– Ergänzungen zu –

Korpuslinguistik

Rainer Perkuhn / Holger Keibel / Marc Kupietz (2012):
Korpuslinguistik. Paderborn: Fink.
(Reihe LIBAC – Linguistik für Bachelor 3433).

Stand: 18. Juni 2012

Inhalt

E8 Sprachliche Strukturen aufspüren

E8.3 Kookkurrenzanalyse des IDS

Im Folgenden beschreiben wir ausführlicher als in unserem Buch den allgemeinen Ablauf des Verfahrens und führen dabei die Parameter ein, die bei Belicas Verfahren eingestellt werden können, jeweils eingebettet in den Schritt, bei dem sich die Auswirkung einer Entscheidung auf das Verfahren und die zugrundeliegende Fragestellung zeigt. Das Verfahren ist in das Recherchesystem des IDS integriert. Sämtliche gezeigten Beispiele basieren auf Kookkurrenzanalysen zu Recherchen im Gesamtarchiv des IDS, Stand Oktober 2010.

Ob Wortform oder Grundform wird für das Bezugswort vorweg bei der *Suche* festgelegt!

Im Standardszenario dreht sich alles um das lexikalische Material, der Gebrauch eines bestimmten Wortes (im Folgenden: Bezugswort; als Wortform oder Grundform, je nach Fragestellung) steht im Mittelpunkt. Ausgehend von diesem Wort wird der Ausschnitt definiert, der für die Auswertung der Kookkurrenz maßgeblich ist. Dazu muss das Analyseverfahren quasi in die Texte hineinschauen, in denen das Wort vorkommt. Unabhängig davon, wie dieser Schritt tatsächlich realisiert ist, kann man sich das am einfachsten so vorstellen, dass nach dem (Bezugs-)Wort gesucht wird und das Verfahren sich die Positionen merkt, an denen das Wort in den Texten / Korpora vorkommt. Ausgehend von diesen Positionen wird nun festgelegt, wie viele Wörter davor und danach für die Fragestellung berücksichtigt werden sollen – der zu untersuchende Kontext wird definiert.

Kontext definieren: wie viele Wörter links? wie viele Wörter rechts? Satzgrenzen beachten?

Neben der Anzahl, wie viele Wörter links und wie viele Wörter rechts von der Position des gefundenen Wortes zu betrachten sind, ist zu hinterfragen, ob das zu untersuchende Phänomen an Satzgrenzen Halt macht oder es womöglich erst zum Ausdruck kommt, wenn die Umgebung eines Wortes über Satzgrenzen hinweg betrachtet wird. Im ersten Fall sollte der zu untersuchende Kontext durch Satzgrenzen zwangsbeschnitten werden, d. h. wenn eine Satzgrenze linkerseits vorkommt, beginnt der Kontext erst danach, auch wenn eigentlich mehr Wörter davor gewünscht wären, wenn eine Satzgrenze rechterseits vorkommt, endet der Kontext spätestens dann, auch wenn danach noch mehr Wörter gewünscht wären (vgl. Abb. 8.1 auf der nächsten Seite).

Das Analyseverfahren wertet in einem ersten Schritt die so definierten Kontexte aus, in dem einfach alle vorkommenden Wortformen gezählt werden, wie oft sie insgesamt in der vorgegebenen Umgebung des Wortes verzeichnet sind (y). Für jede dieser Wortformen ist außerdem bekannt, wie oft sie in dem Gesamtkorpus (der Ausgangstextmenge, in der gesucht wurde) vorkommt (x). Neben diesen beiden Angaben greift nun eine statistische Bewertung auf zwei weitere Größen zurück, die ebenfalls bekannt sind: Die Anzahl der Wortformen im Gesamtkorpus (k) und die Anzahl der Wortformen (l) in der Gesamtheit aller Umgebungen um das gesuchte Wort herum. Die statistische Bewertung versucht nun die Frage zu beantworten: »Zu welchem Grad ist es noch im Rahmen des

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
es	am	dritten	Tag,	sprechen	fast	dieselbe	Sprache	wie	ihre	Vorgänger,	die	Wikinger,	vor	tausend
wären,	hätte	ich	es	im	Aufsichtsrat	zur	Sprache	gebracht.	Ich	habe	mich	darauf	verlassen,	daß
Land,	in	Zusammenhang	mit	dieser	Affäre	zur	Sprache	kamen.						
des	Datenschutzbeauftragten	jedoch	erst	sehr	spät	zur	Sprache.	Tebarth	hielt	sich	zurück	und	verwies	lediglich
den	Wahlen	überrunden.	Denn	er	spricht	die	Sprache	der	kleinen	Leute,	die	sich	immer	unterdrückt
sich	aber	auch	aus	der	Bewegung	heraus.	Sprache	und	Körper	bedingten	einander,	flossen	zusammen	und

Umgebung eines gefundenen Suchwortes

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
es	am	dritten	Tag,	sprechen	fast	dieselbe	Sprache	wie	ihre	Vorgänger,	die	Wikinger,	vor	tausend
wären,	hätte	ich	es	im	Aufsichtsrat	zur	Sprache	gebracht.	Ich	habe	mich	darauf	verlassen,	daß
Land,	in	Zusammenhang	mit	dieser	Affäre	zur	Sprache	kamen.						
des	Datenschutzbeauftragten	jedoch	erst	sehr	spät	zur	Sprache.	Tebarth	hielt	sich	zurück	und	verwies	lediglich
den	Wahlen	überrunden.	Denn	er	spricht	die	Sprache	der	kleinen	Leute,	die	sich	immer	unterdrückt
sich	aber	auch	aus	der	Bewegung	heraus.	Sprache	und	Körper	bedingten	einander,	flossen	zusammen	und

beidseitiger Kontext [-5,5], Satzgrenzen nicht beachten

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
es	am	dritten	Tag,	sprechen	fast	dieselbe	Sprache	wie	ihre	Vorgänger,	die	Wikinger,	vor	tausend
wären,	hätte	ich	es	im	Aufsichtsrat	zur	Sprache	gebracht.	Ich	habe	mich	darauf	verlassen,	daß
Land,	in	Zusammenhang	mit	dieser	Affäre	zur	Sprache	kamen.						
des	Datenschutzbeauftragten	jedoch	erst	sehr	spät	zur	Sprache.	Tebarth	hielt	sich	zurück	und	verwies	lediglich
den	Wahlen	überrunden.	Denn	er	spricht	die	Sprache	der	kleinen	Leute,	die	sich	immer	unterdrückt
sich	aber	auch	aus	der	Bewegung	heraus.	Sprache	und	Körper	bedingten	einander,	flossen	zusammen	und

einseitiger Kontext [-5,0], Satzgrenzen nicht beachten

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
es	am	dritten	Tag,	sprechen	fast	dieselbe	Sprache	wie	ihre	Vorgänger,	die	Wikinger,	vor	tausend
wären,	hätte	ich	es	im	Aufsichtsrat	zur	Sprache	gebracht.	Ich	habe	mich	darauf	verlassen,	daß
Land,	in	Zusammenhang	mit	dieser	Affäre	zur	Sprache	kamen.						
des	Datenschutzbeauftragten	jedoch	erst	sehr	spät	zur	Sprache.	Tebarth	hielt	sich	zurück	und	verwies	lediglich
den	Wahlen	überrunden.	Denn	er	spricht	die	Sprache	der	kleinen	Leute,	die	sich	immer	unterdrückt
sich	aber	auch	aus	der	Bewegung	heraus.	Sprache	und	Körper	bedingten	einander,	flossen	zusammen	und

beidseitiger Kontext [-5,5], Satzgrenzen beachten

Tab. E8.1: Beispiele für unterschiedliche Kontextdefinitionen

x = Anzahl Vorkommen im Gesamtbestand
k = Umfang Gesamtbestand
l = Umfang der (Treffer-)Umgebungen

zufällig Möglichen, dass eine Wortform w, die insgesamt x-mal in k-vielen Wortformen des Korpus vorkommt, y-mal in den l-vielen Wortformen eines Ausschnitts (in diesem Fall definiert über die Umgebung des Wortes v) vorkommt?« Was steckt hinter dieser so kompliziert anmutenden Formulierung? Wenn ich aus einer Menge von Ereignissen eine Stichprobe zufällig herausziehe, dann hat jedes Ereignis eine (zumindest kleine) Chance, mit in die Stichprobe hineinzukommen. Genauso ist es bei dem Verfahren zur Kookkurrenzanalyse auch: Wenn ich eine zufällige Menge Wörter aus unserem Korpus als Stichprobe herausziehe und mir darin die Wörter anschaue, hat jedes Wort aus der Grundgesamtheit eine gewisse Chance in der Menge beobachtet zu werden. Wenn ich also einen Kontext um ein Wort herum definiere, kann im Grunde jedes Wort per Zufall mit einer gewissen Anzahl Vorkommen mit hineinrutschen. Aber denken Sie jetzt noch einmal an das Beispiel mit den Rosinen im Kuchen: Eine bestimmte Anzahl Vorkommen plus einen Toleranzbereich passt mit unserer Erwartungshaltung überein. Die statistische Bewertung versucht auszudrücken, wie weit die beobachtete Anzahl im Kontext von der erwarteten abweicht, ab wann sie überzufällig groß ist und wie zuversichtlich die Statistik in die eigene Bewertung ist.

