verbundprojekt postdemokratie
und neoliberalismus

Matthias Lemke

Frequenzanalyse und Diktionäransatz

1/5 – Serie „Atomenergiediskurs“

Reihe: ePol Text Mining Verfahren (eTMV)

Was ist eTMV?

Die ePol Text Mining Verfahren (eTMV) stellen in kurzen, in sich abgeschlossenen Einheiten die verschiedenen Analyseverfahren des Leipzig Corpus Miners (LCM) vor (Wiedemann/Niekler 2014). Zur Illustration der Verfahren verwenden alle fünf Module dieser Serie Beispiele aus dem Kontext des Diskurses um die Nutzung der Atomenergie in Deutschland. Ziel der eTMV ist es, für interessierte SozialwissenschaftlerInnen eine best practise der Verwendungsmöglichkeiten und -grenzen von Text Mining anhand des LCM zu veranschaulichen – auch wenn bislang keine Vorerfahrungen mit Text Mining bestehen.

Der Datenbestand des ePol-Projekts besteht aus 3.495.822 deutschsprachigen Zeitungsartikeln, die sich wie folgt zusammensetzen: Frankfurter Allgemeine Zeitung (Stichprobe, 1959–2011, 200.389 Artikel), Süddeutsche Zeitung (6027 Ausgaben, 1992–2011, 1.505.714 Artikel), die tageszeitung (7821 Ausgaben, 1986–2012, 1.391.981 Artikel) und Die Zeit (3841 Ausgaben, 1946–2012, 397.729 Artikel). Für die Funktionalität aller Analysen einschlägig ist Wiedemann, Gregor / Niekler, Andreas (2014): Analyse qualitativer Daten mit dem Leipzig Corpus Miner, Hamburg / Leipzig (=Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus Discussion Paper 5), online unter: http://www.epol-projekt.de/discussion-paper/discussion-paper-5/.

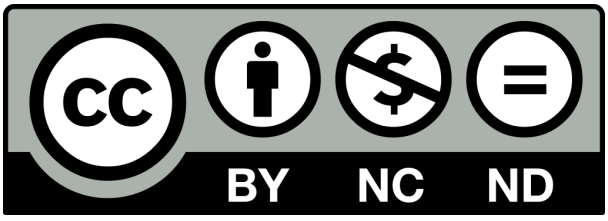
ISSN 2363-6335

Zitationsweise und Lizenz

Lemke, Matthias (2014): Frequenzanalyse und Diktionäransatz, Hamburg/Leipzig (=ePol Text Mining Verfahren, Serie „Atomenergiediskurs“, Modul 1/5).

Rückfragen bitte an die unter **www.epol-projekt.de/kontakt** angegebene Email-Adresse.

Alle Teile der Reihe sind unter **www.epol-projekt.de/etmv** online abrufbar. Das gesamte Material der Reihe ist unter **Creative Commons 4.0 international** als **BY-NC-ND** lizenziert.



Inhalt

1. Definition

„Frequenzanalysen zählen relative und absolute Häufigkeiten von Schlüsselbegriffen im Untersuchungskorpus, was erste Hinweise auf die Verbreitung von bestimmten sprachlichen Mustern geben kann. Über Diktionäre lassen sich mehrere Schlüsselworte zu semantischen Konzepten verdichten und als solche in ihrer Entwicklung im Untersuchungszeitraum beobachten.“ (Wiedemann/Lemke/Niekler 2013: 109).

2. Erkenntnishorizont

Im Sinne des *Blended-Reading-Ansatzes** (Lemke/Stulpe 2015) ist **Text Mining ein modularer Prozess**. Dieser setzt sich aus verschiedenen computergestützten Analyseverfahren und von der Forscherin / dem Forscher zu leistenden Interpretationen von Einzeltexten zusammen. Blended Reading nimmt – was die computergestützten Verfahren anbelangt – eine analytisch-prozedurale Gewichtung vor, die von einfachen, strukturierenden, hin zu komplexeren, inhaltlich tragfähigen Verfahren reicht. Frequenzanalyse und der auf ihr aufbauende Diktionäransatz sind basale Operationen des Text Mining, die am Anfang der Auswertung einer unstrukturierten Textdatenmenge stehen. Indem sie Wortverwendungen in ihrer Häufigkeit abhängig vom Zeitverlauf abbilden, dienen sie dazu, große Textdatenmengen entlang der binären Logik „Wort vorhanden / Wort nicht vorhanden“ zu strukturieren. Somit beinhaltet die Frequenzanalyse zwei strukturierende, im Wesentlichen deskriptive Leistungen:

(1) Sie ermöglicht auf zwei unterschiedliche Arten (Verlaufskurve und *Heatmap**) die Visualisierung des Auftretens eines Wortes im Zeitverlauf und kann – tagesgenau – anzeigen, wann ein Wort oder eine Mehrworteinheit im Korpus vorkommt.

(2) Damit sind erste Rückschlüsse auf die Verwendung eines Wortes oder einer Mehrworteinheit in der politischen Öffentlichkeit möglich: Treten in einem Zeitabschnitt Häufungen auf, so könnte der fragliche Begriff zum entsprechenden Zeitpunkt entweder besonders relevant oder auch besonders umstritten gewesen sein. Im zeitlichen Längsschnitt lassen sich Konjunkturen einer Debatte identifizieren. Es ergeben sich Einstiegspunkte für das Close Reading, die erst durch Text Mining Verfahren sichtbar werden.

Beide Analyseleistungen zusammen erleichtern die (zeitliche) Datenstrukturierung und ermöglichen weitere Analyseschritte.

3. Vorbereitung der Analyse

Zentral für die Vorbereitung der Analyse ist die Überlegung, worauf sich das Erkenntnisinteresse richtet. Dabei sind zwei Fälle zu unterscheiden: Es ist zu prüfen, ob die **Operationalisierung** des Erkenntnisinteresses (1) aus einem einzigen Suchbegriff bestehen kann, oder aber (2) mehrere, gegebenenfalls noch gar nicht bekannte oder noch zusammenzustellende Suchbegriffe enthalten muss.

(1) Wenn sich das Erkenntnisinteresse auf die **Verwendungskonjunktur eines Suchbegriffes** bezieht, der den in Frage stehenden Sachverhalt hinreichend präzise abbildet, dann ist eine einfache Suche nach diesem einen Suchbegriff ausreichend. Meistens ist das bei besonders spezifischen Begriffen der Fall. Um etwa die Verwendungskonjunktur des *Unigramms** „Endlager“ zu bestimmen, wird eine Suche nach genau diesem Begriff ebenso zielführend sein, wie beispielsweise die nach „Atomausstieg“ oder „Tschernobyl“.

