Effiziente halbautomatische Detektion von Neologismuskandidaten

Technical Report IDS-KL-2010-01

Holger Keibel, Sophie Hennig, Rainer Perkuhn
Institut für Deutsche Sprache

2010

(1) Einleitung

Ein zentrales und fortlaufendes Anliegen in der Neologismenforschung und -lexikographie ist es, Neologismen (Neulexeme und Neubedeutungen) zu entdecken. Neologismen können auf verschiedenen Kanälen entdeckt werden: z.B. durch aufmerksames Beobachten als Mitglied der Sprachgemeinschaft oder durch gezielte Lektüre der Medien, denen man hinsichtlich der Abbildung der Allgemeinsprache eine hohe Relevanz zuschreibt. Diese Vorgehensweise ist einerseits unabdingbar, sie läuft andererseits auch Gefahr, viele Neologismen zu übersehen – oder stattdessen große Kosten zu verursachen: Denn eine systematische, lückenlose Lektüre der relevanten Medien wäre nur mit unverhältnismäßig hohem Zeitaufwand zu leisten. Daher ist es sinnvoll, die Auswertung von Medien zusätzlich auch mit automatischen Verfahren durchzuführen. Diese können die relevanten Medien systematisch auswerten und Wissenschaftler/innen auf mögliche Neologismen aufmerksam werden, die ihnen auf anderen Kanälen entgangen wären.

Komplett automatisch Neologismen von Nichtneologismen fehlerfrei zu unterscheiden ist allerdings zumindest mit den heute zur Verfügung stehenden Verfahren unmöglich. Das realistische Ziel bei solchen Verfahren ist daher, eine Kandidatenliste zu generieren, die (i) einen möglichst geringen Anteil nicht relevanter Wörter enthält (d.h. die eine hohe *Precision* aufweist) und die gleichzeitig (ii) möglichst alle genuinen Neologismen umfasst, die in den analysierten Medien vorkommen (d.h., die einen hohen *Recall* aufweist). Diese beiden Kriterien konkurrieren miteinander – der Versuch, ein Verfahren in Bezug auf Precision zu optimieren führt i.A. dazu, dass es in Bezug auf Recall schlechter abschneidet, und analog auch umgekehrt.

Daher wird bei der hier vorgestellten Studie eine halbautomatische Strategie verfolgt: Zunächst wird durch automatische Verfahren eine Kandidatenliste generiert, bei der im Kompromiss zwischen Recall und Precision der Recall stärker gewichtet wird. Recall wird hierbei aber nicht einseitig maximiert, denn sonst wäre die Liste extrem lang und nahezu wertlos. Die automatisch gewonnene Kandidatenliste wird anschließend zügig (und ohne eigentliche Analyse) manuell gesichtet und eindeutige Nichtneologismen werden dabei herausgefiltert. Dadurch wird die Precision erheblich erhöht, während der Recall weitgehend unverändert hoch bleibt. Erst diese gefilterte Liste dient als Input für nähere Expertenanalysen.

Dieser halbautomatische Ansatz besteht insgesamt aus drei Phasen, die im Folgenden näher beschrieben werden. Der Fokus liegt dabei auf Neulexemen – Neubedeutungen werden zwar nicht ausgeschlossen, für die meisten von ihnen ist es jedoch unwahrscheinlich, dass sie mit der hier vorgestellten Methode aufgespürt werden können.

Phase I: Automatische Analysen
In dieser Phase werden sukzessive mehrere Analyseschritte angewendet. Die ersten Schritte ermitteln im gegebenen Korpus Wörter, die eine für Neologismen typische zeitliche Häufigkeitsentwicklung aufweisen. Dabei werden unterschiedliche Oberflächenformen (d.h. flektierte Vollformen und Rechtschreibvarianten) desselben Lexems bewusst als verschiedene Wörter behandelt (vgl. Schritt 0 in Abschnitt 3). Im Folgenden ist mit dem Begriff „Wort“ stets eine Oberflächenform gemeint.

Unter der so generierten initialen Kandidatenliste sind auch viele Wörter, die eindeutig keine Neologismen sind, aber sich aus anderen Gründen quantitativ wie Neologismen verhalten: Die größte Gruppe solcher Wörter sind Eigennamen von Personen, Unternehmen und Orten, die durch außersprachliche Ereignisse ausgelöst mit zunehmender Häufigkeit im öffentlichen Diskurs auftreten.

Daher wird in weiteren Analyseschritten versucht, die größeren solcher Gruppen von eindeutigen Nichtneologismen aus der initialen Liste möglichst trennscharf herauszufiltern. Da eine perfekte Trennschärfe i.A. nicht möglich ist, wird versucht, in Pilotstudien einen guten Kompromiss zu finden zwischen Precision und Recall – also zwischen dem Ziel, unerwünschte Wörter wegzufiltern, und dem Ziel, genuine Neologismen nicht wegzufiltern. Dabei wird tendenziell eher konservativ gefiltert, also Recall stärker gewichtet als Precision. Nach allen Filterschritten verbleibt eine finale Kandidatenliste als Input für Phase II.

Phase II: Manuelles Filtern
Die in Phase I ermittelten und gefilterten Kandidaten werden manuell durchgesehen und eindeutige Nichtneologismen werden dabei herausgefiltert. Das Hauptziel dieser Phase ist es, die Precision der Kandidatenliste mit möglichst geringem Aufwand zu verbessern. Dadurch reduziert sich i.A. der Umfang der Liste deutlich. In früheren Arbeiten hat sich gezeigt, dass in vielen Fällen bereits die Sprachkompetenz ausreicht, um eindeutige Entscheidungen zu treffen. Daher sollten nur unsichere Fälle überhaupt näher betrachtet werden, und dies auch nur mit minimalem Zeitaufwand.

Dieser Schritt kann sehr gut von studentischen Hilfskräften durchgeführt werden. Linguistische Kenntnisse sind vor allem bei der Einarbeitung von Vorteil, aber nicht zwingend erforderlich. Muttersprachliche Kompetenz und eine kurze Trainingsphase sind die einzigen zentralen Voraussetzungen. Für eine zügige Arbeit ist zudem eine hinreichende Aufmerksamkeit für den öffentlichen Diskurs der vergangenen Jahre hilfreich.

Phase III: Expertenanalysen
Die ersten beiden Phasen haben das Ziel, die Arbeit im Bereich der Neologismenforschung und -lexikographie zu unterstützen. Experten/innen aus diesem Bereich können so auf mögliche Neologismen aufmerksam werden, die ihnen auf anderen Kanälen entgangen wären. Für die nach den ersten beiden Phasen verbleibenden Kandidaten ist noch zu überprüfen, ob sie die formalen Neologismuskriterien im Sinne von Steffens (2007) erfüllen und falls ja: in welchem Jahr ihr faktisches Aufkommen jeweils liegt. Diese beiden Fragen können nur durch Expertenanalysen entschieden werden.

