SNBETO	0.9998
SNBEWH	0.9987
SNTH	0.9999
SNTO	0.9992
SNWH	0.9992
THBESN	0.9997
THSN	0.9997

Tabelle 13: Rangkorrelation für Odds Ratio vs. relatives Risisko

2.2.3 Pointwise mutual information

Das Maß Mutual Information stammt aus der Informationtheorie und misst die Stärke des statistischen Zusammenhanges der Zufallsgrößen X und Y. Für die Berechnung im diskreten Fall wird folgende Formel benutzt:

$I(X,Y) = E\left\{\log_2 \frac{p(X,Y)}{p(X)p(Y)}\right\} = \sum_{x\in X,\, y\in Y} p(x,y) \cdot \log_2 \frac{p(x,y)}{p(x)p(y)}$.

Pointwise MI misst die Diskrepanz zwischen der Wahrscheinlichkeit des gemeinsamen Auftretens zweier diskreter Zufallsvariablen und der Wahrscheinlichkeit ihres gemeinsamen Auftretens unter der Unabhängigkeitsannahme.

$MI(X,Y) = \log_2 \frac{p(X,Y)}{p(X)p(Y)}$

Die pointwise MI kann man als den relativen Informationsgehalt des gemeinsamen Auftretens zweier Zufallsvariablen interpretieren. Wenn diese Unabhängig sind ist MI gleich Null. Bei der perfekten Assoziation ist $MI = -\log_2 p(x)$. Wenn die Wahrscheinlichkeit $p(x|y)$ fix ist und die Wahrscheinlichkeit $p(x)$ sinkt, dann steigt die pointwise MI.

Pointwise MI wird in der Sprachwissenschaft oft für die Berechnung der Assoziationsstärke zwischen den grammatikalischen Einheiten benutzt. Für die Basis-Kontingenztafel der Collostruktional analysis wird pointwise MI folgendermaßen gemessen:

$MI(Lexem, Konstrukt) = \log_2 \frac{O11/N}{(C1/N)\cdot(R1/N)} = \log_2 \frac{O11\cdot N}{R1\cdot C1} = \log_2 \frac{O11}{E11}$.

Dieses Maß kann sowohl positive als auch negative Werte annehmen. Positive Werte entsprechen der Anziehung und negative Werte der Abstoßung zwischen Lexem und Konstrukt. Genau wie bei Odds Ratio, hängen die Werte der MI von N ab. Deswegen ist es schwierig diese zu interpretieren. Dabei gilt folgendes:

$\log_2 \frac{10\cdot N\cdot O11}{C1\cdot R1} = \log_2 10 + \log_2 \frac{N\cdot O11}{C1\cdot R1}$.

Die Messungen bei $N = 10^7$ und $N = 10^6$ korrelieren also perfekt miteinander. Die Reihenfolge der Lexeme in einer Rangliste, die mittels MI erstellt wurde, hängt nicht von N ab.

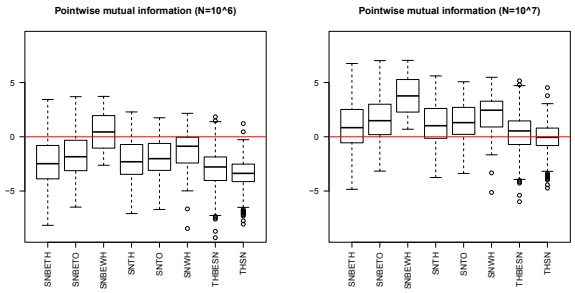

Abbildung 11: pointwise mutual information bei $N = 10^6$ und $N = 10^7$

Wenn R1 und C1 klein im Vergleich zu N sind und O11 klein im Vergleich zu R1 und C1 ist, liefern Odds Ratio und MI ähnliche Ergebnisse, da

$log_2(\text{odds ratio}) = \log \frac{O11\cdot(N-(C1+R1-O11)}{(R1-O11)(C1-O11)} \approx \log_2 \frac{O11\cdot N}{R1\cdot C1}$

Die Abweichungen treten also bei sehr häufigen Lexemen und Konstrukten auf, und wenn Attraction oder Reliance sehr groß sind. Es ist auch zu erwarten, dass bei größeren N die Unterschiede zwischen den beiden Maßen verschwinden (Tabelle 14). Diese Vermutung stimmt für alle Konstrukte außer seltenen SNWH und SNBETO, die die Lexeme mit hohen Reliance und Attraction enthalten.

Es kann also festgestellt werden, dass or, rr und MI sehr ähnliche Ergebnisse liefern und die Reihenfolge ihrer Ranglisten ziemlich resistent gegen einer Veränderung des Stichprobenumfanges ist. Die Abbildung 12 zeigt die Abhängigkeit des logarithmierten Odds Ratio, rr und MI von O11 bei fixen C1, R1 und N. Alle drei Funktionen sind monoton wachsend. Der Verlauf der Kurven von Odds Ratio und rr ist in den meisten Fällen fast identisch.

NAM	N=10^6	N=10^7
SNBETH	0.9999	0.9999
SNBETO	0.9998	0.9999
SNBEWH	0.9987	0.9961
SNTH	0.9999	0.9999
SNTO	0.9992	0.9999
SNWH	0.9992	0.9972
THBESN	0.99972	0.9999
THSN	0.9997	0.9999

Tabelle 14: Rangkorrelation für MI vs. Odds Ratio

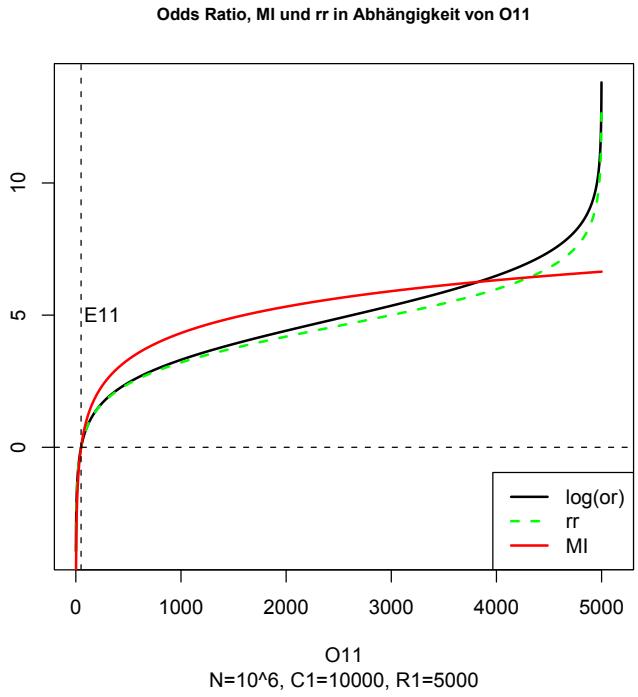

Abbildung 12: Odds Ratio, rr und MI in Abhängigkeit von O11

2.2.4 z-Score

Noch ein Maß, das für die Messung der Assoziationsstärke vorgeschlagen wurde, ist z-Score. Dieses Maß vergleicht die beobachtete Häufigkeit O11 mit der erwarteten Häufigkeit E11 folgendermaßen:

$z = \frac{O11-E11}{\sqrt{E11}}$.

Die Abweichung von der erwarteten Häufigkeit wird hier durch $\sqrt{E11}$ skaliert. Der z-Score nimmt sowohl positive als auch negative **Werte** an. Interpretation der Werte:

$O11 = E11 \Rightarrow z = 0$ (stochastische Unabhängigkeit),
$O11 < E11 \Rightarrow z < 0$ (Abstoßung),
$O11 > E11 \Rightarrow z > 0$ (Assoziation).

Wir vergleichen die z-Scores für unterschiedliche Konstrukte mit Hilfe der Scatterplots (Abbildung 13).

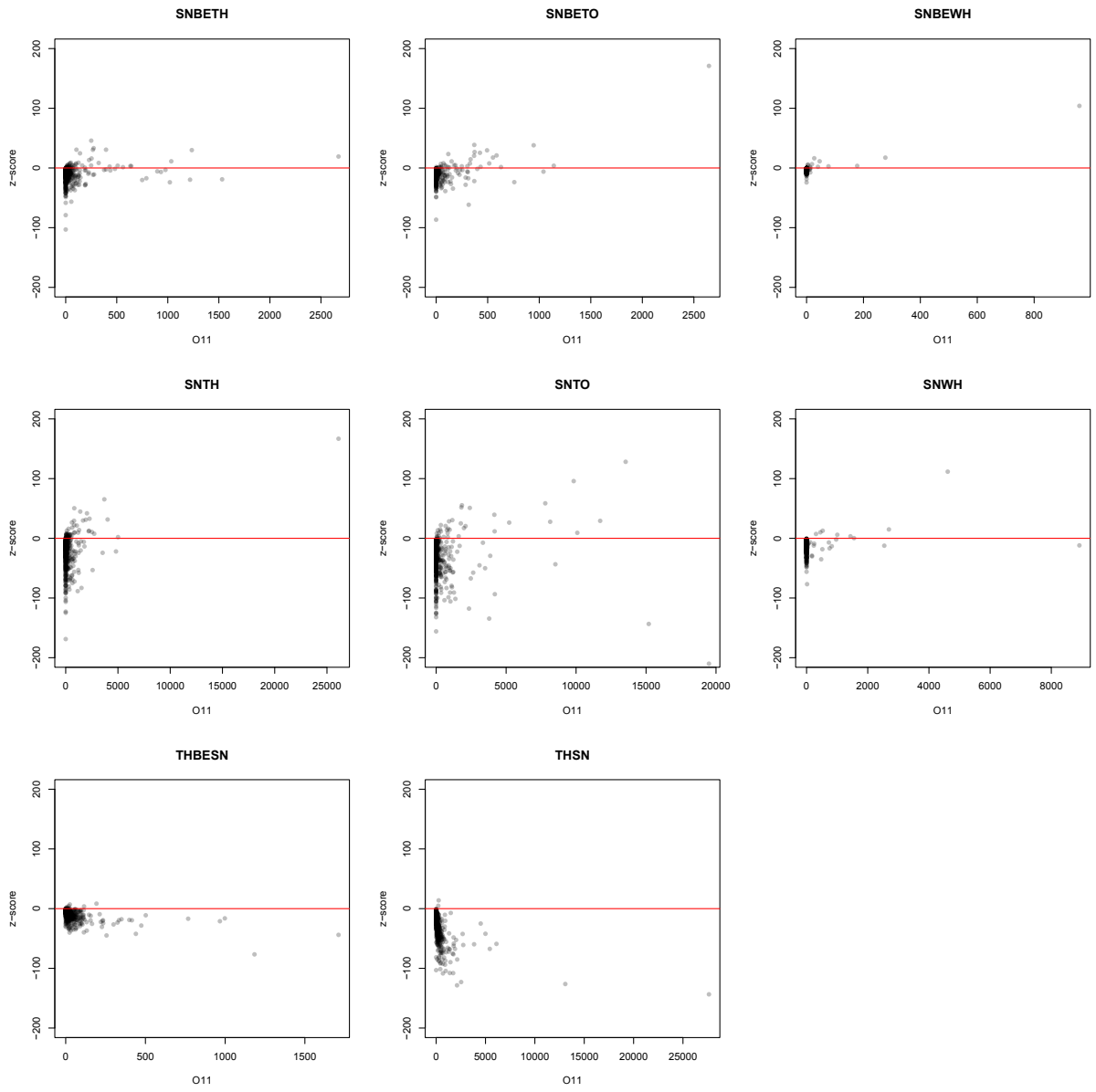

Abbildung 13: z-Score bei $N = 10^6$

Aus der Grafik sieht man, dass die meisten Beobachtungen bei $N = 10^6$ negativ sind, was eine Abstoßung zwischen Lexemen und Konstrukt bedeuten sollte. Mit Steigerung von N sinkt die erwartete Häufigkeit um den Faktor 10 und die z-Score Werte steigen (Abbildung 14). Die meisten davon werden positiv, obwohl sich mit der Veränderung von N nichts in der Beziehung zwischen Lexemen und dem Konstrukt verändern sollte. Eine direkte Interpretation der z-Score-Werte ergibt also wenig Sinn.

Mit Veränderung von N, ändert sich auch die Reihenfolge der Lexeme in den Ranglisten. Der Scatterplot der Messungen bei 10^6 und 10^7 (Abbildung 15) zeigt keinen deutlich monotonen Zusammenhang. Die Korrelation nach Spearman (Tabelle 15) liegt bei den meisten Lexemen um 0.5, ist also relativ schwach. Bei den seltenen Konstrukten mit wenigen Lexemen (SNBEWH, SNBETO, SNBETH) ist die Korrelation hoch bis mittelstark, bei den Konstrukten mit vielen Lexemen (THSN, THBESN) ist die Korrelation sehr schwach.

