Zusammenfassung

Im Projekt »Das sächsisch-magdeburgische Recht als kulturelles Bindeglied zwischen den Rechtsordnungen Ost- und Mitteleuropas«[1] werden Rechtstexte eines breiten sprachlichen, zeitlichen und geographischen Gebietes unter anderem auch linguistisch untersucht. Um in der Menge der Daten relevante Untersuchungseinheiten zu finden, wurden statistische Untersuchungen von n-grams ausgewertet. Doch die damit einhergehende Begrenzung der Daten kann aufgrund von Informationsverlust zu einer Unschärfe in der Datenanalyse führen. Der folgende Text soll – basierend auf einem Vortrag auf der ITUG-Jahrestagung 2016 – die Herangehensweise, die Umsetzung und die Probleme beschreiben.

n-grams

Die Arbeit mit n-grams bietet eine statistische Analysemöglichkeit, um von den einzelnen Teilen einer Gesamtheit an Daten Aussagen über diese Daten machen zu können bzw. um Indizien für eine genauere Betrachtung potenziell linguistisch bedeutsamer Teile zu finden. Die Aussagemöglichkeiten hängen stark von der Mächtigkeit der Daten (dem Umfang an Datenmaterial), möglichen Vergleichsobjekten sowie dem Untersuchungsschwerpunkt ab.

Für die Untersuchung werden alle Einzelteile des Datenmaterials (Mächtigkeit N) auf atomarer Ebene ($n=1$) bzw. in Gruppen von n aufeinanderfolgenden Teilen ($2 \leq n < N$) separiert.

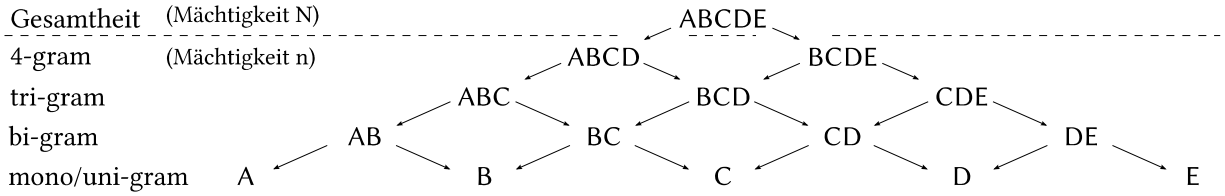

Abbildung 1: vereinfachtes Modell der Hierarchie von n-grams

Die separierten Sequenzen (n-grams) lassen sich nun mittels Häufigkeitsuntersuchung und durch die Erstellung verschiedener Scorewerte (Vergleich berechneter Wahrscheinlichkeiten und tatsächlicher Häufigkeiten) auf Auffälligkeiten hin analysieren. Das Analyseergebnis – eine Art »digitaler Fingerabdruck« – liefert Anhaltspunkte für weitere Untersuchungen.

Beispiele

Ein einfaches Beispiel für die Anwendung von n-gram-Analysen findet sich in der Kryptographie. Verschlüsselungen, welche die Struktur der zu verschlüsselnden Daten nicht verändern, sondern nur die einzelne Teile gleichförmig umwandeln (z. B. Caesar-Verschlüsselung) können mittels Häufigkeitsanalyse der atomaren Teile und Vergleich mit Referenzdaten entschlüsselt werden – wobei die Qualität des Ergebnisses stark von der Mächtigkeit der Gesamtheit abhängig ist. Eine gute Kryptographie versteckt/verändert daher auch die Struktur der Daten. Dies würde zu einem »Fingerabdruck« führen der entweder als »Weißes Rauschen« überhaupt keine statistischen Auffälligkeiten zeigen würde oder eine Struktur, die komplett vom eigentlichen Inhalt losgelöst ist.

In der Sprachstatistik kann anhand eines Vergleichs des »digitalen Fingerabdrucks« eines Textkorpus mit Referenzkorpora auf Textsorte und Textsemantik geschlossen werden. Allgemein lässt eine hohe Frequenz (Häufigkeit) großer Sequenzen (n-grams) z. B. auf einen stark formalisierten Text schließen. Ein Abgleich der Frequenz von Einzeltokens (uni-grams) oder Sequenzen von Tokens (n-grams, $n>1$) mit den Frequenzen in Referenzkorpora lässt Rückschlüsse auf den Inhalt

1 https://www.magdeburger-recht.de

zu. Auch hier ist die Mächtigkeit, sowohl der zu untersuchenden Daten als auch der Referenzdaten, ausschlaggebend für die Qualität der Untersuchungsergebnisse.