Ergebnis dieser ersten Bewertung ist eine Liste von Wörtern aus der Umgebung eines Bezugswortes mit ihrer Häufigkeit und ihrer statistischen Bewertung. Die *auffälligen* nennen wir *primäre Partnerwörter*.

Bezugswort »&rot«		
Wortform	Häufigkeit	statistische Bewertung *Auffälligkeit*
Ampel	6.433	51.273
grünen	3.393	13.074
grüne	3.034	12.132
Fahne	1.901	14.912
Fahnen	1.791	13.254
schreibt	1.774	2.873
Ampeln	1.680	14.833
schreiben	1.215	3.237
geschrieben	1.196	4.845
...	...	...

Abb. E8.2: Ausgewählte statistisch bewertete Kontextwörter

Parameter Lemmatisierung: Ob Wortform oder Grundform wird für die Umgebungswörter bei der *Analyse* festgelegt!

Ob die Wörter im Kontext als Wortformen oder statt ihrer ihre Grundformen ausgewertet werden sollen, ist vom Anfragenden vorzugeben. Im ersten Fall wird jede Wortform für sich gezählt und bewertet, im zweiten Fall werden die Häufigkeiten der beobachteten Wortformen eines Flexionsparadigmas kumuliert, somit werden diese Wortformen gemeinsam bewertet. Beide Entscheidungen lassen sich durchaus begründen, bergen aber eventuell auch kleine Schwierigkeiten. Wortformen sind näher am tatsächlich vorliegenden empirischen Material, eine darauf aufsetzende

Auswertung spiegelt also authentischer wider, wie es sich im Rohmaterial verhält. Eine kleine Gefahr besteht darin, dass einzelne Wortformen so selten auftreten, dass die statistische Bewertung keine Aussage wagt. Außerdem wirkt es natürlich ein wenig lästig, wenn eine Analyse mehrere Ergebnisse hervorbringt, die sich nahezu entsprechen bis auf den Umstand, dass z. B. dasselbe Verb (oder Adjektiv oder Substantiv) in verschiedenen flektierten Formen vorkommt. Hier würde man sich idealerweise wünschen, dass diese ähnlichen Ergebnisse zu einem zusammengefasst würden und statt der verschiedenen flektierten Formen ihre Grundform angegeben wird. So wie man bei einem Lexikoneintrag davon ausgeht, dass dieser in einer Nennform angegeben wird, verhält es sich auch bei Mehrworteinheiten (Redewendungen, Valenzen, Kollokationen): Auch hier ist es durchaus üblich eine kanonisierte Nennform anzugeben. Manchmal kann diese Darstellung aber auch kontraintuitiv wirken, da die Beziehung nur zwischen konkreten Wortformen gilt. Dies kann so stark sein, dass sie zwar auch auf Grundformebene statistisch erfasst werden kann; wenn keine Beziehung zu den anderen Formen des Paradigmas vorliegt, kann die Darstellung eher irreführend sein. Desweiteren kann ein (kleines) verfahrenstechnisches Problem auftreten: die Kumulierung bezieht sich sowohl auf die Seite der Wörter im Kontext (sozusagen positive Beispiele für die Kookkurrenzbeziehung), als auch auf die Seite außerhalb des Kontextes (sozusagen negative Beispiele gegen Kookkurrenzbeziehung). Falls zufällig eine kleine Anzahl eines eigentlich sehr häufigen negativen Beispiels in den Kontext hineingerutscht ist (siehe oben, das ist immer möglich!), wiegen die restlichen zu Recht außerhalb des Kontextes liegenden negativen Beispiele der Wortform plötzlich auch als Gegenbeispiel gegen eine Kookkurrenzbeziehung zum gesamten Paradigma! Eine grundsätzliche Gefahr besteht darin, dass die Unterscheidung zwischen den eigentlichen Daten (soweit es geht roh) und einer Interpretation verloren geht. Denn der Begriff einer *Grundform* ist das Ergebnis einer Interpretation (s. Kapitel Relevanz), die sich nicht vollständig formalisieren und operationalisieren lässt. Es lässt sich nicht vermeiden, dass ein automatisches Verfahren Wortformen anders zu Grundformen zuordnet, als es zu erwarten ist. Selbst mehrere menschliche Bewerter könnten sich in vielen Fällen nicht einigen. Und gerade im Bereich der Neologismen befindet sich noch so viel im Umbruch, dass das Paradigma eines Wortes noch gar nicht festgelegt oder antizipiert werden kann. In diesen Fällen wäre es also auch irreführend von einer Kookkurrenzanalyse sinnvolle Ergebnisse über Grundformen zu erhoffen, solange es kein Verfahren gibt, dass die Grundformen zuverlässig bestimmen kann.

Wortformen eines Paradigmas, die für sich genommen nicht statistisch auffällig waren, fließen bei der Lemmatisierung nicht mit in die kumulative Bewertung ein.

Die Lemmatisierung ist in diesem Zusammenhang auf die Flexion beschränkt. Weitere Wortbildungsformen miteinzubeziehen wäre nur vor dem Hintergrund sehr spezieller Fragestellungen sinnvoll.

Bezugswort »&rot«				
Wort-form	Häufig-keit	Grund-form	kumulierte Häufigkeit	statistische Bewertung
grünen	3.393			
grüne	3.034			
grüner	577	_grün	11.591	39.227
grünem	353			
grünes	447			
Fahne	1.901	_Fahne	3.869	28.948
Fahnen	1.791			
Ampel	6.433	_Ampel	8.655	63.510
Ampeln	1.680			
geschrieben	1.196			
schreiben	1.215			
schreibt	1.774	_schreiben	5.538	9.704
schreibenden	64			
schreibende	54			
schreibe	168			
...	...	...	...	...

Abb. E8.3: Ausgewählte, als Lemma gemeinsam statistisch bewertete Kontextwörter

Wenn wir beliebige Textausschnitte wählen, fast egal wie kurz diese sind, und schauen, welche Wörter dort häufig vorkommen, stellen wir fest, dass eine recht kleine, überschaubare Menge verstärkt anzutreffen ist: Artikel, Konjunktionen, Präpositionen, Pronomen u.Ä. – kurzum die sogenannten Funktionswörter. Das dies so ist, darf aber auch nicht überraschen, denn diese Wörter sind an sich sehr häufig. Genau an diesem Punkt haben die meisten statistischen Bewertungen aber gerade mit gewissen Schwierigkeiten zu kämpfen: Für besonders häufige Wörter ist es sehr schwierig, mit großer Zuversicht eine Aussage zu machen, ob eine bestimmte Anzahl noch im Bereich des zufällig Möglichen liegt oder nicht.

Parameter Funktionswörter

Im Grunde genommen liefern Kookkurrenzen mit Funktionswörtern eine andere Aussicht auf Erkenntnisgewinn als Kookkurrenzen mit Inhaltswörtern: Bei ihnen geht es meistens um morphologische/syntaktische Kongruenzen. Um das Augenmerk des Betrachters stärker auf die inhaltlichen Zusammenhänge, die sich hinter einer Kookkurrenz verbergen, lenken zu können, bietet das Verfahren von Belica die Möglichkeit, Funktionswörter für die weitere Betrachtung auszublenden. An einer späteren Stelle können sie durchaus wieder ins Spiel kommen, aber sie tragen auf diese Weise nicht dazu bei, das Ergebnis einer Analyse unnötig aufzublähen. Manchmal wäre es aber wünschenswert, Funktionswörter differenzierter handhaben zu können: Für manche Fragestellungen

(z. B. bei Untersuchungen zur Valenz) mag es wichtig sein, Präpositionen miteinzubeziehen, andere Funktionswörter aber nicht. Geplant ist, die Möglichkeit zu bieten, interaktiv die Menge der zu ignorierenden Funktionswörter zusammenzustellen, wobei verschiedene Taxonomien berücksichtigt werden sollen und homonyme Formen eine besondere Handhabung erfordern.