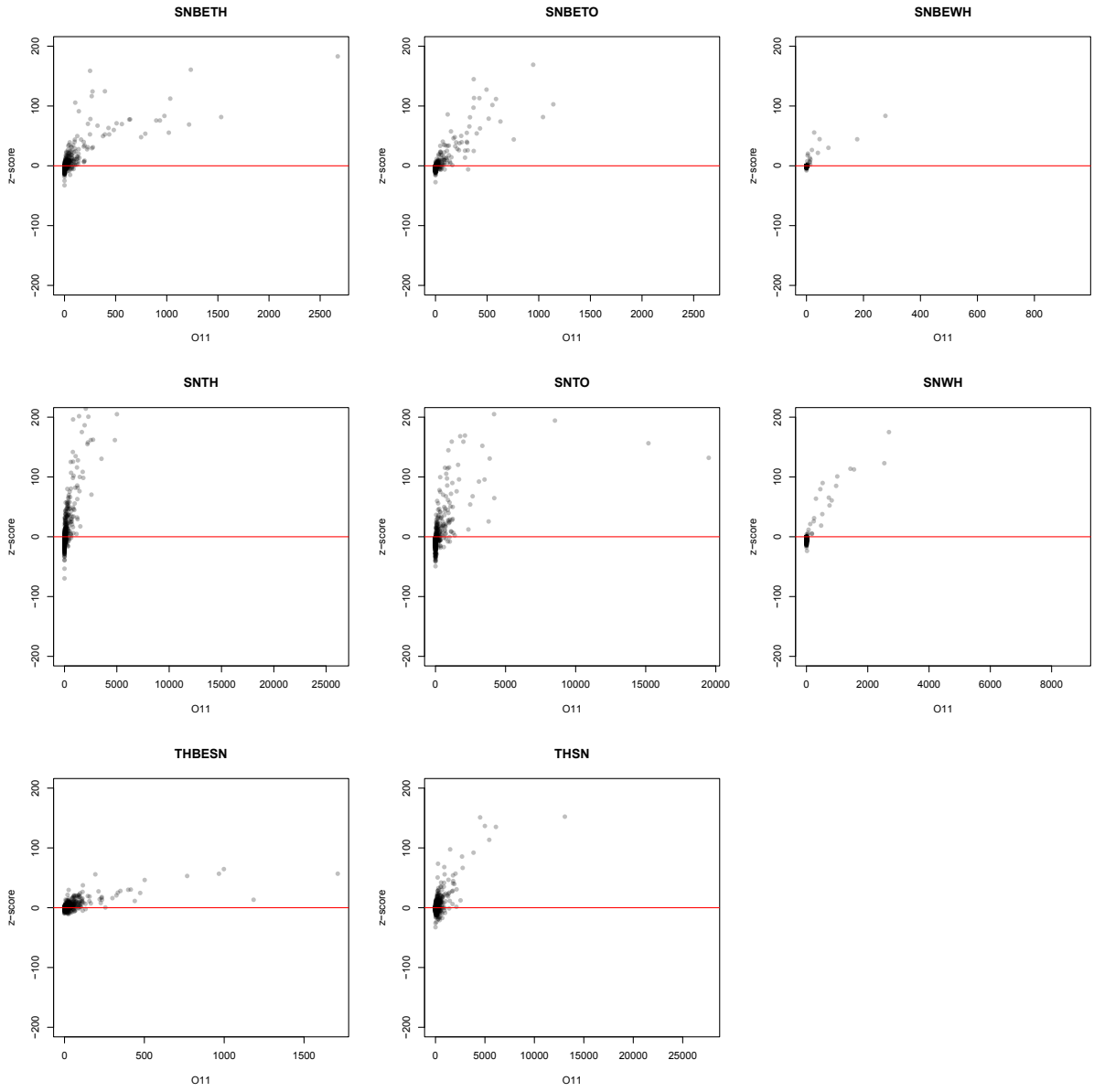

Abbildung 14: z-Score bei $N = 10^7$

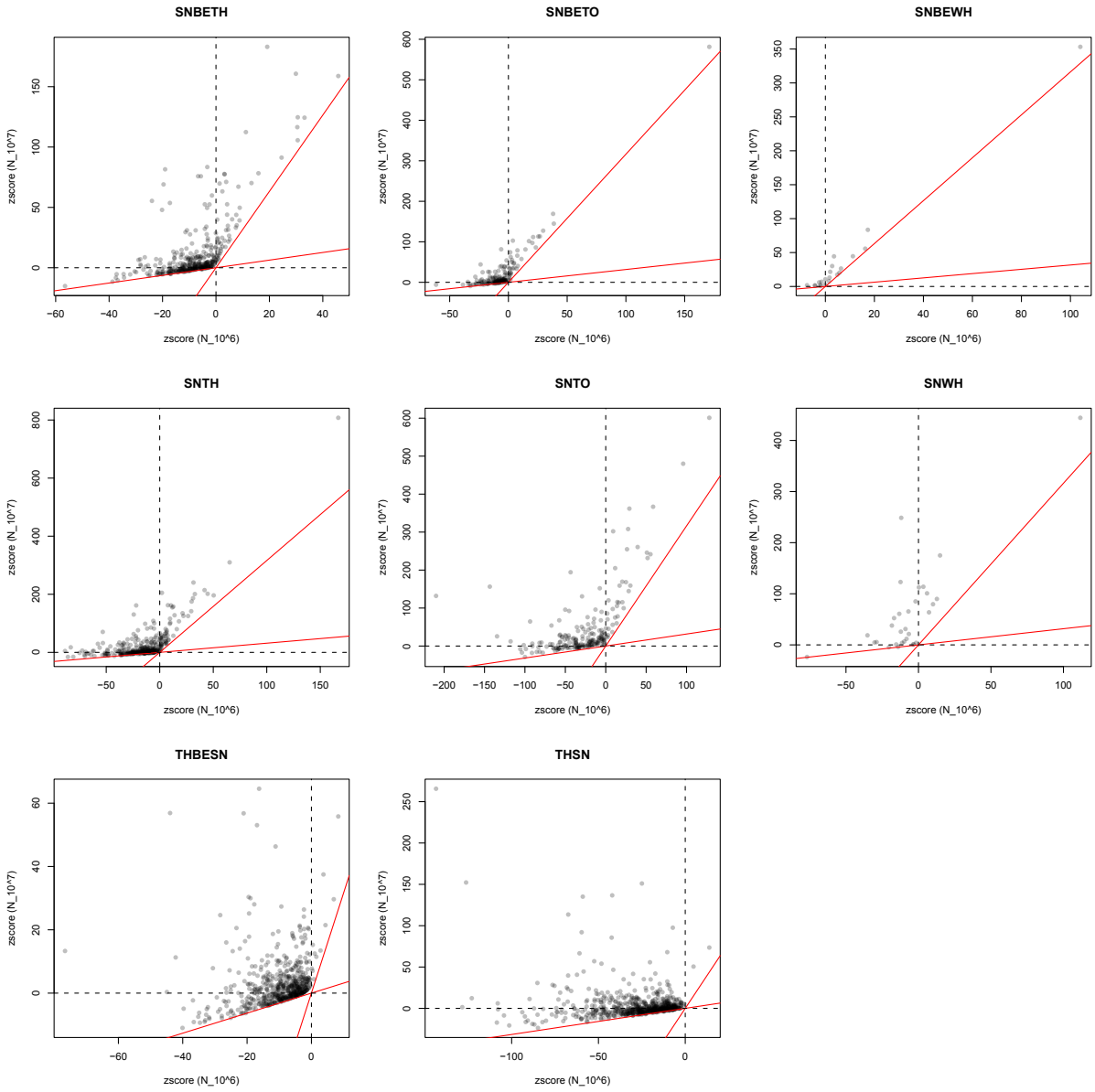

Abbildung 15: Scatterplot der z-Score-Messungen bei $N = 10^6$ und 10^7

Um die Ursache der schwachen Korrelation herauszufinden betrachten wir z-Score als Funktion von O11:

$$z = \frac{1}{\sqrt{E11}} \cdot O11 - \sqrt{E11} = \sqrt{\frac{N}{C1 \cdot R1}} \cdot O11 - \sqrt{\frac{C1 \cdot R1}{N}} \tag{3}$$

Aus der Gleichung (3) sieht man, dass bei fixen R1, C1 und N, z-Score eine lineare Funktion der O11 ist. Dabei wird mit wachsendem R1 die Steigung der entsprechenden Gerade sinken, und die Grafik verschiebt sich nach unten. Die z-Score-Linien für unterschiedliche Lexeme haben also unterschiedliche Steigungen und schneiden die x-Achse an unterschiedlichen Stellen. Wenn N sich vergrößert, wächst die Steigung aller Linien und die Grafik verschiebt sich nach oben. Das führt in manchen Fällen zur Veränderung der Rangfolge der Lexeme.

Beispiel 1: Wir betrachten ein Konstrukt mit $C1 = 186433$ und 2 Lexeme: Lexem 1 mit $R1 = 2000$ und

Konstrukt	Korrelation nach Spearman für z-score bei $N = 10^6$ und $N = 10^7$
SNBETH	0.6444
SNBETO	0.7099
SNBEWH	0.9389
SNTH	0.5943
SNTO	0.6147
SNWH	0.575
THBESN	0.3848
THSN	0.2326

Tabelle 15: Rangkorrelation der z-Score Messungen bei $N = 10^6$ und 10^7

$O11 = 30$ und Lexem 2 mit $R1 = 4000$ und $O11 = 90$. Bei $N = 10^6$ liegen die entsprechenden z-Scores der beiden Lexeme unter der x-Achse, der z-Score des Lexems 1 ist dabei größer als der des Lexems 2. Bei $N = 10^7$ wird der Wert von z-Score des 2.Lexems positiv, während der erste negativ bleibt. Die Reihenfolge der Lexeme in der Rangliste ändert sich (Abbildung 16).

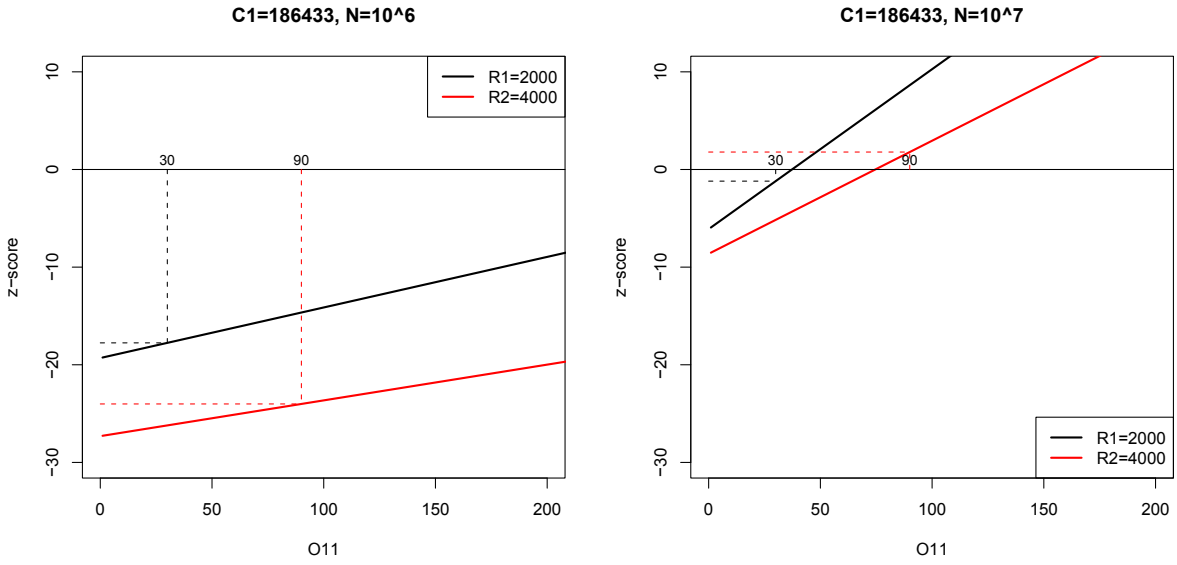

Abbildung 16: Beispiel 1

Beispiel 2. Für das gleiche Konstrukt vergleichen wir das Lexem 1 mit $R1 = 600$ und Häufigkeit $O11 = 20$ und Lexem 2 mit $R1 = 6000$ und $O11 = 200$. Die Reihenfolge der Lexeme verändert sich durch die Veränderung der Steigung der z-Score-Linien (Abbildung 17).