Im Rahmen dieser Studie wurden nur die ersten beiden Phasen umgesetzt (Phase II nur stichprobenweise). Für eine vollständige Evaluation auch der beiden ersten Phasen wäre es jedoch essentiell, auch noch Phase III durchzuführen.

Die allgemeine Strategie wurde im Rahmen der vorliegenden Studie konkret umgesetzt und anhand realer Daten getestet. Das Ziel war dabei explizit, Neulexeme im schriftdeutschen Sprachgebrauch aufzuspüren, die in den 2000er Jahren aufgekommen sind. Ältere Neologismen sollten also als Nichtneologismen behandelt werden.

Insbesondere sollen Probleme aus einer früheren Studie (Keibel 2007) vermieden werden. Die Strategie bei dieser Studie war es, mit einem universellen (wenn auch einfachen) Maß Neologismuskandidaten zu finden. Auch damals sollten gezielt Neologismen mit Aufkommen seit 2000 aufgespürt werden, doch es zeigte sich in anschließenden Expertenanalysen des Projekts „Lexikalische Innovationen“ am Institut für Deutsche Sprache, dass das globale Maß für das Aufkommen unsensibel war und dass die von ihm tatsächlich gefundenen Neologismen bereits in dem vorangegangenen Jahrzehnt aufgekommen waren.[1]

Diesmal wurde das Ziel eines universellen Maßes aufgegeben und stattdessen ein sehr pragmatischer Ansatz verfolgt, um gezielter nach Neologismuskandidaten aus dem gewünschten Jahrzehnt zu suchen. Bei diesem pragmatischen Ansatz musste eine Reihe von Parametern gesetzt werden – diese wurden in Pilotstudien speziell für das verwendete Korpus optimiert. Für andere Korpora und spätere Zeiträume sind diese Parameter nicht direkt verwendbar, sondern müssten in erneuten Pilotstudien auf die anderen Rahmenbedingungen abgestimmt werden.

(2) Korpus

Für diese Studie wurde nach den folgenden Kriterien ein virtuelles Korpus auf der Basis des Deutschen Referenzkorpus (DeReKo) zusammengestellt: Es sollte den Zeitraum 1990-2008 abdecken und ausschließlich Zeitungstexte aus Deutschland enthalten.
Verwendet wurden Texte der folgenden DeReKo-Zeitungskorpora, unterteilt in 4 Regionen:

- NORDEN (Gesamtgröße: 240.070.744 Wörter)
  - Braunschweiger Zeitung (brz): 2005-2008
  - Hamburger Morgenpost (hmp): 2005-2008
  - Hannoversche Allgemeine Zeitung (haz): 2007-2008
  - Der Spiegel (s): 1993-1994
  - Die ZEIT (z): 1994-2004
- OSTEN (Gesamtgröße: 535.370.440 Wörter)
  - Berliner Morgenpost (l):1997-1999
  - Berliner Zeitung (b): 1997-2008
  - die tageszeitung (t): 1990-2008
- SÜDEN (Gesamtgröße: 623.903.617 Wörter)
  - Frankfurter Allgemeine Zeitung (f): 1993, 1995, 1997, 1999, 2001, 2003, 2005[2]
  - Frankfurter Rundschau (r): 1997-1999
  - Nürnberger Nachrichten (nun): 1990-2008
  - Nürnberger Zeitung (nuz): 2002-2008
  - Süddeutsche Zeitung (u): 1995-1999

1 Die Autoren danken Doris Steffens und Doris al-Wadi für dieses Feedback.
2 Pro Jahrgang der F.A.Z. steht in DeReKo nur jeder zweite Monat zur Verfügung.

- SÜDWESTEN (Gesamtgröße: 974.868.553 Wörter)
  - Mannheimer Morgen (m): 1991, 1994-2008
  - Die Rheinpfalz (rhp): 2007-2008
  - Rhein-Zeitung (rhz): 1996-2008

Ideal wäre es, nur Zeitungskorpora zu verwenden, die in dem untersuchten Gesamtzeitraum (1990-2008) nahezu durchgehend verfügbar sind. Durch die Verwendung der teilweise nur über kurze Zeiträume verfügbaren Zeitungskorpora könnten u.U. Artefakte in der Messung entstehen und viele Nichtneologismen quantitativ wie Neologismen aussehen. Um den Recall möglichst hoch zu halten und weil andernfalls auch nur wenige Korpora übrig blieben, wurden diese Zeitungskorpora dennoch einbezogen und eine geringere Precision in Kauf genommen. Über die Filterung von Regionalismen konnte ein Teil dieser Artefakte dennoch herausgefiltert werden.

Tabelle 1 zeigt die Größe des resultierenden virtuellen Korpus, aufgeschlüsselt nach Jahrgängen.

Tabelle 1: Korpusgröße je Jahrgang (in laufenden Textwörtern)

Jahrgang	Korpusgröße
1990	23.506.958
1991	22.971.091
1992	22.802.469
1993	43.133.997
1994	29.802.096
1995	81.577.959
1996	104.465.902
1997	160.884.515
1998	181.429.709
1999	226.531.196
2000	116.886.753
2001	151.273.858
2002	129.531.259
2003	154.444.919
2004	149.832.230
2005	197.721.816
2006	190.240.811
2007	197.421.723
2008	189.754.093
Gesamt	**2.374.213.354**

Idealerweise sollte die Korpusgröße für jeden Jahrgang etwa gleich groß sein. Das hätte in diesem Fall jedoch bedeutet, die Daten jedes Jahrgangs auf die Korpusgröße des kleinsten Jahrgangs (1992) zu reduzieren und damit also insbesondere in den relevantesten Jahrgängen (2000-2008) einen sehr großen Teil der verfügbaren Daten zu ignorieren. Auch auf diese Maßnahme wurde zugunsten eines möglichst hohen Recalls verzichtet.

Nachfolgend werden wir das hier beschriebene Korpus nur noch kurz als *das Korpus* bezeichnen.

(3) Vorgehensweise

Es wurden nacheinander die folgenden Schritte durchgeführt. Dabei sind die Schritte 0 bis 5 allesamt automatische Schritte, sie gehören also zur oben beschriebenen Phase I.

0. Ausgangsliste mit Worthäufigkeiten bestimmen
In diesem Schritt wurden zunächst jahresweise die absoluten Korpusfrequenzen aller im Korpus vorkommenden Wörter bestimmt. Diese Frequenzen dienen als primäre Grundlage für die quantitativen Analysen.

