Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora

Alexander Geyken – Berlin-Brandenburgische Akademie der Wissenschaften

Abstract

Das DWDS-Wortprofil ist das Ergebnis einer automatischen syntaktischen und statistischen Analyse sehr großer Korpora. Es liefert einen kompakten Überblick über die statistisch signifikanten syntagmatischen Beziehungen eines Wortes. Beispiele dieser sogenannten syntaktischen Relationen sind Attribut-Nomen Verbindungen wie *schöne Bescherung* oder Verb-Objekt Beziehungen wie *Flasche entkorken*. Die Darstellung der Relationen erfolgt in Form einer Schlagwortwolke oder in Tabellenform. Die Berechnung des DWDS-Wortprofils erfolgt in drei Etappen: Festlegung der zu extrahierenden syntaktischen Relationstypen, Extraktion der Relationen mittels einer automatischen syntaktischen Analyse und Bewertung der statistischen Signifikanz der extrahierten Relationen. Der derzeitige Prototyp des DWDS-Wortprofils beruht auf einer Mischung eines Referenz- und eines Zeitungskorpus, dem DWDS-Kernkorpus und dem ZEIT-Archiv (1946-2009), und hat eine Gesamtgröße von 500 Millionen laufenden Textwörtern. Aus dem Korpus wurden etwa 90.000 Lemmata mit 2.000.000 Relationen extrahiert. Erste Auswertungen zeigen, dass bereits der gegenwärtige Prototyp eine große Reichhaltigkeit von Wortverbindungen enthält, die ihn auch im Vergleich mit großen einsprachigen Wörterbüchern interessant machen. Für den an der Textproduktion interessierten Nutzer bietet das Wortprofil aber einen weit über das gedruckte Wörterbuch hinausgehenden Mehrwert: Jede Relation des Wortprofils ist direkt mit den dazugehörigen Satzkontexten im Korpus verknüpft. Der Nutzer erhält somit unmittelbar einen Überblick über die weiteren semantischen und pragmatischen Kontexte in authentischen Texten. Dies kann gerade bei sprachlichen Unsicherheiten von erheblicher Hilfe sein. Das DWDS-Wortprofil ist über die Web-Plattform www.dwds.de abfragbar.

1. Einleitung

Elektronische Textkorpora werden seit über 20 Jahren in der Wortschatzforschung und für das Sprachlernen eingesetzt (z.B. Sinclair 1991, Braun et al. 2006). Beispiele aus Korpora haben dabei im Unterschied zu Kompetenzbeispielen den Vorzug, dass sie die verschiedenen Verwendungen des Wortes auf der Grundlage realer Textbeispiele widerspiegeln. In der Regel basieren elektronische Plattformen, die diese Korpora analysieren, bei der Anzeige der Suchergebnisse auf Konkordanzen (s. Abb. 1). Damit können die Gebrauchskontexte eines Wortes auf übersichtliche Weise darge-

stellt werden. Mit wachsender Größe der Korpora - gängig sind mittlerweile Korpora einer Größe von mehreren hundert Millionen bis hin zu einigen Milliarden Textwörtern - steigt nicht nur der Variantenreichtum der Gebrauchskontexte, sondern auch die Anzahl der Belege für ein Suchwort bzw. einen Suchausdruck. In der Regel sind daher die Konkordanzanzahlen für eine Wortsuche zu groß für eine manuelle Auswertung. Beispielsweise haben die Wörter *Feindbild* oder *grau* im Archiv der ZEIT (1946-2009) Frequenzzahlen von 2364 bzw. 21656 Treffern (s. www.dwds.de).

Abb. 1: Konkordanzzeilen für das Substantiv Feindbild

In Abb. 1 sind die ersten 20 der 2364 Konkordanzzeilen für das Substantiv *Feindbild* dargestellt. Das Wort *Feindbild* kommt hier in sehr unterschiedlichen syntagmatischen Kontexten vor. Bei der Frage jedoch, ob und welche dieser Kontexte typisch sind, geben Konkordanzzeilen keine unmittelbare Hilfe. Wie kann man beispielsweise durch das Lesen von Konkordanzzeilen herausfinden, mit welchen Adjektiven das Substantiv *Feindbild* typischerweise verwendet wird oder umgekehrt, ob bzw. wie gebräuchlich in Abbildung 1 die Adjektive *perfekt* (Treffer 14), *dämonisierend* (15), *neu* (17) oder *klar* (19) für das Substantiv *Feindbild* sind? Ähnliche Fragen stellen sich für verbale Konstruktionen wie *auf einem Feindbild beharren* (12,13), *auf Feindbilder angewiesen sein* (2), oder *ein Feindbild verlieren* (5) bzw. *ein Feindbild abgeben* (14). Um diese Fragen „per Kopf" zu beantworten, müsste man die über 2300 Konkordanzzeilen durchlesen und Strichlisten über die entsprechenden Konstruktionen führen. Es ist unmittelbar einleuchtend, dass die Beantwortung dieser Fragen bei noch häufigeren Wörtern oder bei noch größeren Korpora noch aufwändiger zu beantworten wäre.

