# Methoden zum qualitativen Vergleich von Signifikanzmaßen zur Kollokationsidentifikation[1]

Stefan Evert, Ulrich Heid, Wolfgang Lezius • Institut für Maschinelle Sprachverarbeitung • Universität Stuttgart

## Zusammenfassung

Zur Extraktion von Kollokationen und idiomatischen Wendungen aus Textkorpora werden in der Literatur zahlreiche Signifikanzmaße vorgeschlagen. Doch statistisch fundierte und vielfach eingesetzte Maße wie etwa der $\chi^2$-Test schneiden bei der Kollokationsidentifikation eher schwach ab. Daher besteht insbesondere aus linguistischer und lexikographischer Sicht Bedarf an einem aussagekräftigen empirischen Vergleich von Signifikanzmaßen. In diesem Artikel zeigen wir die Nachteile der gängigen Vorgehensweise für solche Vergleiche auf (manuelle Bewertung der besten $n$ Kandidaten für jedes untersuchte Maß). Wir schlagen anschließend eine alternative Vorgehensweise vor, die auf der Betrachtung von *Precision* und *Recall* in Abhängigkeit von $n$ basiert. Wir präsentieren erste Ergebnisse unserer Experimente und interpretieren diese im Vergleich zu den herkömmlichen Verfahren.

## 1 Kontext und Zielsetzung

Die typische Vorgehensweise zur Gewinnung von Kollokationen aus Textkorpora besteht darin, zunächst durch (morpho-)syntaktische Verfahren[2] homogenes Rohmaterial zu gewinnen. Anschließend kommen statistische Signifikanzmaße (z.B. Mutual Information (MI), t-Score, $\chi^2$, common-birthday oder Log-Likelihood[3]) zum Einsatz, die einen Schätzwert für den Korrelationsgrad (und damit die Relevanz) der Kandidatenpaare liefern. Diese Maße werden zum Umsortieren und Filtern der Ausgangslisten herangezogen.[4]

Bislang gibt es wenige Verfahren zur vergleichenden Beurteilung der Güte von Signifikanzmaßen hinsichtlich der Identifikation von Kollokationen. Abschnitt 3 stellt die Probleme publizierter Evaluierungsstrategien zusammen. In Abschnitt 4 beschreiben wir eine neue Vorgehensweise, die den objektiven Vergleich von Signifikanzmaßen ermöglichen soll. Als exemplarische Anwendung haben wir aus einem Korpus deutscher Gesetzestexte (813 483 Wortformen) Adjektiv+Nomen-Paare extrahiert und diese mit den eingangs genannten Signifikanzmaßen bewertet. Durch Vergleich mit einer manuell erstellten vollständigen Referenzmenge wurden für jedes der fraglichen Maße Precision- und Recall-Kurven berechnet (siehe hierzu Abschnitt 5.2).

[1] Wir danken Brigitte Krenn für ihre Kommentare zu einer früheren Version dieses Artikels, Daniela Knoche und Ciprian Vasii-Gerstenberger für ihre Unterstützung bei der Erstellung der Referenzliste.

[2] Extraktion auf der Basis von Wortklasse, Adjazenz, Kasus, Determination, etc.; ggf. (partielles) Parsen.

[3] Siehe etwa [13] oder [5] für eine Übersicht dieser und ähnlicher Verfahren. Im Gegensatz zu praktisch allen anderen Verfahren basiert MI nicht auf einem statistischen Test. Wir behalten dennoch den hier nicht ganz zutreffenden Begriff „Signifikanzmaß“ seiner Prägnanz wegen bei.

[4] Vgl. [11].

## 2 Zum Kollokationsbegriff

Der Begriff „Kollokation“ ist in der Linguistik und Lexikographie unterschiedlich definiert worden, und zum Teil werden ganz verschiedene Phänomene als „Kollokationen“ bezeichnet. Diese unterschiedlichen Definitionsansätze wirken sich natürlich auch bei der manuellen Sichtung von Kollokationskandidaten aus.