Unterschiedliche Oberflächenformen (d.h. flektierte Vollformen und Rechtschreibvarianten) desselben Lexems werden dabei bewusst als verschiedene Wörter behandelt. Eine automatische Lemmatisierung oder Schreibweisenharmonisierung wäre in dieser Studie nicht sinnvoll. Da es hier darum geht, neue Wörter aufzuspüren, die bisher noch nicht als solche erfasst sind, ist es äußerst unwahrscheinlich, dass bestehende Lemmatisierer diese Wörter bereits kennen. Ein Lemmatisierer würde die beobachtete Oberflächenform eines neuen Wortes also i.A. entweder gar nicht auf eine Grundform abbilden oder – falls die beobachtete Form identisch ist mit einer Form eines bereits etablierten anderen Wortes – auf eine falsche Grundform abbilden. Im ersten Fall wäre der Lemmatisierer keine Hilfe, im zweiten Fall sogar ein Problem, denn er würde einen genuinen Neologismus als ein etabliertes Wort maskieren, so dass der Neologismus unentdeckt bliebe. Daher ist es bei neuen Wörtern nicht ratsam, sich auf Lemmatisierer und andere morphologische Werkzeuge zu stützen. Analoges gilt für Werkzeuge zur Rechtschreibkorrektur und -harmonisierung.

Es ist daher wahrscheinlich, dass genuine Neologismen in mehreren Varianten in der finalen Liste auftreten. Eine Lemmatisierung oder Schreibweisenharmonisierung sollte nur manuell durch einen kompetenten Sprecher erfolgen (und bei Bedarf nach Sichtung eine Stichprobe von Korpusbelegen). Im Rahmen der hier vorgestellten Vorgehensweise kommt hierfür also entweder Schritt 6 oder Schritt 7 infrage.

Sofern nichts anderes dazugesagt wird, ist in diesem Paper mit dem Begriff „Wort“ stets eine Oberflächenform gemeint.

1. Wörter wegfiltern, die höchstwahrscheinlich keine lexikalischen Wörter sind
Hierbei sind v.a. Zeichenfolgen gemeint, die ungewöhnliche Zeichen enthalten (URLs, Code, ...)

2. Quantitatives Kriterium I
Betrachte nur Wörter, die im Korpus (also im gesamten Zeitraum 1990-2008) mindestens 20 Mal vorkommen. Motivation: Bei Wörtern mit weniger Vorkommen ließe sich die zeitliche Entwicklung nicht sinnvoll auswerten.

3. Nach weiteren quantitativen Kriterien filtern
Anschließend wurden die folgenden sehr groben Filterkriterien angewendet. Diese wurden durch stichprobenartige Auswertungen schrittweise festgelegt und optimiert. Die Strategie hierbei war, mit den verschiedenen Ausschlusskritierien jeweils einen Ausschnitt aus der Kandidatenliste zu entfernen, in dem sich nach unseren Erfahrungen mit den Stichproben fast ausschließlich eindeutige Nichtneologismen befinden. So sollte die Precision auf ein erträgliches Maß gebracht werden, ohne

dabei den Recall zu stark zu beeinträchtigen.

1. Betrachte nur Wörter, deren durchschnittliche Absolutfrequenz seit dem (berechneten) Aufkommen mindestens 7 Vorkommen pro Jahr beträgt.
   Dabei wird "Aufkommen" hier formal definiert als das erste Jahr mit der Eigenschaft, dass in diesem und dem Vorjahr jeweils mindestens 2 Vorkommen gefunden wurden. Dieses Konzept ist nur ein technisches Hilfsmittel und i.A. nicht deckungsgleich mit dem Begriff "Aufkommen", der Neologismus-Wörterbüchern zugrunde liegt.
   (Durch dieses Kriterium soll der Vorteil von früh aufgekommenen Wörtern bei Schritt 2 neutralisiert werden.)

2. Betrachte nur Wörter, deren Gesamtfrequenz mindestens 15 x die Anzahl Jahre seit ihrem maximalen Jahr (also dem Jahr mit den meisten Vorkommen) beträgt. Bei dieser Jahresanzahl ist das maximale Jahr eingeschlossen.
   (Dadurch werden viele Wörter weggefiltert, die Eintagsfliegen waren und sich nach ihrem größten Frequenz-Ausschlag offenbar nicht im allgemeinen Sprachgebrauch festsetzen konnten.)

3. Betrachte nur Wörter, die im Zeitraum 1990-1999 insgesamt höchstens 5 Vorkommen haben, und dabei in keinem einzigen Jahr mehr als 4 Vorkommen haben.

4. Betrachte nur Wörter, die im Zeitraum 2000-2008 insgesamt mindestens 5 Vorkommen haben.
   (Durch dieses und das vorherige Kriterium soll sichergestellt werden, dass es sich um Neologismen aus dem Zeitraum 2000-2008 handelt. Im Jahrzehnt 1990-1999 null Vorkommen zu verlangen hat sich als ein zu scharfes Kriterium herausgestellt, weil es einen großen Anteil genuiner Neologismuskandidaten aus dem gewünschten Zeitraum wegfiltern würde. Es kann viele Gründe geben, warum ein Neologismus bereits vor seinem in den späteren Expertenanalysen (Schritt 7) festgestellten Aufkommen im Korpus vorkommt – in vielen solchen Fällen handelt es sich um einen Schreibfehler oder eine andere Verwendung des Wortes, manchmal aber durchaus um dasselbe Wort aus einer Zeit, in der es z.B. Teil einer fachspezifischen oder gruppenspezifischen Sprachvarietät war und noch keinen Eingang in die Allgemeinsprache gefunden hatte.)

Es wurde bewusst nicht verlangt, dass ein Neologismuskandidat nach seinem vermuteten Aufkommen in *jedem* Jahr mit einer gewissen Mindestfrequenz im Korpus vorkommen muss (Bei Filterkriterium 3 in dieser Aufzählung ging es ja nur um eine durchschnittliche Anzahl Vorkommen). Durch die notwendigerweise immer lückenhafte Korpuszusammenstellung – und bei einigen Begriffen auch durch eine Abhängigkeit von außersprachlichen Anlässen, um diese Begriffe zu verwenden – kann ein genuiner Neologismus nach seinem Aufkommen im Korpus recht volatil sein.

4. Wörter wegfiltern, die vorwiegend in einer bestimmten Region auftreten

Viele Nichtneologismen, die sich im gesamten Korpus wie Neologismen verhalten, kommen v.a. in einer bestimmten Region oder sogar nur in einer einzigen Zeitung vor: Dazu zählen Regionalismen, redaktioneller Code (Autorenkürzel, Rubrikenüberschriften, etc.), Nichtstandard-Schreibweisen etablierter Wörter und bereits sehr viele Eigennamen (Orte, Firmen, Personen), usw. Um einen Großteil dieser Wörter effizient wegfiltern zu können, wurde das Korpus auf vier Regionen (Norden, Osten, Süden, Südwesten) aufgeteilt (vgl. Abschnitt 2).

Als globales Maß für die regionale Unausgewogenheit eines Wortes wurde Gries' *Dispersionsmaß DP* (*deviation of proportions*) verwendet (Gries 2008). Neben diesem globalen Maß wurde durch zwei weitere Maße außerdem bestimmt, welche der vier Regionen die dominante ist (d.h., in welcher Region das Wort am häufigsten verwendet wird).