Church & Hanks (1989) waren die ersten, die zeigten, dass lexikalische Statistiken sinnvoll sind, um Konkordanzzeilen auf typische Verwendungsweisen hin zu analysieren. Sie verwenden dabei das statistische Maß der Mutual Information, das ver-

einfacht gesprochen, Worte in einer gewissen Nachbarschaft des Suchwortes (in der Regel zwischen 5 und 10 Wörtern oder auch innerhalb des ganzen Satzes) daraufhin bewertet, ob sie statistisch gesehen überproportional häufig mit dem Suchwort erscheinen. Eine Reihe von Datenbanken verwenden diese oder vergleichbare Kookkurrenzstatistiken, um die Gebräuchlichkeit von Wortverbindungen zu bewerten. Ein gutes Beispiel hierfür liefert das weit verbreitete Portal „Deutscher Wortschatz", welches zu einem Suchwort die statistisch signifikanten Kookurrenzen aufführt und darüber hinaus auch die signifikanten linken und rechten Nachbarn (s. Abb. 2)

Signifikante Kookkurrenzen für Feindbild:

ein (99), als (73), das (71), eins (68), neues (65), klares (54), Nummer (49), dient (45), Saudische (45), neoliberales (43), islamischen (42), irrsinnigen (42), auserkoren (41), Islam (40), : (40), geworden (38), Türk (35), Globetrotter (35), , (34), zum (33), Steuerstaates (32), Schmähgesänge (30), Juden (30), aufbaue (26), Westen (26), aufzubauen (25), gemeinsames (25), herhalten (25), dargestellt (24), willkommenes (24), Bayern-Fans (23), abgenutzt (23), löst (23), Jude (22), Bruchhagen (22), SPD-Arbeitsmarktexperte (21), Bill (21), Aug (20), Verhöhnung (20), aufgeheizten (20), Gates (19), Massen (19), Zelt (19), Terror (19), sein (19), zurecht (18), Sarkozy (17), Profilierung (17), Staatsmacht (17), entdeckt (17), beliebtes (16), ins (16), 68er (16), Amerika (16), Bushs (15), Brandner (15), Wanken (15), Israel (14), rückt (14), geliefert (14), USA (14), Husseins (14), entzündet (14), schlechthin (14), ist (14), Kuwait (13), Vernichtung (13), alte (12), Klientel (12), Mittleren (12), Kalten (12), parat (12), Gegners (12), größtes (12), vorm (12), Kapitalismus (11), Boll (11), Naturschutz (11), ausmachen (11), wilden (11)

Signifikante linke Nachbarn von Feindbild:

das (187), als (149), zum (149), neues (122), klares (95), ein (91), neoliberales (60), gemeinsames (48), alte (46), willkommenes (36), kein (35), sein (32), beliebtes (28), Das (28), gutes (26), größtes (24), wilden (23), großes (15), weiteres (15), ihr (12), " (11), dem (7), ins (6), vom (4)

Signifikante rechte Nachbarn von Feindbild:

Nummer eins (141), Nummer (113), herhalten (43), Islam (41), Bruchhagen (39), herhalten müssen (37), zurecht (36), geworden (36), entdeckt (35), auserkoren (30), USA (20), Israel (19), geliefert (19), geriet (19), dient (19), vieler (17), des (17), ab (17), " (15), Westen (14), Staat (13), gefunden (12), eines (10), der (9), vom (8), : (7), , (5), war (4)

Abb. 2: http://wortschatz.uni-leipzig.de (Abfrage vom 30.11.2010)

Zwar liefert eine rein statistische Methode einen sehr schnellen Überblick über die signifikanten Wortpaare innerhalb eines Satzes. Die Reihenfolge scheint jedoch, was den syntaktischen Bezug zum Suchwort *Feindbild* angeht, zu einem gewissen Grade beliebig. So lässt sich zwar vermuten, dass die Artikel *ein*, *das* und die Adjektive *neues* und *klares* in Zeile (1) der Abbildung 2 einen klaren lokalen Bezug zum Suchwort haben, z.B. *ein Feindbild*, *neues Feindbild*, *klares Feindbild*. Schon weniges klar sind aber beispielsweise die Kookurrenzen in Zeile (3): *Türk*, *Globetrotter* oder *Schmähgesänge*. Die syntagmatischen Bezüge lassen sich hier nur erraten. Im Beispiel *Feindbild* ist diese syntaktische Unordnung noch durch etwas mehr Leseaufwand in den Griff zu kriegen, da es sich hier nur um etwa 50 Kookurrenzen handelt. Wirklich problematisch wird es jedoch bei Suchwörtern, die mehrere hundert oder sogar über 1000 Kookurrenzen haben. Dies ist kein so seltener Fall, wie in Abschnitt 4.2 erläutert wird. Man wird daher für die Beantwortung der oben gestellten Fragen nach typischen syntaktisch motivierten Wortverbindun-

gen andere Verfahren benötigen, die auch syntaktische Informationen nutzen.