Aus lexikographischer wie linguistischer Sicht ist es sinnvoll, Kollokationen als teilidiomatisierte Verbindungen von zwei Lexemen aufzufassen[5]. Die teilweise Idiomatisierung drückt sich beispielsweise in Einschränkungen der kompositionellen Analysierbarkeit aus, die zum Teil als Tests für das Vorliegen von Kollokationen benutzt werden können: in vielen Nomen+Verb-Kollokationen ist die Bedeutung des Verbs „abgeschwächt“ gegenüber der (oder den) Bedeutung(en) desselben Verbs außerhalb der Kollokation: ein Paradebeispiel sind die Funktionsverbgefüge, wo die Bedeutung des Verbs ggf. bis zur Operatorfunktion und/oder zur Aktionsartmarkierung reduziert ist. In vielen Nomen+Verb-Kollokationen verhält sich die Nominalgruppe morphosyntaktisch anders als eine Nominalgruppe in kompositionell analysierbaren Syntagmen (Einschränkungen hinsichtlich Determination, Modifizierbarkeit, referentieller Verfügbarkeit des Nomens).

In einem solchen Ansatz stehen Kollokationen in einer Reihe mit anderen (partiell) idiomatisierten Wortverbindungen.

Im Information Retrieval und in bestimmten korpuslinguistischen Ansätzen wird ein weiter gefaßter Kollokationsbegriff verwendet, der die Häufigkeit von Wortkombinatio-

[5] Zwei Lexeme plus closed class-Items, vgl. z.B. [4]; für das Deutsche bedeutet dies: Nomen+Verb, Adjektiv+Nomen, Verb+Adverb, Adjektiv+Adverb.

nen sehr stark berücksichtigt und typische, häufige Kombinationen mit einschließt.[6] Zum Bereich „Kollokation“ werden die für „flüssige Sprache“ benötigten typischen Wortkombinationen gerechnet, selbst wenn keine semantische oder morpho-syntaktische Besonderheit vorliegt. In Fachsprachen stellt sich dann das Problem der Abgrenzung zwischen Kollokation und Terminus (d.h. lexikalisierter Benennung für einen Gegenstand oder Sachverhalt). Aus dieser Sicht stehen Kollokationen in einer Reihe mit typischen Selektionsphänomenen.[7]

Statistische Verfahren zur Extraktion von Zweiwortkombinationen, wie sie hier evaluiert werden sollen, extrahieren verschiedene Typen von Wortkombinationen ohne Unterschied: typischerweise enthält das Extraktionsergebnis Beispiele für Kollokationen im engeren, linguistischen Sinne ebenso wie typische Wortverbindungen, die auf der Basis von Selektionsangaben beschrieben werden können oder Kombinationen, die einen in der Domäne häufigen Sachverhalt bezeichnen.

Adjektiv+Nomen-Verbindungen stellen in dem Sinne ein zusätzliches Problem dar, als hier, anders als bei Nomen+Verb-Kollokationen, eine (teil-)automatische Subklassifizierung der Typen von Wortverbindungen nur eingeschränkt möglich ist (vgl. Abschnitt 4.2).

# 3 Die Bewertung von Signifikanzmaßen

Die uns bekannten Evaluierungen von Signifikanzmaßen für die Kollokationsextraktion beruhen auf einer manuellen Beurteilung der $n$ besten Kandidaten. Dabei wird für die 50, 100 oder mehr Kandidatenpaare mit den höchsten Maßzahlen der Anteil der „true positives“ (TPs) gemäß der Kollokationsintuition des Evaluators bestimmt, die sogenannte Precision. Auf diese Weise evaluieren u.a. [3][8] und [9] ihre Kollokationssuche, und so vergleicht [11] verschiedene Verfahren; eine Ausnahme ist [7].

Untersuchungen der $n$ besten Kandidaten sind geeignet, um einen groben Überblick über die Resultate eines Verfahrens zu geben. Für lexikographische Anwendungen oder den Aufbau von NLP-Lexika ist dies aber nicht ausreichend:

- Die Evaluierung bestimmt nur die Precision, nicht den Recall der Signifikanzmaße (der Recall gibt an, welcher prozentuale Anteil der insgesamt im Korpus vorkommenden, manuell ausgewählten Kollokationen durch die $n$ besten Kandidaten abgedeckt wird).
- Die Ergebnisse beziehen sich auf Stichproben: unklar bleibt, ob sich das Bild, das die 100 besten Kandidaten geben, auf die Analyse größerer Teile des Korpus übertragen läßt. Insbesondere kann ein TP, den Maß $A$ unter die ersten 100 Kandidaten einordnet, bei Maß $B$ auf Rang 101 stehen und so nicht mehr gewertet werden.
- Jede Änderung einer Berechnungsformel oder die Aufnahme eines neuen Signifikanzmaßes in den Vergleich erfordert zusätzliche manuelle Arbeit.
- Der qualitative Vergleich der Ergebnisse verschiedener Maße[9] ist mühevoll.