Die Vertauschung der Lexeme in den Ranglisten erfolgt in der Nähe vom Schnittpunkt mit der x-Achse. Die Ranglisten, die mittels z-Score erstellt wurden sind also empfindlich gegen eine Veränderung des Stichprobenumfanges, was für Collostructional analysis unerwünscht ist.

2.3 p-Werte als Assoziationsmaße

Eine besondere Gruppe der Maße, die zur Messung der Assoziation zwischen einem Lexem und einem Konstrukt vorgeschlagen wurden sind die p-Werte der Nullhypothesentests. Ein p-Wert ist die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in Richtung der Alternative extre-

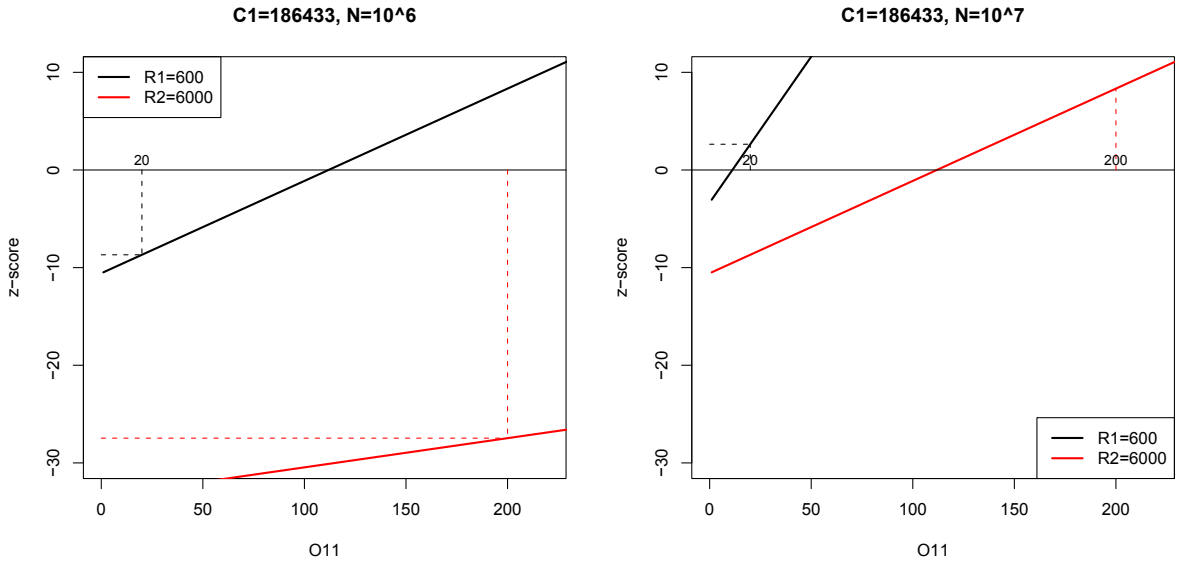

Abbildung 17: Beispiel 2

meren Wert zu erhalten. Wir betrachten die logarithmierten p-Werte des exakten Test nach Fisher und des χ^2-Tests.

2.3.1 p-Wert des exakten Tests nach Fisher

Der p-Wert des exakten Tests nach Fisher (FET) wurde als Assoziationsmaß von Stefanowitsch und Grieß besonders empfohlen [5], da dieser keine Verteilungsannahmen verlangt und alle Informationen, die in der Häufigkeitstabelle enthalten sind, berücksichtigt.

Bei der Berechnung der p-Werte des exakten Tests nach Fischer stellt man folgende Nullhypothese: $O11 = E11$, d.h. Lexem L und Konstrukt C sind stochastisch unabhängig. Die Alternative lautet $O11 \neq E11$, L und C sind stochastisch abhängig. Unter H0 kann man $R1$ als eine Stichprobe betrachten die aus der Grundgesamtheit mit zwei Merkmalsausprägungen: C und $\neg C$ gezogen wird. Dann wird $O11$ als die Anzahl der Elemente mit der Ausprägung C in der Stichprobe $R1$ betrachtet. Diese muss unter H0 hypergeometrisch verteilt sein mit $N = C1 + C2$, $M = C1$ und $n = R1$.

Für die Messung der Assoziation zwischen Lexem und Konstrukt werden die p-Werte des einseitigen Fisher-Tests betrachtet, mit H0: $O11 > E11$ und H1: $O11 \leq E11$. Die Prüfgröße dieses Tests sieht folgendermaßen aus:

$$\sum_{k=O11}^{\min\{R1,C1\}} \frac{\binom{C1}{k} \cdot \binom{C2}{R1-k}}{\binom{N}{R1}}$$

Die Berechnung dieser Größe ist sehr rechenintensiv. Die p-Werte sind oft sehr klein, deswegen nimmt man die logarithmierten p-Werte. Da der p-Wert zeigt, wie hoch die Wahrscheinlichkeit ist, unter der Nullhypothese den Beobachteten oder den extremeren Wert in Richtung der Alternative zu betrachten, geht man davon aus, dass, je kleiner der p-Wert, desto höher die Assoziation zwischen Lexem und Konstrukt ist.

In dem Programmpaket R wurde dafür folgender Befehl verwendet:

`ft<-log(sum(dhyper(O11:min(R1,C1]),C1,N-C1,R1))+exp(-300))`.

Um keine NAs beim Logarithmieren der p-Werte, die gleich Null sind zu erhalten, wurde zu dem p-Wert ein sehr kleiner Wert `exp(-300)` addiert.

Es ist bekannt, dass die p-Werte vom Stichprobenumfang abhängig sind. Abbildung 18 stellt die Verteilung der Werte des Fisher-Tests für die Stichprobenumfänge $N = 10^6$ und $N = 10^7$ dar. Die Verteilungen sind linkssteil. Tabelle 16 vergleicht die entsprechenden Medianen.

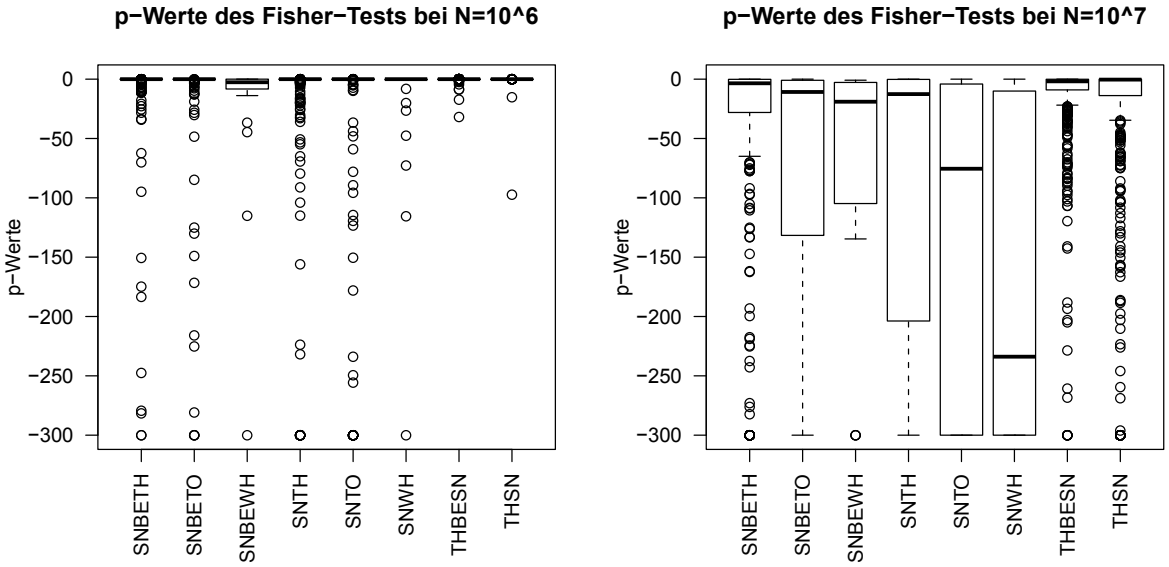

Abbildung 18: p-Werte des exakten Test nach Fisher

Konstrukt	Median der p-Werte des FET $N = 10^6$	Median der p-Werte des FET ($N = 10^7$)
SNBETH	0.1742	1.7970
SNBETO	-4.99e-16	-10.7919
SNBEWH	-2.6163	-19.0164
SNTH	-1.1102e-16	-12.5999
SNTO	0	-75.4967
SNWH	-5.55e-16	-233.8630
THBESN	-2.22e-16	-1.8410
THSN	0	-0.3914

Tabelle 16: Vergleich der Medianen des ft für $N = 10^6$ und $N = 10^7$

Die p-Werte für $N = 10^7$ sind wesentlich kleiner als die p-Werte für $N = 10^6$. Es ist leicht erklärbar, wenn wir uns erinnern, dass Odds Ratio im Fall $N = 10^6$ bei der überwiegender Mehrheit der Beobachtungen unter 1 lag. In diesem Fall sind die p-Werte groß. Die nächste Grafik zeigt den Zusammenhang zwischen den p-Werten und Odds Ratio. Wir fixieren die Häufigkeiten R1 und C1 und vergrößern O11. Dadurch vergrößert sich auch der Odds Ratio. Die schwarze Linie entspricht den logarithmierten p-Werten bei $N = 10^6$ und die rote bei $N = 10^7$. Die gestrichelten Linien zeigen die Häufigkeit O11 wenn $OddsRatio = 1$.

Aus dem oberen Bild der Abbildung 18 (mit $C1 = 141476$ und $R1 = 8000$) sieht man, dass im Bereich zwischen ca. O11=100 und O11=2000 die rote Linie weit unter der schwarzen liegt. Die gleiche Situation kann man im unteren Bild (mit $C1 = 23000$ und $R1 = 200$) beobachten. Es fällt dabei auf, dass die Form der Linien sich verändert hat. Die Schnelligkeit des Fallens der p-Werte bei steigendem O11 hängt also nicht nur von N sondern auch von C1 und R1 ab. Diese Tatsache verursacht zwei negative Effekte bei der Bildung der Ranglisten.

Der erste negative Effekt: Die Messungen des ft für $N = 10^6$ und $N = 10^7$ korrelieren ziemlich schwach miteinander (siehe Abbildung 20 und Tabelle 17).

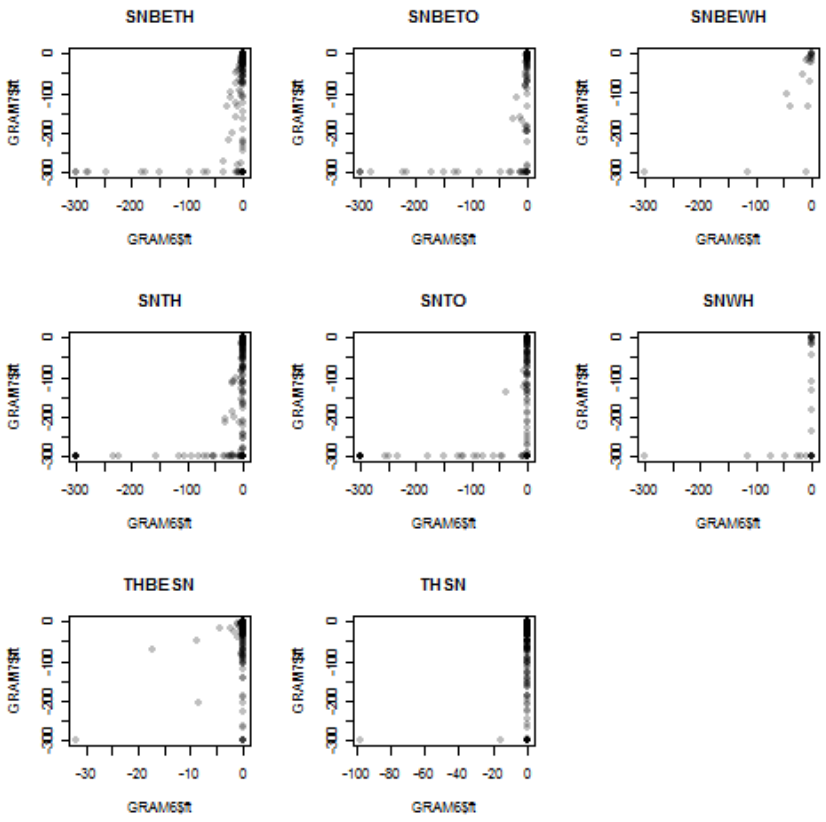

Abbildung 19: Scatterplot des ft für $N = 10^6$ und $N = 10^7$

Es fällt auf, dass das Konstrukt SNBEWH mit ziemlich großen Odds Ratios eine relativ hohe Korrelation aufweist (0.85), das Konstrukt THSN mit sehr kleinen Odds Ratios jedoch keine (0.07). Um herauszufinden, weswegen die Konstrukte mit den kleinen Odds Ratios so unterschiedliche Ergebnisse bei der Ranglistenbildung liefern, betrachten wir folgendes Beispiel.