Die Liste der Wörter in der Umgebung des vorgegebenen Bezugswortes (eventuell zu Grundformen zusammengefasst und um Funktionswörter reduziert) kann jetzt absteigend nach der Stärke der statistischen Auffälligkeit (LLR, Log-Likelihood-Ratio) sortiert werden. Die obersten Einträge stellen die auffälligsten Verbindungen dar. Je weiter man die Liste aber nach unten durchgeht, desto deutlicher wird, dass eigentlich immer unklarer wird und der Betrachter sich immer unsicherer wird, worin eine Auffälligkeit eigentlich begründet sein soll. Im Grunde geht es der Statistik genauso: Die Zuversicht der statistischen Bewertung, dass es sich tatsächlich um ein auffälliges gemeinsames Vorkommen handelt, nimmt mit abnehmenden Werten ab. Diese Zuversicht verbindet man in der Statistik mit entsprechenden Signifikanzniveaus. Es ist z. B. üblich, von einem Signifikanzniveau von 99%, von 99,9% oder 99,99% auszugehen. Welche LLR-Werte und welche Niveaus tatsächlich zum Tragen kommen, sollte für einen Betrachter unerheblich sein. Es ist aber durchaus sinnvoll, zum einen eine Sortierung nach LLR-Werten anzubieten und auswählen zu lassen, welche Signifikanzniveaus vorgegeben werden kann. Die Sortierung nach LLR ist bei Belicas Verfahren die Voreinstellung. Auch wenn die Werte für sich betrachtet absolut wenig aussagekräftig erscheinen (im Laufe der Zeit wird man zumindest eine Gefühl für die Größenordnungen bekommen), so sind doch die Verhältnismäßigkeiten und die Abstände zueinander ein Indikator für unterschiedlich starke Auffälligkeiten. Die Vorgabemöglichkeit verschiedener Signifikanzniveaus ermöglicht die Entscheidung, ob lieber eine möglichst umfangreiche Menge von Wortverbindungen gewünscht ist, die allerdings mit mehr Aufwand durchgearbeitet werden muss und in der sicher auch einige unzuverlässige Ergebnisse sein können – oder ob man sich auf eine kleinere Menge beschränken kann, die zwar zuverlässiger gute Ergebnisse liefert, dabei allerdings in Kauf nehmend, dass gewollte, sinnvolle Ergebnisse herausfallen, nur weil unglückliche Umstände zusammengekommen sind. Über den Parameter *Zuverlässigkeit* kann also eingestellt werden, ob beim Identifizieren von primären Partnerwörtern eher Wert auf höhere Präzision oder auf einen größeren Recall gelegt werden soll (vgl. S. 40 in unserem Buch).

LLR = Log-Likelihood-Ratio

Parameter Zuverlässigkeit

Bei der untersten Stufe (»analytisch«) werden genaugenommen noch weitere Bedingungen aufgeweicht. Statt der für LLR üblichen Mindestanzahl in Grundgesamtheit bzw. im Ausschnitt wird eine geringere Anzahl Vorkommen gefordert. Dies hat den Effekt, dass die Liste nicht nur nach unten länger wird, sondern sich auch in den oberen Bereichen Einträge einschieben können, die es bei den anderen Einstellungen nicht gegeben

Bezugswort »&Spott«					
Wort	Häufigkeit	statistische Bewertung (LLR)	Zuverlässigkeitsstufe		
Hohn	2.660	39.298	hoch	normal	analytisch
...	...	...			
manchmal	35	40			
Angst	49	39	ignoriert		
...	...	...			
zwangsläufig	6	6			
abbekommen	2	5		ignoriert	
...	...	...			
zusehends	2	1			
herablassende	...	...			ignoriert
...	...	...			
(»statistisch unspezifisch«)	...	...			

Tab. E8.2: Verschiedene Einstellungen des Parameters *Zuverlässigkeit*

hat. Es wird also insgesamt mehr Kandidaten eine Chance eingeräumt, ins Blickfeld der Betrachtung zu kommen – auch solche, die für die anderen Einstellungen zu selten sind. Jedoch sind diese neuen Kandidaten sehr mit Vorsicht zu genießen, da bei ihnen Annahmen, die die statistische Bewertung macht, vielleicht gar nicht gelten. Insofern sollte man diesen Parameter nur dann wählen, wenn ein Versuch mit einem härteren Parameter ein wenig aussagekräftiges Bild geliefert hat. Und auch dann sollte man die Vorschläge noch defensiver als bei den anderen Parametern nur als Hinweise deuten und sich stets vergegenwärtigen, dass die statistische Bewertung vielleicht daneben gelegen haben kann, weil von ihr mehr gefordert wurde als sie eigentlich bieten kann.

Als Zwischenergebnis liegt jetzt quasi eine Liste von signifikanten, sogenannten primären Partnerwörtern vor. Mit Hilfe dieser Auswahl wird gezielter weiter geschaut, welche komplexeren Wortverbindungen sich als auffällig erweisen. Für die weitere Betrachtung werden nunmehr die Wechselwirkungen zwischen diesen Wörtern bewertet. Für jedes Partnerwort ist bekannt, wie oft es in dem betrachteten Textfenster (Kontext um das Suchwort herum) vorkommt. Die Statistik hat dann diese Häufigkeit mit der Häufigkeit in der Grundsamtheit verglichen und bewertet. In einem weiteren Schritt stellen wir nun eine leicht veränderte Frage. Für jedes gegebene Partnerwort möchten wir nun wissen, ob es Zufall sein kann, dass ein weiteres Partnerwort mit einer bestimmten Häufigkeit in dessen Nähe innerhalb der Treffermenge beobachtet wurde – wobei die Nähe zum Suchwort damit implizit einhergeht, da sich ja darüber die Treffermenge erst ergeben hat. Dabei soll die Statistik nun die Häufigkeiten der anderen Partnerwörter in der Nähe eines gegebenen Partnerwortes zu deren Häufigkeit in der Treffermenge vergleichen und bewerten. Die Tabelle

illustriert dies an dem Beispiel »Schaden«, das als Partnerwort zu »&Spott« bereits erkannt wurde. Jeder Eintrag zu jedem der anderen Partnerwörter spiegelt durch eine Anzahl Sternchen die Stärke der Wechselwirkung wieder, zu sich selbst ist die Wechselwirkung natürlich quasi unendlich. Wir verzichten hier (und ebenso die Implementierungen des Verfahrens etwa in den COSMAS-Systemen) auf die Angabe konkreter Zahlenwerte, da die Erklärung ihrer Berechnung den Rahmen dieser Einführung sprengen würde und sie auch nur sehr schwierig zu interpretieren sind. Die Größenordnungen der Grundgesamtheiten und der Häufigkeiten sind deutlich kleiner als bei der Bewertung des primären Partners. Obwohl sich dafür für die für uns zur Verfügung stehenden sehr großen Datenmengen LLR als prädestiniert herausgestellt hat, muss bei den weiteren Partnern den kleineren Zahlen Rechnung getragen werden. Deshalb wird hier ein gewichtetes Mittel aus LLR und *mutual information* bestimmt.

Wechsel-wirkung	Schaden	…	Zum	…	kommt	…	noch	…	Wer	…	kam	…
Schaden	∞	*	* * * * * *	*	**	*	* * *	*	* * * * * **	*	* * * *	*

Tab. E8.3: Verschieden starke Wechselwirkungen zwischen dem primären Partnerwort »Schaden« und anderen primären Partnerwörtern

In Belicas Verfahren kann analog zur Zuversicht für das statistische Maß für den primären Partner eine Zuversicht in dieses gewichtete Maß für die weiteren Partner als Granularität eingestellt werden. Für dieses sind zur Zeit vier Stufen vorgesehen von sehr grob über grob und mittel bis hin zu fein, bei denen jeweils ansteigend umfangreichere Mengen von potenziellen weiteren Partnern in Betracht gezogen werden. Dadurch ergibt sich, dass die groben Einstellungen nur kurze, schlagwortartige Verbindungen aufdecken, während die feineren Einstellungen auch längere Formulierungen auftun – wobei die Zuversicht, dass die vorgeschlagenen Bestandteile der Wortverbindungen auch relevante Bestandteile sind, abnimmt. Bitte beachten Sie, dass die Komplexität der verschiedenen Wortverbindungskandidaten auch (indirekt) von den Einstellungen zur Zuverlässigkeit und den Funktionswörtern abhängen!