# 4 Vorgehensweise und Experimente

## 4.1 Methodik

Um diese Nachteile auszugleichen, haben wir für unsere Experimente die folgende Vorgehensweise gewählt (vgl. **Abbildung 1**).

1. Aus dem zugrunde gelegten Korpus (hier 813 483 Wortformen) werden lemmatisierte Kandidatenpaare extrahiert (hier adjazente Adjektiv+Nomen-Paare).[10] Um den manuellen Arbeitsaufwand in akzeptablen Grenzen zu halten, werden Wortpaare mit der Frequenz 1 eliminiert.[11] Wir bezeichnen die resultierende Liste als Grundmenge (sie enthält in unserem Experiment 4652 Kandidatenpaare).
2. Durch manuelle Sichtung der Grundmenge wird eine Referenzmenge erstellt.[12] Im wesentlichen Unterschied zu anderen Evaluierungsmethoden wird das Ausgangsmaterial vollständig gesichtet. In unserem Experiment wurden 618 Kandidatenpaare akzeptiert, was einem Anteil von 13.28% TPs entspricht.
3. Die Kandidatenpaare der Grundmenge werden nach verschiedenen Signifikanzmaßen bewertet und sortiert. Für jedes Signifikanzmaß entsteht eine sogenannte Signifikanzliste. Die ersten $n$ Einträge jeder Signifikanzliste entsprechen den $n$ nach dem jeweiligen Maß besten Kandidatenpaaren. In unserem Experiment wurden bislang die folgenden Verfahren untersucht: Mutual Information (MI), $\chi^2$, t-Score, Entropie, Poisson-Verteilung, Log-Likelihood (vgl. [3]), Common Birthday (vgl. [10]) und Mutual Expectation (vgl [2]).

[6] In der Lexikographie und Korpuslinguistik z.B. [14].

[7] Diese Unterscheidung diskutiert [12] ausführlich im Zusammenhang mit der Reihenbildung bei Nomen+Verb-Kollokationen in der Informatik-Fachsprache.

[8] Dunning gibt hierbei keine genauen Precision-Werte an.

[9] Welche Phänomene werden in welcher Statistik nicht ausreichend berücksichtigt? Welches Maß liefert gute Ergebnisse für häufige Kandidatenpaare, welches für seltene?

[10] Der Extraktionsschritt wurde in unseren Experimenten mit Hilfe der Korpusanfrage-Software CQP durchgeführt (siehe [6]).

[11] Über derart seltene Ereignisse lassen sich kaum sinnvolle statistische Aussagen treffen. So enthält die nach dem heutzutage gebräuchlichen Log-Likelihood-Maß sortierte Liste aller Kandidatenpaare unter den ersten 2240 Einträgen kein einziges Adjektiv+Nomen-Paar mit der Frequenz 1.

[12] Jeder Teil der Grundmenge wird hierbei unabhängig von mindestens zwei Annotatoren bearbeitet (siehe hierzu auch Abschnitt 4.3).

4. Anhand der Referenzmenge können Precision und Recall der Signifikanzlisten für beliebige Werte von $n$ ermittelt werden. Bei graphischer Darstellung dieser Werte relativ zum prozessierten Anteil der jeweiligen Signifikanzliste ergeben sich Precision- und Recall-Kurven (siehe Abschnitt 5.2). Zusätzlich können Schaubilder für verschiedene Frequenzschichten oder für einzelne Annotatoren erzeugt werden.

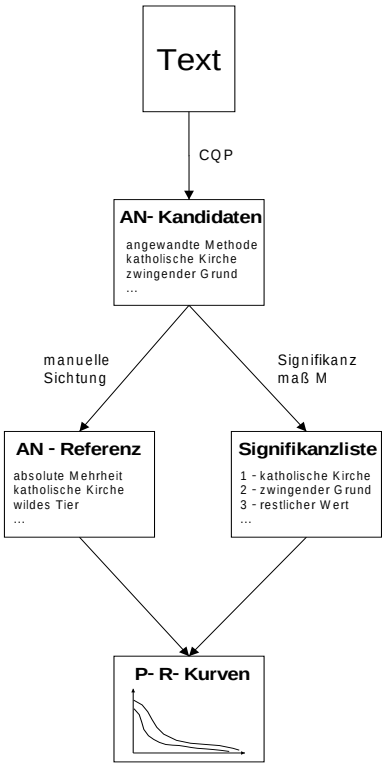


Abbildung 1: Vorgehensweise

## 4.2 Linguistische Aspekte der Erstellung des Referenzmaterials

Für unsere Experimente wurden Adjektiv+Nomen-Kombinationen ausgewählt, weil solche Daten auf sehr einfache Weise vollständig aus Corpora extrahiert werden können. Andererseits ist die Klassifikation von Adjektiv+Nomen-Kombinationen in Kollokationen, „typische Wortverbindungen“ und andere, nicht relevante Adjektiv+Nomen-Kombinationen in manchen Fällen problematisch.