Extraktion von n-grams

Für die Extraktion von n-grams aus einer Gesamtheit von Daten lassen sich prinzipiell zwei Wege gehen[2]. 1) Der top-down-Ansatz geht von der Gesamtheit der Daten aus und bildet jeweils zwei Gruppen der Mächtigkeit $n = N - 1$ indem der erste bzw. letzte atomare Teil[3] aus der Gesamtheit entfernt wird – das Prinzip ist in Abbildung 1 skizziert. Jede dieser entstehenden Sequenzen wird wiederum auf dem gleichen Wege weiter reduziert, bis in der letzten Ebene nur noch atomare Teile übrig sind. Ob dieser Ansatz performant ist, also in einer vernünftigen Korrelation zwischen Ergebnis und Rechenaufwand steht, hängt stark von der Struktur der Daten ab. Eine starke Formalisierung sich wiederholender Einheiten mag hier zu einer statistischen Häufung von Sequenzen großer Mächtigkeit führen. Im Allgemeinen werden Auffälligkeiten aber eher in n-grams geringerer Mächtigkeit zu finden sein.

2) Entsprechend ist ein bottom-up-Ansatz eher zielführend. Hierfür wird die Gesamtheit der Daten in ihre atomaren Bestandteile zerlegt. Aus diesen wiederum werden Sequenzen der gewünschten Mächtigkeit n zusammengestellt. Die Mächtigkeit n der Sequenzen ist vor allem von der Struktur der Daten abhängig und kann stark variieren.

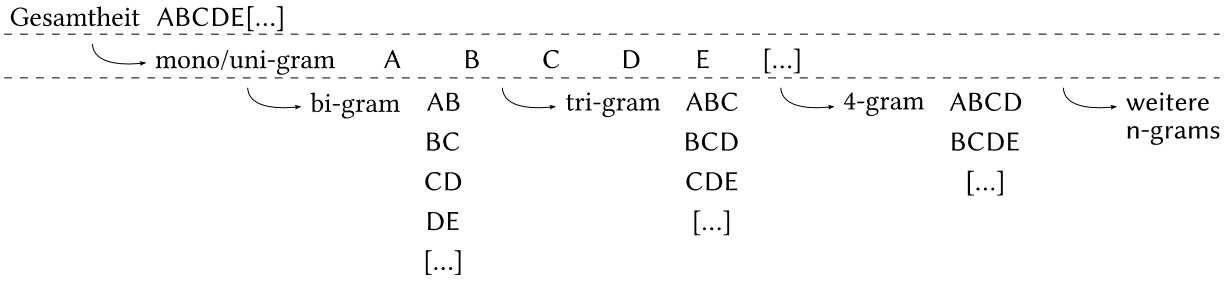

Abbildung 2: bottom-up-Ansatz zur Erstellung von n-grams

Die Ergebnisse dieser n-gram-Extraktion können als Auflistung sowohl aller Sequenzen einer bestimmten Mächtigkeit (Abb. 3, Fig. α) als auch der anhand des Starttokens der Sequenz geordneten Sequenzen (Fig. β) betrachtet werden.

α)

Gesamtheit (Mächtigkeit N)	ABCDEFG
4-gram_1	ABCD
4-gram_2	BCDE
4-gram_3	CDEF
4-gram_4	DEFG

β)

Gesamtheit (Mächtigkeit N)	ABCDEFG
6-gram	ABCDEF
5-gram	ABCDE
4-gram	ABCD
3-gram	ABC

Abbildung 3: α) Auflistung der Sequenzen einer bestimmten Mächtigkeit ($n = 4$),
β) Ordnung anhand des Starttokens der Sequenz (A, $3 \leq n \leq 6$)

Frequenz von n-grams

Neben der Mächtigkeit n der extrahierten Sequenzen interessiert uns vor allem die Frequenz der einzelnen Sequenzen. Statt der großen Menge von unauffälligen Sequenzen, welche nur ein

2 Den Weg jeweils Gruppen von Daten mit einer festgelegten Mächtigkeit von n aus dem Datensatz einzeln »auszuschneiden« möchte ich an dieser Stelle aufgrund schlechter Performance ausblenden. Dieser Weg ist nur zielführend, wenn Gruppen einer bestimmten, möglichst großen Mächtigkeit n gesucht sind.