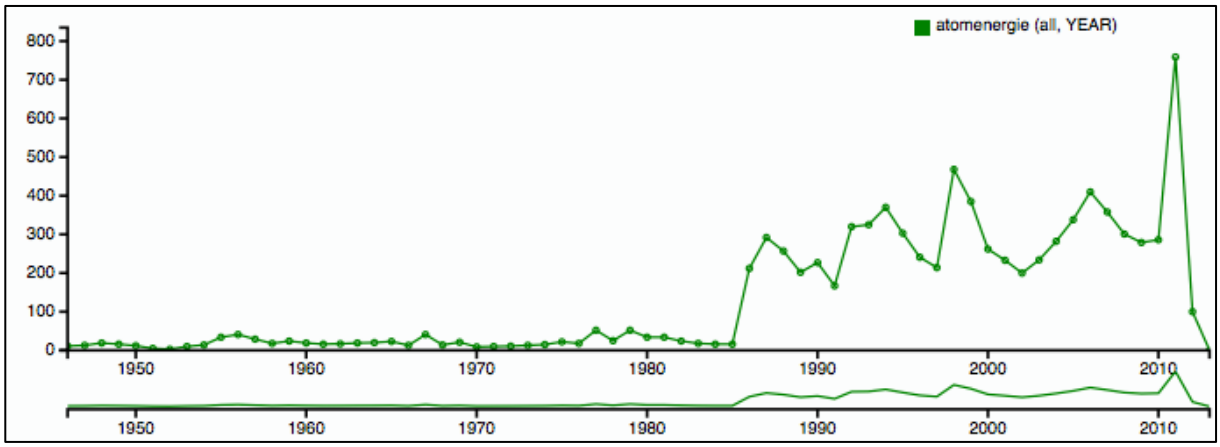
(2) Wenn sich das Erkenntnisinteresse auf einen **abstrakteren Begriff** bezieht, der als Suchanfrage nicht die Trefferdokumente generiert, die er inhaltlich beschreibt, dann ist die Erstellung einer konkreten Wortliste für die Suchanfrage erforderlich. Diese Wortliste muss den Begriff inhaltlich so füllen, dass sie bei einer Suchanfrage passende Trefferdokumente generiert. So ist es beispielsweise wenig hilfreich, angesichts eines Erkenntnisinteresses über die Ökonomisierung der Gegenwartsgesellschaften nach „Ökonomisierung“ zu suchen. Stattdessen können Einzelbegriffe, wie Staat, Markt, Marktwirtschaft, Kapital, Konsument, Geld, Geiz etc. geeignete Worte sein, um den Begriff „Ökonomisierung“ zu erschließen. Oder wenn es um die Frage nach der „Religiosität“ in Deutschland geht, sind neben der Einschränkung „deutsch“ bei der Suchanfrage Worte wie „Kirche“, „Austritt“, „evangelisch“, „katholisch“, „Bischof“, „Papst“ hilfreich. Die **Generierung von Wortlisten**, die abstrakte Begriffe füllen, kann auf zweierlei Weisen erfolgen: Induktiv, also unter Berücksichtigung des Datenmaterials und seiner algorithmischen Auswertung, etwa durch Topic Modelle; oder aber deduktiv, durch Verwendung von externen Quellen oder Synonymwörterbüchern.
Im Zusammenhang mit der Etablierung von Konzepten ist insbesondere auf die Gefahr des Vorliegens **polyvalenter Suchworte** zu achten. Diese liegen vor wenn die Elemente des zu

suchenden Konzepts einerseits Bestandteil der Alltagssprache sind, andererseits aber auch auf das zu suchende Konzept referieren. Etwa im Fall der sogenannten „TINA-Rhetorik“ (auch: „Alternativlosigkeitsrhetorik“): Ein für die Beschreibung des Konzepts geeignetes Wort wäre etwa das Adjektiv ‚alternativlos’. Analyse und Interpretation müssen in Rechnung stellen, dass das Wort sowohl im Kontext der TINA-Rhetorik Verwendung finden kann, als auch in beliebigen alltagssprachlichen Kontexten, die keine TINA-Rhetoriken repräsentieren. In diesem Fall ist, wie im Rahmen der Strategie des *Blended Reading** gefordert, die händische Überprüfung ausgewählter Einzeltexte erforderlich, um mit Blick auf den zu analysierenden Dokumentenbestand einschätzen zu können, ob die Kombination von Suchworten, also das *Diktionär**, tatsächlich auch die gewünschten Trefferdokumente generiert.

4. Frequenzanalyse und Diktionäransatz im Kontext des „Atomenergiediskurses“

Eine Frequenzanalyse ermöglicht im sinne des Blended Reading Ansatzes den Einstieg in die Analyse des Atomenergiediskurses. Sie kann etwa vom Suchwort „Atomenergie“[1] ausgehen. Ohne Einschränkung des Suchumfeldes[2] ergibt sich im Gesamtkorpus folgende absolute Häufigkeit im Zeitverlauf:

Abb. 4.1 – Frequenzanalyse „Atomenergie“ (n=8.840 Dokumente; absolut).

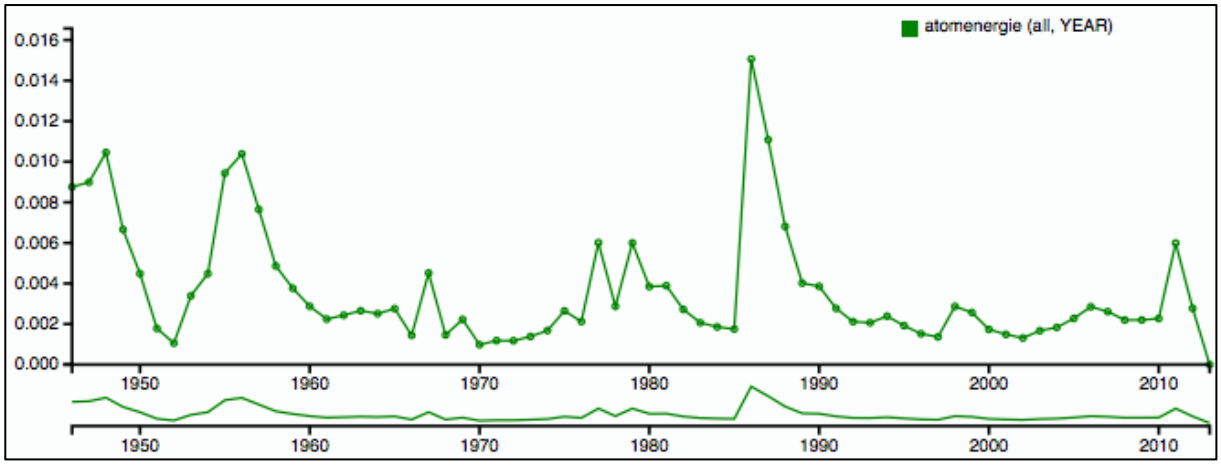


Abgebildet werden alle Artikel eines jeden Jahres des vom Gesamtkorpus abgedeckten Zeitraums, in denen das Suchwort auftritt. In standardisierter Häufigkeit, also unter grundsätzlicher Berücksichtigung der auf das Jahr gesehen ungleichen Verteilung der Artikel, sieht die Visualisierung der Trefferdokumente im Zeitverlauf anders aus:

[1] Mit Blick auf die Kontextualisierung der Suchbegriffe wäre zu berücksichtigen, dass der begriff „Kernkraft“ eher positiv konnotiert ist und von den Befürwortern der Kernenergie verwendet wird, während der Begriff „Atomkraft“ bzw. „Atomenergie“ von den Gegnern eingesetzt wird. Vgl. hierzu auch den Graphen unter Abb. 4.4 und 4.5.