> $Maß_1$: der Anteil der Region an allen Vorkommen des Wortes
>
> $Maß_2$: der (bzgl. Korpusgröße) korrigierte Anteil der Region an allen Vorkommen des Wortes. Genauer: $Maß_2$ ist wie $Maß_1$ mit dem Unterschied, dass bei $Maß_2$ die beobachteten Vorkommen in jeder Region bzgl. der Größe der regionsspezifischen Teilkorpora korrigiert werden.
>
> Bei beiden Maßen bedeutet z.B. der Wert 0,7 eines Wortes für eine bestimmte Region, dass von allen Vorkommen dieses Wortes im (bei $Maß_2$ korrigierten) Korpus 70% in dem Teilkorpus zu dieser Region gefunden wurden.

In einer einfachen Pilotstudie hat sich gezeigt, dass für die einzelnen Regionen unterschiedliche Kriterien formuliert werden müssen, um jeweils eine gute Trennschärfe zu erreichen. Es wurden für das hier verwendete Korpus folgende Schwellenwerte bestimmt:

Dominante Region gemäß $Maß_2$:

dominante_Region=Norden	UND	($Maß_2$ > 0,9 ODER DP >0,6)
dominante_Region=Osten	UND	($Maß_2$ > 0,9 ODER DP >0,75)
dominante_Region=Süden	UND	($Maß_2$ > 0,75 ODER DP >0,55)
dominante_Region=Südwesten	UND	($Maß_2$ > 0,85 ODER DP >0,55)

Dominante Region gemäß $Maß_1$:

dominante_Region=Norden	UND	$Maß_1$ > 0,8
dominante_Region=Osten	UND	$Maß_1$ > 0,9
dominante_Region=Süden	UND	$Maß_1$ > 0,8
dominante_Region=Südwesten	UND	$Maß_1$ > 0,9

Ein Wort wurde weggefiltert, wenn mindestens eines dieser Kriterien zutraf. Um allerdings nicht zu streng wegzufiltern, wurden diese in der Pilotstudie bestimmten Schwellenwerte bei der finalen Analyse um einen Sicherheitspuffer angehoben (konservatives Filtern). Dieser Puffer lag bei 25% der zu 1,0 = 100% fehlenden Proportion. Beispiel: Ein Pilot-Schwellenwert von 0,7 wurde transformiert zu 0,7 + 0,25 * (1,0 - 0,7) = 0,775.

5. Eigennamen wegfiltern

Eigennamen (Personen, Orte, Unternehmen, Institutionen, usw.) sind der häufigste Typ von Wörtern, die hinsichtlich der Häufigkeitsentwicklung typischen Neologismen sehr ähneln, obwohl sie keine Neologismen sind. Anders als Regionalismen i.w.S. (Schritt 4) können Eigennamen jedoch nicht über einfache quantitative Kriterien grob weggefiltert werden. Stattdessen sind hier andere Verfahren notwendig.

Isolierte Wörter (Types, d.h. Types von Oberflächenformen) sind mit automatischen Verfahren nicht zuverlässig als Eigennamen zu erkennen. Nur für Wörter (Tokens von Oberflächenformen), die in einen Satzkontext (oder zumindest einem hinreichend großen Satzfragment) integriert sind, kann mit automatischen Verfahren festgestellt werden, ob es sich bei ihnen (wahrscheinlich) um einen

Eigennamen handelt. Um auch für die Wörter (Types) aus der Kandidatenliste eine solche Aussage machen zu können, wurde zu jedem Kandidaten eine Zufallsstichprobe von 200 KWIC-Zeilen aus dem Korpus[3] gezogen, sofern vorhanden. Jede einzelne KWIC-Zeile wurde mit einem Eigennamenerkenner analysiert (dieser Eigennamenerkenner wurde von dem IDS-Projekt "Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen" zur Verfügung gestellt). Zwar analysierte der Eigennamenerkenner dabei die gesamte KWIC-Zeile hinsichtlich möglicher Eigennamen, relevant war in unserem Kontext jedoch nur das Ergebnis für das Token des betreffenden Neologismuskandidaten. Aus Sicht des Eigennamenerkenners war dies für jedes einzelne Token eine Ja/Nein-Entscheidung. Darüber hinaus unterschied der Eigennamenerkenner drei verschiedene Arten von Eigennamen: Namen von Personen, Orten und Institutionen.

Um von den Tokens eines Neologismuskandidaten zu einer globalen Aussage über den Kandidaten als Type zu kommen, wurde nun über alle 200 zugehörigen KWIC-Zeilen der Anteil *P* von Fällen bestimmt, in denen der Eigennamenerkenner den betreffenden Token als Eigennamen klassifiziert hat.

In einer Pilotstudie wurde festgestellt, ob dieser Anteil P als grobes Maß für Eigennamen taugt: d.h., ob tatsächliche Eigennamen tendenziell einen hohen P-Wert haben und Nicht-Eigennamen tendenziell einen niedrigen. In derselben Pilotstudie wurde zudem ein empirischer Schwellenwert (55%) ermittelt, oberhalb dessen man recht zuverlässig auf einen Eigennamen schließen kann.
In Schritt 5 der finalen Auswertung wurden Wörter mit einem P-Wert oberhalb dieses Schwellenwerts weggefiltert. Ausnahmen: Wörter, für die weniger als 25 Vorkommen analysiert werden konnten, verblieben in der Liste. Ebenso blieben Wörter in der Liste, bei denen der Schwellenwert nur deshalb überschritten wurde, weil der Eigennamenerkenner besonders oft auf den Namen einer Institution geschlossen hat – denn bei solchen Klassifikationen ist der verwendete Eigennamenerkenner laut "Wechselwirkungen“-Projekt oft unzuverlässig. Weitaus zuverlässiger hingegen ist er bei Personennamen und Ortsnamen. Konkret operationalisiert wurde diese Einschränkung wie folgt: Wörter, bei denen der Anteil von Personennamen und Ortsnamen an den erkannten Eigennamenvorkommen unter 50% lag, wurden nicht weggefiltert.

6. Manuelles Filtern

Die Schritte 1 bis 5 der hier vorgestellten Vorgehensweise wurden dann in der abschließenden Anwendung komplett automatisch durchgeführt. Erst in Schritt 6 kamen manuelle Arbeiten ins Spiel: In diesem Schritt wurden die nach Schritt 5 verbliebenen Neologismuskandidaten manuell von einer studentischen Hilfskraft durchgesehen und annotiert. Drei mögliche Annotationswerte waren vorgesehen: Eindeutiger Nichtneologismus (n), vermuteter Neologismuskandidat (x) und unsichere Fälle (?).