3 Was jeweils »atomarer«, also kleinster Teil ist, muss abhängig von Untersuchungsgegenstand und -frage festgelegt werden. In Texten kann dies der einzelne Buchstabe oder auch ein Token sein.

einzelnes Vorkommen, also eine Frequenz $f = 1$ haben, wird bei der Analyse eine minimal zu untersuchende Frequenz festgelegt. Zu beachten ist, dass durch den Informationsverlust (Einschränkung von f und n der sequenzierten Daten) auch Unsicherheiten auftreten können. Dies muss bei der abschließenden Auswertung beachtet werden.

n-grams als Indiz in korpuslinguistischen Untersuchungen

Für unsere Untersuchungen mittelalterlicher Rechtstexte setzen wir die Analyse von n-grams zur Identifizierung der sich wiederholenden und dadurch für diese Texte signifikanten Sequenzen ein. Diese bieten Anhaltspunkte, um die Entwicklung einer formalisierten Rechtssprache und die Beeinflussung zwischen in Kontakt stehenden Sprachen innerhalb eines sprachlichen Registers zu untersuchen. Die reine Betrachtung separierter n-grams scheint jedoch für die Analyse nur bedingt geeignet. Als aufschlussreicher erwies sich die Erstellung paradigmatischer Beziehungen zwischen den n-grams.

Werkzeuge

Anfangs haben wir für unsere Untersuchung auf bestehende Werkzeuge zurückgegriffen.[4] Dies hat sich aber aufgrund folgender Probleme als nicht praktikabel erwiesen:

- Die Daten werden als Ganzes interpretiert. Eine Vorseparierung der Daten in kleinste zu betrachtende Einheiten muss außerhalb der Werkzeuge erfolgen. Dies hat wiederum zur Folge, dass die n-grams jeweils nur für den einzelnen Textabschnitt eruiert werden und im Nachgang erneut zusammengefasst werden müssen.
- Die vorgefundenen Werkzeuge können nur bedingt mit umfangreicheren Zeichensystemen umgehen. Da wir mit mehrsprachigen Daten arbeiten, muss dies aber gewährleistet sein. Bei einem Werkzeug werden Ergebnisse auch falsch kodiert wieder ausgegeben und müssen nachträglich korrigiert werden.
- Die Berechnung der ermittelten Scorewerte ist nur bedingt nachvollziehbar, da die zugrunde liegenden Formeln für die Berechnung oft unzureichend dokumentiert werden.
- Der Arbeitsablauf kann aufgrund fehlender Schnittstellen nicht automatisiert werden. Die Daten müssen jeweils von Hand in die Werkzeuge kopiert bzw. in diesen als Datei ausgewählt werden.

Hinzu kommt, dass wir uns gegen eine reine Betrachtung aufgelisteter Sequenzen einer gemeinsamen Mächtigkeit n und für die Betrachtung der paradigmatischen Beziehungen zwischen den n-grams entschieden haben und sich dies mit nach eigenen Vorstellungen entwickelten Werkzeugen besser realisieren lässt.

n-gram-Separierung mittels TUSCRIPT

Unsere Daten werden mittels TUSTEP[5] verwaltet. Somit lag es nahe auch die Separierung der n-grams mittels TUSCRIPT – einer Programmiersprache innerhalb von TUSTEP – umzusetzen. Das für diesen Zweck erstellte Makro NGRAM_MAK.TF ist in drei Segmente unterteilt.

Das Segment A_NGRAM erwartet beim Aufruf den Namen der zu verarbeitenden Datei. Die Minimum- und Maximumwerte für n (n_{min}/n_{max}) sind mit 2 und 12 vorbelegt, können beim Aufruf aber auch abgeändert werden.[6] Gleiches gilt für die Minimalfrequenz f_{min}, welche im Standard auf 2

4 Omer 2006, Greaves & Warren 2010 und Kopaczyk 2013.

5 http://tustep.wikispaces.com/TUSTEP

6 Eine Einschränkung von $2 \leq n \leq 12$ hat sich für unsere Korpora bewährt. Mächtigere n-grams ($n = 21$) mit $f_{min} = 2$ haben sich nur in einem Subkorpus gefunden. Dabei handelt es sich um eine Auflistung von Namen.

eingestellt ist. Weitere Optionen beeinflussen die Ausgabe zusätzlicher Informationen – so lassen sich z. B. Listen mit n-grams je Untersuchungseinheit ausgeben oder die Behandlung von Groß- und Kleinschreibung beeinflussen. Im ersten Teil des Makros wird der Text in einzelne Untersuchungseinheiten separiert. Bei Texten ist dies normalerweise der Satz als syntaktische Einheit. Da eine Festlegung der Satzgrenzen in mittelalterlichen Texten nicht immer eindeutig möglich ist und der Text insgesamt nach anderen als den gegenwärtigen Regeln segmentiert wurde, haben wir als kleinste Einheit den Paragraphen gewählt. Je nach vorheriger Annotation des Textes kann das Makro an dieser Stelle aber auch leicht an andere Wünsche angepasst werden (Tags, Satzzeichen etc.).