Parameter Granularität

In weiteren Runden wird für das bisher schon erkannte bewertet, ob weitere Wörter sich überzufällig zu dem schon festgehaltenen dazugesellen, und so immer komplexere Wortverbindungen bilden. So ist z. B. das Wort »glänzt« in der Nähe des Wortes »&Gold« auffällig, in der Nähe dieser beiden wird zusätzlich das Wort »alles« ausgemacht. Dies wird jeweils mit dem Ergebnis des vorhergehenden Schritte wiederholt, bis keine weiteren Wörter mehr hinzukommen. So treten zu der bisher erkannten Partnerwortfolge nacheinander »was« sowie noch weitere Wörter hinzu.

<table>
<tr><td colspan="5">allgemeines Schema</td></tr>
<tr><td>gemittelte Bewertung (LLR und MI)</td><td colspan="4">Granularitätsstufe</td></tr>
<tr><td>* * * * * * * * * * * * * * * * * * *
* * * * * * * * * * **</td><td>sehr grob</td><td rowspan="2">grob</td><td rowspan="3">mittel</td><td rowspan="4">fein</td></tr>
<tr><td>* * * * * * **
* * * * **</td><td rowspan="4">ignoriert</td></tr>
<tr><td>* * **
* * *</td><td rowspan="3">ignoriert</td></tr>
<tr><td>**
**</td><td rowspan="2">ignor.</td></tr>
<tr><td>*
*</td><td>ign.</td></tr>
</table>

Tab. E8.4: Verschiedene Einstellungen des Parameters *Granularität*

Auch wenn die Reihenfolge zunächst etwas verwirrend anmuten mag: Wenn wir die Wörter »&Gold, glänzt, alles, was« ein wenig im Kopf jonglieren lassen, kommen wir schnell auf die Redewendung »Es ist nicht alles Gold was glänzt«. Diese Fügung hat durch ihren regelmäßigen wiederholten Gebrauch dieser Wörter in Kombination dazu geführt, dass die statistische Bewertung diese als auffällig ausgemacht hat.

<table>
<tr><td colspan="8">Bezugswort »&Gold«</td></tr>
<tr><td>primärer Partner</td><td>statistische Bewertung</td><td>sekundärer Partner</td><td>statistische Bewertung</td><td>tertiärer Partner</td><td>statistische Bewertung</td><td>… Partner</td><td>statistische Bewertung</td></tr>
<tr><td rowspan="12">glänzt</td><td rowspan="12">10.415</td><td rowspan="6">alles</td><td rowspan="6">> grob</td><td rowspan="5">was</td><td rowspan="5">> grob</td><td>Skythen</td><td>> fein</td></tr>
<tr><td>echtes</td><td>> fein</td></tr>
<tr><td>schimmert</td><td>> fein</td></tr>
<tr><td>silbern</td><td>> mittel</td></tr>
<tr><td>China</td><td>> fein</td></tr>
<tr><td>funkelt</td><td>> mittel</td><td></td><td></td></tr>
<tr><td>Silberschmuck</td><td>> fein</td><td>funkelt</td><td>> mittel</td><td></td><td></td></tr>
<tr><td>was</td><td>> grob</td><td>funkelt</td><td>> mittel</td><td></td><td></td></tr>
<tr><td>echtes</td><td>> fein</td><td></td><td></td><td></td><td></td></tr>
<tr><td>glitzert</td><td>> mittel</td><td></td><td></td><td></td><td></td></tr>
<tr><td>funkelt</td><td>> mittel</td><td></td><td></td><td></td><td></td></tr>
<tr><td>China</td><td>> fein</td><td></td><td></td><td></td><td></td></tr>
</table>

Tab. E8.5: Komplexe Wortverbindungen zu Bezugswort »&Gold« und primärem Partner »glänzt«

Komplett sähe das bisherige Zwischenergebnis schematisch so aus: eine Liste von Folgen von primären und weiteren Partnerwörtern mit ihren jeweiligen statistischen Bewertungen. Der Übersichtlichkeit halber (und wegen der noch schwierigeren Interpretierbarkeit) werden die statistischen Bewertungen der nicht-primären Partner aber nicht präsentiert.

Bezugswort »&Gold«				
primärer Partner	statistische Bewertung	sekundärer Partner	tertiärer Partner	... Partner
glänzt	10.415	alles	was	Skythen
glänzt	10.415	alles	was	echtes
glänzt	10.415	alles	was	schimmert
glänzt	10.415	alles	was	silbern
glänzt	10.415	alles	was	China
glänzt	10.415	alles	funkelt	
glänzt	10.415	Silberschmuck	funkelt	
glänzt	10.415	was	funkelt	
glänzt	10.415	echtes		
glänzt	10.415	glitzert		
glänzt	10.415	funkelt		
glänzt	10.415	China		

Tab. E8.6: Komplexe Wortverbindungen zu Bezugswort »&Gold« und primärem Partner »glänzt«

Die einzigen Unterschiede dieser letzten Form zu den Darstellungsarten, wie Sie sie über die verschiedenen Nutzeroberflächen erfahren, besteht darin, dass die statistische Bewertung des primären Partnerwortes üblicherweise vorangestellt wird, gefolgt von einem Feld, in dem die Folge der Partnerwörter eines Syntagmas nur durch Leerzeichen getrennt aufgelistet werden.

Achtung!
Die angegebene statistische Bewertung bezieht sich stets auf das primäre Partnerwort!

Je nach eingestellter Granularität ergeben sich unterschiedlich komplexe Strukturen, die dem Anfragenden präsentiert werden. In der Tabelle E8.7 sind spaltenweise die entsprechenden Sichten auf die obige Gesamtstruktur dargestellt. Idealerweise sollten bereits mit den Einstellungen »grob« oder ggf. auch »mittel« die deutlichsten Anzeichen für feste Wortverbindungen aufgedeckt werden. Die beiden anderen Einstellungen sollten nur dann ins Spiel kommen, wenn sich für eine konkrete Analyse die Erstgenannten als nicht hilfreich erwiesen haben.

Die erste Nennung eines primären Partners wird typographisch hervorgehoben (z. B. fett), um den Übergang von einem Teilergebnis der Analyse zum nächsten zu markieren.

Die Texte, in denen signifikante Partnerwörter erkannt wurden, werden jetzt diesen Folgen zugeordnet. Aufgrund der in einem konkreten Kontext vorliegenden Wörter kann es jedoch sein, dass dieser Text mehreren Folgen zugeordnet werden könnte. Dies ist auch ein durchaus plausibles Vorgehen, hat allerdings den kleinen Nachteil, dass Einheiten entstehen, die für sich betrachtet eigentlich keine Daseinsberechtigung haben,

Partnerwortfolgen zu Bezugswort »&Gold« und primären Partner »glänzt«			
fein	mittel	grob	sehr grob
glänzt alles was Skythen			
glänzt alles was echtes			
glänzt alles was schimmert			
glänzt alles was silbern	**glänzt** alles was silbern		
glänzt alles was China			
glänzt alles was	glänzt alles was	**glänzt** alles was	
glänzt alles funkelt	glänzt alles funkelt		
glänzt alles	glänzt alles	glänzt alles	
glänzt Silberschmuck funkelt			
glänzt Silberschmuck			
glänzt was funkelt			
glänzt was	glänzt was	glänzt was	
glänzt echtes			
glänzt glitzert			
glänzt funkelt			
glänzt China			
glänzt	glänzt	glänzt	**glänzt**

Tab. E8.7: Komplexe Wortverbindungen zu Bezugswort »&Gold« und primärem Partner »glänzt« – für verschiedene Granularitäten zeilenweise aus Tabelle E8.5 herausgelesen

n-Tupel ↓

Partner zu v: (a b c)	
. . . b v c a . . .	Cluster
. . b . v c . a . .	
. . a b v . . c . .	

weil sich diese erst aus den komplexeren Verbindungen ergibt. Unschön ist weiterhin, dass sich Häufigkeiten übergeordneter Einheiten nicht mehr als Summe der Häufigkeiten der untergeordneten Einheiten ergeben, was bei der Einschätzung dieser Angabe manchmal irritierend sein mag. Als Alternative, die diese beiden Unannehmlichkeiten verhindert, könnten die Texte auch folgendermaßen eindeutig zugeordnet werden: Absteigend nach statistischer Bewertung und Komplexität wird eine Textstelle der Folge zugewiesen, die als erstes passt. Damit ist die Textstelle sozusagen verbraucht und kann an keiner anderen Stelle mehr zugewiesen werden.