Ein für die syntagmatische Analyse weitergehender Ansatz stellt die von Adam Kilgarriff eingeführte Sketch Engine dar (Kilgarriff 2004). Die Sketch Engine geht über reine Kookkurrenzstatistiken insofern hinaus, als sie nur diejenigen Kookurrenzen berücksichtigt, die in einer vordefinierten syntaktischen Relation stehen. Solche Relationen können beispielsweise Adjektiv-Nomen, Verb-Objekt, Genitivattribute von Nomen oder Verb-Präpositionalphrasen Verbindungen sein. Sketch Engine Plattformen gibt es für so verschiedene Sprachen wie Englisch, Tschechisch, Japanisch oder Chinesisch. Eine einfache Übertragung des Sketch-Engine Ansatzes z.B. vom Englischen auf das Deutsche, ist aus wenigstens den beiden folgenden Gründen schwierig: die freie Wortstellung im Deutschen und der Kasussynkretismus führen dazu, dass eine Extraktion von syntaktischen Relationen auf der Basis von Wortarten und darauf basierenden Satzmustern, anders als im Englischen, zu keinen befriedigenden Ergebnissen führt. So haben Experimente mit der Sketch Engine für das Deutsche gezeigt, dass je nach Parametrisierung der Regeln, entweder die Analysegenauigkeit unzureichend ist oder aber die Abdeckung, d.h. der Anteil des analysierbaren Texts, zu gering ist (Kilgarriff 2004, Ivanova et al. 2008). Aus diesem Grund beruhen die beiden existierenden Ansätze für das Deutsche zur Extraktion von syntaktischen Relationen aus großen Textkorpora auf einem allgemeineren Formalismus, der syntaktische Satzfunktionen erkennen und lokale Mehrdeutigkeiten auflösen kann. Der erste, an der Universität Stuttgart entwickelte Ansatz zur Extraktion „signifikanter Wortpaare als Webservice“ (Fritzinger et al. 2009), beruht dabei auf dem Dependenzparser FSPAR (Schiehlen 2003), der zweite an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) entwickelte Prototyp, das DWDS-Wortprofil (Geyken et al. 2009), basiert auf Syncop (Didakowski 2007), einem auf syntaktischem Tagging beruhenden Parsing-Formalismus.

In diesem Beitrag soll das DWDS-Wortprofil aus Nutzersicht erläutert werden. Dazu werden wir die Abfragefunktionalität, so wie sie sich für den Nutzer auf der Web-Plattform des DWDS präsentiert, am Beispiel des eingangs erwähnten Wortes *Feindbild* illustrieren (Abschnitt 4.1) und in Abschnitt 4.2 eine erste Einschätzung der Qualität des Wortprofils anhand des Vergleichs mit einem großen einsprachigen Wörterbuch gegeben werden. Abschnitt 5 gibt einen Ausblick auf die Potentiale des Wortprofils. Zuvor werden in Abschnitt 2 die Methode zur Erstellung von Wortprofilen und in Abschnitt 3 das derzeit verwendete DWDS-Wortprofil sowie dessen Datengrundlage beschrieben.

2. Das DWDS-Wortprofil

Das DWDS-Wortprofil ist das Ergebnis einer automatischen syntaktischen und statistischen Analyse sehr großer Korpora. Es liefert einen kompakten Überblick über die statistisch signifikanten syntagmatischen Beziehungen eines Wortes mit anderen Wörtern in Form einer Schlagwortwolke oder in Tabellenform. Die Berechnung des

DWDS-Wortprofils erfolgt in drei Etappen, die in den folgenden Abschnitten näher beschrieben werden: Festlegung der zu extrahierenden syntaktischen Relationstypen, Extraktion der Relationen mittels einer automatischen syntaktischen Analyse und Bewertung der statistischen Signifikanz der extrahierten Relationen.

2.1 Festlegung der syntaktischen Relationen

Die Festlegung der syntaktischen Relationstypen erfolgt außerhalb der Berechnung der Relationen durch den Syntaxparser Syncop (s. Abschnitt 2.2), da die Relationstypen bestimmen, welche Analysen des Parsens in das Wortprofil einfließen. Die Liste der Relationstypen beruht auf den in Foth (2005) verwendeten Etiketten zur Markierung von Dependenzrelationen, die wiederum von SynCop verwendet werden. Typische im Wortprofil verwendete Etiketten sind ATTR oder OBJD. ATTR steht für eine Attribut-Relation zwischen Adjektiv und Nomen, OBJD für eine Dependenzrelation zwischen einem Verb und einer Nominalphrase im Dativ. Mit OBJD werden aus Gründen der Einfachheit für die Notation sowohl Dativobjekte als auch fakultative Ergänzungen oder freie Angaben zusammengefasst, da der Syncop über keine Verbsemantik verfügt und somit nicht entscheiden kann, ob es sich bei der dependenten Nominalphrase um ein Objekt oder nur eine fakultative Ergänzung oder Angabe handelt. Nicht alle für das Parsing verwendete Etiketten sind für den Benutzer des Wortprofils relevant. Dies gilt beispielsweise für die Relation DET, die einen Artikel mit einem Nomen verbindet, die Relation APP, die festlegt ob eine Abfolge von Wörtern zu einer Nominalphrase gehören oder die Relation NEB, die das Verb eines Nebensatzes mit dem übergeordnete Wort verbindet. Diese Relationen sind zwar für den Parsing-Prozess wichtig, stellen aber für das Wortprofil nur Zwischenergebnisse dar.