Nach der Separierung werden die Daten der einzelnen Untersuchungseinheiten an das Segment FKT_NGRAM übergeben. Dort werden Satzzeichen, Tags etc. eliminiert und die Daten unter Gleichbehandlung von Groß- und Kleinschreibung in atomare Teile (Tokens) separiert. Diese werden im Anschluss zu n-grams entsprechend der Vorgaben zusammengesetzt und wieder an A_NGRAM zurückgegeben.

Wenn alle Untersuchungseinheiten diesen Schritt durchlaufen haben, werden im letzten Segment FKT_FREQUENZ die Häufigkeitsanalysen vorgenommen und die n-grams als Listen im Format FT oder TXT zur weiteren Bearbeitung ausgegeben.

Beziehungseruierung zwischen n-grams mittels Shell-Script

Das zweite Werkzeug wurde anfangs nur zu Testzwecken als Shell-Skript in bash erstellt, um das Prinzip zu verdeutlichen.[7] Dieses Script kann innerhalb von TUSTEP aufgerufen werden und untersucht die einzelnen n-grams auf Beziehungen untereinander. Die Ergebnisse werden tabellarisch als HTML-Datei sowie als Plaintext (TXT) ausgegeben (Abb. 4). In der HTML-Datei sind zusätzliche Funktionalitäten enthalten, welche das Aus- und Einblenden einzelner Arten von Beziehungen, Gruppen von n-grams und Einträgen bestimmter Frequenz erlauben.

```
4-16  3  TM-X                    přijdu oba před pravo
      2  EK       [7-2]          ranita a přijdu oba před pravo a
      3  TMEX 1   [6-0]          a přijdu oba před pravo a
      2  TM       [6-5]          ranita a přijdu oba před pravo
      3  TM-X     [5-3]          přijdu oba před pravo a
      3  TM-X     [5-6]          a přijdu oba před pravo
      3  TM-X           [3-132]  přijdu oba před
      3  TM-X           [3-137]  oba před pravo
     46  TMEX 12        [2-8]    před pravo
      4  TMEX 1         [2-436]  oba před
      3  TM-X           [2-697]  přijdu oba
```

Abbildung 4: Beispiel für die Ausgabe der ermittelten Beziehungen zwischen den n-grams; Spalten: Nummer des n-grams zur Referenzierung, ermittelte Frequenz, Beziehung (s. unten), eigene Vorkommen bei TMEX (s. unten), Referenznummer bei n-grams größerer Mächtigkeit, ~ kleinerer Mächtigkeit, Sequenz des n-grams

Innerhalb des Scripts werden die Listen der n-grams eingelesen und anschließend die einzelnen n-grams mit n-grams höherer Mächtigkeit verglichen. Dadurch können folgende Beziehungen aufgezeigt werden:

- **Endknoten (EK):** Sequenzen, welche im Datensatz keine Teilmengen größerer n-grams sind,
- **»reine« Teilmengen (TM):** Sequenzen, welche in einem größeren n-gram (EK) als Sequenz enthalten sind und dieselbe Frequenz aufweisen (s. Abb. 5),

7 Dies ist keineswegs die Programmiersprache der Wahl um große Textmanipulationen und Analysen vorzunehmen. Daher soll das Skript in einer zukünftigen Version in TUSCRIPT umgesetzt werden.