Parameter Zuordnung

Die Verteilung übernimmt ein sogenanntes Clusteringverfahren je nach Einstellung des Parameters *Zuordnung*. Die entstandene Einheit aus Textstellen, Folge der Partnerwörter und den weiteren quantitativen Angaben nennen wir *(Text-)Cluster*.

Partner zu v: (a b)
. . b . v . a . .
. . a . v b . . .
. . a b v . . c . . (*)

(*) nur bei *mehrfach*

Die eindeutige Zuordnung birgt die kleine Gefahr, dass theoretisch denkbare und sinnvolle Einheiten unterdrückt werden, da alle die Wortverbindung enthaltenden Textstellen bereits für andere Einheiten verbraucht wurden. Dadurch kann es auch passieren, dass Textcluster aufgeführt werden, die weniger Textstellen umfassen, als als Schwellwert für die entsprechende Zuverlässigkeitsstufe erforderlich gewesen wären. Das erklärt sich ganz einfach: Bei der statistischen Bewertung wurden noch alle Textstellen berücksichtigt, beim Clustering wurde aber ein Teil der Textstellen bereits an anderer Stelle verbucht.

Ein Textcluster fasst also eine Menge von Texten zusammen, die eine bestimmte Menge von Wörtern innerhalb eines Textfensters gemeinsam haben, die in ihrer Kombination und im Vergleich zur Grundgesamtheit überzufällig gemeinsam vorkommen. Ein Cluster wird mit der Aufzäh-

Cluster zu Bezugswort »&Gold« und primären Partner »glänzt«					
eindeutige Zuordnung	Textstelle	mehrfache Zuordnung			
glänzt was	Manchmal ist es Gold, was glänzt ist das wohl Gold, was darin so glänzt Es ist nicht immer Gold, was glänzt		glänzt was		glänzt
glänzt alles was	hier ist wirklich alles Gold, was glänzt Nicht alles ist Gold, was glänzt Es ist nicht alles Gold, was glänzt Hier hat alles Gold, was im Glase glänzt Nicht alles Gold, was glänzt Keineswegs glänzt alles, was Gold sein soll	glänzt alles was		glänzt alles	
glänzt alles	nicht alles so Gold ist, wie es derzeit glänzt bis alles glänzt wie Gold alles glänzt bereits in herbstlichem Gold				
glänzt	Warum glänzt Gold gelb ...				

Tab. E8.8: Unterschiedliche Zuordnung von Textstellen

lung dieser auffälligen Wörter benannt (sogenannte n-Tupel), wobei diese wohlgemerkt an unterschiedlichen Stellen in den Texten vorkommen können. Je nach Einstellung des Parameters *Zuordnung* umfasst ein Cluster alle (Einstellung: mehrfach) oder nur einen Teil (Einstellung: eindeutig) der verantwortlichen Textstellen. Dabei entstehen (im zweiten Fall nicht immer) Gruppen von Texten, deren n-Tupel quasi Fortsetzungen voneinander sind. Über diese Fortsetzungsrelation lassen sich die Cluster hierarchisch anordnen.

hierarchische Anordnung:

```
      (a b)
      /   \
(a b c)   (a b d)
```

Falls Sie beim Betrachten eines Textclusters sich bei einzelnen Textstellen wundern sollten, warum diese in dieses Cluster und nicht in ein anderes eingeordnet worden sind, so sollten Sie zunächst drei Punkte überprüfen:

- Wurden Wortformen bei der Auswertung zu Grunde gelegt und, wenn ja, stimmt bei allen Parnterwörtern die Groß- bzw. Kleinschreibung?
- Befinden sich alle Wörter, die Sie als Partnerwörter erwarten, tatsächlich im vorher festgelegten Kontext?
- Wird der gesamte Kontext im KWIC angezeigt oder wird er an der Seite abgeschnitten?

Wie viele Textstellen das Clusteringverfahren aufgrund der Zuordnungseinstellung zu den jeweiligen Clustern zusammengefasst hat, wird schließlich als Anzahl angezeigt. Diese Angabe bezieht sich also jeweils auf eine Zeile. Bei der Mehrfachzuordnung sind die Zahlen zu den kürzeren Partnerwortfolgen größer; je nach Festigkeit der Wortverbindung kann es bei der eindeutigen Zuordnung sich durchaus komplementär verhalten.

Anzahl

Bei der Erläuterung zu dem statistischen Maß für die Bewertung der primären Partner sollte ein Zusammenhang noch präsent sein: Zwei wichtige Größen sind die Anzahl der Vorkommen in dem betrachteten Ausschnitt und der Umfang des Ausschnitts. Die statistische Bewertung fällt höher

Bezugswort »&Gold«, Z: hoch; G: mittel; F: +; AF: -		
statistische Bewertung	Anzahl	Partner
10415	2	glänzt alles was silbern
10415	955	glänzt alles was
10415	1	glänzt alles funkelt
10415	966	glänzt alles
10415	1	glänzt was funkelt
10415	989	glänzt was
10415	3	glänzt glitzert
10415	3	glänzt funkelt
10415	1243	glänzt

Tab. E8.9: Komplexe Wortverbindungen zu Bezugswort »&Gold« und primärem Partner »glänzt«

aus, wenn

- bei gleich großem Ausschnitt mehr Vorkommen gezählt werden oder
- wenn gleich viele Vorkommen in einem kleineren Ausschnitt gezählt werden.

Ausschnittumfang
(5 + 5) ∗ 1.000 = 10.000

x x x x x v x x x x x
x x x x x v x x x x x
…
x x x x x v x x x x x
x x x x x v x x x x x

Ausschnittumfang
(5 + 5) ∗ 500 = 5.000

x x x x x v x x x x x
…
x x x x x v x x x x x

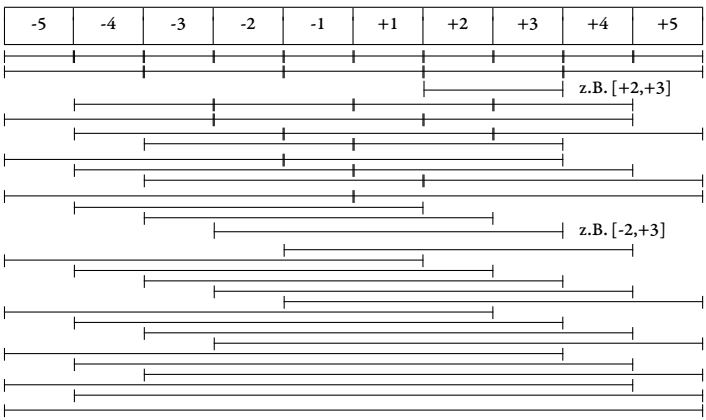
Ausschnittumfang
(5 + 0) ∗ 1.000 = 5.000

x x x x x v
x x x x x v
…
x x x x x v
x x x x x v

Wenn z. B. ein (Bezugs-)Wort gesucht wurde und 1.000 Mal gefunden wurde, der Kontext auf 5 Wörter rechts und links eingestellt ist, dann enthält der Ausschnitt 10.000 Wörter. Ein Partnerwort, dass in dem Ausschnitt 50 Mal beobachtet wurde, wird als weniger auffällig bewertet als ein Partnerwort, dass in dem Ausschnitt 100 mal beobachtet wurde (gegeben die gleiche Häufigkeit in der Gesamtmenge). Ein Partnerwort, das 50 Mal in einem kleineren Ausschnitt beobachtet wurde, würde auch höher bewertet (ebenfalls gegeben die gleiche Häufigkeit in der Gesamtmenge). Ein kleinerer Ausschnitt kann dadurch zustande gekommen sein, dass das Bezugswort seltener gefunden wurde. Gab es etwa nur 500 Treffer, dann bewertet die Statistik das Vorkommen von 50 Wörtern in 5.000 höher als im Vergleich zu vorher 50 Wörter in 10.000. Einen vergleichbaren Effekt kann man natürlich die Verkleinerung des Kontextes herbeiführen: Hätte man nur 5 Wörter links und kein Wort rechts eingestellt, würde dieselbe höhere Bewertung herauskommen. Der Haken bei der Sache ist nur, dass man für gemeinhin nicht vorher weiß, an welchen Positionen ein Partnerwort bevorzugt auftritt. Gerade die deutsche Sprache ist in der Hinsicht schwieriger zu handhaben als manche andere. Man müsste also eigentlich die Analysen mit möglichst vielen verschiedenen Kontexten durchführen und dann die Ergebnisse untereinander vergleichen. Abgesehen davon, dass dieses Vorgehen sehr mühsam und aufwändig wäre, gibt es noch einen weiteren (zugegeben etwas pragmatischen) Punkt, der eine elegantere Lösung wünschen lässt: Die Kontextdefintion, so wie sie hier eingeführt