Derzeit werden die folgenden Relationstypen für das DWDS-Wortprofil verwendet:

Relationstypen innerhalb von Phrasen

- Adjektiv-Nomen (Etikett: ATTR): *klares Feindbild, schöne Bescherung*
- Nomen-Koordination-Nomen (CJ): *grün und blau, Kopf oder Zahl*
- Nomen-Nomen (im Genitiv) (GMOD): *Abbau* des *Feindbildes, Abbau* von *Vorurteilen*
- Adverb-Adjektiv (ADVA): *sehr intelligent*, *hoch erfreut*

Phrasenübergreifende Relationen

- Nomen -Verb (SUBJ): das *Feindbild verblasst,* das *Badewasser läuft aus*
- Nomen-Verb (PSUBJ): die *Flasche* wurde *entkorkt*
- Nomen -Verb (OBJA): ein *Feindbild abbauen,* eine *Rede halten*
- Nomen -Verb (OBJD): dem *Publikum verkünden*
- Verb-Präposition-Nomen (V PP): *zur Verfügung stehen*, *auskommen ohne Feindbild*
- Verb-Verb (INFOBJ): *aufgehen sehen,* (auf etw.) *zu sprechen kommen*
- Adverb Verb (VADV)*: schallend lachen, freimütig zugeben*

2.2 Extraktion der syntaktischen Relationen

Wie bereits in Abschnitt 1 erwähnt, ist es für das Deutsche im Unterschied zum Englischen notwendig, die Extraktion syntaktischer Relationen auf der Basis eines syntaktisch voranalysierten Texts durchzuführen. Die syntaktische Analyse kann dabei aufgrund der speziellen Charakteristika des Deutschen (lange Abhängigkeiten und relative freie Wortstellung), nicht auf „flachen" Sequenzabfolgen von Wortarten beruhen, sondern muss in der Lage sein, Dependenzinformationen für phrasenübergreifende Relationen aufzubauen. Ein solches Verfahren, welches sowohl syntaktische Satzfunktionen erkennen und lokale Mehrdeutigkeiten auflösen kann, ist das für das DWDS-Wortprofil verwendete Analysewerkzeug Syncop. Hier soll die Grundidee des Verfahrens dargestellt werden. Eine vollständige technische Beschreibung von Syncop findet sich in Didakowski (2007).

Das Parsing-Verfahren von Syncop beruht auf fünf Modulen, die sequenziell auf die Korpustexte angewandt werden:

(1) Satzsegmentierung und morphologische Analyse mit der TAGH-Morphologie (Geyken & Hanneforth 2005)

(2) Ermittlung der Phrasen und Annotation der darin enthaltenen Modizierer/Koordinationsfunktionen

(3) Ermittlung von dependenten bzw. koordinierenden Phrasen

(4) Ermittlung von Nebensätzen und Annotation der darin enthaltenen syntaktischen Köpfe

(5) Ermittlung von Hauptsätzen und Annotation der darin enthaltenen syntaktischen Köpfe

Bestandteile dieser Module sind außerdem:

- Die Auflösung von Kasus-Numerus-Genus Kongruenz. Dies ist für die Zuordnung von Attributivphrasen genauso wie für Subjekt-Verb Relationen von Bedeutung,
- Die Erkennung von Verb-Partikeln, die für die korrekte Lemmatisierung komplexer Verben in Verbzweitstellung notwendig ist,
- Präferenzregeln zur Auflösung mehrdeutiger globaler Satzanalysen,
- die Möglichkeit, syntaktische Regeln zu verletzen, um auch halb- oder ungrammatische Korpussätze zumindest partiell analysieren zu können.

Wie eingangs erwähnt, basieren Wortprofile auf sehr großen Korpora mit „real existierenden", d.h. im Allgemeinen auch sehr komplexen oder unvollständigen Sätzen. Das Parsing-Verfahren zur Extraktion von Wortprofilen muss somit sowohl robust als auch effizient sein. Die Robustheit bedeutet in diesem Zusammenhang, dass das Verfahren nicht abbricht, wenn es bei einem komplexen Satz keine vollständige Satzanalyse liefern kann, sondern sich gegebenenfalls auch damit begnügt, partielle Analysen

von Phrasen oder unterspezifizierte Satzfunktionen als Analyse zurückzugeben. Die Effizienz des Verfahrens wird dadurch gewährleistet, dass Syncop nicht beliebig eingebettete Sätze extrahiert – dies ist nur mit rekursiven Verfahren möglich –, sondern nur Sätze mit einer beschränkten Einbettung analysiert. Für praktische Zwecke der Korpusannotation ist dies aber durchaus ausreichend (Koskenniemi 1990).

Wie sich dies auf die Satzanalyse sowie die daraus zu extrahierenden syntaktischen Relationen auswirkt, soll an dem folgenden im DWDS-Kernkorpus enthaltenen Satz illustriert werden. Schwierigkeiten dieses Satzes sind die „langen" Abhängigkeiten der passivischen Subjektfunktion (*Aspekt, beleuchtet*), die Passivierung des Gesamtsatzes sowie die komplexen Koordinationen (*Aspekt, Folgen, Auswirkungen*).

„Jeder Aspekt des Vertrags von Rom .. und alle Folgen und Auswirkungen, die ein britischer Beitritt nach sich ziehen dürfte, sind von allen Seiten beleuchtet worden." Hervorhebung (A.G.) (Archiv der Gegenwart 36, 1966)

Von Syncop wird dieser Satz folgendermaßen analysiert (die Analyse beschränkt sich dabei auf die für die Extraktion wichtigen Satz- und Koordinationsrelationen).