Im Folgenden stellen wir zusammen, nach welchen Kriterien die TPs aus den Adjektiv+Nomen-Kombinationen manuell ausgefiltert wurden:

Nicht zu den Kollokationen gezählt werden:

- Kombinationen mit Adjektiven mit subkategorisierten Komplementen: *der [auf x] entfallende Anteil, [darüber] hinausgehende Auskünfte*;
- Kombinationen mit Ordinalia: *der siebte Abschnitt*; Ausnahmen: *das 18. Lebensjahr, das 65. Lebensjahr*;
- Kombinationen mit textdeiktischen Adjektiven (*erwähnt, folgend*), allgemeinen verweisenden Adjektiven (*derartig, ähnlich, entsprechend, übrig, . . .*) bzw. mit „distributiven“ Adjektiven (*jeweilig, einzeln, gewisse, . . .*);
- Teile von idiomatisierten Präpositional- bzw. Nominalphrasen: *auf absehbare Zeit, auf eigene Gefahr, [. . . ] gleichen Inhalts, [. . . ] höherer Ordnung.*

Folgende teilweise strittigen Fälle werden zu den TPs gerechnet:

- Von Verben „ererbte“ Kombinationen mit Partizipien: *abgegebene Erklärung, abzugebende Erklärung, ergehendes Urteil*;[13]
- „Typische“ Selektionsbeispiele[14]: *angemessen + [Entschädigung | Entgelt | Ausgleich | Frist...]*. Analog bei Komposita: *gesetzliche (Kündigungs)frist*.
- Durch Gebrauchskonvention idiomatisierte Adjektiv+Nomen-Kombinationen, z.B. in Namen: *Deutscher Bundestag, Katholische Kirche, . . .* (vgl. dazu auch [11]).

## 4.3 Probleme der manuellen Annotation

Bei der manuellen Sichtung der Daten treten Abweichungen zwischen verschiedenen Annotatoren auf: Gründe hierfür liegen in den unterschiedlichen Kollokationsauffassungen von Annotatoren (vgl. Abschnitt 2), dem Fehlen exakter Richtlinien für die Auswahl aus Adjektiv+Nomen-Kombinationen, sowie in der angestrebten Anwendung: z.B. haben [1] festgestellt, daß erhebliche Diskrepanzen zwischen Terminologen, Juristen, Lexikographen und Linguisten auftreten, die juristisches Textmaterial hinsichtlich der Frage sichten, welche Kombinationen wörterbuchrelevant sind.

In unseren Experimenten steht eine parallele Auswahl von Kollokationskandidaten durch mehrere Personen (und damit die Möglichkeit, Abweichungen quantitativ zu erfassen) noch aus. Für den Moment wird in das Referenzmaterial jede Kombination aufgenommen, die von mindestens einer Person als lexikonrelevant klassifiziert worden ist.

[13] Solche Formen können vorab identifiziert und dem Nomen+Verb-Kollokationen-Material zugeschlagen werden. Dann würden für die Auszählung von Adjektiv+Nomen-Kollokationen nur Fälle übrigbleiben, wo die Partizipialform lexikalisiert ist: *abgeschlossenes Hochschulstudium, gesprochenes Wort*, d.h. wo keine Nomen+Verb-Kollokation verfügbar ist.

[14] Frequenzinformation wird den manuellen Annotatoren nicht verfügbar gemacht, wohl aber Information über die prozentuale Verteilung verschiedener Formen einer Kombination.

# 5 Auswertung und Ergebnisse

In diesem Abschnitt stellen wir die Precision- und Recall-Kurven für die bislang untersuchten Signifikanzmaße dar und erläutern die sich bietenden Interpretationsmöglichkeiten. Im Vergleich zu den statistischen Verfahren wird zusätzlich eine lediglich nach der absoluten Häufigkeit (Frequenz) der Kandidatenpaare sortierte Kandidatenliste betrachtet.