wurde, erlaubt nur zusammenhängende Kontexte über das Bezugswort hinweg. Oder anders ausgedrückt: Es ist nicht möglich, einen Kontext zu einem Wort zu definieren, bei dem nur das vierte und fünfte Wort links von diesem Wort betrachtet werden soll. Vielleicht stellt sich ja gerade auf diesen beiden Positionen ein auffälliger Zusammenhang heraus, der bei einer weiteren Betrachtungsweise nicht aufgefallen wäre? Um genau dies zu überprüfen, bietet Belicas Verfahren einen zusätzlichen Schritt an, bei dem alle Unterkontexte zu einem vorgegebenen definierten Kontext ausgewertet werden. Der vorgegebene definierte Kontexte ist dann quasi nur als maximal möglicher Kontext vorgegeben. Diese zusätzliche Auswertung wird durch die Option *Autofokus* aktiviert. Es werden dann alle denkbaren zusammenhängenden Unterkontexte für die statistische Bewertung zugrunde gelegt (mit den entsprechend angepassten Werten für Umfang des Ausschnitts und Anzahl Vorkommen im Ausschnitt). Der Preis, den der Betrachter dafür zahlt, ist die längere Zeit, die für die Analyse benötigt wird. Der Gewinn ist eine zusätzliche Information über die positionellen Präferenzen der primären Partnerwörter. Bei kontrastiven Analysen ist darauf zu achten, dass diese zusätzliche Information eine neue, zu kontrollierende Variabilität in das Analyseergebnis mit hineinbringt.

Parameter Autofokus



Tab. E8.10: Autofokus: Mögliche zusammenhängende Subkontexte

Der Autofokus-Mechanismus geht also der Reihe nach alle Unterkontexte durch und ermittelt ihre statistischen Bewertungen. Die Grenzen des Kontextes, der die höchste statistische Auffälligkeit verbuchen konnte, werden als zusätzliche Information betrachtet, in welchem Bereich sich das primäre Partnerwort besonders gerne aufhält. Wäre im Beispiel in Tabelle E8.11 der LLR-Wert für das grau hinterlegte Feld maximal für die Dreiecksmatrix, so hätte das Verfahren das Intervall [x,y] als bevorzugte Positionen für das Partnerwort relativ zum Bezugswort ermittelt.

bevorzugtes Intervall

Bei der Anzeige werden bei aktiviertem Autofokus dann die Werte der linken und rechten Intervallgrenze zu einem Cluster angegeben, wobei

Sub-Kontext-grenzen	r	...	y	...	x	...	l
l (z.B. = -5)	LLR für [l,r]	...	LLR für [l,y]	...	LLR für [l,x]	...	LLR für [l,l]
...	...	...	...	...	...	...	
x (z.B. = -2)	LLR für [x,r]	...	LLR für [x,y]	...	LLR für [x,x]		
...	...	...	...	...			
y (z.B. = +3)	LLR für [y,r]	...	LLR für [y,y]				
...	...	...					
r (z.B. = +5)	LLR für [r,r]						

Tab. E8.11: Autofokus: Statistische Auffälligkeiten für mögliche Subkontexte

die Notation mit vorangestelltem Minus- bzw. Pluszeichen zwischen linkerhand bzw. rechterhand des Bezugswortes unterscheidet. Vergleichen Sie z. B. in der Tabelle E8.10 die Intervalle [-2,+3] und [+2,+3] miteinander. Zu beachten ist hier wiederum, dass – genauso wie die statistische Bewertung – die Autofokus-Angabe sich auf das primäre Partnerwort bezieht: Nur genau dieses findet sich verstärkt auf den durch das Intervall eingeschränkten Positionen. Über die Anordnung der anderen Partnerwörter sagt die Angabe nichts aus.

Intervallangabe bezieht sich nur auf primäre Partner!

Bezugswort »&Gold«, Z: hoch; G: grob; F: +; AF: +			
statistische Bewertung	Anzahl	Autofokus Intervall	Partner
10468	45	[+1,+3]	glänzt nicht ist längst
10468	720	[+1,+3]	glänzt nicht ist
10468	54	[+1,+3]	glänzt nicht längst
10468	860	[+1,+3]	glänzt nicht
10468	34	[+1,+3]	glänzt ist Nicht
10468	858	[+1,+3]	glänzt ist
10468	56	[+1,+3]	glänzt Nicht
10468	1243	[+1,+3]	glänzt

Tab. E8.12: Komplexe Wortverbindungen zu Bezugswort »&Gold« und primärem Partner »glänzt«

Die Reihenfolge der Partnerwörter in der Liste der primären Partner, aber auch innerhalb der Beschreibung der komplexeren Clusterstrukturen (die über den primären Partner hinausgehenden Beziehungen/Kohäsionen zu weiteren Wörtern) hat sich bisher vor allem an dem Maß bzw. den Maßen der statistischen Auffälligkeit orientiert. Dies setzt wiederum nur auf den quantitativen Eckdaten auf (Größe der Gesamtheit, Größe des Ausschnitts, Vorkommenshäufigkeit in der Gesamtheit, Vorkommenshäufigkeit im Ausschnitt). Die Reihenfolge, in der eine Partnerkaskade eines Bezugswortes ermittelt wurde, spiegelt nur in Ausnahmefällen die Rei-

henfolge wieder, in der die Wörter für gemeinhin in einem Satz gebraucht werden. Man könnte versuchen, durch geschickte Wahl der Analyseeinstellungen (Wahl des Suchwortes und Definition des Kontextes), sich dieser Reihenfolge zu nähern, wird aber stets entweder (1) durch zu rigide Vorgaben bestimmte relevante Aspekte des untersuchten Wortes aus dem Blick verlieren, oder (2) durch das nicht antizipierbare Verhalten des Wortes Abweichungen in der Reihenfolge in Kauf nehmen müssen. Besser wäre es natürlich, die Analyseeinstellungen so zu belassen, wie es für die eigentliche Fragestellung am sinnvollsten ist – und erst im Nachhinein zu schauen, in welcher Reihenfolge sich die ermittelten kohäsiven Wörter üblicherweise manifestieren. In dem von Belica (2003) noch einen Schritt weiter gedachten Konzept eines *syntagmatischen Musters* werden aus den Texten, die einer bestimmten Kookkurrenzrelation zugrundeliegen, d. h. einem Textcluster, folgende Angaben ermittelt:

syntagmatisches Muster

1. die an der Kohäsionsrelation beteiligten Wörter (blau gekennzeichnet) in ihrer häufigsten Reihenfolge
2. die prozentuale Häufigkeit der Treffer mit dieser Reihenfolge

xy%	← …	**wort**$_1$	…	**wort**$_2$	…	vgl. 1.+2.
	↑ **füller**$_1$ ↓		↑ [**füller**$_2$] ↓		↑ **füller**$_3$ | **füller**$_4$ ↓	vgl. 3.-5.
xy%	**füller**$_1$	**wort**$_1$	[**füller**$_2$]	**wort**$_2$	**füller**$_3$ | **füller**$_4$	Gesamtmuster

Tab. E8.13: Schematischer Aufbau eines syntagmatischen Musters

3. Lücken zwischen den beteiligten Wörtern
   a) entweder unbestimmt (als »… «)
   b) oder gefüllt durch weitere Wörter, die zwar nicht als kohäsiv erkannt wurden, aber an bestimmten Stellen besonders häufig in den Texten verzeichnet sind
4. sowie Aufhellungen bei der Darstellung der Punkte (2) gemäß des Wertes und (3a) und (3b) gemäß ihres anteiligen Vorkommens
5. wobei Lücken auch als optional (mittels »[ ]«) oder als alternativ zueinander (mittels »|«) erkannt und gekennzeichnet werden können

Die Prozentangabe bezieht sich auf die Häufigkeit der Reihenfolge der konkreten Wortformen, die bei der Kookkurrenzanalyse beteiligt waren (auch bei lemmatisiertem Bezugswort!). Falls eine Lemmatisierung der Partnerwörter gewählt wurde, wird kein syntagmatisches Muster angezeigt.