Als Entscheidungsgrundlage diente dabei primär die Sprachkompetenz der Hilfskraft. Bei Zweifelsfällen sollte die Hilfskraft zugehörige KWICs sichten, die bewusst aus dem gesamten

3 Aus technischen Gründen konnten diese KWIC-Zeilen nur auf aus den über COSMAS II recherchierbaren Korpora erzeugt werden, während die Kandidatenliste direkt aus DeReKo (bzw. aus dem oben definierten virtuellen Korpus) generiert wurde. Da neue Zeitungsjahrgänge aus DeReKo immer erst mit etwas Zeitverzögerung auch in COSMAS II verfügbar sind, konnte für Wörter mit sehr spätem Aufkommen (in oder kurz vor dem Jahr 2008) meist nur wenige KWIC-Zeilen extrahiert und analysiert werden. Daher sind die Eigennamen-Filterkriterien für diese Wörter weniger zuverlässig als für Wörter mit früherem Aufkommen. In extremen Fällen, bei denen die Diskrepanz DeReKo vs. COSMAS II besonders groß war, wurden die oben beschrieben Eigennamen-Filterkriterien außer Kraft gesetzt. Solche Wörter blieben also in der Liste.

DeReKo entnommen wurden, nicht nur aus dem verwendeten virtuellen Korpus. Die Maxime bei diesen Entscheidungen war, konservativ zu filtern, also möglichst nur eindeutige Fälle herauszufiltern.

Im Vorfeld dieses manuellen Schritts wurde zunächst die gesamte Kandidatenliste nach folgenden Kriterien sortiert:

- nach absteigendem "technischem Aufkommen" im Zeitraum 1990-2008
- bei Gleichheit: nach absteigendem Jahr mit maximaler Frequenz
- bei weiterer Gleichheit: nach absteigender Gesamtfrequenz im Zeitraum 1990-2008
- bei weiterer Gleichheit: nach absteigendem M-Score[4]
- bei weiterer Gleichheit: nach absteigendem LLR-Wert (log-likelihood ratio)[5]
- bei weiterer Gleichheit: alphabetisch

Um die hier beschriebene Vorgehensweise zu evaluieren, wurde der manuelle Schritt 6 im Rahmen dieser Studie nur für einen Ausschnitt der gesamten nach Schritt 5 verbliebenen Kandidatenliste durchgeführt. Die Ergebnisse dieser Filterung sind im Abschnitt (4) beschrieben.

7. Expertenanalysen

Für jeden der nach der ersten manuellen Filterung (Schritt 6) verbliebenen Neologismuskandidaten müsste anschließend in Expertenanalysen untersucht werden, ob er die formalen Neologismuskriterien gemäß Steffens (2007) erfüllt und falls ja: in welchem Jahr sein faktisches Aufkommen liegt.

Dieser Schritt ist deutlich aufwändiger als Schritt 6 und wurde im Rahmen dieser Studie bislang noch nicht umgesetzt. Für eine abschließende Evaluation der gesamten hier beschriebenen Vorgehensweise wäre es hilfreich zu sehen, wie viele faktische Neologismen sich unter den nach Schritt 6 verbliebenen Kandidaten befinden.

(4) Ergebnisse der manuellen Filterung

Der manuelle Filterschritt 6 wurde er noch nicht auf die gesamte automatisch erstellte Kandidatenliste (also noch nicht auf alle 11.842 nach Schritt 5 verbliebenen Kandidaten) angewendet, sondern nur auf drei Teilbereiche:

---

4 Der M-Score ist ein interner Score, der einige der o.g. quantitativen Filterkriterien (Schritte 2 und 3) in sich vereint. Der M-Score selbst wurde bei den Filterkriterien jedoch nicht explizit verwendet, sondern hat primär deskriptive Zwecke. Formal definiert ist der M-Score als:

$$M = M_1 / c$$

mit $M_1 = (freq_B + 1) / (freq_A + 1)$

$c = (size_B + 1) / (size_A + 1)$

wobei $freq_A$ = Gesamtfrequenz des Wortes im Zeitraum 1990-1999

$freq_B$ = Gesamtfrequenz des Wortes im Zeitraum 2000-2008

$size_A$ = Korpusgröße im Zeitraum 1990-1999

$size_B$ = Korpusgröße im Zeitraum 2000-2008

5 Der LLR-Wert wird hier bestimmt für die beiden Zeiträume 1990-1999 und 2000-2008. Ein hoher LLR-Wert drückt in diesem Szenario also aus, dass das betreffende Wort überproportional häufig in einem der beiden Zeiträume vorkam (aufgrund der Filterkriterien ist dies bei den verbliebenen Wörtern stets der Zeitraum 2000-2008). Insofern leistet der LLR-Wert etwas ähnliches wie der M-Score: Er fasst verschiedene Filterkriterien zusammen. Wie der M-Score wurde jedoch auch der LLR-Wert bei den Filterkritierien nicht explizit verwendet.

A. Alle 1.915 Kandidaten mit "technischem Aufkommen" im Jahr 2006 (Zeilen 4744-6659 in der sortierten Kandidatenliste)
B. Die ersten 603 Kandidaten mit "technischem Aufkommen" im Jahr 2007 (Zeilen 3159-3761 in der sortierten Kandidatenliste)[6]
C. Die letzten 1.843 Kandidaten mit "technischem Aufkommen" im Jahr 2002 oder früher (Zeilen 10.000-11.842 in der sortierten Kandidatenliste)[7]

Die nach diesem Filtern verbliebenen Kandidaten sind in den drei Anhängen aufgelistet. Quantitativ ergab sich folgendes Bild wie in Tabelle 2 dargestellt, wobei hier nur die positiv markierten Kandidaten (also die vermuteten Neologismuskandidaten) gezählt werden.

Tabelle 2: Reduktionseffekte durch manuelles Filtern

Kandidaten-gruppe	Kandidaten vor Filtern	Verblieben nach Filtern	Quote
A	1.915	334	17,4%
B	603	93	15,4%
C	1.843	246	13,3%
Gesamt[8]	**4.361**	**673**	**15,4%**

Von dieser Stichprobe (4.361 manuell gesichtete Kandidaten) hochgerechnet auf alle 11.842 nach Schritt 5 verbliebenen Kandidaten würde man also bei einer vollständigen Filterung dieser Liste insgesamt ca. 1.800 verbleibende Kandidaten erwarten. Diese 1.800 hochgerechneten Kandidaten würden mit Sicherheit noch immer viele Nichtneologismen enthalten, die aber nur noch durch Expertenanalysen verlässlich weiter gefiltert werden können.

In der nach Schritt 6 verbliebenen Liste kommen einige potenzielle Kandidaten mehrfach vor, Beispiele finden sich auch bereits in der bearbeiteten Stichprobe im Anhang:

- verschiedene flektierte Wortformen: z.B. „Juniorprofessor“ vs. „Juniorprofessorin“
- verschiedene Schreibweisen: z.B. „youtube“ vs. „YouTube“

Verschiedene Oberflächenformen wurden in der automatischen Analyse als verschiedene Wörter behandelt (vgl. hierzu die Aussagen in Abschnitt 3 unter Schritt 0). In Schritt 6 wurden diese verschiedenen Oberflächenformen bislang noch nicht zusammengefasst. Grundsätzlich wäre dies aber durchaus eine Aufgabe, die sich zum Abschluss von Schritt 6 umsetzen ließe, ggf. unterstützt durch eine alphabetische Sortierung der Kandidaten.