## 5.1 Top-$n$-Listen

Nach der traditionellen Vorgehensweise ergeben sich für die ersten $n = 100$ bzw. $n = 400$ Kandidatenpaare der Signifikanzlisten die in **Tabelle 1** aufgeführten Precision-Werte.[15] Die früher oft angewandten Maße MI und $\chi^2$ schneiden erwartungsgemäß[16] deutlich schlechter ab als die lediglich nach Frequenz sortierten Listen. Während die Werte für $n = 100$ zunächst die in vielen Arbeiten beobachtete Präferenz für das Log-Likelihood-Maß bestätigen, wird diese bei $n = 400$ aufgehoben und es liegt sogar der Schluß nahe, daß ausgefeilte statistische Methoden wie Log-Likelihood gegenüber der Sortierung nach Frequenz keine wesentliche Verbesserung erzielen.

| | $n = 100$ | $n = 400$ |
|---|---|---|
| Mutual Expectation | 54.00% | 39.00% |
| Common Birthday | 53.00% | 38.00% |
| Poisson | 41.00% | 38.00% |
| t-Score | 51.00% | 37.75% |
| Log-Likelihood | 57.00% | 37.25% |
| Entropie | 49.00% | 37.00% |
| $\chi^2$ | 33.00% | 29.50% |
| Mutual Information | 19.00% | 19.75% |
| Frequenz | 46.00% | 35.50% |

Tabelle 1: Precision-Werte für die ersten $n$ Kandidaten.

## 5.2 Precision- und Recall-Kurven

Aussagekräftiger ist hingegen die Betrachtung des gesamten Verlaufs der Precision-Kurven[17] in den **Abbildungen 2 und 3**. In Abbildung 2 werden die Signifikanzlisten für MI und $\chi^2$ mit der nach Frequenz sortierten Kandidatenliste verglichen. Hier bestätigen sich die Beobachtungen von Abschnitt 5.1: die nach Frequenz sortierte Kandidatenliste erzielt bis zu einem $x$-Wert von knapp 50% (was einer Top-2000-Liste entspricht) durchweg bessere Ergebnisse als die Listen für MI bzw. $\chi^2$.

Aus dem Verlauf der Precision-Kurven in Abbildung 3 wird hingegen deutlich, daß die fast identische Güte der Bewertung nach Log-Likelihood im Vergleich zur Frequenzliste bei $n = 400$ (8.7%) als „Ausreißer“ einzustufen ist. Für andere Werte von $n$ erzielt Log-Likelihood deutlich bessere Werte, so daß wir die im vorigen Abschnitt geäußerte Vermutung, daß statistische Methoden keinen wesentlichen Vorteil bringen, verwerfen müssen. Die Precision-Kurven für Common Birthday, Mutual Expectation, sowie die aus Platzgründen nicht dargestellten Maße t-Score, Entropie und Poisson unterscheiden sich nicht wesentlich von den Ergebnissen, die Log-Likelihood erzielt.

Bei genauer Betrachtung von Abbildung 3 fällt auf, daß die aus den Top-$n$-Listen für $n = 100$ und $n = 400$ abgelesenen Precision-Werte in den erheblichen Schwankungen unterworfenen Anfangsbereich der Kurven fallen. Ein Vergleich der Kurven scheint folglich nur dann sinnvoll, wenn mehr als 10% der Signifikanzlisten berücksichtigt werden.

Die gepunktete horizontale Linie spiegelt den Anteil von TPs in der Grundmenge wieder. Dieser entspricht dem zu erwartenden Precision-Wert bei zufälliger Auswahl der Kandidaten (vgl. die Zufallsliste in Abbildung 2). Ein Vergleich der Precision-Kurven ist daher nur für solche $n$ interessant, für die die erzielten Werte erheblich über diesem Grundniveau liegen (in unserem Beispiel liegt die intellektuell bestimmte Grenze bei ca. 30%, was $n \approx 1500$ entspricht).