Betrachtet werden dafür Wortformen, unabhängig davon, ob bei der Suche eine Lemmatisierung gewählt wurde!

Dadurch, dass die Lückenfüller nur gezählt, aber nicht statistisch bewertet werden, haben häufige Wörter eine höhere Chance im Muster zu erscheinen – auch wenn sie statistisch unauffällig oder als Funktionswort von der statistischen Analyse ausgeschlossen worden waren (vgl. dazu auch Bemerkung am Ende von S. E8-6).

Die Beispiele für syntagmatische Muster in Tabelle E8.14 lassen sich etwa wie folgt umschreiben: Bezogen auf die Anzahl der Textstellen, die

1.	97%	**Morgenstund hat Gold im Mund**
2.	95%	**Morgenstund hat Gold im Mund**
3.	97%	**sein|ihr zweites […] Gold**

Tab. E8.14: Beispiele für syntagmatische Muster

zu dem jeweiligen Textcluster beitragen, wird beim zweiten Beispiel in 95% der Fälle das Wort »Gold« (zwei Positionen) vor »Mund« beobachtet. In dem dritten Beispiel befindet sich das Wort »zweites« (ein oder zwei Positionen) vor »Gold«. Das erste Beispiel zeigt drei kohäsive Elemente, auf deren Gesamtanordnung sich die Prozentzahl bezieht; dieses speziellere Cluster dokumentiert, dass auch »Morgenstund« als Bestandteil einer komplexeren übliche Wortverbindung erkannt wird. Die nicht-blauen Bestandteile kommen mit einem ihrer Helligkeit entsprechendem Anteil auf ungefähr dieser Position vor, wie z. B. »Morgenstund« zwei Positionen und »hat« eine Position vor »Gold« und »im« zwischen »Gold« und »Mund«. Ob aber alle diese Lückenfüller bevorzugt gemeinsam auftreten, ist nicht unbedingt zu schließen – es könnten sich an dieser Stelle auch verschiedene Formulierungen überlagern.

Im zweiten Beispiel findet sich unmittelbar vor »zweites« keine dominierende Besetzung: Neben der recht häufigen Form »sein« findet sich auf derselben Position mit ebenfalls nicht zu vernachlässigender Häufigkeit »ihr«. Die mögliche alternative Besetzung der Position erfasst also beide Varianten »sein zweites«, so wie auch »ihr zweites«. Die eckigen Klammern des Ausdrucks zwischen »zweites« und »Gold« deutet an, dass die Berechnung ergeben hat, dass die beiden Wörter nicht zwingend nacheinander folgen müssen – es kann sich auch sozusagen optional ein Ausdruck dazwischen schieben. Die Helligkeit der Markierung zeigt allerdings, dass dies sehr selten der Fall ist. Innerhalb der eckigen Klammern könnte eine Wortform (oder eine Folge oder eine Alternative von Wortformen) angegeben werden, die in den Fällen, in denen tatsächlich eine Lücke existiert, diese häufig füllt. In dem Beispiel hat sich aber kein Lückenfüller aufgedrängt, so dass mit den Punkten die Unspezifiziertheit der Lücke angedeutet wird: Es gibt zwar manchmal eine Lücke (genauer: einen Abstand zwischen den angegebenen Bestandteilen), aber es gibt an der Stelle nichts, was sinnvollerweise als Lückenfüller angegeben werden könnte.

Im Fall fester, etwa idiomatischer Prägungen, bei denen die längeren Partnerwortfolgen fast alle kürzeren dominieren, erleben wir bei den syntagmatischen Mustern Effekte, die mit den Häufigkeitsangaben korrelieren: Da die syntagmatischen Muster die Texte auswerten, die zu einem Cluster zusammengefasst wurden, werden bei der Mehrfachzuordnung auch die Cluster mit den kürzeren Partnerwortfolgen von der festen Fügung beherrscht (vgl. Bsp. in Tab. E8.14). Das syntagmatische Muster spiegelt dann bei fast allen Clustern schon diese Fügung wieder – nur eben mit einigen Bestandteile nicht blau gefärbt (als über die Kookkurrenzanalyse berechnete Partnerwörter), sondern in verschiedenen Grauschattierungen (als über die Ermittlung des syntagmatischen Musters

erkannt). Bei der eindeutigen Zuordnung trifft dies in der Form nicht zu. Dabei werden die syntagmatischen Muster eher die Abwandlungen der Fügung oder andere Zusammenhänge erfassen.

Eine Kookkurrenzanalyse fragt: Ist die Anzahl beobachteter Wörter in der Gesamtheit der definierten Umgebungen um ein vorgegebenes Wort herum in der Größenordnung dessen, was aufgrund ihrer Häufigkeiten in der Gesamtmenge, dem Gesamtkorpus, zu erwarten wäre? Wenn ein Wort häufiger als erwartet erscheint, dann haben wir einen Hinweis darauf, dass das Wort nicht zufällig in der Umgebung gelandet ist, sondern dass vermutlich irgendetwas dahintersteckt. Das *irgendetwas* kann eine sprachliche Affinität zum vorgegebenen Wort sein. Wir können aber nicht ausschließen, dass es Interferenzen mit anderen Phänomenen (wie etwa des Weltgeschehens) gibt.

Analysewort: rot, Analysetyp 0

+	1	1	73578	**Zahlen** gerutscht	301	98% in die roten Zahlen gerutscht
+	1	1	73578	Zahlen schreiben	443	63% rote [...] Zahlen [...] schreiben
+	1	1	73578	Zahlen schreibt	574	72% schreibt [...] rote Zahlen
+	1	1	73578	Zahlen	6584	58% in die|den roten [...] Zahlen
+	1	1	51846	**Faden** zieht	643	62% zieht sich wie ein roter Faden durch die
+	1	1	51846	Faden	3511	56% wie ein roter [...] Faden durch die
+	1	1	39533	**Karte** sah Tätlichkeit	21	52% nach einer Tätlichkeit an die rote Karte sah
+	1	1	39533	Karte sah	376	52% sah ... die rote [...] Karte
+	1	1	39533	Karte gezeigt	141	100% die rote Karte [...] gezeigt
+	1	1	39533	Karte Tätlichkeit	66	77% nach|wegen einer Tätlichkeit ... die rote Karte sah
+	1	1	39533	Karte	3435	77% die rote [...] Karte
+	1	1	25330	**Teppich** ausgerollt langen	3	100% langen roten Teppich [...] ausgerollt
+	1	1	25330	Teppich ausgerollt	211	50% der rote [...] Teppich [...] ausgerollt
+	1	1	25330	Teppich ausrollen langen	1	100% langen roten Teppich ausrollen
+	1	1	25330	Teppich ausrollen	67	97% den|einen roten Teppich [...] ausrollen
+	1	1	25330	Teppich langen	26	100% einen|Meter langen roten Teppich
+	1	1	25330	Teppich	1927	64% auf dem|den roten [...] Teppich
+	1	1	15344	**Blutkörperchen** Blut Hämoglobin	1	100% Blut ... Hämoglobin ... roten Blutkörperchen
+	1	1	15344	Blutkörperchen Blut	62	38% Anzahl|Anteil der roten Blutkörperchen im Blut und
+	1	1	15344	Blutkörperchen Blutfarbstoff Häm	5	20% Blutkörperchen ... Hämoglobin ... roten Blutfarbstoff
+	1	1	15344	Blutkörperchen Blutfarbstoff	9	44% Blutfarbstoff ... der roten Blutkörperchen
+	1	1	15344	Blutkörperchen Hämoglobin	19	52% das Hämoglobin in den der roten Blutkörperchen
+	1	1	15344	Blutkörperchen	1032	66% der roten [...] Blutkörperchen
+	1	1	13504	**Tuch**	1096	72% ein rotes [...] Tuch
+	1	1	12865	**Laterne** trägt	61	52% die|Die rote Laterne [...] trägt
+	1	1	12865	Laterne trug	9	55% trug ... die rote Laterne

Abb. E8.4: Kookkurrenzprofil des Wortes »&rot« (Ausschnitt)

Versuchen wir an dieser Stelle, kurz zusammenzufassen, was bisher vermittelt werden sollte: Wir brauchen große Datenmengen, um Strukturen hervortreten zu lassen, wir brauchen aber auch Instrumente, gedankliche und operationale, um uns in den großen Datenmengen zurechtzufinden. Nur nach etwas zu suchen, bringt nur zufällig Erkenntnisse über kaum mehr als eine Handvoll Beispiele.