[2] Denkbar sind hier insbesondere zeitliche Einschränkungen, die nur bestimmte Jahre – etwa um den GAU in Tschernobyl oder in Fukushima-Daichi – berücksichtigen.

Abb. 4.2 – Frequenzanalyse „Atomenergie“ (n=8.840 Dokumente; standardisiert).



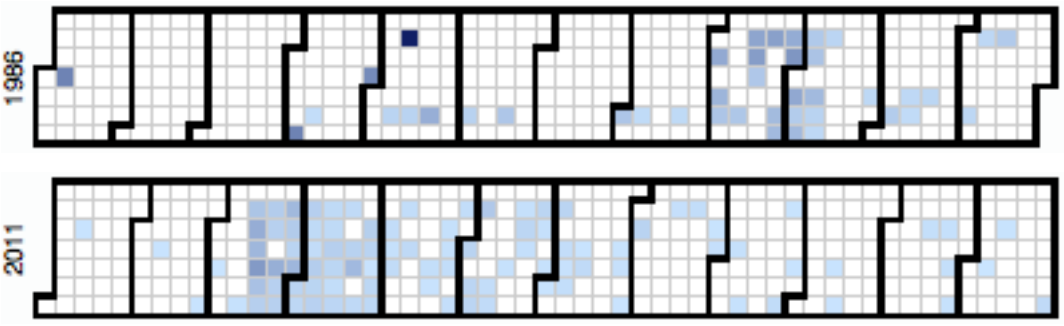
Bei der Betrachtung sowohl der absoluten[3] wie auch der standardisierten[4] Visualisierung der Wortfrequenz ist zu beachten, dass die im ePol-Korpus verfügbaren Dokumente über die Zeit nicht gleich verteilt sind. Während für den Zeitraum bis 1986 nur die Artikel aus der FAZ und der Zeit vorliegen (insgesamt ca. 600.000 Artikel), steigt ab 1986 mit dem Hinzukommen der Artikel aus der taz und dann ab 1992 mit den Artikeln aus der SZ die Gesamtsumme der verfügbaren Texte dramatisch an. Gerade die Übergänge an diesen Schlüsseljahren bedürfen daher einer **besonderen Sorgfalt bei der Interpretation der Daten**: Während vor 1986 pro Tag mitunter nur einige wenige Artikel vorliegen können, liegen ab 1992 in der Regel mehrere hundert Artikel pro Tag vor. Da für die Jahrgänge bis 1986 nur eine sehr geringe Anzahl der tatsächlich veröffentlichten Artikel überhaupt in die Auswertung einbezogen werden kann, ist ein durchgängiger Rückschluss in der Form „hoher Prozentwert = hochfrequentes Thema = wichtiges Thema“ nur bedingt möglich. Genau anders herum stellt sich der Sachverhalt bei den absoluten Zahlen dar: eine geringe Anzahl von Trefferdokumenten muss nicht zwangsläufig bedeuten, dass das vom Suchbegriff repräsentierte Thema in der fraglichen Zeit wenig relevant gewesen wäre – gegebenenfalls ist lediglich die Stichprobe zu klein. Hier werden die Erkenntnisgrenzen der Frequenzanalyse deutlich: Während in beiden Graphen die Jahre 1986 und 2011 – sowie auch das Jahr 1998 – aus ihrem Umfeld herausstechen und sich damit für eine weitere, vertiefte Analyse unter Hinzuziehung anderer Verfah-

[3] x-Achse: Zeit in Jahren; y-Achse: Anzahl der Dokumente.
[4] x-Achse: Zeit in Jahren; y-Achse: Anteil der Trefferdokumente an der Gesamtzahl der Dokumente pro Jahr.

ren anbieten, ist die inhaltliche Aussagekraft der Graphen jenseits dieser **grundlegenden Strukturierungsleistung** gering.

Neben der Abbildung als Frequenzgraph ist auch die Visualisierung in Form einer *Heatmap** möglich. Die beiden folgenden *Heatmaps** zweigen die Suchergebnisse zum *Unigramm** „Atomenergie“ für die Jahre 1986 und 2011, in denen sich die Reaktorkatastrophen in Tschernobyl und in Fukushima-Daichi ereigneten.

Abb. 4.3 – Frequenzanalyse „Atomenergie“ (standardisiert), als Heatmap 1986, 2011.[5]



Auch ohne Kenntnis der konkreten Daten lässt sich aus den Visualisierungen ersehen, dass der atomare Unfall (GAU) in Tschernobyl im Jahr 1986 (26.4.) und der in Fukushima-Daichi im Jahr 2011 (11.3.) stattgefunden haben muss.

Das Unigramm „Atomenergie“ könnte sich jedoch als nicht ausreichende Suchanfrage erweisen, wenn es um die Abbildung der Konjunkturen des Diskurses um die Verwendung atomarer Energie geht. So werde naheliegende und häufig gebräuchliche Synonyme[6] – wie etwa „Kernenergie“, „Kernkraft“, „Atomkraft“, „Atomstrom“ etc. – bei der ausschließlichen Verwendung des Unigramms „Atomenergie“ als Suchbegriff nicht aufgefunden. Dass damit möglicherweise relevante Texte nicht gefunden werden, wird deutlich, wenn man die relativen Frequenzen der beiden Unigramme „Atomenergie“ und „Kernenergie“ in einem Diagramm gemeinsam abbildet.

[5] Ein Kasten steht für ein Kalenderjahr, der wiederum in zwölf größere Kästchen (Monate, von Januar, links, nach Dezember, rechts) unterteilt ist. Diese wiederum sind in einzelne Quadrate unterteilt, von denen jedes einen Tag repräsentiert. Je dunkler die Kästchen, desto mehr Treffer liegen für die Suchanfrage am fraglichen Tag vor. Im LCM kann zudem durch Anklicken eines Quadrats ein Fenster geöffnet werden, das eine Liste der Überschriften der für diesen Tag vorliegenden Treffertexte enthält. Durch anklicken der Überschriften kann auf den Einzeltext zugegriffen werden.