Wir fixieren ein Konstrukt mit $C1 = 141746$ (wie SNTH) und betrachten drei Lexeme mit $R1 = 765$, $R1 = 250$, $R1 = 123$. Danach stellen wir grafisch die Abhängigkeit der p-Werte von der Häufigkeit O11 dar (Abbildung 21). Nehmen wir an, dass die drei Lexeme folgende Häufigkeiten $O11$ haben: 150, 80 und 70. Gestrichelte Linien in der Grafik zeigen die entsprechende Stellen. Aus der Grafik sieht man, dass bei $N = 10^6$ das "grüne" Lexem den kleinsten p-Wert hat, danach kommt das "rote" und zuletzt das "schwarze" Lexem. Bei $N = 10^7$ steht an der 1.Stelle das "schwarze" Lexem, danach das "grüne" und als letztes das "rote". Die Reihenfolge dieser Lexeme in der Rangliste hat sich also umgedreht. Ein Grund dafür ist die unterschiedliche Steigung der Kurven in dem Punkt, wo Odds Ratio den Wert eins annimmt. Wenn also die absolute Häufigkeiten sich in diesem Bereich befinden, werden ihre Reihenfolgen in Ranglisten bei $N = 10^6$ und $N = 10^7$ wenig Übereinstimmungen zeigen.

Der zweite negative Effekt bei der Bildung der Ranglisten besteht darin, dass die seltenen Lexeme weniger Chancen haben hoch in die Rangliste zu kommen, sogar wenn diese sehr stark vom Konstrukt angezogen sind. Als Illustration dafür nehmen wir ein Konstrukt mit C1=22000 und zwei Lexeme mit R1=5000 und R1=150. Als Stichprobenumfang nehmen wir $N = 10^7$ (Abbildung 22).

Den gleichen p-Wert -150 erhält das Lexem 2 bei $O11 = 40$ und das Lexem 1 bei $O11 = 105$. Tabelle 18 zeigt Attraction, Reliance und Odds Ratio der beiden Lexeme.

Es ist offensichtlich, das das Lexem 2 stärker zu dem Konstrukt angezogen ist als Lexem 1. Trotzdem

Konstrukt	Median(odds ratio) bei $N = 10^7$	Median(odds ratio) bei $N = 10^6$	Korrelation zwischen den p-Werten bei $N = 10^6$ und $N = 10^7$
SNBETH	1.79	0.17	0.5375
SNBETO	2.80	0.27	0.5922
SNBEWH	14.30	1.37	0.8593
SNTH	2.06	0.17	0.4272
SNTO	2.57	0.19	0.3794
SNWH	5.62	0.52	0.3901
THBESN	1.45	0.14	0.3071
THSN	0.96	0.07	0.0737

Tabelle 17: Medianenvergleich für ft bei $N = 10^6$ und $N = 10^7$

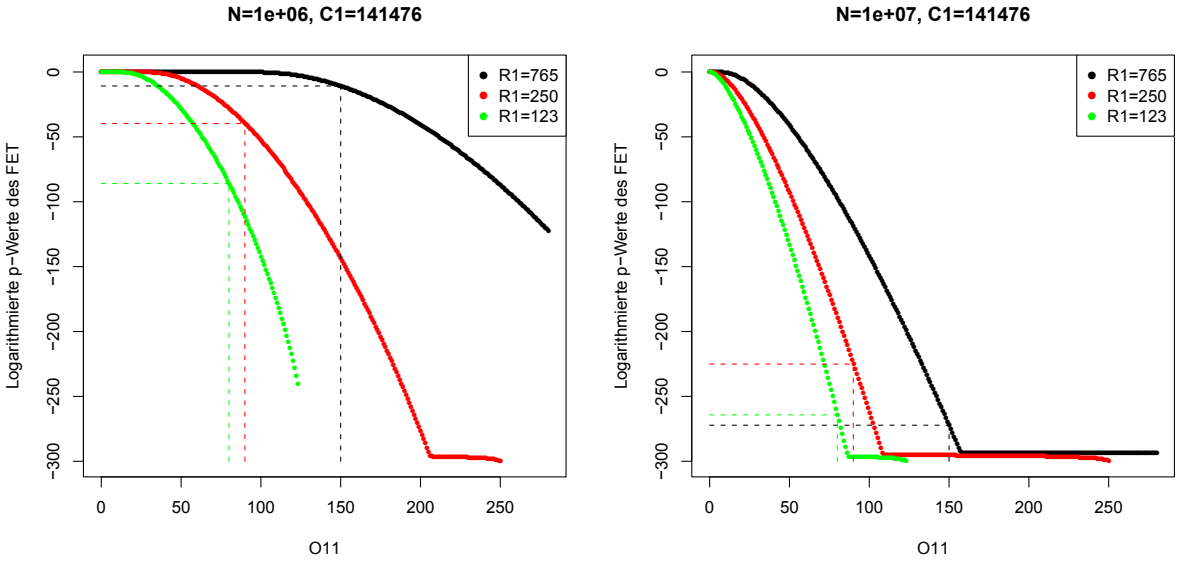

Abbildung 20: Beispiel des Reihenfolgewechsels für ft

haben die beiden den gleichen p-Wert für FET, sie stehen also in der Rangliste nebeneinander.

Die beiden geschilderten Probleme stellen die Verwendbarkeit der p-Werte des Fisher-Tests als Assoziationsmaß in Frage.

2.3.2 p-Werte des χ^2 -Tests

Der nächste p-Wert, der als Assoziationsmaß benutzt wird ist der p-Wert des χ^2 Unabhängigkeitstests. Die Nullhypothese des Tests lautet: Das Konstrukt C und das Lexem L sind stochastisch unabhängig. Die Prüfgröße für die 2x2 Basis-Kontingenztafel wird aus folgenden Überlegungen abgeleitet: Die Grundgesamtheit enthält Elemente mit dem Merkmal "Konstrukt C". Der Anteil dieser Elemente $p = \frac{C1}{N}$ ist ein

	ft	R1	O11	Reliance	Attraction	odds ratio
Lexem 1	-150	5000	105	0.021	0.0047	9.77
Lexem 2	-150	150	40	0.26	0.0018	165.22

Tabelle 18: Vergleich der Messungen für Lexem 1 und Lexem 2 aus dem Beispiel

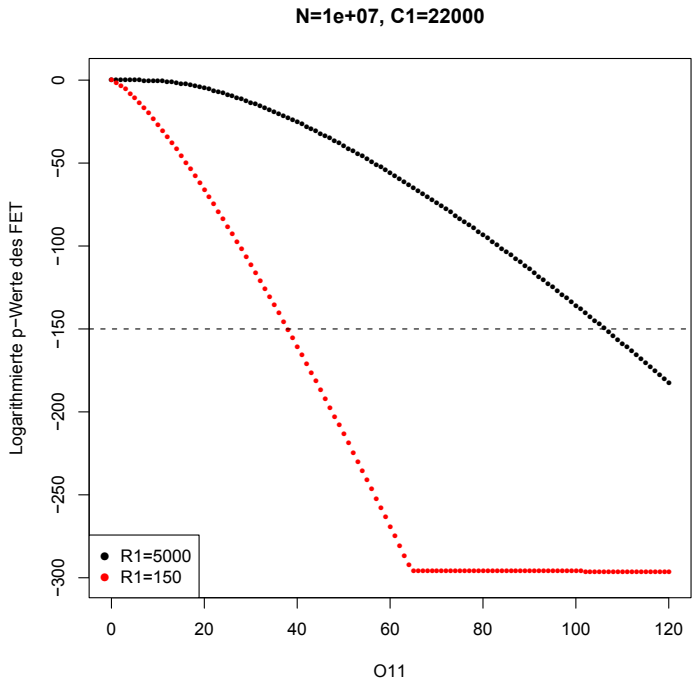

Abbildung 21: ft-Werte für Lexeme mit unterschiedlichen R1

Schätzer für die Wahrscheinlichkeit ein Element mit diesem Merkmal zu erhalten. Unter H0 wird diese Wahrscheinlichkeit sowohl für die Menge, die Lexem L enthält, als auch für die Menge ohne Lexem L gleich sein. Deswegen $O11 \sim B(R1, p)$ und $O21 \sim B(R2, p)$. Man kann zeigen, dass unter H0 die Prüfgröße des Tests

$$\chi^2 = \frac{N(O11 \cdot O22 - O12 \cdot O21)^2}{R1 \cdot R2 \cdot C1 \cdot C2} \sim^{approx} \chi^2_{(1)}.$$

Die Approximation gilt, wenn $p = \frac{C1}{N}$ groß genug ist und R1 und R2 auch relativ groß sind. Dies wird stimmen, wenn nach Faustregel jede der erwarteten Haufigkeiten mindestens 5 und keine dieser Häufigkeiten Null ist. Für kleinere Werte muss die Kontinuitätskorrektur nach Yates verwendet werden. In unserem Fall wird diese Bedingung für die meisten Lexeme und Konstrukte erfüllt.

Konstrukt	Anteil der Lexeme mit $E11 > 5$ bei $N = 10^6$	Anteil der Lexeme mit $E11 > 5$ bei $N = 10^7$
SNBETH	93%	56%
SNBETO	97%	67%
SNBEWH	62%	9%
SNTH	98%	82%
SNTO	99%	94%
SNWH	100%	90%
THBESN	93%	54%
THSN	99%	88%

Tabelle 19: Erfüllung der Approximationsbedingungen

Die Prüfgröße nimmt **Werte** von 0 bis Unendlich an. Je größer die Prüfgröße, desto weiter ist die Datenlage von der Nullhypothese entfernt. Der p-Wert fällt mit steigendem Wert der Prüfgröße, da die Dichte der χ^2-Verteilung mit einem Freiheitsgrad eine monoton fallende Funktion ist. Die Korrelation nach Spearman

zwischen der Prüfgröße und dem entsprechenden p-Wert beträgt -1.

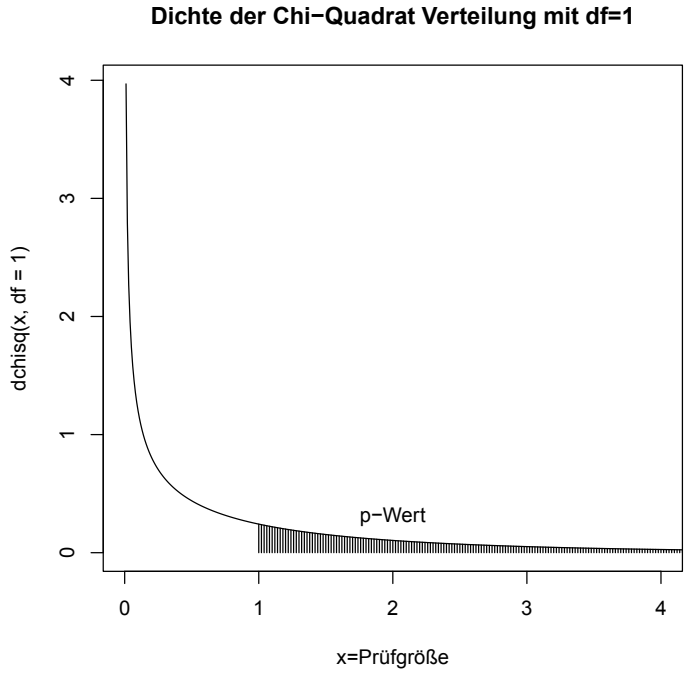

Abbildung 22: Dichte der χ^2 verteilung mit df=1

Im Weiteren betrachten wir nur die Prüfgröße chi. Da der χ^2-Test zweiseitig ist, ist es nicht klar, was ein großer Wert der Prüfgröße bedeutet - Abstoß oder Anziehung zwischen Lexem und Konstrukt. Deswegen ist die Verwendung dieser Statistik zur Erstellung der Assoziationslisten fraglich. Wir ziehen zusätzlich einen künstlich erzeugten p-Wert des einseitigen Tests chi1 in Betracht. Dieser wird nach folgender Regel gemessen:

$chi1 = \frac{1}{2}\cdot$ p-Wert des zweiseitigen Tests, wenn $O11 > E11$,

$chi1 = 1 - \frac{1}{2}\cdot$ p-Wert des zweiseitigen Tests, wenn $O11 \leq E11$.