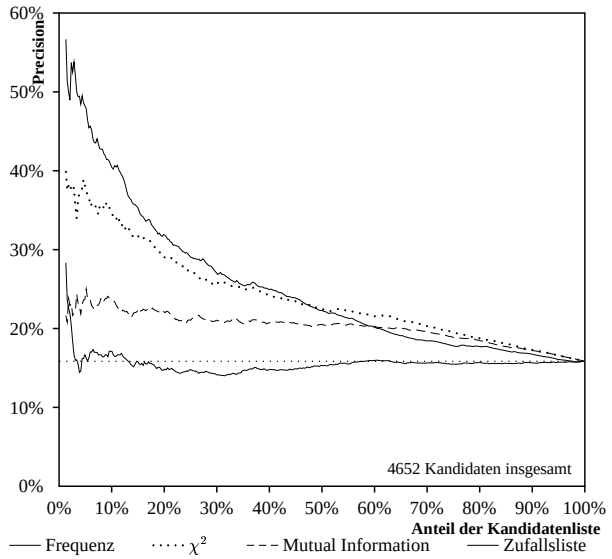


Abbildung 2: Precision-Kurven (1)

Für die Kollokationsextraktion aus kleinen Korpora und für terminologische Anwendungen ist auch der meist nicht berücksichtigte Recall-Wert relevant. **Abbildung 4** zeigt

[15] Die Einträge der Tabelle sind nach den für $n = 400$ erzielten Precision-Werten geordnet.

[16] Vgl. u.a. [11].

[17] Die $x$-Achse gibt hierbei an, welcher prozentuale Anteil der nach dem jeweiligen Signifikanzmaß sortierten Signifikanzliste betrachtet wird. Auf der $y$-Achse ist der zugehörige Precision-Wert abgetragen. Die Precision-Werte für $n = 100$ und $n = 400$ sind demnach bei $x = 2.2\%$ und $x = 8.7\%$ abzulesen.

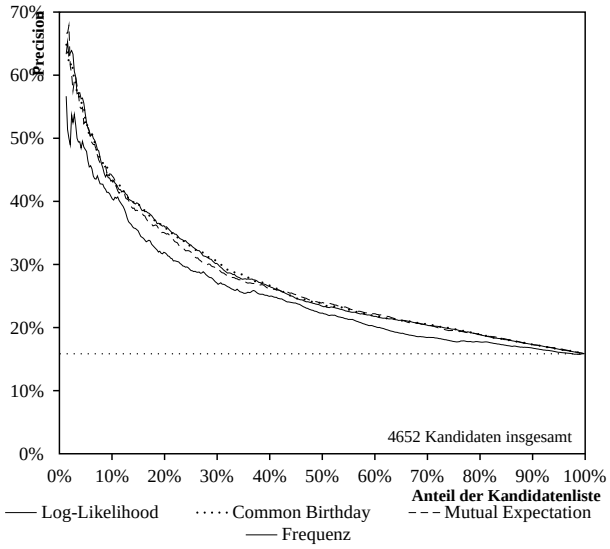


Abbildung 3: Precision-Kurven (2)

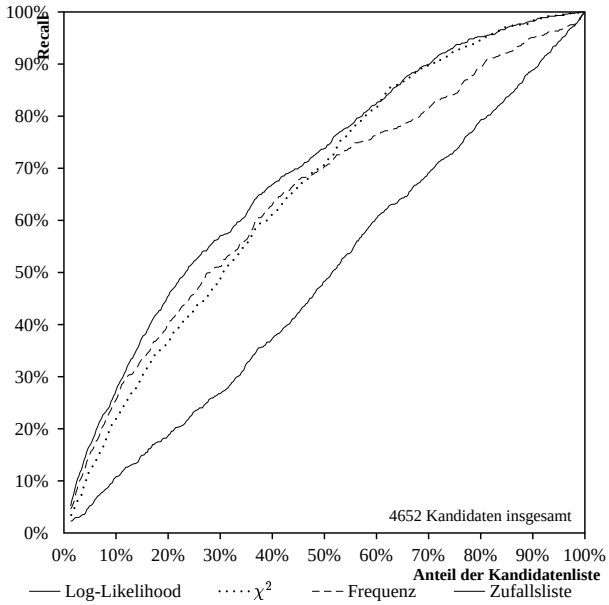


Abbildung 4: Recall-Kurven

die Recall-Kurven für Frequenz, $\chi^2$ und Log-Likelihood.[18] Auch hier erzielt Log-Likelihood über den gesamten Wertebereich die besten Ergebnisse. Eine nähere Betrachtung der Abbildung zeigt jedoch, daß selbst im günstigsten Fall deutlich mehr als die Hälfte der Signifikanzliste gesichtet werden muß, um eine 80-prozentige Abdeckung der Referenzliste zu erreichen. Nach den Abbildungen 2 und 3 liegt in diesem Bereich die Precision bereits unter 20%, so daß die Anwendung statistischer Verfahren keine wesentlichen Vorteile mehr bietet. Interessant ist, daß das ansonsten erheblich schlechter abschneidende $\chi^2$-Maß gleichauf zu Log-Likelihood liegt und bessere Ergebnisse liefert als die Sortierung nach Frequenz. Ist hingegen nur eine 50-prozentige Abdeckung der Referenzliste verlangt, so zeigen sich deutliche Unterschiede zwischen den Signifikanzmaßen.