[6] Zur Erschließung einschlägiger Synonyme vgl. die Datenbank des Portals Deutscher Wortschatz der Universität Leipzig unter http://wortschatz.uni-leipzig.de.

Abb. 4.4 – Frequenzanalyse „Atomenergie“ und „Kernenergie“ standardisiert (oben) und absolut (unten).

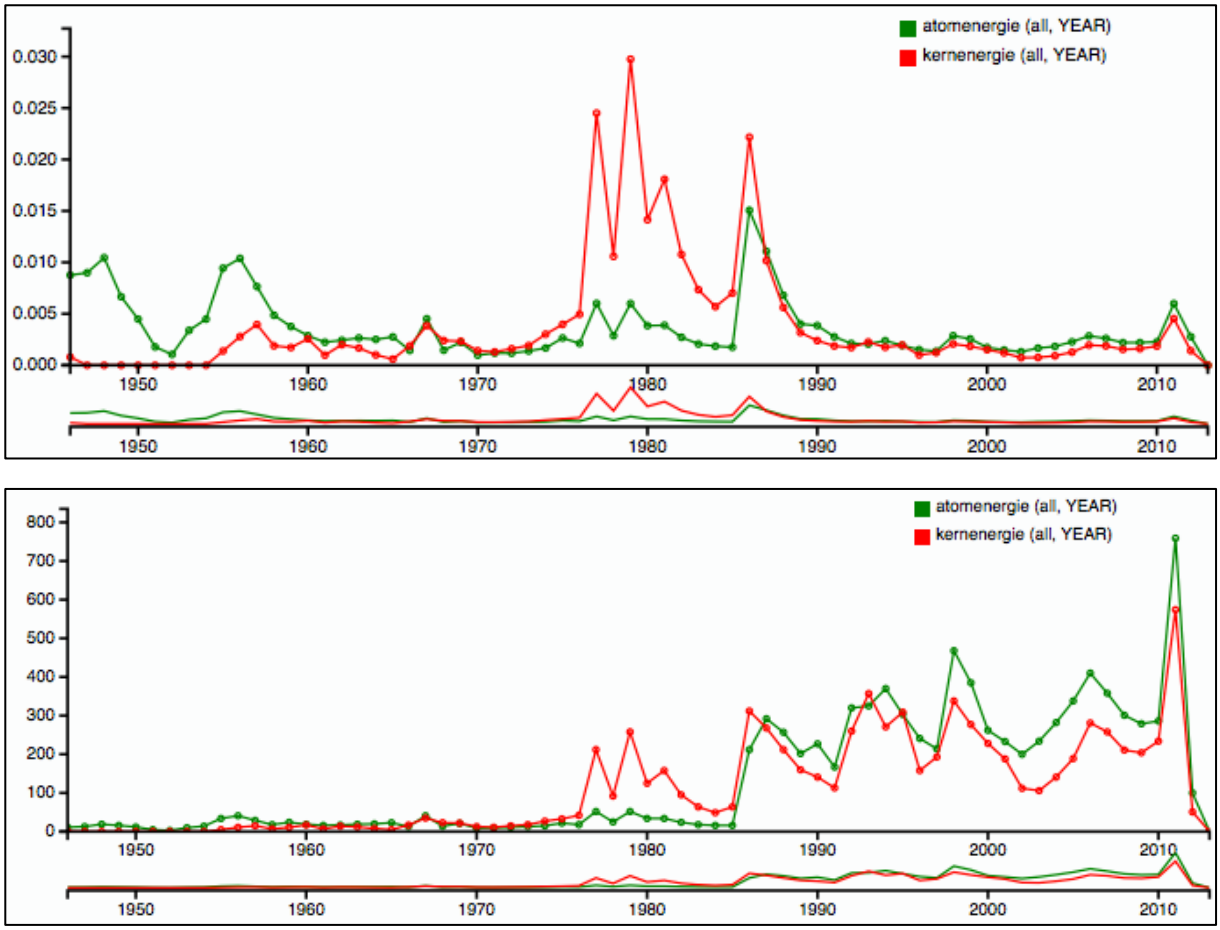
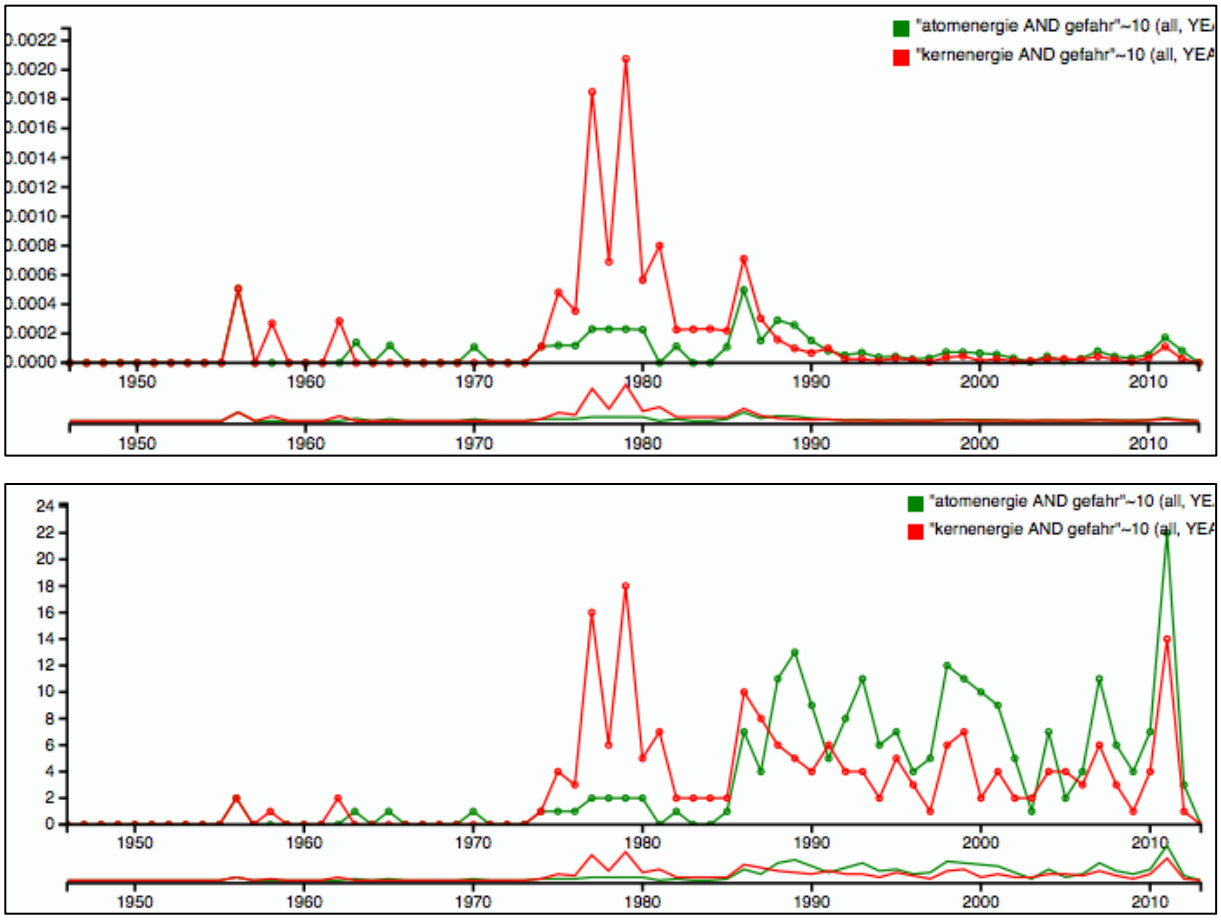