Ein kleinerer p-Wert chi1 bedeutet eine stärkere Assoziation zwischen Lexem und Konstrukt. Im Weiteren betrachten wir die logarithmierten p-Werte.

Mit wachsendem N steigt die Prüfgröße chi und sinkt der p-Wert chi1. Untersuchen wir ob das eine Auswirkung auf die Reihenfolge der Lexeme in der Rangliste hat. Aus der Tabelle 20 sieht man, dass die Korrelation nach Spearman für Chi ziemlich schwach und manchmal sogar negativ ist, die Korrelation für chi1 ist etwas größer:

Die Ursache der schwachen Korrelation für die Prüfgröße chi bei $N = 10^6$ und $N = 10^7$ ist leicht erklärbar. Die Prüfgröße kann man folgendermaßen darstellen:

$\chi = \frac{N\cdot(O11\cdot N - R1\cdot C1)^2}{R1\cdot C1\cdot(N-R1)\cdot(N-C1)}$

Bei den fixen N,C1, und R1 ist es eine quadratische Funktion von O11 mit dem Scheitel im Punkt $(\frac{R1\cdot C1}{N}, 0) = (E11, 0)$. Vergrößerung von N verkleinert E11, was zur Verschiebung der Parabel entlang der x-Achse führt. Zusätzlich verändert sich die Steigung der Parabel (siehe Abbildung 24).

Einige Häufigkeiten, die bei $N = 10^6$ links von E11 liegen (die Parabel fällt hier), befinden sich bei $N = 10^7$ rechts vom neuen E11 (hier steigt die Funktion). Deswegen kommt die Reihenfolge der Lexeme

Konstrukt	Korrelation nach Spearman für chi bei $N = 10^6$ und $N = 10^7$	Korrelation nach Spearman für chi1 bei $N = 10^6$ und $N = 10^7$
SNBETH	-0.0670	0.6238
SNBETO	-0.1089	0.6902
SNBEWH	0.5259	0.9389
SNTH	-0.1152	0.5851
SNTO	-0.1901	0.6056
SNWH	-0.0068	0.5205
THBESN	0.1436	0.3626
THSN	0.4050	0.1659

Tabelle 20: Rangkorrelation zwischen Messungen bei $N = 10^6$ und $N = 10^7$

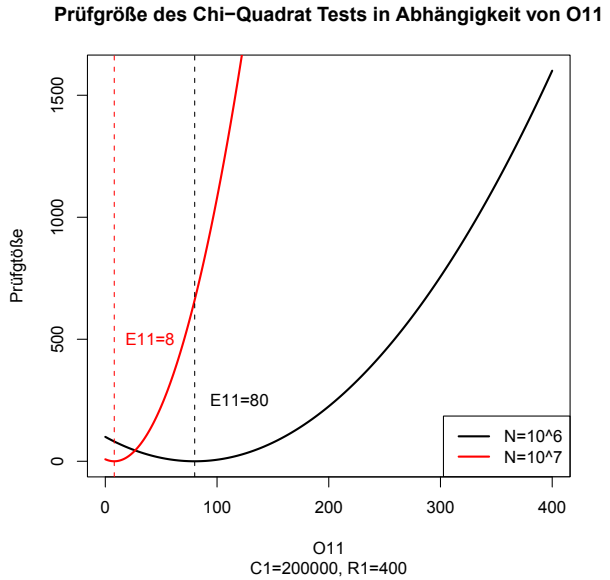

Abbildung 23: Beispiel

in der Rangliste bei der Veränderung von N durcheinander. Nur wenn sich der Großteil der Lexeme rechts von den beiden erwarteten Häufigkeiten befindet, liefert χ^2- Statistik ähnliche Ergebnisse. Der p-Wert des einseitigen χ^2-Tests verhält sich ähnlich wie der p-Wert des Fisher-Tests und hat die gleichen negativen Eigenschaften bei der Bildung der Ranglisten (Beispiel: Abbildung 25). Wir haben also festgestellt, dass die Reihenfolge der Lexeme in den Listen, die mittels der p-Werte erzeugt wurden, sehr vom Stichprobenumfang abhängig ist. Daher passen die p-Werte der einseitigen Tests nicht optimal für die Messung der Assoziationsstärke im Rahmen der Collostructional analysis. Die p-Werte der zweiseitigen-Tests sind für diese Aufgabe ungeeignet.

In der Literatur findet man noch andere Kritikpunkte zur Verwendung der p-Werte. H.-J. Schmidt und H.Küchenhoff [1] nennen folgende Argumente dagegen:

1. Ein logisches Problem: Ein statistischer Test wurde so gestaltet, dass die Nullhypothese verworfen wird, wenn der p-Wert kleiner als eine Grenze α ist, die durch Signifikanzniveau, (z.B. 0.95, 0.90) festgesetzt wurde. Wenn der p-Wert größer als α ist, können wir keine Aussagen über die strenge der Assoziation treffen. In diesem Fall ist der p-Wert nicht klein genug um die Nullhypothese zu verwerfen. Dies ist aber nicht hinreichend für die Behauptung, dass die Assoziation zwischen dem Wort und dem Konstrukt schwach ist.

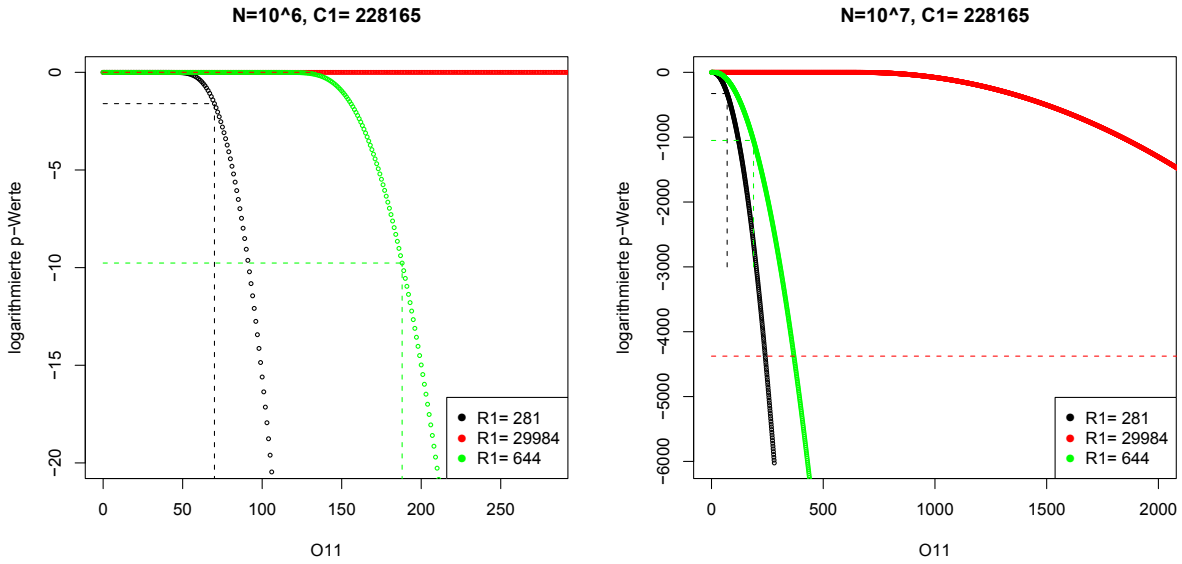

Abbildung 24: Beispiel der Reihenfolgeveränderung für chi1

2. Ein statistisches Problem: Sowohl der Fisher-Test als auch die anderen Nullhypothesen-Tests vergleichen die beobachtete Häufigkeit mit der erwarteten unter der Annahme, dass die Nullhypothese stimmt. Die wichtige Voraussetzungen für einen korrekten Nullhypothesentest sind: a) Unter der Nullhypothese dürfen keine Beziehungen zwischen dem Wort und der Konstruktion statt finden, b) Die Daten sollen zufällig verteilt sein. Diese Voraussetzungen werden im Fall des Korpus verletzt, da eine Sprache ohne Zweifel bestimmte Muster enthält. Der "Collostructional Analysis" selbst geht davon aus, dass die Sprache aus bestimmten Mustern besteht. Außerdem lässt die Art der Erstellung eines Korpus nicht annehmen, dass sein Inhalt einen zufälligen Charakter hat.

3. Problem des Stichprobenumfangs: Es ist bekannt, dass die p-Werte von dem Stichprobenumfang abhängig sind. Diese Tatsache erschwert den Vergleich der Assoziationsmaße für Korpora mit unterschiedlichem Stichprobenumfang. Zugleich liefern die Tests bei einem ausreichend großen Stichprobenumfang signifikante p-Werte, sogar wenn die Daten rein zufällig sind. Dieses Problem ist besonders zu beachten, da die meisten Korpora einen sehr großen Umfang haben.

4. Mit dem dritten Problem eng verbunden ist das Problem der 4.Zelle des Kontingenztafeln. Es ist relativ klar, wie man die Häufigleiten O11, O12 und O21 errechnet. Doch die Berechnung der Häufigkeit O22 (Anzahl aller Konstrukte außer C, die alle Lexeme außer L erhalten), kann sehr willkürlich sein. Diese hängt davon ab, wie der Spezialist das Wort "andere" interpretiert. Doch diese Häufigkeit bestimmt den Stichprobenumfang N und folglich die p-Werte des Tests.

5. Problem der Null-p-Werte: Da der Stichprobenumfang in der Regel sehr groß ist, werden die p-Werte sehr klein. Ab einer bestimmten Grenze (bei R ist es 10^{-300}) werden diese als Null wiedergegeben. Wenn also der Korpus groß ist, bekommt man im Output viele p-Werte, die gleich Null sind, und dies verhindert des Erstellen der Ranglisten der Lexeme nach Grad der Assoziationsstärke mit einer Konstruktion.

2.4 Likelihood der Poissonverteilung

Das letzter Maß, das wir betrachten ist Poisson-Likelihood. Diese wird aus folgenden Überlegungen verwendet. Nehmen wir an, dass das Lexem L und das Konstrukt C stochastisch unabhängig sind. Die

Wahrscheinlichkeit für Merkmal C ist dann $p \approx \frac{C1}{N}$. Die Anzahl der Elemente mit dem Merkmal C in der Stichprobe R1 wird unter der Unabhängigkeitsannahme binomialverteilt mit dem Erwartungswert $E(O11) = n \cdot p = R1 \cdot \frac{C1}{N} = E11$.Wenn p klein ist und n groß, kann man die Binomialverteilung durch Poissonverteilung mit $\lambda = E11$ approximieren. Dann wird die Wahrscheinlichkeit den gegebenen Wert O11 zu beobachten folgendermaßen berechnet:

$P(O11) = e^{-E11} \frac{(E11)^{O11}}{O11!}$

Je kleiner die Wahrscheinlichkeit ist, desto weniger entspricht die Annahme über stochastische Unabhängigkeit des Konstrukts und des Lexems der Datenlage.

Die Poisson-Likelihood hat folgende Nachteile: 1) Die kleine Wahrscheinlichkeit gibt uns keine Information darüber, ob $O11 > E11$ oder $O11 < E11$ ist. Deswegen eignet sich dieser Maß nicht für die Erstellung der Ranglisten nach Assoziationsstärke. 2) Nach der Faustregel ist die Approximation verwendbar, wenn $n \cdot p = E11 \leq 10$. In dem gegebenen Datensatz wird diese Bedingung nur selten erfüllt.

Konstrukt	Anteil der Lexeme mit $E11 < 10$ bei $N = 10^6$	Anteil der Lexeme mit $E11 < 10$ bei $N = 10^7$
SNBETH	8%	47%
SNBETO	3%	34%
SNBEWH	42%	90%
SNTH	1%	18%
SNTO	0.5%	6%
SNWH	0%	9%
THBESN	7%	49%
THSN	0.6%	13%

Tabelle 21: Erfüllung der Approximationsbedingung

Bei einem größeren Erwartungswert, nähert sich die Poissonverteilung der Normalverteilung. Die Grafik nähert sich also einer Glockenform mit Symmetriezentrum in $E11$. Das erklärt eine hohe negative Korrelation nach Spearman mit der χ^2-Teststatistik (Tabelle 22).