[[Jeder*@DetN* Aspekt*@HEAD*]$_{np}$*@SUBJ* [des*@DN* Vertrags*@HEAD*]$_{np}$*@GN* ... und*@CC* [alle*@DN* Folgen*@HEAD* und*@CC* Auswirkungen*@HEAD*]$_{np}$*@SUBJ* sind*@FAUXV* ... beleuchtet*@FMAINV* worden*@FAUXV* .]$_{cl\_passive}$

In dieser Klammernotation werden syntaktische Funktionen (Kopf- und Modifiziererrelationen) durch das Symbol @ markiert. Im Beispiel steht dabei @HEAD für die Kopffunktion, @Subj für die Subjektrelation, @CC für die Koordination, die anderen Funktionen stehen für Modifiziererrelationen: *@DN* für eine Nomen-Determinierer Relation, *@GN* für die Nomen-Genitiv Relation, *@FMAINV* für das Hauptverb, *@FAUXV* für die Verb-Auxiliar-Relation. Die Markierung der Sätze und Phrasen werden mit eckigen Klammern notiert: Im Beispiel gibt es einen Satz, der als Passiv-Klausel markiert (cl_passive) ist und drei Phrasen, die jeweils als Nominalphrase (NP) annotiert werden. Zur einfacheren Lesbarkeit ist die Klammernotation in Abbildung 4 in der Dependenzansicht dargestellt.

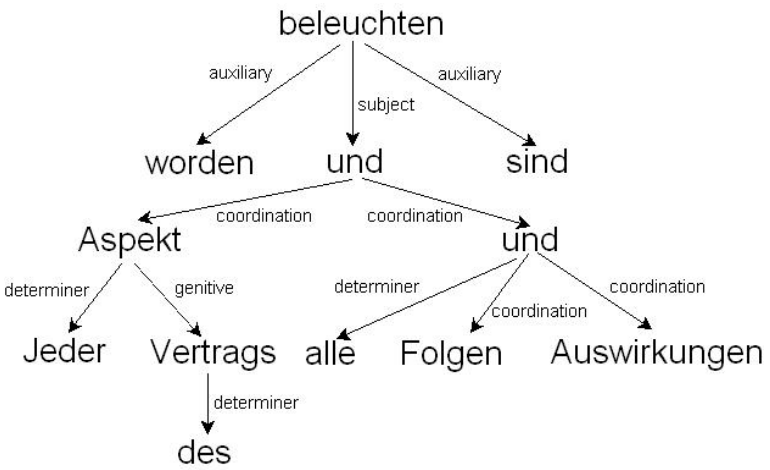

Abb. 3: Dependenzbaum

Aus der Satzanalyse des Beispielsatzes durch den Syncop-Parser lassen sich die folgenden syntaktischen Relationen (s. 2.1) für das Wortprofil extrahieren, wie sich dem Dependenzbaum ablesen lässt:

- PSUBJ: (Aspekt , beleuchten)
- GMOD: (Aspekt, Vertrag)
- CJ: (Aspekt, Folge)
- CJ: (Aspekt, Auswirkungen)
- PSUBJ: (Folge, beleuchten)
- PSUBJ: (Auswirkung, beleuchten)

Zusammenfassend lässt sich somit festhalten, dass das Analysesystem Syncop in der Lage ist, Phrasen, Hauptsätze (Klauseln) und Nebensätze (Sub-Klauseln) eines Satzes zu markieren sowie die dazugehörigen syntaktischen Funktionen (Kopf- und Modifiziererrelationen) zu ermitteln. Dabei gilt folgende Einschränkung: Bei komplexen oder unvollständigen Sätzen versucht Syncop nicht, eine vollständige Syntaxanalyse des Satzes zu erzeugen, sondern gibt auch partielle Analysen zurück. Aus den Analysen wiederum werden die syntaktischen Relationen extrahiert.

2.3 Berechnung der statistischen Salienz

Die extrahierten syntaktischen Relationen werden mittels eines statistischen Verfahrens bezüglich ihrer statistischen Signifikanz gewichtet. Dafür werden die aus den Texten extrahierten Relationen aufsummiert und danach bewertet, ob sie überproportional häufig vorkommen. Wichtig hierbei ist, dass für die Zählung weder einfache Wortformen noch Lemmaformen, sondern Lemma-Wortart-Paare verwendet werden. Ein Lemma-Wortart-Paar ist beispielsweise (*grau*, Adjektiv) wie in *graue Eminzenz* oder (*grau*, Modifizierer) wie in *grau meliert*. Der Grund, weshalb man Lemma-Wortart-Paare und nicht einfach Lemmata verwendet, besteht darin, dass Lemmata bezüglich ihrer Wortart mehrdeutig sein können und somit andere Relationstypen besitzen. Beispielsweise ist (*grau*, Adjektiv) Attribut eines Nomens, wohingegen (*grau*, Modifizierer) Attribut eines Adjektivs ist. Diese Vermischungen umgeht man, indem man für die Lemma, Wortart – Paare unterschiedliche Relationstabellen aufbaut.

Das im DWDS-Wortprofil verwendete Maß der Salienz (Lin 1998) ist eine Erweiterung der Mutual Information (MI). Bei der Salienz wird wie für die MI berechnet, ob eine Wortkombination im Korpus „häufiger als erwartet“ vorkommt. Allerdings bezieht die Salienz hierbei nicht die relativen Häufigkeiten über das gesamte Korpus mit ein, sondern schränkt diese auf den spezifischen syntaktischen Relationstyp ein, über den die Wortkombination in Verbindung steht. Formal ausgedrückt wird die Salienz eines Tripels (w, r,w‘) wie folgt definiert:

$$Sal(w,r,w') = \frac{\|w,r,w'\| \times \|*,r,*\|}{\|w,r,*\| \times \|*,r,w'\|} \quad (1)$$