Abb. 4.5 – Frequenzanalyse „Atomenergie AND Gefahr“ und „Kernenergie AND Gefahr“ standardisiert (oben) und absolut (unten), mit Einschränkung des Suchumfeldes auf 10 Worte.[7]

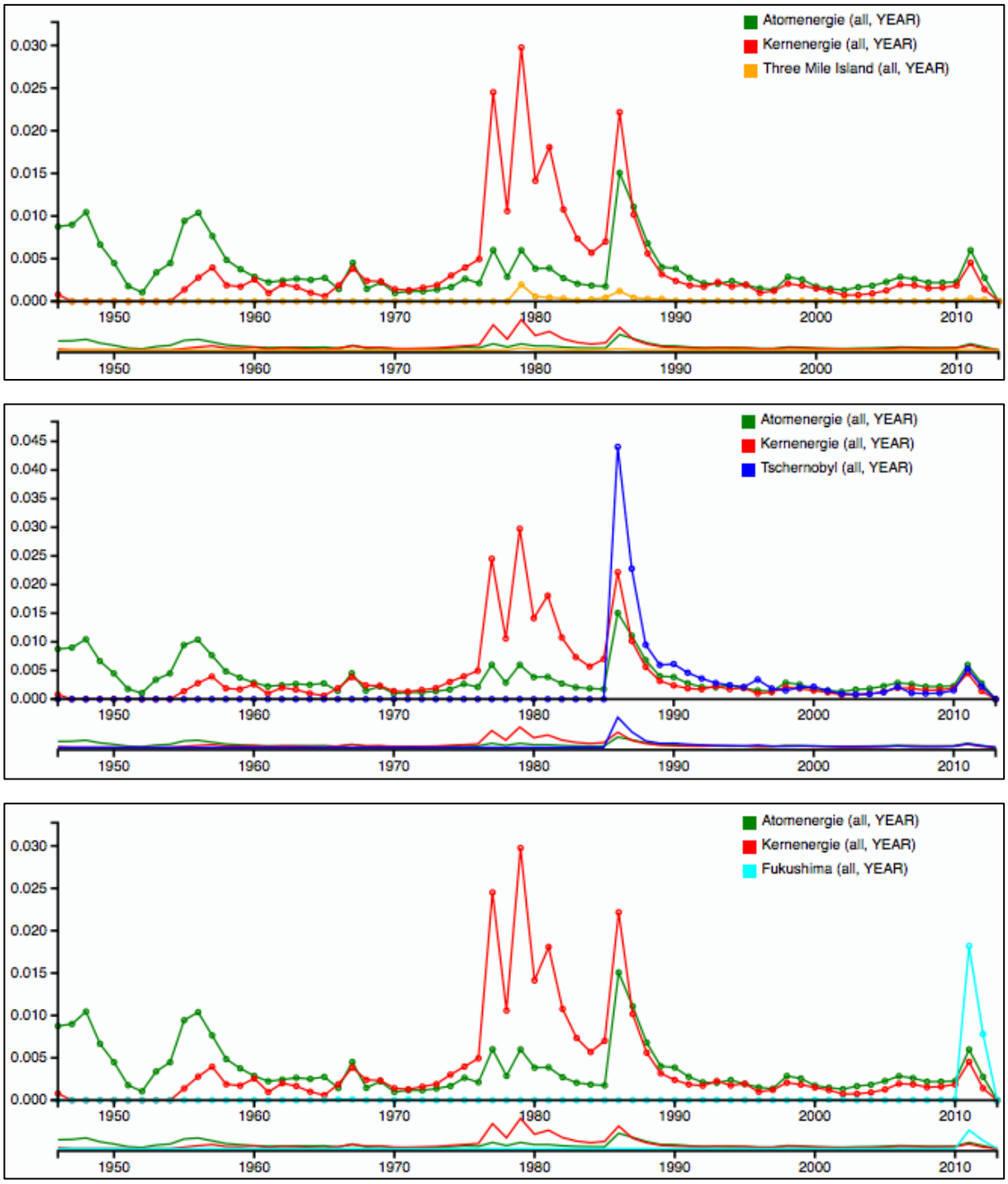


Die Graphen aus den Abbildungen 4.4 und 4.5 können – unterstellt man den affirmativen Kontext von „Kernenergie“ und den eher kritischen von „Atomenergie“ – dahingehend interpretiert werden, dass das Wort „Atomenergie“ alleine den Diskurs in der Tat nicht hinreichend abzudecken vermag. Schaut man auf die starke Konjunktur des Begriffes „Kernenergie“ besonders zu Beginn der 1980er, dann wird zweierlei deutlich: Die Konjunkturen in den Graphen von 4.4 und 4.5 ähneln sich sehr, so dass offenbar sowohl bei Atom- als auch bei Kernenergie häufig von Gefahrenkontexten gesprochen wird. Zudem gelingt es dem kritisch konnotierten Begriff der Atomenergie erst zum Ende der achtziger Jahre des 20. Jahrhunderts, den Begriff der Kernenergie als höherfrequenten Begriff abzulösen. Dabei beginnt eine

[7] Die Einschränkung des Suchumfeldes erfolgt in der Lucene Query Syntax durch Einfügung des Tilde-Zeichens („~“) unmittelbar nach den Suchbegriffen, gefolgt von der Zahl der Worte, die als Suchumfeld definiert werden (etwa „10“). Die Einschränkung erreicht, dass nur Dokumente aufgelistet werden, in denen die Suchbegriffe im definierten Abstand zueinander im Text auftreten.

hochfrequente Verwendung insbesondere des Begriffs Kernenergie schon zum Ende der siebziger Jahre des 20. Jahrhunderts.

Abb. 4.6 – Frequenzen (standardisiert) von „Three Mile Island“ (oben), „Tschernobyl“ (Mitte) und „Fukushima“ (unten) in Relation zu „Atomenergie“ und „Kernenergie“.