Konstrukt	Korrelation nach Spearman zwischen pois und chi für $N = 10^6$	Korrelation nach Spearman zwischen pois und chi für $N = 10^7$
SNBETH	-0.9908	-0.9765
SNBETO	-0.9884	-0.9828
SNBEWH	-0.9584	-0.9597
SNTH	-0.9912	-0.9873
SNTO	-0.9903	-0.9895
SNWH	-0.9794	-0.9927
THBESN	-0.9972	-0.9641
THSN	-0.9978	-0.9810

Tabelle 22: Rangkorrelation nach Spearman für pois und chi

Aus der Abbildung 26 sieht man, dass bei den fixierten $C1$, $R1$ und N die χ^2-Statistik mit wachsendem $O11$ genau dann steigt, wenn die Poisson-Likelihood senkt und umgekehrt.

Die Poisson-Likelihood liefert also ähnliche Ergebnisse wie χ^2-Teststatistik, ist aber für die Messung der Assoziation zwischen Lexem und Konstrukt nicht geeignet.

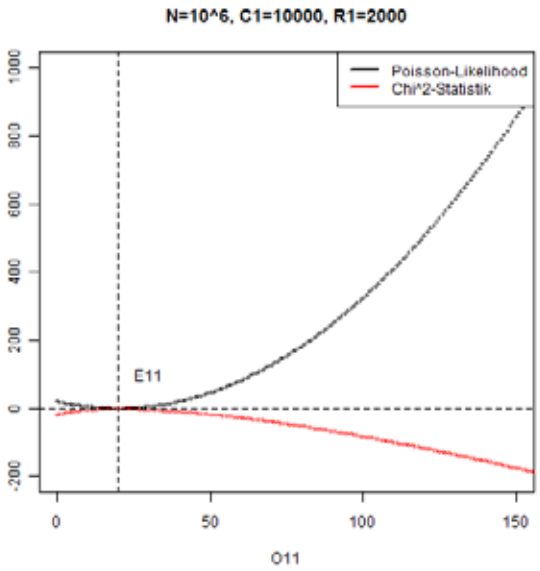

Abbildung 25: Vergleich von pois und χ^2-Statistik

3 Clusteranalyse der Assoziationsmaße

Im Kapitel 2 wurden 11 Assoziationsmaße der Collostructional analysis betrachtet und analysiert. Die Maße wurden dabei intuitiv nach ihre Herkunft in folgende Gruppen untterteilt: 1) Assoziationsmaße die von N unabhängig sind: MS, Jaccard und L2, 2) Assoziationsmaße die von N abhängig, aber keine p-Werte sind: Odds Ratio, relatives Risiko, MI und z-Score. 3) die p-Werte: der p-Wert des exakten Test nach Fisher und der p-Wert des χ^2-Tests, 4) Poisson-Likelihood.

Es hat sich folgendes herausgestellt:

- Grundsätzlich sind die Werte der Assoziationsmaße außerhalb der Konstrukte nicht vergleichbar.
- Die Werte der Assoziationsmaße sind wegen der Willkürlichkeit der Auswahl von N nicht interpretierbar. Eine Ausnahme bilden die Maße aus der 1. Gruppe, die von N unabhängig sind.
- Dem Vergleich der Assoziationsstärke der Lexeme zu Konstruktn dienen also nur die Ranglisten, die mit Hilfe der Assoziationsmaße hergestellt wurden.
- Den p-Wert und die Teststatistik des zweiseitigen χ^2-Tests sowie die Poisson-Likelihood kann man für die Bildung der Ranglisten nicht verwenden, da diese die Richtung der Entfernung von der Nullhypothse (evtl. Erwarteten Häufigkeit) nicht angeben.
- Die Ranglisten der Maße aus der 1.Gruppe hängen nicht von N ab, die Koeffizienten aus der zweiten Gruppe sind auch ziemlich robust gegen Veränderung des Stichprobenumfanges. Die p-Werte reagieren auf die Vergrößerung von N sehr sensibel.
- Die Ähnlichkeit der Maße hängt vom Konstrukt ab: von seiner Häufigkeit, von der Anzahl der Lexeme, die in diesem Konstrukt vorkommen und von der Verteilung der Häufigkeiten O11.

3.1 Clusteranalyse: Methode

Den Vergleich der Maße miteinander führen wir mit Hilfe der Clusteranalyse durch. Als Basis dafür nehmen wir die Korrelationsmatrix nach Spearman. Zwei Maße: χ^2-Teststatistik und Poisson-Likelihood werden aus der Analyse ausgenommen. Die Clusteranalyse wird getrennt für jedes Konstrukt bei $N = 10^6$ und $N = 10^7$ durchgeführt.

Im Gegensatz zu den anderen Maßen interpretiert man die kleineren p-Werte des Fischer-Tests und des einseitigen χ^2-Tests als ein Zeichen für größere Assoziation zwischen Lexem und Konstrukt. Deswegen wurden die Vorzeichen der entsprechenden Korrelationskoeffizienten umgekehrt.

Vor der Analyse wird die Korrelationsmatrix in eine Distanzmatrix umgewandelt. Da ein Korrelationskoeffizient die Werte zwischen -1 und 1 animmt, wird dafür folgende Transformation verwendet:

$d_{nm} = \sqrt{2 * (1 - s_{nm})}$.[4, S.442]

Diese Distanz nimmt Werte von 0 bis 2 an.

Die Clusteranalyse wird nach agglomirativen Verfahren mit Hilfe der R-Funktion `agnes()` aus dem Paket `cluster` geführt. Für die Berechnung der Distanzen zwischen den Clustern wird das average-Verfahren verwendet: der Abstand zwischen den Clustern wird als Durchschnitt aller Distanzen zwischen Objekten dieser Klassen berechnet:

$D(C_k, C_j) = \frac{1}{n_k n_j} \sum_{n \in C_k} \sum_{m \in C_j} d_{nm}$.

Fusioniert werden in jedem Schritt die Cluster mit der minimalen Distanz. Dadurch entstehen die Cluster mit den Elementen, die im Mittel hinreichend ähnlich sind.

Bemerkung: Es wurden auch die anderen Verfahren (complete und single Linkage) ausprobiert. Diese liefern in den meisten Fällen die gleiche Ergebnisse. Nur bei den untypischen Konstrukten SNWH und SNBEWH gibt es leichte Abweichungen.

Die Ergebnisse der Clusteranalyse und die entsprechenden Bannerplots und Dendrogramme werden uns helfen, die Ähnlichkeit der Assoziationsmaße zu analysieren. Eine Vorstellung über die Struktur der Daten bekommen wir mit Hilfe des agglomerativen Koeffizienten

$ac = \frac{1}{N} \sum_{i=1}^{N} 1 - \frac{d_{i,1}}{s_{i,l}}$,

mit i - jede Beobachtung, $d_{i,1}$- Abstand zum 1.Cluster, mit dem das Element i fusioniert wird und $d_{i,l}$ ist der Abstand zu dem Cluster, der im letzten Schritt fusioniert wird. $1 - ac$ zeigt also das durchschnittliche Verhältnis für $\frac{d_{i1}}{d_{il}}$. Je kleiner dieses ist desto stärker unterscheiden sich die Daten voneinander.

Meistens teilen sich die Koeffizienten in 3 Gruppen. Um die Qualität dieser Unterteilung zu überprüfen, vergleichen wir den minimalen Abstand zwischen den Zentren der Cluster mit dem maximalen durchschnittlichen Abstand der Elemente zu den Zentren ihrer Cluster (Das entsprechende Verhältnis nennen wir K). Wenn dieser viel größer ist, dann sind die Gruppen in sich homogen und die Cluster sind untereinander heterogen. Um die nötigen Größen zu berechnen, verwenden wir die R-Funktionen `kmeans()` und `diss()` aus dem Package `cluster`.

Die Clusteranalyse wurde für alle Konstrukte durchgeführt. Die Ergebnisse für alle Konstrukte außer SNWH und SNBEWH unterscheiden sich kaum. Deswegen betrachten wir nur zwei Beispiele davon: für die Konstrukte SNTO und SNBETH. Die Konstrukte SNWH und SNBEWH betrachten wir extra.

3.2 Konstrukte SNTO und SNBETH

Das **Konstrukt SNTO** ist häufig ($C1 = 228165$). Es treten aber nur 30% der Lexeme mit diesem Konstrukt zusammen auf. Die Korrelation nach Spearman für $N = 10^6$ und $N = 10^7$ ist in den Tabellen 23 und 24 berechnet.

	L2	or	rr	MI	MS	jac	zscore	ft	chi1
L2	1.000	0.996	0.999	0.999	0.3087	0.334	0.811	−0.6000	−0.771
or	0.996	1.000	0.999	0.999	0.2854	0.312	0.830	−0.5975	−0.773
rr	0.999	0.999	1.000	1.000	0.2981	0.324	0.821	−0.5986	−0.772
MI	0.999	0.999	1.000	1.000	0.2981	0.324	0.821	−0.5986	−0.772
MS	0.308	0.285	0.298	0.298	1.0000	0.999	−0.135	−0.0946	−0.127
jac	0.334	0.312	0.324	0.324	0.9990	1.000	−0.106	−0.1071	−0.144
zscore	0.811	0.830	0.821	0.821	−0.1354	−0.106	1.000	−0.6160	−0.796
ft	−0.600	−0.597	−0.598	−0.598	−0.0946	−0.107	−0.616	1.0000	0.800
chi1	−0.771	−0.773	−0.772	−0.772	−0.1270	−0.144	−0.796	0.8003	1.000

Tabelle 23: Rangkorrelationsmatrix für SNTO, $N = 10^6$

	L2	or	rr	MI	MS	jac	zscore	ft	chi1
L2	1.000	0.999	0.999	0.999	0.308	0.334	0.922	−0.884	−0.922
or	0.999	1.000	0.999	0.999	0.300	0.327	0.917	−0.880	−0.917
rr	0.999	0.999	1.000	1.000	0.298	0.324	0.916	−0.879	−0.916
MI	0.999	0.999	1.000	1.000	0.298	0.324	0.916	−0.879	−0.916
MS	0.308	0.300	0.298	0.298	1.000	0.999	0.583	−0.590	−0.583
jac	0.334	0.327	0.324	0.324	0.999	1.000	0.607	−0.615	−0.607
zscre	0.922	0.917	0.916	0.916	0.583	0.607	1.000	−0.983	−0.999
ft	−0.884	−0.880	−0.879	−0.879	−0.590	−0.615	−0.983	1.000	0.983
chi1	−0.922	−0.917	−0.916	−0.916	−0.583	−0.607	−0.999	0.983	1.000

Tabelle 24: Rangkorrelationsmatrix für SNTO, $N = 10^7$

Die Tabellen 23 und 24 zeigen starke Korrelation zwischen MS und Jaccard, und zwischen L2, Odds Ratio, relativem Risiko und MI. Es gibt auch eine mittelstarke Korrelation zwischen z-Score und L2, sowie mit den p-Werten. Bei $N = 10^7$ erhöht sich wesentlich die Korrelation zwischen z-Score, p-Werten und den anderen Maßen außer MS und Jaccard. Die Clusteranalyse liefert folgende Ergebnisse für $N = 10^6$.

Das linke Dendrogramm zeigt, dass die untersuchten Koeffizienten sich visuell in drei Gruppen teilen:

GRUPPE 1: L2, rr, MI, or
GRUPPE 2: MS und Jaccard
GRUPPE 3: z-Score, chi1, ft

Die Gruppen 1 und 2 sind sehr homogen und weit voneinander entfernt. Die dritte Gruppe ist weniger homogen. Bei N^7 nähert sich die 3.Gruppe der 1.Gruppe. Die Messungen innerhalb der 3.Gruppe werden homogener. Insgesamt steht diese Klassifikation in Übereinstimmung mit dem Ergebnis von D.Wiechman.

Der agglomerative Koeffizient vergrößert sich mit wachsendem N von 0.87 bis 0.98, was eine stärkere Struktur in den Daten bedeutet. Der Abstand zwischen den Gruppen vor der letzten Fusion wird dabei etwas kleiner (von 2.91 zu 2,6).