(5) Filterschritte in Zahlen

In der folgenden Tabelle 3 ist dargestellt, wie viele Kandidaten die Liste nach dem jeweiligen Filterschritt umfasst. Von anfangs über 9 Millionen Wörtern konnte die Liste durch die automatischen Schritte sukzessive auf 11.842 Neologismuskandidaten reduziert werden, und durch den einfachen manuellen Filterschritt 6 würde sich diese Liste voraussichtlich nochmals auf ca.

6 In den Zeilen 3762-4743 befinden sich weitere Kandidaten, deren "technisches Aufkommen" im Jahr 2007 liegt.
7 In den Zeilen 9657-9999 befinden sich weitere Kandidaten, deren "technisches Aufkommen" im Jahr 2002 liegt.
8 Diese Zeile gibt die Zahlen für die gesamte in Schritt 6 manuell gesichtete Stichprobe an.

1.800 Kandidaten reduzieren, die anschließend nur noch durch Expertenanalysen weiter reduziert werden können.

Tabelle 3: Umfang der Kandidatenliste nach den einzelnen Filterschritten

Schritt	Filterkriterium	Verbliebene Kandidaten	
0.	– (Ausgangsliste: alle Wörter)	9.117.247	
1.	nichtlexikalische Wörter	8.853.946	
2.	seltene Wörter (Gesamtfreq. <20)	952.014	
3.	weitere quantitative Kriterien	29.593	
4.	Regionalismen i.w.S.	15.914	
5.	Eigennamen	11.842	
6.	manuell: eindeut. Nichtneologismen	ca.1.800	(Hochrechnung)
7.	Expertenanalysen	?	

Der Effekt der Eigennamenfilterung (Schritt 5) erscheint hier gering. Tatsächlich befanden sich anfangs jedoch äußerst viele Eigennamen in den Kandidatenliste. Die meisten von ihnen wurden bereits in den vorangegangenen Filterschritten herausgefiltert, insbesondere auch bei der Filterung von Regionalismen i.w.S. (Schritt 4). Wegen seines großen Rechenaufwands hat es sich jedoch bewährt, die Eigennamenfilterung als letzten der automatischen Filterschritte durchzuführen.

(6) Fazit

Die hier vorgestellte Methodik erscheint vielversprechend, weil – vorbehaltlich der noch ausstehenden Expertenanalysen – zum einen viele relevante Kandidaten gefunden wurden und zum anderen die ursprüngliche Liste durch die automatischen Filterschritte 1 bis 5 auf 0,13% seines Volumens kondensiert werden konnte und durch das einfache manuelle Nachfiltern gemäß Schritt 6 nochmals eine erhebliche Kondensierung zu erreichen ist (laut Hochrechnung auf ca. 15%).

Um das volle Potenzial dieses Ansatzes evaluieren zu können, müssten Filterschritt 6 vervollständigt und anschließend die Expertenanalysen für alle nach Schritt 6 verbliebenen Kandidaten durchgeführt werden.

Unabhängig von der Evaluation der Methodik können die im Anhang aufgelisteten Kandidaten unmittelbar als Quelle für konkrete lexikographische Projekte im Bereich Neologie dienen. Durch Vervollständigen von Schritt 6 könnte man diese Quelle natürlich nochmal deutlich vergrößern.

Literatur

Gries, Stefan Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics 13:4*, 403-437.

Keibel, Holger (2007). *Automatische Detektion von Neologismuskandidaten mit anschließender manueller Aufbereitung.* Internal Progress Report (IDS-KL-2007-01). Mannheim: Institut für Deutsche Sprache.

Steffens, Doris (2007). Von „Aquajogging“ bis „Zickenalarm“. Neuer Wortschatz im Deutschen seit den 90er Jahren im Spiegel des ersten größeren Neologismenwörterbuches. *Der Sprachdienst 4/07*, 146-159.

Anhang
In den folgenden drei Anhängen werden die nach dem manuellen Filtern (Schritt 6) verbliebenen Neologismuskandidaten der drei Kandidatengruppen A-C (vgl. S. 10) aufgelistet.

Anhang 1:
Kandidatengruppe A ("technisches Aufkommen" im Jahr 2006)

Nach dem manuellen Filtern verbliebene Kandidaten: 334
Agern, Airboat, Akkulaufzeiten, Alltagsbegleiter, Allwetterjacken, Alphatierchen, Alus, Amarettini, Angrillen, Antiterrordatei, ARGEn, Armutsgefährdung, Audiokommentare, Aufstallungsgebot, Aufstallungspflicht, Autogasanlage, Autogasfahrzeuge, Babyprämie, Bachelorstudiums, Beamershow, beknabbern, Bekos, Bergpartei, Berufsnavigator, Bespitzelungsaffäre, Bigpoint, bildungsfern, Biobewegung, Biokraftstoffindustrie, Biomarken, Bioobst, Bioöl, Blackberrys, blog, Bloggers, bloggt, Blogosphäre, blogspot, Blueliner, Bluetec, Breitbandinternet, Breitbandversorgung, Brennprogramm, Brennpunktviertel, Bringdienstes, Britrocker, Brüllfalle, Bündelangebote, Bürgernetzwerk, Bürogolf, Caches, Chartspitze, Chillis, Choreografieren, cim, Computerspielsucht, Datenklaus, Datenturbo, dazubuchen, Deleted, Demenzgarten, Dialogbeauftragte, Dialogbeauftragter, Dieselklau, Dinnerkrimi, Distanzregelung, dschihadistischen, durchgefeierten, durchklingeln, Durchregieren, durchregieren, Düsterrocker, eCall, EcoFuel, Ecofuel, Ehrenmordprozess, Einführungsworkshop, Eingliederungsmanagement, Einstiegsbenziner, Einstiegsqualifizierungsjahr, Ekelfleisch, Elternförderung, Energieausweis, Energieausweise, Energieausweises, EPAs, ePass, Episodendrama, Erlebnismuseums, Exzellenzclusters, Exzellenzwettbewerb, Exzellenzwettbewerbs, Fanmeile, Feinstaubalarm, Feinstaubbelastungen, Feinstaubemissionen, Feinstaubfilter, Feinstaubkonzentration, Feinstaubrichtlinie, Feinstaubverordnung, Feldbefreiung, Ferienresorts, Fernsehköchen, Feuerkörben, Fightnight, Firmware, Fitnessklub, Fitnesslevel, Flachbildfernsehern, flickr, Fotofalle, Frühchenstation, Galão, Gammelfleisch, Gammelfleischskandal, Gammelfleischskandale, Gänsehautfaktor, gedaddelt, Geheimflüge, Gemeinschaftsbaum, Gendreck, Genmaisfeld, Geomarketing, Gerechtigkeitsdefizit, Getreideheizung, Gewichtsmanagement, Gleitsicht, Gleitsichtbrillen, googlen, Greeters, Halloweenpartys, Handydaten, Handyfoto, Handyfotos, Handykamera, Handytarif, Handyverbindungen, Handywerk, Himmelslichter, Holzpelletanlagen, Homejacking, Hundeflüsterer, Hybridmodell, Hybridmodelle, Hybridversion, Identitätsdiebstahl, Indieszene, Industrieschnee, Internetgemeinschaft, Internetlexikon, Internettelefonie, Internetwache, Internetwahl, Internetzensur, Jobbers, Kaffeepads, Kandidatenwatch, kandidatenwatch, Kinderwunschzentrum, Kitesurf, Kitesurfer, Klickzahlen, Klimafalle, Klimaskeptiker, Knochenflicker, Kochschau, Kombiklasse, Kompetenzagentur, Kooperationsschulen, krawallige, Kreativwirtschaft, Kreditverkäufe, Krumping, Kundenhotline, Kurzzeitarbeitslose, Kurzzeitarbeitslosen, Lachyoga, Lärmkartierung, Laubpuster, Leasingautos, Lernlandschaft, Lernpartnerschaft, Lernpartnerschaften, Lichtmikroskopie, Lohas, Malware, Masterclasses, Masterstudiums, Mautausweichverkehr, Mautflucht, Mautflüchtlinge, Mehrgenerationenhäuser, Mehrgenerationenhäusern, Mehrgenerationenhauses, Mehrgenerationenwohnen, Mentalist, Mikrobizide, Mitternachtsshopping, Modesünden, Molekularküche, Mondscheinkinder, Musikhandys, Müze, Nachholfaktor, Naherholungseinrichtungen, Navigationsgeräts, Netzentgelt, Nichtraucherlokal, Niedrigseilgarten, Noroviren, Notfallkarte, Onlinebuchung, Onlineforen, Onlineforum, Onlinejournalismus, Onlineportals, Optionstickets, Oxyfuel, Palliativversorgung, Pandemieplan, Partyfotos, pestet, Petaflop, Pflegeassistenten, Pflegeassistentin, Pocketbikes, Podcast, Podcasting, Podcasts, Postpoint, Präventionssport, Preisrallye, Produktkommunikation, Rapex, Raucherkneipen, Rauschbrille, Rauschbrillen, Rechtsklick, Reggaeton, Registerkarte, Reiseportale, Rendition, Rezertifizierung, Riesterverträge, Rockolymp, Röhrenfernseher,