In der Formel (1) sind w und w' Lemma-Wortart-Paare, r ein syntaktischer Relationstyp. $\|w,r,w'\|\|w,r,w'\|$ steht für die Anzahl der Vorkommen des Tripels (w,r,w') (w,r,w') im annotierten Korpus. ' * ' * ' ist ein Platzhalter für eine beliebiges Lemma-Wortart-Paar, und $\|w,r,*\|$ $\|w,r,*\|$ ist definiert als die Summe der Häufigkeiten über alle Lemmata $wj'wj'$ mit $\|w,r,w_j'\|$ $\|w,r,w_i'\|$. Entsprechend wird $\|*,r,*\|$ $\|*,r,*\|$ definiert als die Summe über alle Tripel $(w,r,w_j')(w,r,w_j')$, die miteinander über die Relation r in Verbindung stehen. Diese Formel entspricht der Formel, die Church & Hanks (1989) für die MI vorgeschlagen haben mit dem zusätzlichen Faktor $\|*,r,*\|\|*,r,*\|$. In Übereinstimmung mit Kilgarriff (2004) lässt sich feststellen, dass das Maß der Salienz im Vergleich zum Maß der MI den Vorteil hat, Tripel mit geringer Häufigkeit nicht überzubewerten.

3. Datengrundlage und Parametrisierung des DWDS-Wortprofils

Die Korpora, die als Datengrundlage des DWDS-Wortprofils dienen, sind grundsätzlich frei wählbar. Die Zusammensetzung und Größe der Korpora spielen für das Wortprofil jedoch eine wichtige Rolle.

Die Zusammensetzung der Korpora ist insofern relevant, als die extrahierten syntaktischen Relationen die im Korpus vorkommenden syntaktischen Nachbarn des Wortes widerspiegeln. Daher erhöht ein breit gestreutes, nach Textsorten ausgewogenes Korpus, ein sogenanntes allgemeinsprachliches Referenzkorpus, die Qualität des Wortprofils in Bezug auf die allgemeinsprachliche Aussagekraft. Spezialkorpora oder spezielle Zeitungskorpora werden somit andere Wortprofile liefern als Referenzkorpora.

Auch die Korpusgröße hat einen großen Einfluss auf die Wortprofile, denn Wortprofile sind in der Regel nur aussagekräftig, wenn das Lemma wenigstens 500, besser jedoch 1000 Mal im Korpus auftaucht (siehe auch Ivanova et al. 2008). Unter dieser Zahl ist die Aussagekraft eines Wortprofils nur begrenzt, da viele syntaktische Relationen dann in der Regel nur ein oder zwei Mal vorkommen und somit kaum nachweisbar ist, dass es sich bei den extrahierten syntaktischen Relationen um typische Beispiele und nicht um Zufallsfunde handelt. Insofern spielt auch die absolute Korpusgröße eine Rolle, als sich mit wachsender Korpusgröße auch die Anzahl der verschiedenen Wörter erhöht, die hochfrequent im Korpus vorkommen. Dabei stellt sich heraus, dass eine Korpusgröße von 100 Millionen laufenden Textwörtern zu klein ist, um eine für die zu erwartende Benutzungssituation ausreichende Anzahl von Wortprofilen zu extrahieren. So gibt es beispielsweise im 100 Millionen Textwörter umfassenden DWDS-Kernkorpus (s. unten) nur etwa 5.000 Lemmata, die mehr als 1.000 Mal vorkommen. Daraus ergibt sich, dass man weitaus größere Korpora benötigt, wenn man qualitativ ausreichende Wortprofile für eine Stichwor-

tanzahl erstellen möchte, die zumindest der eines Kompaktwörterbuchs entspricht.

Für das Deutsche liegen bislang keine Referenzkorpora geschriebener deutscher Standardsprache mit einer nach Textsorten ausgewogenen Verteilung vor[1]. Bei der Erstellung eines hinreichend großen Wortprofils ist man somit entweder auf Spezialkorpora oder auf eine Mischung aus Referenz und Zeitungskorpora angewiesen. Der derzeitige Prototyp des DWDS-Wortprofils (fortan kurz: Wortprofil) beruht auf einer Mischung eines Referenz- und eines Zeitungskorpus mit einer Gesamtgröße von 500 Millionen Tokens: dem DWDS-Kernkorpus, einem nach Textsorten ausgewogenen und zeitlich gleichmäßig gestreuten Referenzkorpus der deutschen Sprache des 20./21. Jahrhunderts (Geyken 2007, www.dwds.de/textbasis), ca. 110 Millionen Tokens) und dem ZEIT-Archiv (1946-2009, ca. 400 Millionen Tokens, www.zeit.de). Daraus wurden für etwa 90.000 Lemmata Wortprofile extrahiert mit etwa 2.000.000 syntaktischen Relationen. Knapp ein Fünftel aller Wortprofile basiert auf Lemmata, die mindestens 1000 Mal im Korpus vorkommen, also gemäß des o.g. Schwellwerts ausreichend für ein umfassendes Wortprofil sind. Als Schwellwert für die Mindestanzahl des Vorkommens einer syntaktischen Relation wurde die Zahl 4 gewählt. Damit soll verhindert werden, dass okkasionelle Verbindungen fälschlicherweise in das Wortprofil aufgenommen werden. Ob der Schwellwert von 4 zu hoch ist, lässt sich nur experimentell und in Abhängigkeit von einer Fragestellung beantworten. In der Folge soll die Nutzung sowie die Qualität des gegenwärtigen Wortprofils anhand einiger Beispiele illustriert werden.