Dieser und folgende Peaks gehen – so zeigen die Graphen der Abbildungen 4.6 – mit den Reaktorkatastrophen in Three Mile Island, Tschernobyl und Fukushima-Daichi einher, die sich als realweltliche Großereignisse offenkundig und wenig überraschend katalytisch auf die Frequenz in der Medienberichterstattung auswirken.
Dass das Wort „Atomenergie“ schon vor der Reaktorkatastrophe von Tschernobyl, und auch noch vor der von Three Mile Island (Harrisburg, Pennsylvania, 28.3.1979) auftritt, kann im übrigen auch das Aufkommen ökologischer Bewegungen – in Deutschland etwa die Partei „Die Grünen“ – anzeigen, die dann in Folge eines konkreten Atomunfalls zunehmend ihren Themen in die öffentliche Debatte platzieren können. Diese Interpretation der Wortfrequenzen bedarf jedoch einer weiteren Absicherung durch Kookkurrenzanalysen (vgl. eTMV 2/5 – Serie „Atomausstieg“) oder Topic Modelle (vgl. eTMV 3/5 – Serie „Atomausstieg“).

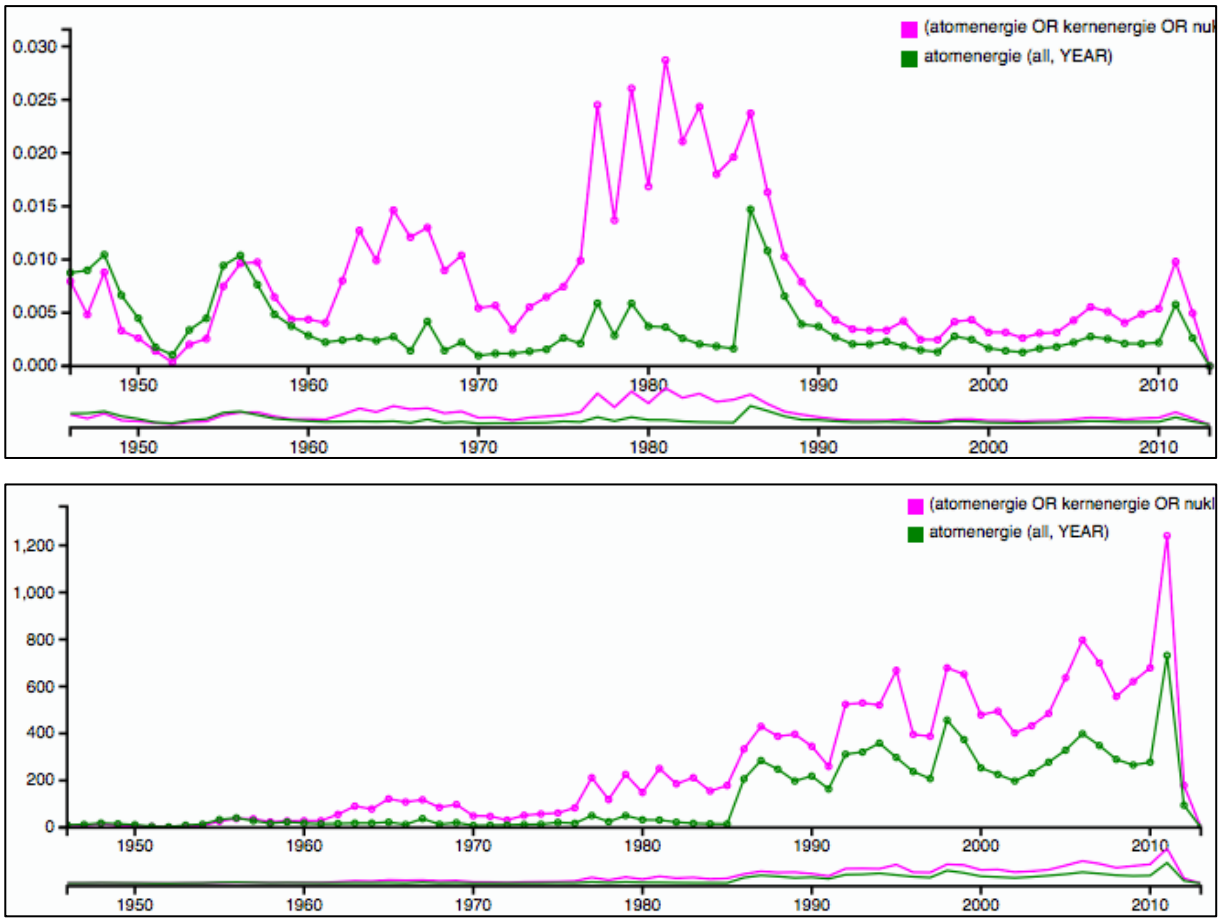
Um den kompletten Diskurs über die Verwendung der Atomenergie in Deutschland adäquat in den Texten auffinden zu können, so zeigen bereits diese wenigen Beispiele, ist **eine Kombination mehrerer Suchwörter, also von Dictionaries**, zielführend. Atomenergie erweist sich damit als ein – sozialwissenschaftlich ausgedrückt – Konzept[8], das nicht unmittelbar und aus sich selbst heraus den Sachverhalt beschreibt, für den es steht. Ähnlich wie bei anderen Konzepten, wie Ökonomisierung, Neoliberalisierung oder Arbeitslosigkeit, bedarf ein Konzept einer Beschreibung durch inhaltstragende Begriffe. Eine solche Liste inhaltstragender Begriffe wird als *Diktionär** bezeichnet. Es enthält mindestens zwei bis – theoretisch – unendlich viele Begriffe, die allesamt ein Konzept inhaltlich konkretisieren.
Für das Konzept „Atomenergiediskurs“ könnten etwa folgende Begriffe zu einer inhaltlichen Bestimmung beitragen: „Atomenergie“, „Atomkraft“, „Kernenergie“, „Kernkraft“, „Atomstrom“, „Atomkraftwerk“, „Tschernobyl“, „Fukushima“, „Brückentechnologie“. Im LCM kann nach solchen, auf Diktionären basierenden Konzepten gesucht werden, indem im Suchfeld ‚Custom’ eine entsprechende Suchanfrage eingegeben wird, die alle für den Begriff relevanten Einzelworte mit einer Oder-Verknüpfung (‚OR’) verbindet. Im Falle des Atomenergiediskurses könnte eine solche Suchanfrage etwa

„(atomenergie OR kernenergie OR nuklear* OR kernkraft) AND deutsch*“

[8] Sprachwissenschaftlich würde man von einem Begriff sprechen.

lauten, wobei die nachgestellte Einschränkung darauf hinwirken soll, gezielt Texte mit einem Deutschlandbezug zu finden. Für die relative und absolute Frequenz dieser komplexen Suchanfrage (n=17.331 Trefferdokumente) ergibt sich – im Vergleich zur Einzelsuchabfrage „Atomenergie“ – folgendes Bild:

Abb. 4.7 – Frequenz des Diktionärs (standardisiert, oben; absolut, unten) im Vergleich zur Frequenz des Einzelwortes „Atomenergie“.