## 5.3 Aufteilung in Frequenzschichten

Log-Likelihood wird oft als ein Signifikanzmaß beschrieben, das sich besonders für seltene Ereignisse eignet. Wir haben in einem weiteren Experiment die Grundmenge in Kandidatenpaare mit niedriger Häufigkeit (weniger als 5 Vorkommen) und solche mit relativ hoher Häufigkeit (5 oder mehr Vorkommen) aufgeteilt.[19] **Abbildungen 5 und 6** zeigen, daß Log-Likelihood in beiden Fällen die besten Ergebnisse erzielt. Überraschend ist hingegen, daß $\chi^2$ und MI bei seltenen Paaren besser abschneiden als die Sortierung nach Frequenz und bereits nach 30% der Signifikanzlisten gleichauf mit Log-Likelihood liegen. Die wesentlich besseren Ergebnisse von Log-Likelihood und Frequenz für die gesamte Grundmenge erklären sich dadurch, daß Kandidatenpaare mit hoher Frequenz bevorzugt werden[20]: für diese ist das Grundniveau an TPs mit 23.44% mehr als doppelt so hoch wie bei den seltenen Kandidatenpaaren (9.43%, vgl. Abbildungen 5 und 6).

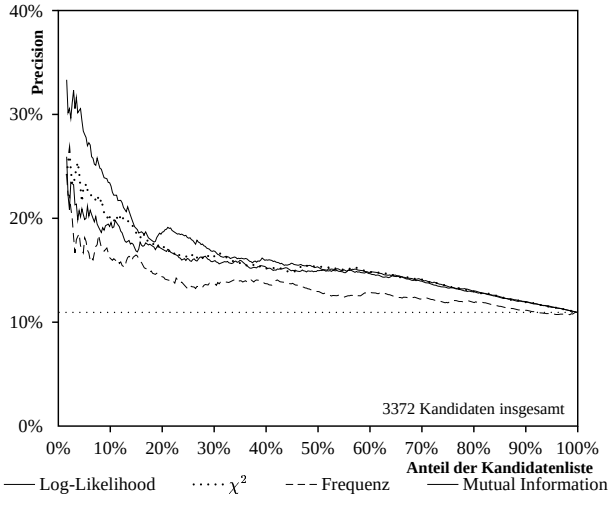


Abbildung 5: Precision-Kurven für seltene Paare

[18] Im Vergleich hierzu ist die Recall-Kurve der zufällig sortierten Liste dargestellt, die entlang der Bilddiagonalen verläuft.

[19] Der Grenzwert $f = 5$ ist willkürlich gewählt. Die 1280 Adjektiv+Nomen-Paare mit hoher Häufigkeit machen kaum mehr als ein Viertel der Grundmenge aus, so daß es nicht sinnvoll wäre, den Grenzwert höher anzusetzen.

[20] Bei der Signifikanzliste für Log-Likelihood finden sich unter den ersten 1000 Einträgen nur 172 Paare mit weniger als 5 Vorkommen.