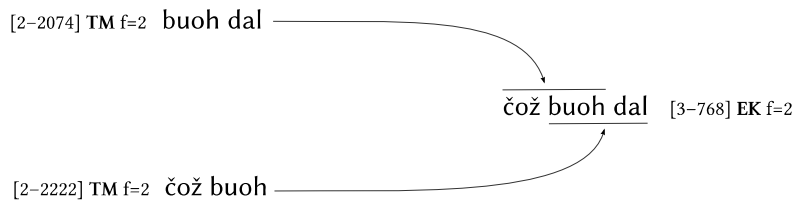

Abbildung 5: Beziehung zwischen Endknoten (EK) und »reinen« Teilmengen (TM) [Schreiber A]

- **»reine« multiple Teilmengen (TM-X):** Sequenzen, welche in mehreren größeren n-grams (EK) enthalten sind und deren Häufigkeit mit der Summe der Häufigkeiten der EK übereinstimmt,

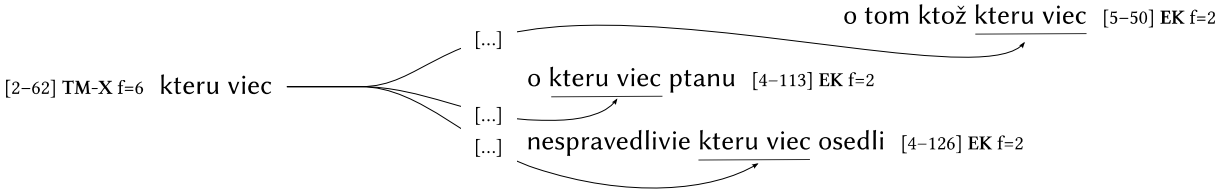

Abbildung 6: eine »reine« Teilmenge mehrerer Endknoten (TM-X) [Schreiber B]

- **Teilmengen, welche gleichzeitig Endknoten sind (TMEX):** Sequenzen, welche sowohl Teilmenge eines oder mehrerer größerer n-grams sind, aber auch eine höhere Frequenz haben und somit auf Fundstellen hinweisen, welche durch keine größeren n-grams aufgezeigt werden. Die Anzahl dieser Fundstellen wird ausgegeben.

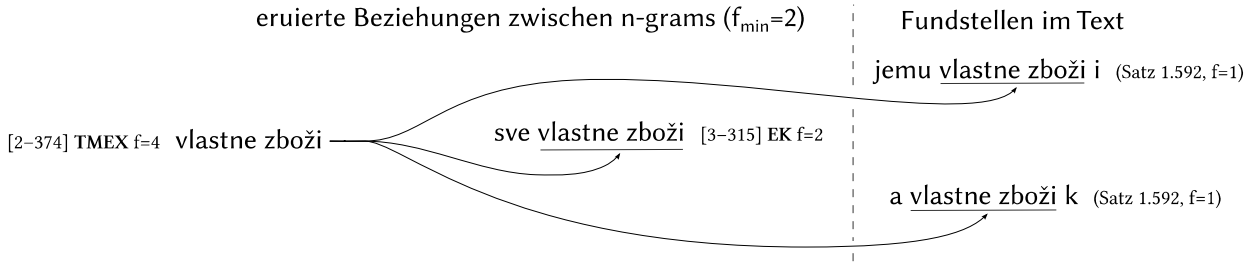

Abbildung 7: eine Sequenz ist in zwei Fundstellen durch einen Endknoten (EK, f=2) belegt und hat zusätzlich zwei Fundstellen, die aufgrund geringer Frequenz (f=1) nicht als höheres n-gram belegt sind [Schreiber A]

Probleme bei der Eruierung der Beziehungen

Bei der Separierung der n-grams wird durch einzelne Vektoren (n_{min}/n_{max}, f) Einfluss auf die auszuwertenden Daten genommen. Dies führt zu einer Steigerung der Performance, da nur Daten ausgegeben und weiterverarbeitet werden, die innerhalb eines gegebenen Rahmens liegen.[8] Dieser gewollte Ausschluss von Informationen kann zu Unsicherheiten in der Eruierung der Beziehungen zwischen den n-grams führen. Dieser Fall ist gegeben, wenn zwei Endknoten sich eine Fundstelle teilen, diese aber aufgrund eines einmaligen Vorkommens nicht mehr in den selektierten Daten vorhanden ist. Dadurch kann es in Sequenzen, welche in beiden Endknoten enthalten sind, zu Ungenauigkeiten in der Berechnung der eigenen Vorkommen als Endknoten (TMEX) kommen. Das Diagramm in Abbildung 8 erläutert dieses Problem.

8 In einem Datensatz sind z. B. 21 122 bi-grams separiert worden. Nach Normalisierung mehrfacher Vorkommen und Extrahierung von Einzelvorkommen sind noch 2250 bi-grams gelistet. Deutlicher wird dies mit größeren Sequenzen. So werden 18 312 12-grams extrahiert. Doch nur eine Sequenz weist eine Frequenz f>1 auf, ist also mehrfach (f=2) vorhanden.