Rollkoffern, Rucksackbomber, Ruheoasen, Rundmails, Runflat, Saunaland, Schachboxen, Schadsoftware, Schaumkuss, Scheinväter, Schmerzmanagement, Schrottwichteln, Schweinerocker, Screamo, Seilgartens, Singlebörse, Soccerhalle, Sozialkaufhäuser, Speichersticks, Sprachförderkraft, Sprengstoffweste, Spritsorte, Spritspartraining, Stacking, Stevia, Stilikonen, Stippen, Studienkredit, Studienkredite, Suchmaschinenbetreibers, Sudoku, Sudokus, Supernanny, Tasern, Tauchcontainer, teambildende, Terrordrama, Terrorfahndern, Terrorlisten, Themendienst, Todespfleger, Traceure, Tributeband, Tropho, Trutsche, Tuss, Tütenträger, Twincharger, Überhit, umetikettieren, umgeparkt, Umweltplakette, Umweltplaketten, Umweltzone, Umweltzonen, unterbekommen, Unterschichtenfernsehen, Verbraucherbildung, Verbraucherportal, verfrachte, verknickt, Vogelgrippefälle, Vorrangprüfung, Vorratsdatenspeicherung, Wahlomat, Wahlstift, Wärmenetze, Waterboarding, waterboarding, Webradios, wegschenken, Wellfood, Wellnessbehandlung, Werbefolien, Whirlwanne, Whistleblowern, WiFi, Winterchecks, Wissensfabrik, Wochenbasis, WoW, Zahlungsraum, Zauberfuß, zauseliger, Zedernrevolution, Zeitarbeitsverhältnisse, Zukunftskonvent, Zukunftszentrum

Nach dem manuellen Filtern verbliebene Kandidaten mit ?-Markierung: 7
Amphe, Assir, Cowon, Dixg, Humvees, Spänner, Workingman

Anhang 2:
Kandidatengruppe B ("technisches Aufkommen" im Jahr 2007)

Nach dem manuellen Filtern verbliebene Kandidaten: 93
Abschiedswald, Agrarboom, anpassbar, Aufstocker, Biosprits, Bloggerin, Bodypainterin, Bonuskonto, Budgetnehmer, Castingband, computersüchtig, Datenspionage, Dinnershows, Dokukanal, doppisch, eGK, Einbürgerungstest, Erwerbsminderungsrentner, Facebook, Fashionweek, Feinstaubplakette, Flaschensammler, Freundesliste, Funktionsjacke, Gebrauchtkäufer, Gehaltsspirale, Grundsicherungsniveau, Gruselkick, Haushaltsassistenz, hochzuladen, Hotelbewertungen, hübscht, Integrationspolitiker, Internetriese, Internetriesen, Internetspiele, Jobaufbau, Kassenlandschaft, Konkurrenzformat, Kreativbranche, Lernbegleitung, Lernküche, ligatauglich, MacBook, Massenüberwachung, Matcha, Myspace, Naturerlebniszentrum, Netzwerkseiten, Nichtraucherkneipen, Niedriglohnschwelle, Ökostromanbieters, Paintballer, Plattformbetreiber, podcast, Profilseite, randarf, Raucherkneipe, Raucherlokal, Raucherlounge, Regenradar, Röhrenhose, runtergekühlt, Schadprogrammen, Schrottimmobilie, Schuldenatlas, Schülerrichter, Schulhelfer, Simulationsbrille, Skimmer, Sofortkauf, Sommermärchens, Splashdiving, Sprachförderkräfte, Sprachförderkräften, Sprachpaare, Spurhalteassistent, StudiVZ, Tafelarbeit, Ticketportal, Väterzeit, verkrumpelt, Videoblog, Wahlcomputern, Wellnessreise, Werkstorprinzip, Wiimote, Wohlfühlbad, Wohnfabrik, Wunschgroßeltern, Youtube, YouTube, Zertifizierten

Nach dem manuellen Filtern verbliebene Kandidaten mit ?-Markierung: 110
albopictus, Andasol, Ärztenetzwerk, berufundfamilie, Bestellbar, Bewegungspädagogik, Bildungsrepublik, Bindungsstörungen, Biokunststoffe, Bope, Chandrayaan, Dämpfungsfaktoren, Doppelsechs, dreizügiges, Durchgezogen, Erprobungsbetrieb, Erstversorger, Flächenkonkurrenz, Flächenpool, Ganztagsgymnasien, Gasrebellen, Gastchoreograf, Gebäudeenergieausweis, Gepanschter, Glühpunsch, Golddorf, googlemail, Grundversorgungstarif, Hauptumsatzbringer, Hobbyturnieren, Höhenkammer, Immobilienportal, Implantologen, Internetversorgung,