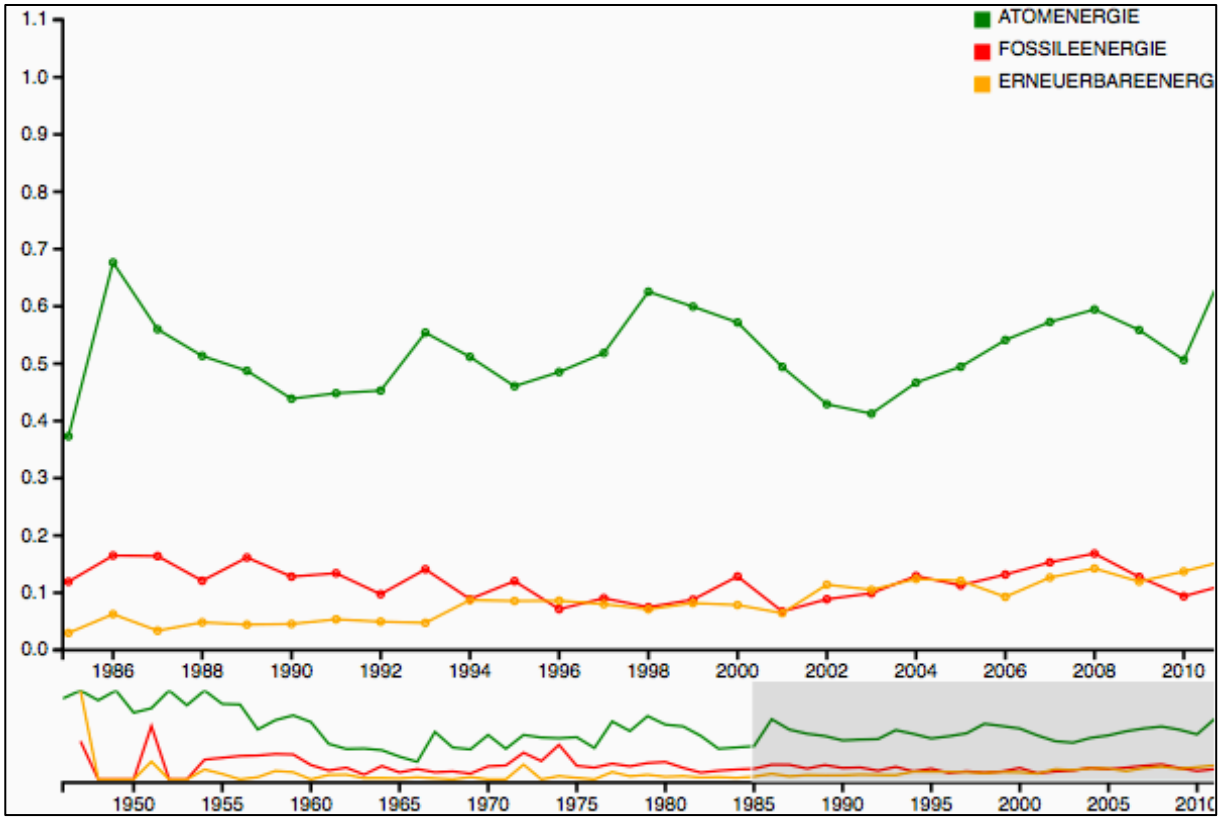


Im Unterschied zur Einzelsuche zeigt die Verlaufskurve an, dass bei einer auf einem Diktionär beruhenden Suche mehr Dokumente aufgefunden werden. Das bedeutet, dass das für die Analyse zur Verfügung stehende Material komplexer wird und tendenziell eher den gesamten Diskurs einfangen dürfte, als das bei einer Einzelwortsuche der Fall ist.

Der auf Basis der Suche mit dem Diktionär generierte *Subkorpus** mit seinen 17.331 Trefferdokumenten kann im LCM separat abgespeichert werden. Das ist die Voraussetzung für die Anwendung nachgeordneter Analyseverfahren (etwa Kookkurrenzen oder Topic Modelle). Zudem kann innerhalb des Subkorpus – im LCM über den Reiter ‚Collection Worker’, dort

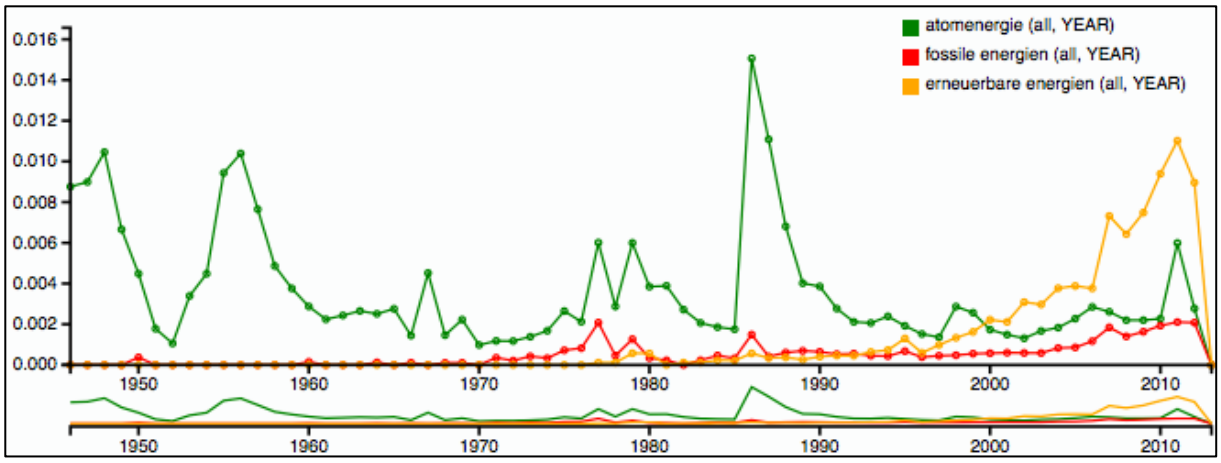
unter dem Menüpunkt ‚Dictionary Frequency Analysis’ – nach einem oder mehreren Konzepten anhand weiterer Diktionäre gesucht werden. Zwei Analyseleistungen stechen dabei hervor: So kann etwa die **Kontextualisierung bestimmter Begriffe** bestimmt werden: Begriffe, deren Kontextualisierung festgestellt werden soll, werden in einer Liste erfasst. Anhand von Kontextlisten (die etwa eine positive/negative, im Prinzip aber jede beliebige Kontextualisierung ausdrücken können) werden dann Textstellen gesucht, die jeweils Begriffe der vordefinierten Listen enthalten. Die gefundenen Textstellen werden entsprechend der Listeneigenschaft (positiv/negativ etc.) kategorisiert und dann im Zeitverlauf abgebildet. Auch lässt sich feststellen, welche **beliebigen weitern Konzepte im Subkorpus** enthalten sind und wie sie sich über die Zeit verteilen. So ließen sich etwa innerhalb des Atomenergiediskurses Referenzen zum Thema „erneuerbare Energien“ ebenso wie zum Thema „fossile Energien“ über ein Dictionary auffinden und hinsichtlich ihrer Frequenz miteinander vergleichen:

Abb. 4.8 – Frequenz von „erneuerbare Energien“ und „fossile Energien“ im Atomenergiediskurs (standardisiert), fokussiert auf die Jahre 1985–2011.



Erst ab der Mitte der 2000er Jahre werden im Subkorpus zum Atomenergiediskurs die „erneuerbaren Energien“ höherfrequent verwendet, als die „fossilen Energien“. Beide reichen in ihrer Frequenz aber nicht an die Verwendungshäufigkeit von „Atomenergie“ heran, die ja als Suchbegriff – zusammen mit anderen – konstitutiv für die Erstellung des Subkorpus war. Such man „erneuerbare Energien“ im Vergleich zur Frequenz von „fossiler Energie“ und „Atomenergie“ im Gesamtkorpus, ergibt sich folgendes Bild:

4.9 – Frequenz von „erneuerbare Energien“, „fossile Energien“ und „Atomenergie“ im Gesamtkorpus (standardisiert).



Hier wird – mit Blick auf eine breitere politische Öffentlichkeit – deutlich, dass das Thema der erneuerbaren Energien bereits seit Mitte der 1990er Jahre eine höhere Frequenz aufweist, als das Thema fossile Energien. Ab Beginn der 2000er Jahre überholt seine Frequenz auch diejenige von Atomenergie – bis hin zu einem Peak im Jahr 2011, dem Jahr der Reaktorkatastrophe von Fukushima. Der gesamtgesellschaftliche Stellenwert der erneuerbaren Energien ist also wesentlich höher als derjenige der anderen Energieträger. Dementsprechend kann der Graph aus Abbildung 4.8 auch als Indiz dafür gelesen werden, wie gut es innerhalb des Atomenergiediskurses gelingt, die Optionen alternativer Energienutzung zu einem Randthema zu machen. Diese Hypothese auf Basis der Frequenzanalyse formulierte Hypothese müsste nun – Stichwort *Blended Reading** – in einem Close Reading exemplarischer Texte aus einzelnen Zeitabschnitten validiert werden.

5. Glossar

Begriff	Erläuterung
Blended-Reading-Ansatz	Der Blended-Reading-Ansatz (Lemke/Stulpe 2015) verweist mit Blick auf die von Franco Moretti etablierte Unterscheidung von Close und Distant Reading auf die Notwendigkeit, beide Ebenen im Forschungsprozess integriert zu behandeln. Eine bestimmte Verfahrensfolge im Text Mining (von der Verwendungskonjunktur zum Verwendungskontext) ist durch den jederzeit möglichen Rückgriff auf die der Sprachstatistik zugrunde liegenden Einzeltexte zu begleiten.
Diktionär	Ein Diktionär ist eine beliebig lange, mindestens aber zwei Worte umfassende Liste von Worten, die ein abstraktes oder komplexes Konzept inhaltlich bestimmt und für eine Suchanfrage erschließt. Abstrakte oder komplexe Konzepte, wie etwa Ökonomisierung oder Atomenergiediskurs, die nicht aus sich selbst heraus den Inhalt repräsentieren, für den sie stehen, bedürfen konkretisierender Begriffe, die im Rahmen von Suchanfragen diejenigen Texte auffinden, die das Konzept am besten abbilden.
Heatmap	Eine Heatmap ist eine Form der Visualisierung von Daten, deren abhängige Werte in unterschiedlichen Farbschattierungen repräsentiert werden.
Subkorpus	Ein Subkorpus ist eine Teilmenge von Artikeln aus dem Gesamtkorpus, die auf Basis einer Artikelsuche (im Modus Simple, Detailed oder Custom) generiert wurde und Gegenstand weiterer Analyseverfahren ist.
Text Mining	Nach Gerhard Heyer et al. (2006: 3) bezeichnet der Begriff „[...] computergestützte Verfahren für die semantische Analyse von Texten [...], welche die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Mengen von Texten, unterstützen.“
Unigramm	Im Rahmen der Zerlegung von Texten in einzelne Fragmente, etwa Worte, bezeichnen N-Gramme die Anzahl der betrachteten Wortgruppen. Ein Uni- oder auch Monogramm ist dabei die Bezeichnung für eine Einworteinheit. Eine Zweiworteinheit wird als Bigramm, eine Dreiworteinheit dementsprechend als Trigramm bezeichnet.

6. Verwendete Literatur

Heyer, Gerhard et al. (2006): Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse, Bochum.

Lemke, Matthias / Stulpe, Alexander (2015): Text und soziale Wirklichkeit. Theoretische Grundlagen und empirische Anwendung durch Text Mining Verfahren am Beispiel des Bigrams ‚soziale Marktwirtschaft', erscheint in Zeitschrift für Germanistische Linguistik, Themenheft „Automatisierte Textanalyse“, hg. von Noah Bubehofer und Joachim Scharloth.

Wiedemann, Gregor / Lemke, Matthias / Niekler, Andreas (2013): Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentation in der Bundesrepublik Deutschland 1949–2011. Ein Werkstattbericht. In: Zeitschrift für Politische Theorie 4 (1), 99–115.

Wiedemann, Gregor / Niekler, Andreas (2014): Analyse qualitativer Daten mit dem Leipzig Corpus Miner, Hamburg / Leipzig (=Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus Discussion Paper 5).

7. Weiterführende Literatur

Manning, Christopher D. / Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing, Cambridge (MA).

Saussure, Ferdinand de (1967): Grundfragen der allgemeinen Sprachwissenschaft, Berlin.

Zipf, George K. (1935): The Psychobiology of Language. An Introduction to Dynamic Philology, Boston.

Zipf, George K. (1945): The meaning – frequency relationship of words. In: Journal of General Psychology, 33, 251–266.

eTMV-Serie „Atomenergiediskurs“

1 – Frequenzanalyse und Diktionäransatz
2 – Kookkurrenzanalyse
3 – Topic Modelle
4 – Annotation und Klassifikation
5 – Sentimentanalyse

www.epol-projekt.de/etmv