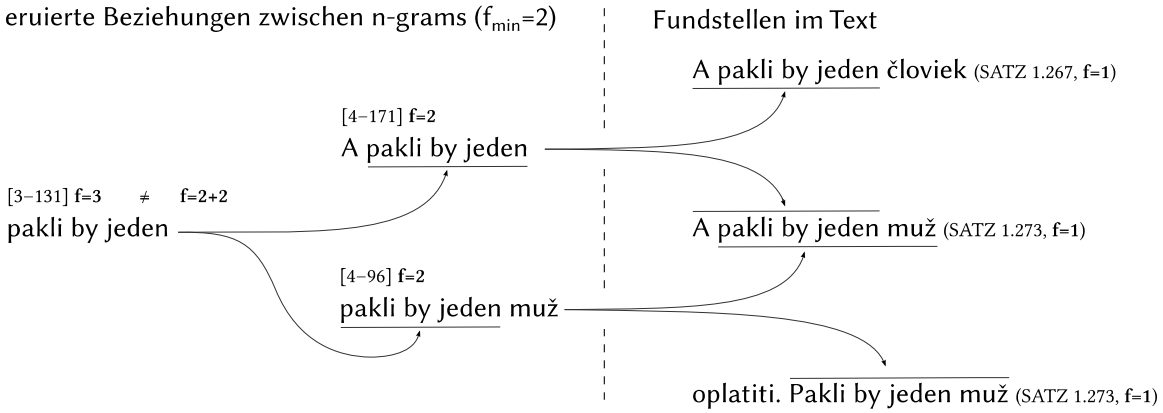

Abbildung 8: fehlende Information führt zu Ungenauigkeiten in den eruierten Beziehungen [Schreiber A]

Um diese Unschärfe im Datenmaterial wieder zu kompensieren wird ein performanter Ansatz benötigt. Anfällig für unser Problem sind all die n-grams, welche Teilmengen von mindestens zwei Endknoten sind und dort den Anfang bzw. das Ende der Sequenz bilden (Bsp. in Abb. 8: ›a pakli by jeden‹ und ›pakli by jeden muž‹). Aus beiden EK zusammen kann nun eine Sequenz gebildet werden, bei der unser ursprüngliches n-gram (›pakli by jeden‹) die verbindende Folge von Token in der Mitte der Sequenz bildet (›a pakli by jeden muž‹). Diese Sequenz muss nun auf ihr Vorkommen im Ursprungstext hin untersucht werden.[9]

Je kleiner die Sequenz ist und je größer die Häufigkeit der Sequenz sowie die Anzahl vorhandener Endknoten in Bezug zu dieser Sequenz sind, um so größer kann die mögliche Abweichung der Daten sein. Ob diese Aussage generalisierbar ist, müssen weitere Untersuchungen zeigen.

Ausblick

Als nächster Schritt soll der bisher in bash umgesetzte Algorithmus zur Eruierung der paradigmatischen Beziehungen zwischen den n-grams ebenfalls in TUSCRIPT umgesetzt werden. Dies ermöglicht gleichzeitig den einfachen Abgleich problematischer Fundstellen mit dem Text (s. vorhergehender Abschnitt). Weitere Wünsche sind die Einbindung in eine webbasierte Anwendung um den Zugang zu erleichtern sowie Werkzeuge zur Visualisierung von Ergebnissen. Wünschenswert sind folgende Darstellungen und weiterführende Analysen:

- Netzwerkmodell der Beziehungen,
- Anzeige der n-grams im Kontext der Ursprungskorpora,
- Ermittlung verschiedener Scorewerte,
- graphische Darstellung statistischer Daten (Vergleich).

Schwierigkeiten bietet noch die fehlende Möglichkeit für kompliziertere mathematische Funktionen (Berechnung von Logarithmen) in TUSTEP/TUSCRIPT die für die Ermittlung verschiedener Scorewerte benötigt wird und z. Zt. noch den Rückgriff auf externe Werkzeuge nötig macht.

9 Ein mögliches Vorkommen kann sich wortwörtlich nur in *einer* Fundstelle zeigen. Bei zwei Vorkommen wäre die Sequenz ja ansonsten Teil unserer extrahierten n-grams und die mangels fehlender Informationen generierte Unschärfe nicht vorhanden. Momentan erfolgt diese Kontrolle noch händisch. Eine Implementierung dieser Überprüfung soll in der nächsten Version des Scriptes erfolgen.