4. Beispiele einiger Wortprofile

4.1 Feindbild

Das DWDS-Wortprofil kann über die Webseite des lexikalischen Informationssystems DWDS (www.dwds.de) genutzt werden. Im DWDS-Informationssystem werden für ein Suchwort, welches der Benutzer in die Suchmaske eingibt, die Ergebnisse in mehreren Fenstern (panels) präsentiert. Es gibt drei Typen von Fenstern: Wörterbuch-, Korpus- und Statistikfenster. Die Fenster lassen sich frei zu sogenannten Sichten (views) zusammenstellen. Das DWDS-Wortprofil ist eines der Fenster und Teil der Standardsicht des DWDS-Informationssystems.

Anhand des bereits in Abschnitt 1 erwähnten Wortes *Feindbild* sollen die verschiedenen Recherchemöglichkeiten mit dem Wortprofil im DWDS-Informationssystem illustriert werden. Die Standardansicht des Wortprofils erfolgt in Form einer Schlagwortwolke (Abb. 4). Man sieht, wie bei Schlagwortwolken üblich, signifikante Begriffe in einer größeren Schrift als weniger signifikante. Dabei basiert die Signifikanz auf der in Abschnitt 2.3 erläuterten Salienz: je salienter eine syntaktische Relation, desto größer die Schriftart. Im Unterschied zu einer klassischen Schlagwortwolke werden hier

1 Die Situation ist anders, wenn man Webcorpora betrachtet, also Korpora, die mit Crawling-Techniken aus dem Web gewonnen wurden. Dafür gibt es für das Deutsche bereits ein 1,6 Milliarden Token großes Korpus (Pomikalek et al. 2009). Allerdings bleibt zu zeigen, dass dessen Aussagekraft vergleichbar mit einem Korpus ist, dessen Quellen sich aus gesicherten, redaktionell bearbeiteten Texten speist.

somit die Begriffe über die reine statistische Signifikanz auch nach der syntaktischen Relevanz gewichtet (s. Abschnitt 1).

Für das bereits weiter oben erwähnte Wort *Feindbild* werden im DWDS-Wortprofil 32 verschiedene syntaktische Relationen mit insgesamt 384 Vorkommen extrahiert. Die Voraussetzung für die Aufnahme einer Relation in das Wortprofil ist, dass dafür wenigstens vier Belege im Korpus vorkommen. Abgedruckt ist die im Korpus am häufigsten verwendete flektierte Form.

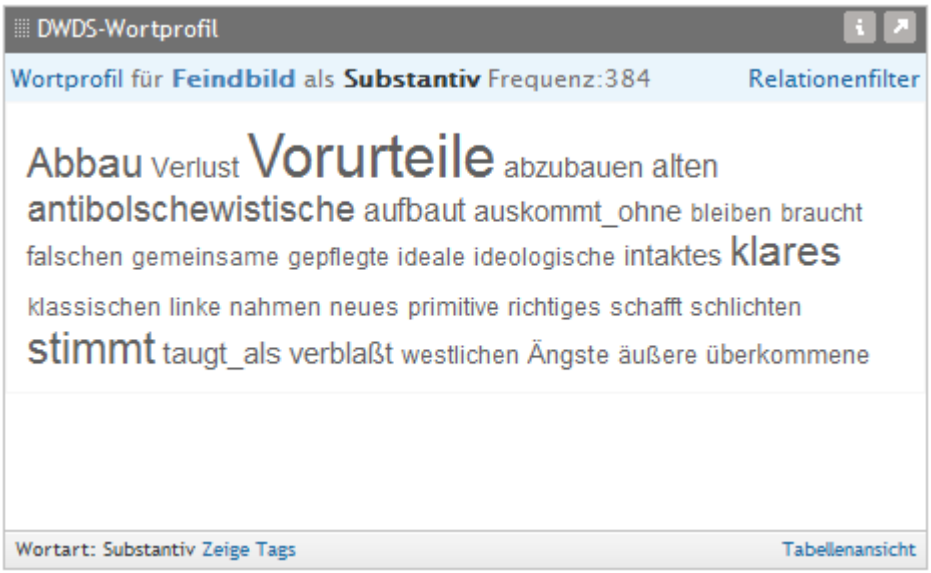

Abb. 4: Wortprofil in Schlagwortwolkenansicht

Neben der Darstellung in Form der Schlagwortwolke kann das Wortprofil auch in der klassischen Tabellenansicht präsentiert werden (Abb. 5).

DWDS-Wortprofil

relevanteste syntaktische Relationen für **Feindbild** als **Substantiv** Frequenz:384 — Relationenfilter

Stammform	Wortart	syntaktische Relation	Salience	Frequenz
Vorurteil	Substantiv	Feindbild hat die Beiordnung Vorurteile	14.65	15
stimmen	Verb	stimmt hat Aktiv-Subjekt Feindbild	11.42	18
klar	Adjektiv	klares ist Attribut von Feindbild	10.66	39
Abbau	Substantiv	Abbau hat Genitivattribut Feindbild	10.39	9
antibolschewistisch	Adjektiv	antibolschewistische ist Attribut von Feindbild	9.02	4
alt	Adjektiv	alten ist Attribut von Feindbild	7.71	103
aufbauen	Verb	aufbaut hat Akkusativobjekt Feindbild	7.09	8
taugen_als	PP	taugt hat Vergleichskonjunktion Feindbild	6.7	4
verblassen	Verb	verblaßt hat Aktiv-Subjekt Feindbild	6.37	4
auskommen_ohne	PP	auskommt hat Präpositionalphrase ohne_Feindbild	5.74	4

Abb. 5: Wortprofil in Tabellenansicht

In der Tabellenform werden in der ersten Spalte alle statistisch signifikanten Wortformen aufgeführt (Spalte Stammform). Spalte 2 verzeichnet die Wortart, bei mehreren Wörtern auch den Phrasentyp (beispielsweise PP, s. letzte Zeile der Abbildung). In Spalte 3 ist der syntaktischen Relationstyp beschrieben. Die Wortformen in Spalte 1 sind nach absteigender Salienz (s. Spalte 4) geordnet. Durch Klick auf die Spaltenüberschrift „Frequenz" (Spalte 5) wäre auch eine Sortierung nach der Frequenz möglich.