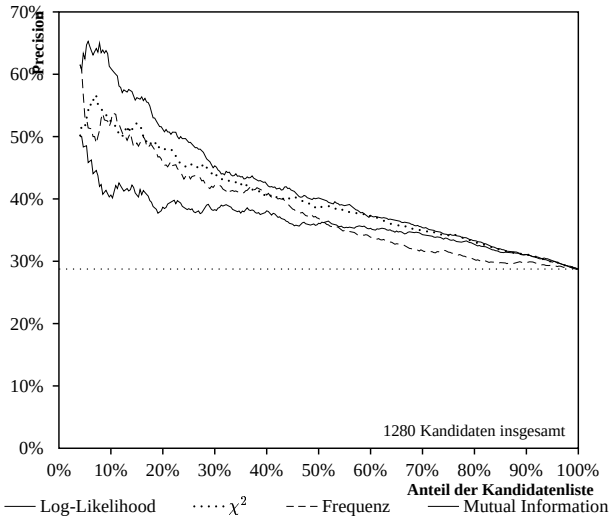


Abbildung 6: Precision-Kurven für häufige Paare

# 6 Zusammenfassung und Ausblick

Unser Experiment bestätigt die hohe Güte des Log-Likelihood-Maßes im Vergleich etwa zu MI oder $\chi^2$. Eine Gruppe weiterer Maße (Common Birthday, Mutual Expectation und Poisson) liefert fast identische Ergebnisse zu Log-Likelihood, dicht gefolgt von Entropie und t-Score. Überraschend ist das gute Abschneiden der rein nach Häufigkeit sortierten Kandidatenliste. Sie wirft insbesondere bei ausschließlicher Betrachtung der 400 am höchsten bewerteten Paare die Frage auf, ob die Verwendung statistischer Verfahren überhaupt Vorteile mit sich bringt (vgl. [11]). Anhand unserer Precision- und Recall-Kurven kann diese Frage generell mit einem klaren Ja beantwortet werden, jedoch nicht für die immer noch verbreiteten MI- und $\chi^2$-Maße.

Aus den Schaubildern wurde ersichtlich, daß in unseren Experimenten die besten bekannten Signifikanzmaße für die 30% am höchsten bewerteten Kandidaten wesentliche Vorteile gegenüber ungewichteten Kandidatenlisten bieten und dabei mehr als die Hälfte der manuell erstellten Referenzmenge abdecken.

Wir schlagen vor, anstatt der bisher üblichen Precision-Werte für $n$-Besten-Listen (mit willkürlich gewähltem $n$) kombinierte Precision/Recall-Diagramme zur vergleichenden Beurteilung von Signifikanzmaßen heranzuziehen. Mit ihrer Hilfe läßt sich insbesondere je nach Ausgangssituation[21] zwischen Precision und Recall abwägen.

[21] Gewünschte Qualität der Signifikanzlisten sowie zu erzielende Abdeckung.

# Literatur


[1] Maria Teresa Cabré, Rosa Estopà: “On the units of specialized meaning used in professional communication”, ms., 10pp., Barcelona: Universitat Pompeu Fabra, IULA, 1999.

[2] Gael Dias et al.: “Benefitting from multidomain corpora for extracting terminologically relevant multiword lexical units,” in: *Proceedings of the* EURALEX *International Congress 2000*, Stuttgart, 2000, to appear.

[3] Ted Dunning: “Accurate Methods for the Statistics of Surprise and Coincidence,” *Computational Linguistics*, **19**/1, 61–74.

[4] Franz Josef Hausmann: “Le dictionnaire de collocations”, in: Hausmann et al. (Hrsg.): *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch*, Berlin: de Gruyter, 1989, 1010–1019.

[5] Adam Kilgariff: “Which words are particularly characteristic of a text? A survey of statistical approaches,” in: *Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex University, 1996, 33–40.

[6] Esther König, Oliver Christ, Bruno M. Schulze, Anja Hofmann: *CQP User’s Manual*, Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1999.
`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench`

[7] Brigitte Krenn: “Empirical implications on association measures,” in: *Proceedings of the* EURALEX *International Congress 2000*, Stuttgart, 2000, to appear.

[8] Brigitte Krenn: “Collocation mining: Exploiting corpora for collocation identification and representation,” in: *Akten der KONVENS 2000 (Ilmenau)*, VDE-Verlag, 2000.

[9] Lucie Langlois, Pierre Plamondon: “Le repérage automatique de collocations équivalentes à partir de bitextes”, in: *Proceedings of the* EURALEX *International Congress 1998*, Liège, 1998.

[10] Martin Läuter, Uwe Quasthoff: “Kollokationen und semantisches Clustering,” in: J. Gippert, P. Olivier (Hrsg.): *Multilinguale Corpora – Codierung, Strukturierung, Analyse – 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung*, Prag: Enigma Corporation, 1999.

[11] Wolfgang Lezius: “Automatische Extrahierung idiomatischer Bigramme aus Textkorpora,” in: *Tagungsband des 34. Linguistischen Kolloquiums*, Germersheim, 1999.

[12] Marie-Claude L’Homme: “Charactérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale”, in: *Proceedings of the* EURALEX *International Congress 1998*, Liège, 1998.

[13] Christopher D. Manning, Hinrich Schütze: *Foundations of Statistical Natural Language Processing*, Cambridge: MIT Press, 1999.

[14] John Sinclair: “Corpus, Concordance, Collocation,” in: *Describing English Language*, Oxford: OUP, 1991.