Berechnen wir für $N = 10^6$ die euklidischen Abstände zwischen den Clustern:

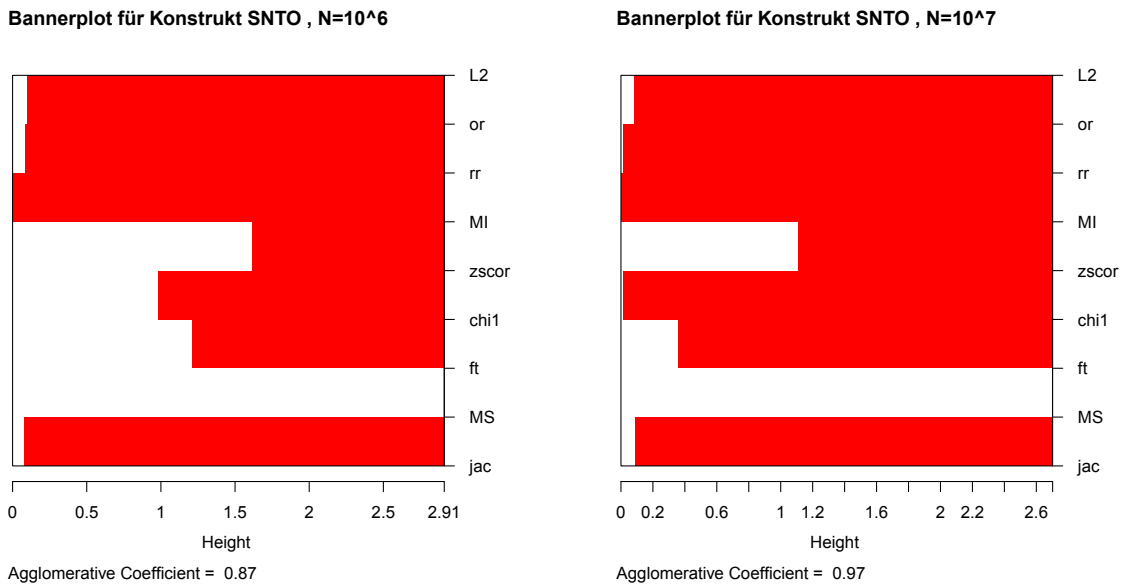

Abbildung 26: SNTO: Bannerplots der Clusteranalyse

	1	2
2	3.049	
3	1.494	2.647

Mittlere euklidische Abstände von Zentrum innerhalb der Cluster:

Gruppe	Mittlerer Abstand zum Zentrum
1	0.026
2	0.029
3	0.383

$K = \frac{\min(3.049,1.494,2.647)}{\max(0.026,0.029,0.383)} = 3.901$

Für $N = 10^7$ sehen die entsprechenden Zahlen folgendermaßen aus. Distanzen zwischen den Clustern:

	1	2
2	2.287	
3	1.096	2.934

Mittlere euklidische Abstände von Zentrum innerhalb der Cluster:

Gruppe	Mittlerer Abstand zum Zentrum
1	0.097
2	0.032
3	0.018

$K = 11.275$

Sowohl im Fall $N = 10^6$ als auch im Fall $N = 10^7$ gibt es eine klare Trennung zwischen den Clustern. Diese wird bei größerem N deutlicher.

Der **Konstrukt SNBETH** ist relativ häufig ($C1 = 30992$) und kommt zusammen mit 55% der Lexeme vor. Die Clusteranalyse liefert für dieses Lexem folgende Ergebnisse (Abbildung 28): Aus den Bildern sieht man, dass der Abstand zwischen den Maßen etwas kleiner als bei SNTO ist. Der agglomerative

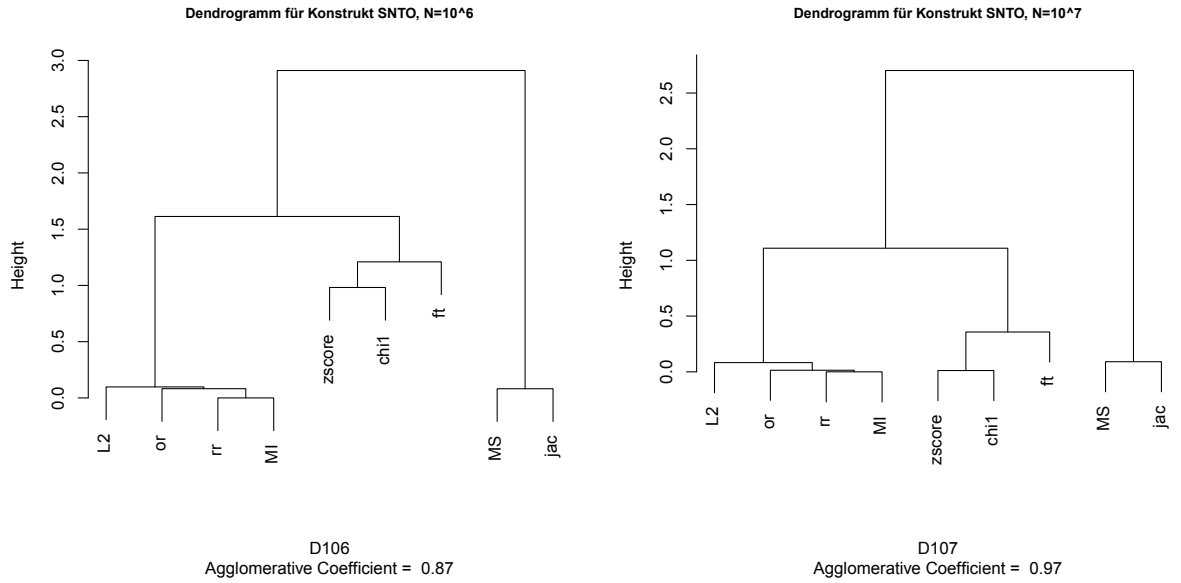

Abbildung 27: SNTO: Dendrogramm der Cluasteranalyse

Koeffizient ist bei $N = 10^6$ höher und bei $N = 10^7$ kleiner als bei dem Konstrukt SNTO. Die Messungen von L2 unterscheiden sich stärker von den Messungen anderer Maße aus der 1.Gruppe. Trotzdem liegen sie sehr nah beeinander. Die erste Gruppe und die dritte Gruppe sind etwas näher aneinander als bei SNTO. Grundsätzlich sind aber die Ergebnisse ziemlich ähnlich.

Distanzen zwischen den Clustern bei $N = 10^6$.

	1	2
2	2.460	
3	2.689	2.934

Mittlere euklidische Abstände von Zentrum innerhalb der Cluster:

Gruppe	Mittlerer Abstand zum Zentrum
1	0.047
2	0.083
3	0.22

$K = 7.38$

Distanzen zwischen den Clustern bei $N = 10^6$.

	1	2
2	2.330	
3	0.771	2.085

Mittlere euklidische Abstände vom Zentrum innerhalb der Cluster:

Gruppe	Mittlerer Abstand zum Zentrum
1	0.044
2	0.093
3	0.045

$K = 8.27$

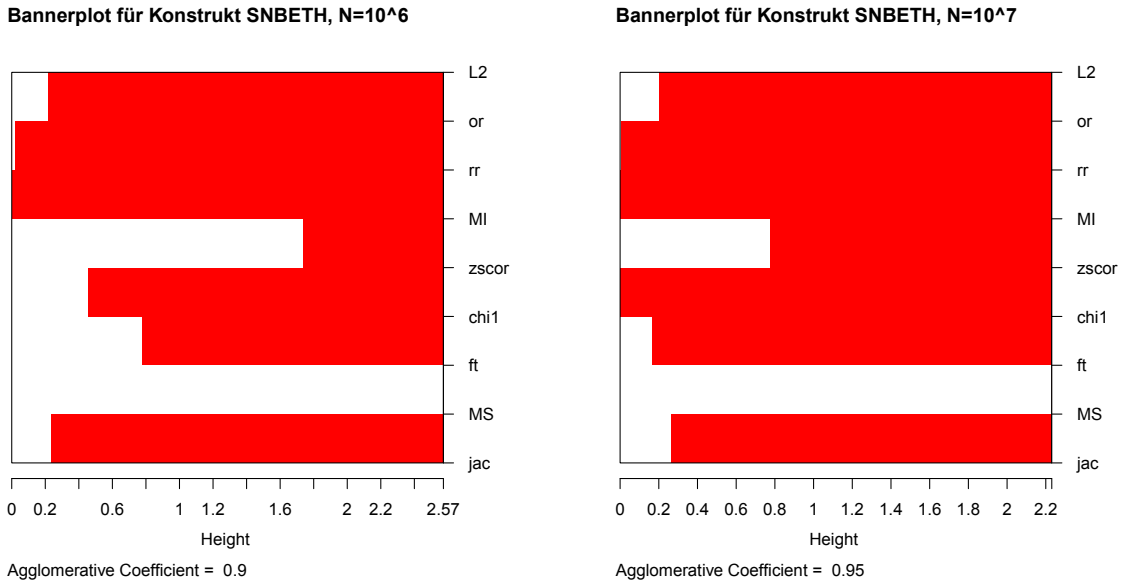

Abbildung 28: SNBETH: Bannerplots

3.3 Konstrukte SNBEWH und SNWH

Die Ergebnisse der Clusteranalyse für diese zwei Konstrukte unterscheiden sich stark von den anderen. Das Konstrukt SNBEWH ist selten (C1=1712), das Konstrukt SNWH ist relativ häufig. Die beiden kommen aber nur mit wenigen Lexemen vor: 3% und 5%.

Konstrukt SNWH ($C1 = 29492$, 31 Lexeme)

Korrelationsmatrix nach Spearman für $N = 10^6$ ist in der Tabelle 26 berechnet,für $N = 10^7$ in der Tabelle 27 :

	L2	or	rr	MI	MS	jac	zscore	ft	chi1
L2	1.000	0.926	0.931	0.931	0.932	0.966	0.590	−0.511	−0.550
or	0.926	1.000	0.999	0.999	0.848	0.872	0.797	−0.700	−0.744
rr	0.931	0.999	1.000	1.000	0.852	0.877	0.788	−0.688	−0.735
MI	0.931	0.999	1.000	1.000	0.852	0.877	0.788	−0.688	−0.735
MS	0.932	0.848	0.852	0.852	1.000	0.987	0.485	−0.417	−0.475
jac	0.966	0.872	0.877	0.877	0.987	1.000	0.504	−0.436	−0.481
zscore	0.590	0.797	0.788	0.788	0.485	0.504	1.000	−0.884	−0.914
ft	−0.511	−0.700	−0.688	−0.688	−0.417	−0.436	−0.884	1.000	0.915
chi1	−0.550	−0.744	−0.735	−0.735	−0.475	−0.481	−0.914	0.915	1.000

Tabelle 25: Rangkorrelation für SNWH, $N = 10^6$

Ergebnisse der Clusteranalyse (Abbildungen 31 und 32):

Die Messungen der Koeffizienten sind ähnlicher als bei den anderen Konstrukten. Die Struktur der Daten ist weniger ausgeprägt: der agglomerative Koeffizient beträgt 0.83 für $N = 10^6$ und $N = 10^7$, ist also relativ klein verglichen mit den anderen Konstrukten. Die Verteilung in den Clustern ändert sich. Jetzt liefern L2, MS und Jaccard ähnliche Ergebnisse. Ein Grund dafür sind die vergleichbaren Größen der Mengen C1 und R1 bei dem Konstrukt SNWH. Bei $N = 10^7$ liefern L2, z-Score und chi1 überraschenderweise