Investitionspakt, Kaffeespezialität, Kappenfest, Kaudroge, Kletterelemente, Klimaschutzauflagen, Klimaverträglichkeit, Klubnacht, Kochfans, Königsblauer, Konvergenzklausel, Kopi, Kunstreferentin, kuschele, Labormobil, Landschaftsführerin, Linkspolitikerin, Lirex, Lösekrug, Luwak, Markttreff, Marmal, MeinProf, Metalldieben, mitfahrzentrale, Modepolizei, Morrinho, Naturweg, Neubeschäftigten, Neuverschuldungsverbot, Nichtrauchergesetze, Nichtraucherschutzgesetze, Objektbeschichter, Offiziersanwärterin, Oldboy, Paketboxen, Phorms, Picasa, Piratenpartei, pleitegeht, Postdemokratie, Raumgestalterin, Regionaltöchter, Schlussmachen, Schooling, Schulbusbegleitern, Schuldnerquote, Schuldnerquoten, Schulhund, Schulkompromiss, Skysails, Slipstream, Smokys, Solitärpflanzen, Stadionluft, Stadtteilmanagerin, Strandsport, StuSta, Tanzbereiche, Tepee, Überlastungsanzeigen, Überleitungspflege, Unternehmensjurist, Verbrauchsausweis, Vertikalspiel, Vierspartenhaus, Waldsofas, Watcha, WeGebAU, Weiterbildungstages, Weltreiter, Wemm, youtube, Zaunfahne, Zielabweichung, Zweitteam

Anhang 3:
Kandidatengruppe C ("technisches Aufkommen" im Jahr 2002 oder früher)

Nach dem manuellen Filtern verbliebene Kandidaten: 246
Abnutzungskampf, Afterwork, Alternastage, Analoga, Antiterrorkampf, Antiterrorkampfes, Antiterrorkrieg, Antiterrormaßnahmen, Antiterrorpolitik, Antivirenprogramm, aufgehübschten, aufhübschen, Babyklappe, Babyklappen, Bachelorstudiengang, Bachelorstudium, Bandcontest, Beraternetzwerk, Bewohnerparken, Bezahlsenders, Bezahlstudium, Bezahlstudiums, Bezahlsystem, bildungsferner, Bildungskredit, Bildungskredite, Bildungsnetzwerk, Billigairline, Billigfliegers, Biokraftstoffe, Biosupermärkte, Biowärme, Blutdiamanten, Boulderwand, Breitbandanschlüsse, Breitbandanschlüssen, Brötchentaste, Caipi, Chancentod, chillige, Communitys, Dateianhang, Dateianhänge, Digicam, Dokusoap, Dönerfleisch, Dosenpfandes, Dosenpfands, Dotcom, Drogenscreenings, Dschihadis, Dschihadisten, Dunkelbar, durchzuschnaufen, Ehrenmord, Ehrenmorde, Einwegpfand, eLearning, Elternzeit, Energiewirte, Ersti, Essbereich, Eventmanagerin, Fallmanagerin, Fallmanagern, Familienmanagerin, Fanpaket, Festnetzsparte, Filesharing, Fitnesscoach, Fitnesskaufmann, Flatrates, Flatscreen, fluffig, Funkchip, Gänsehautatmosphäre, Gänsehautfeeling, Gänsehautgefühl, Gendoping, Genreis, Geopark, Gerichtsshows, Glamourfaktor, grenzdebilen, Hackschnitzelheizungen, Handykarten, Handykosten, Handykunden, Handyladen, Heroinambulanz, Heroinambulanzen, hochladen, Hochschulraum, Holzpellet, Informationsportal, Interimslager, Internetauftritts, Internetauktion, Internetauktionen, Internetchat, Internetfernsehen, Internetforen, Internetkamera, Internetkriminalität, Internetplattformen, Internetportale, Internetportals, Internetrecherchen, Internetshop, Internetspiel, Internetzeitung, Islamophobie, Juniorprofessoren, Juniorprofessorin, Juniorprofessuren, Kletterpark, Klingeltönen, Klonbaby, Klonbabys, Klonforscher, Kompetenznetz, Kompetenzteams, Kostenfallen, Lebensmitteldiscountern, Lernplattform, Leuchtturmprojekte, Lipgloss, Livestream, Liveticker, Mauterfassung, Mautsystems, Megacitys, Metrorapid, Microbus, Milchaustauscher, Mineralität, Mixtape, Monstertruck, Multiversum, Musikportal, Nanomaterialien, Navigationsleiste, Netzeitung, Netzwerkern, Nickname, Onlineausgabe, Onlinehändler, Onlinemagazin, Onlineshop, Onlineshops, Parkchip, Partnerfiliale, Partykrachern, Patchworkfamilien, Pelletheizung, Pelletkessel, Pendlerpauschale, Pilates, Plasmabildschirm, Praxisgebühr, Prekarität, Premiumprodukt, Presslinge, Publikumsjoker, Putztruppe, Quadfahrer, Quadrathlon, Rabattaktionen, Rabattschlacht, Realityshow, Regelenergie, Regelkinder, Retrowelle, Roadmap, Rotbusch, Ruheoase, Rundmail, Scheininnovationen, Schlafmünzen, Schwedenfeuern, Selbstmordbomber, Selbstmordpiloten, Separatorenfleisch, Servicehotline, Shootern, Shorttracker, simsen, Smartphone, Smartphones, Solarparks, Sparfuchs, Spartentarifvertrag, Spaßbremsen, Sportbars, Sprengstoffgürtel, Spyware,

Stammzellenforscher, Stammzellenforschung, Stammzellspende, Stammzellspender, Stammzelltherapie, Starterkits, Terroralarm, Terrorangriffs, Terrorangst, Terrorausbildung, Terrorbedrohung, Terrorfinanzierung, Terrorgefahren, Terrornetzwerk, Terrornetzwerke, Terrorverdächtige, terrorverdächtigen, Teuro, Themenpfad, Thementafeln, Todespilot, Todespiloten, toppten, Transfergesellschaften, Triband, Verkaufssender, verlinken, Videobotschaften, VoiceStream, Wachstumskerne, Wahlautomaten, Walkerin, Walkerinnen, Weblogs, Webportal, Weggucker, Wellnesscenter, Wellnessoase, Wellnessurlaub, Werbeflyer, Wesenstest, wirkungsgleich, Wissensshow, Wohlfühlprogramm, Wraps, Zickenalarm, Zickereien, Zielabweichungsverfahren, Zielabweichungsverfahrens, Zwangsheiraten

Nach dem manuellen Filtern verbliebene Kandidaten mit ?-Markierung: 3
Babytalk, Realisierungsgesellschaft, Tiefenhirnstimulation