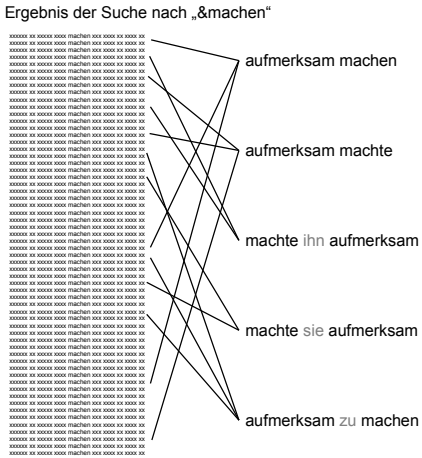


Abb. E8.5: Treffer unsortiert

Ein erster Weg könnte sein, die Treffer zu sortieren, wofür wir (in der Cosmas II-/Recherchesitzung) verschiedene Möglichkeiten kennengelernt haben.

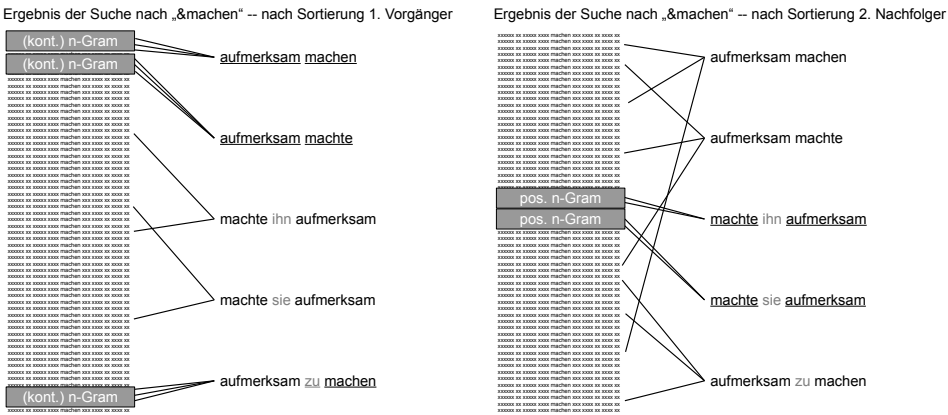


Abb. E8.6: Treffermenge sortiert, positionell gebunden

Aber auch hierbei kommen wir den Phänomenen nur ein kleines Stückchen näher, da viele Entscheidungen beim Sortieren schon Annahmen über die Natur des Phänomens vorwegnehmen müssen.

Sortieren/Ordnen anhand Kookkurrenzen

Mit dem Clustering über die Kookkurrenzanalyse haben wir ein sehr mächtiges Instrument an der Hand, quasi ein sehr luxuriöses Sortierverfahren, dass Variabilität in Formen und Positionen selber handhaben kann.

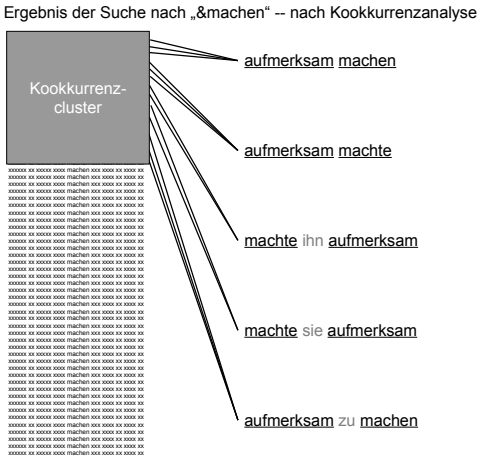


Abb. E8.7: Treffer geclustert

Als Ergebnis liefert es eine Zusammenfassung von Textstellen, die sich in gewisser Hinsicht ähneln und für die es interessant sein könnte, sie gemeinsam betrachten zu können.

Abb. E8.8: Einblick in die dem Cluster zugrundeliegenden Textstellen

Partnerwortfolgen und syntagmatische Muster sind sozusagen erste Label für diese Zusammenfassungen, die die Deutung ihres inneren Zusammenhalts erleichtern sollen.

E8.4 Aufgaben zum theoretischen Teil

Versuchen Sie zunächst, die ersten vier Aufgaben zu bearbeiten, ohne eine praktische Sitzung durchzuführen.

1. Welche Auswirkungen haben bei der Kookkurrenzanalyse (vermutlich) verschieden eingestellte Kontexte auf das Analyseergebnis, je nach Wortart des untersuchten Wortes? Finden sich etwa bestimmte Arten von Partnerwörtern in bestimmten Unterkontexten? Worin besteht der (methodische) Unterschied, ob ich diesen zu erwartenden Unterkontext bereits als zu untersuchenden Kontext vorgebe oder ihn mir durch den Autofokusmechanismus ermitteln lasse?
2. Diskutieren Sie kurz, ob die *häufigsten* Wortverbindungen zu einem Wort durch die Kookkurrenzanalyse ermittelt werden.
3. Welche Arten von Wortverbindungen kann die Kookkurrenzanalyse aufdecken? Versuchen Sie zunächst hypothetisch eine intuitive Charakterisierung. Können Sie Ihre Typen zu traditionell linguistischen Kategorien z. B. aus der Phraseologie in Beziehung setzen?
4. Für welche Fragestellungen lässt sich die Kookkurrenzanalyse sinnvoll anwenden?

E8.5 Praktische Sitzung

Und nun sollten Sie in *medias res* gehen … Starten Sie eine Sitzung mit dem Recherchesystem des IDS wie in Abschnitt 2.9 unseres Buches bzw. im Begleitmaterial angedeutet. Formulieren Sie geeignete Suchanfragen und starten Sie die Kookkurrenzanalyse. In einem Dialogfenster können Sie dann alle der oben beschriebenen Einstellungen festlegen. Belassen Sie es zunächst bei der Standardeinstellung und lassen Sie die Analyse ausführen … bitte etwas Geduld, je nach Datenmenge und Einstellungen kann dieser Vorgang einige Zeit in Anspruch nehmen. Nach Abschluss dieses Vorgangs wird Ihnen das Ergebnis in der in diesem Kapitel beschriebenen Form präsentiert. Neben der Gesamtstruktur bietet Ihnen die grafische Oberfläche eine interaktive Möglichkeit an, die Textstellen einzusehen, die zur Bildung des jeweiligen Kookkurrenzclusters geführt haben.

Neben den hier gestellten Aufgaben gilt ansonsten generell: Ausprobieren! Ausprobieren! Ausprobieren! Wählen Sie unterschiedliche Korpora, unterschiedliche Suchanfragen, unterschiedliche Einstellungen und schauen Sie sich die Analyseergebnisse an. Versuchen Sie, die Unterschiede zu ergründen und zu dem in diesem Kapitel Gelernten in Beziehung zu setzen.

E8.6 Aufgaben zum praktischen Teil

1. Führen Sie mithilfe des IDS-Recherchesystems eine Kookkurrenzanalyse mit den Standardeinstellungen durch. Wählen Sie dazu ggf. ein kleineres Korpus und lassen Sie eine Zufallsauswahl aus der Treffermenge bilden. Besonders geeignet sind z. B. Farbadjektive sowie Bezeichnungen von Körperteilen und Tierarten. Ordnen Sie ausgewähl-

te Kookkurrenzen linguistischen Kategorien zu und beschreiben Sie ggf. deren Besonderheiten in eigenen Worten. Versuchen Sie dabei, mal mehr, mal weniger, sich an Formulierungen zu orientieren, die Ihnen aus Wörterbuchartikeln vertraut erscheinen.

2. Variieren Sie nun insbesondere die Parameter Granularität und Zuordnung. Schauen Sie sich die syntagmatischen Muster an und versuchen Sie, die Unterschiede zu ergründen.
3. *(Eine etwas umfangreichere Aufgabe, als Hausarbeit ausbaubar:)* Wählen Sie einen Wortartikel mittlerer Länge aus einem DaF-Lernerwörterbuch aus und betrachten Sie die Angaben zu typischen Verwendungsweisen (und ggf. Kollokationen – sofern ausgezeichnet). Überprüfen Sie stichprobenartig, welche Angaben auch mit der Kookkurrenzanalyse (mit welchen Einstellungen?) hätten aufgespürt werden können. Vergleichen Sie dazu auch typische Verwendungsweisen, die die Kookkurrenzanalyse ermittelt, die aber in dem Wortartikel nicht angegeben sind.