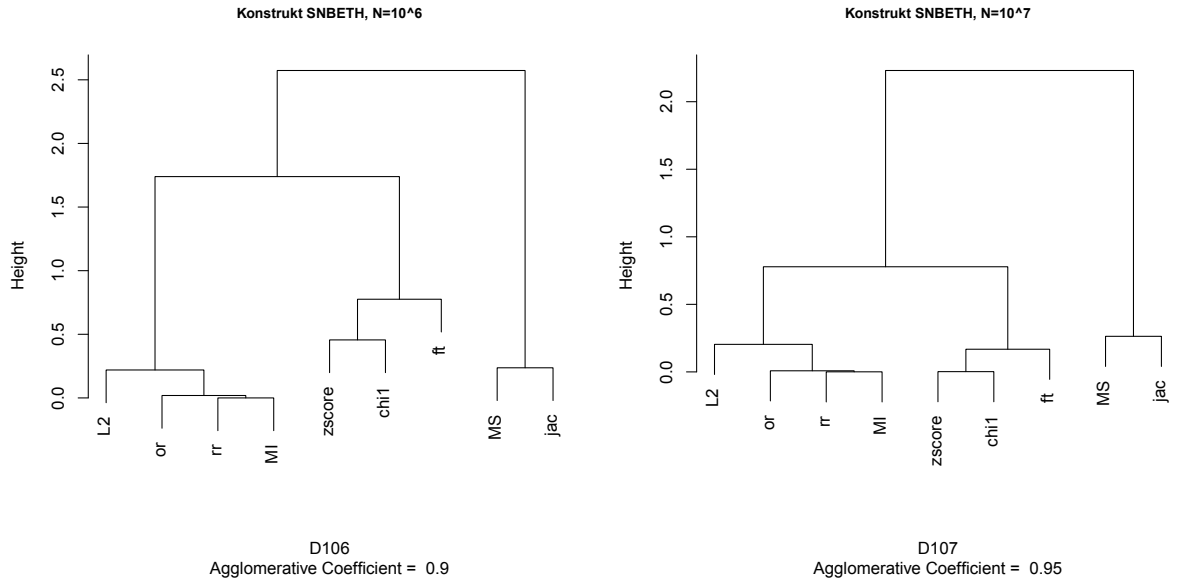

Abbildung 29: SNBETH: Dendrogramme

	L2	or	rr	MI	MS	jac	zscore	ft	chi1
L2	1.000	0.940	0.931	0.931	0.932	0.966	0.988	−0.931	−0.988
or	0.940	1.000	0.997	0.997	0.853	0.883	0.924	−0.892	−0.924
rr	0.931	0.997	1.000	1.000	0.852	0.877	0.917	−0.887	−0.917
MI	0.931	0.997	1.000	1.000	0.852	0.877	0.917	−0.887	−0.917
MS	0.932	0.853	0.852	0.852	1.000	0.987	0.961	−0.929	−0.961
jac	0.966	0.883	0.877	0.877	0.987	1.000	0.984	−0.932	−0.984
zscore	0.988	0.924	0.917	0.917	0.961	0.984	1.000	−0.940	−1.000
ft	−0.931	−0.892	−0.887	−0.887	−0.929	−0.932	−0.940	1.000	0.940
chi1	−0.988	−0.924	−0.917	−0.917	−0.961	−0.984	−1.000	0.940	1.000

Tabelle 26: Korrelationstabelle für SNWH, $N = 10^7$

ähnliche Ergebnisse, während der Fisher-Test sich Jaccard und MS nähert.

GRUPPEN für $N = 10^6$:
1. Odds Ratio, MI, rr
2. ft, MS, Jaccard
3. z-Score, chi1, L2

Distanzen zwischen den Clustern bei $N = 10^6$.

	1	2
2	0.98	
3	1.696	1.946

Gruppe	Mittlerer Abstand zum Zentrum
1	0.021
2	0.161
3	0.216

$K = 4.98$

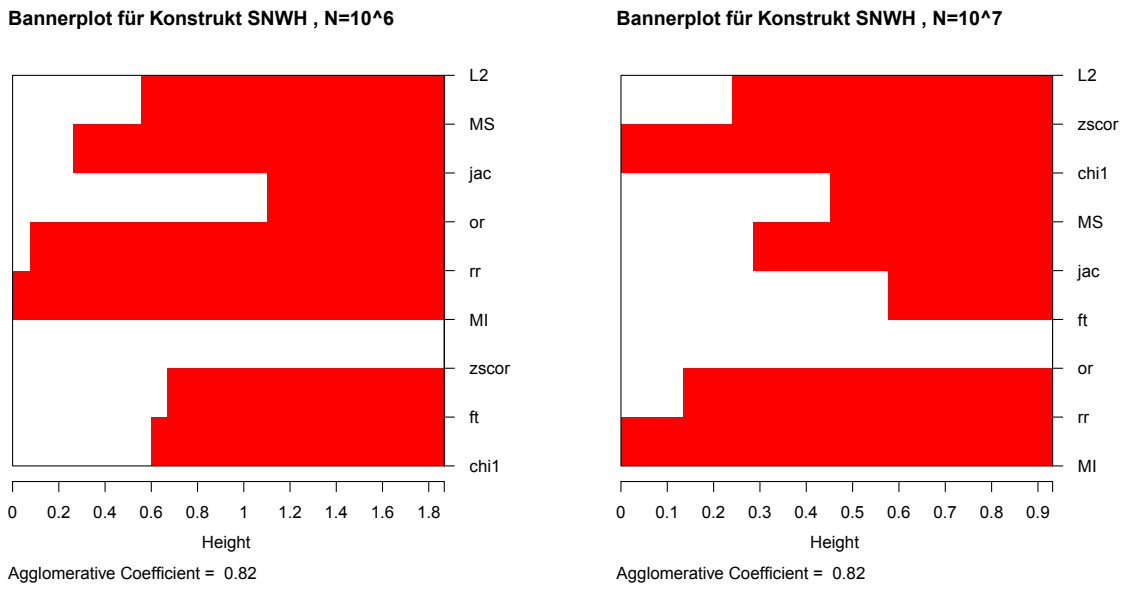

Abbildung 30: SNWH: Bannerplots

Obwohl sich die Distanzen zwischen den Cluster verkleinert haben, ist der Abstand zwischen den Clustern immer noch wesentlich größer als die Abstände innerhalb der Cluster.

GRUPPEN für $N = 10^7$:
1. Odds Ratio, MI, rr
2. MS, Jaccard, ft
3. L2, z-Score, Chi1

Distanzen zwischen den Clustern bei $N = 10^7$.

	1	2
2	0.987	
3	0.917	0.460

Gruppe	Mittlerer Abstand zum Zentrum
1	0.037
2	0.175
3	0.081

$K = 2.62$

Clusteranalyse für SNBEWH liefert ähnliche Ergebnisse.

4 Zusammenfassung

In der vorliegenden Arbeit wurden 11 Koeffzienten der Collostructional analysis statistisch untersucht. Es hat sich folgendes herausgestellt:

1) Die Koeffizienten MS und Jaccard liefern ähnliche Ergebnisse und wurden bei der Clusteranalyse immer in die gleiche Gruppe klassifiziert. Allerdings unterscheiden Messungen dieser Gruppe sich sehr stark von

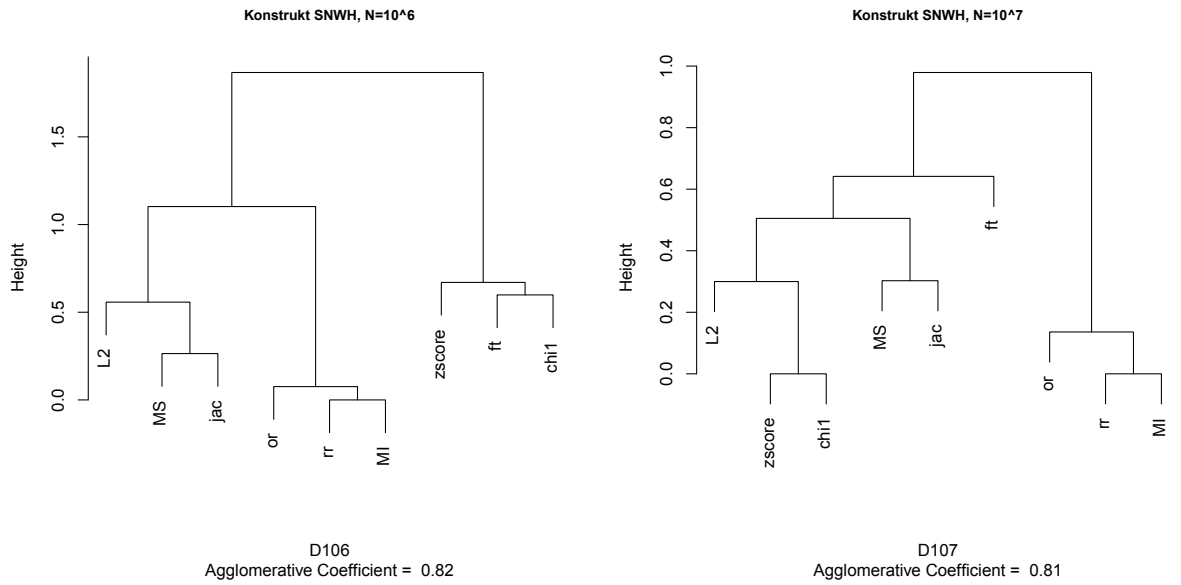

Abbildung 31: SNWH: Dendrogramme

den Ergebnissen der anderen Koeffizienten. Ein Grund dafür ist die Tatsache, dass die beide Koeffizienten nicht alle Informationen aus der Basis-Kontingenztafel berücksichtigen.

2) Der Vorschlag, Assoziation mittels des zweidimensionalen Vektors (Reliance, Attraction) zu beschreiben lässt die Beziehungen zwischen Lexemen und Konstrukten anschaulich darstellen. Die euklidische Norm L2 dieses Vektors bewährt sich als ein vernünftiges Assoziationsmaß, das einerseits für seine Berechnung keinen Stichprobenumfang benötigt, andererseits sich den Koeffizienten Odds Ratio, MI und rr nähert, die alle Informationen in der Basistabelle benutzen. Im Gegensatz zu Odds Ratio, MI und rr hängt L2 eindeutig positiv sowohl von Reliance als auch Attraction ab.

3) Die Ranglisten, die mittels Odds Ratio, rr und MI erstellt wurden sind für jedes Konstrukt fast identisch und resistent gegenüber einer Veränderung des Stichprobenumfanges. Allerdings bestehen Zweifel, dass diese in allen Fällen wirklich die Assoziation zwischen Konstrukt und Lexem messen. Der Grund ist ein zweideutiger (positiver und negativer) Einfluss von Attraction auf diese Koeffizienten. Die Werte der Koeffizienten hängen von N ab und können daher nicht direkt interpretiert werden.

4) z-Score und der p-Wert des einseitigen χ^2-Tests liefern sehr ähnliche Ergebnisse. Der p-Wert des exakten Tests nach Fisher liefert etwas andere Ergebnisse aber liegt trotzdem sehr nah bei den beiden. Allerdings hängen diese Messungen sehr stark vom Stichprobenumfang ab. Besonders wenn die beobachteten Häufigkeiten in der Nähe der erwarteten Häufigkeiten liegen, kommt es zur Veränderung der Reihenfolge der Lexeme in den Ranglisten für $N = 10^6$ und $N = 10^7$. Bei $N = 10^7$ korrelieren die Messungen dieser drei Maßen stärker mit den Messungen von Odds Ratio u.ä. Wegen der Empfindlichkeit gegenüber der Veränderung von N sind diese Maße nicht sehr gut für die Collostructional analysis geeignet. Außerdem neigen die p-Werte dazu, die Assoziationsstärke bei den seltenen Lexemen zu unterschätzen.

5) Der p-Wert des zweiseitigen χ^2-Tests und die Poisson-Likelihood sind für die Messung der Assoziationsstärke nicht geeignet, da diese die Richtung der Abweichung von der erwarteten Häufigkeit nicht identifizieren.

6) Insgesamt nähern sich alle Messungen außer MS und Jaccard bei steigendem Stichprobenumfang einander an und werden homogener. Allerdings verhalten sich die Maße bei den seltenen Konstrukten und bei den Konstrukten, die mit wenigen Lexemen auftreten, untypisch. Gründe dafür wurden in der vorliegenden Arbeit im Kapitel 2 analysiert.

5 Literatur

1. H.-J. Schmid, H. Küchenhoff. Lookong behind the scenes of collostructional analysis (unveröffentlicht)
2. D.Wiechmann. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias: Corpus Linguistics and LinguisticTheory 42 (2008), 253290.
3. W. Voß. u.a. Taschenbook der Statistik. Fachbuchverlag Leipzig. 2004.
4. L.Fahrmeier u.a. Multivariate statistische Verfahren: Walter de Gruyter. Berlin. New York, 1996.
5. S. Th. Gries and A. Stefanowitsch. Extending collostructional analysis. A corpus-based perspective on alternations: International Journal of Corpus Linguistics 9:1 (2004), 97-129.
6. A.Stefanowitsch. Quantitative Korpuslinguistik und sprachliche Wirklichkeit: s. 141-155
7. S.Th.Gries. Useful statistics for corpus linguistics.
8. S. Th. Gries and A. Stefanowitsch. Collostructions: Investigating the interaction of words and constructions: International Journal of Corpus Linguistics 8:2 (2003), 209-243.