Die syntaktisch relevanten Nachbarn von *Feindbild* sind in den folgenden syntaktischen Relationstypen zu finden:

- Adjektiv-Nomen (Etikett: ATTR): *altes, antibolschewistisches, gemeinsames, gepflegtes, ideales, ideologisches, intaktes, klares, klassisches, linkes, neues, primitives, richtiges, schlichtes, verblasstes, westliches, äußeres, überkommenes Feindbild*
- Nomen-Nomen (im Genitiv) (GMOD): *Abbau*, *Verlust*
- Nomen-Koordination-Nomen (CJ): *Vorurteil*
- Nomen-Verb (SUBJ): *bleiben*, *stimmen*, *verblassen*
- Nomen-Verb (OBJA): *abbauen*, *aufbauen*, *brauchen*, *nehmen*, *schaffen*
- Verb-Präposition-Nomen (V PP): *auskommen_ohne*, *taugen_als*

Die Relationstypen lassen sich über das Wortprofil-Fenster ansteuern, indem man den Relationenfilter anklickt (Abb. 4). Es werden dann die syntaktischen Relationstypen aufgeklappt (Abb. 6). Klickt man dann auf eine der Relationstypen, beispielsweise auf „Attribut", erhält man alle Wortformen, die in einer Attributrelation (Adjektiv-Nomen) zum Wort *Feindbild* stehen. Diese Filter können bei hochfrequenten Wortprofilen sehr nützlich sein. Beispielsweise hat das geläufige Substantiv *Haar* 13509 Relationen (davon 532 verschiedene) im DWDS-Wortprofil. Eine Darstellung als Schlagwortwolke wäre hier sehr unübersichtlich. Durch das Filtern nach einzelnen Relationstypen hingegen erhält man homogene Listen und handhabbare Größen.

Abb. 6: Filtern nach Relationstyp Attribut

Ein wesentlicher Mehrwert des Wortprofils besteht darin, dass alle extrahierten Relationen stets mit den dazugehörigen Satzkontexten im Korpus verknüpft sind und somit einen Überblick über den Verwendungszeitraum und die semantischen und pragmatischen Kontexte ermöglichen, in denen die syntaktische Relation verwendet wird. Klickt mal beispielsweise in Abbildung 5 auf das erste Wort *Abbau*, welches zum Suchwort *Feindbild* in einer Attributrelation im Genitiv steht (Relation GMOD), gelangt man zu den insgesamt neun Kontexten, die vom Analysesystem Syncop aus dem Wortprofil-Korpus extrahiert wurden (Tabelle 1). Die Beispiele sind nummeriert und nach absteigendem Datum geordnet: der jüngste Beleg stammt aus dem Jahre 2000, der älteste von 1982. Die möglichen Textsorten (Belletristik, Gebrauchstext, Wissenschaft, Zeitung) sind in der dritten Spalte aufgeführt; diese ist mit einem Doppelpunkt von der Belegquelle getrennt. In der 4. Spalte findet sich der Kontext für die syntaktische Relation, wobei die gesuchten Wörter farblich hervorgehoben sind.

1 2000-06-29 Zeitung: ZEIT Wir brauchen einen Abbau der Feindbilder von den „illegalen Einwanderern“, bei denen in der Regel Täter und Opfer verwechselt werden.

2 1998-04-23 Zeitung: ZEIT Wir brauchen einerseits einen Abbau des Feindbilds von den illegalen Einwanderern, bei denen ohnehin in der Regel Täter und Opfer verwechselt werden.

3 1990-04-20 Zeitung: ZEIT Und was den Abbau der Feindbilder betrifft, die ja schon vor 1945 von Deutschland gegen die Sowjetunion und von der UdSSR gegen Deutschland geschaffen wurden:

4 1990-02-23 Zeitung: ZEIT Der Abbau der Feindbilder sollte den Sowjetbürgern eine europäische Identität und der Perestrojka westlichen Elan bringen.

5 1989-05-26 Zeitung: ZEIT Er mahnte aber auch die Bundesrepublik, in der Abrüstung nicht hinter der DDR zurückzubleiben (die gerade Truppen einseitig reduzierte) und wünschte sich einen Abbau des Feindbildes auch in der Bundesrepublik.

6 1989-05-05 Zeitung: ZEIT Bobowikows Beitrag ließ auch über die Wahlen hinaus unverblümt die ideologische Kriegserklärung gegen die Grundwerte der Perestrojka erkennen — gegen Glasnost im Inneren und gegen den Abbau der Feindbilder.

7 1988-10-21 Zeitung: ZEIT Sie verfolgt jedoch nicht mehr die Illusion des Ausspielens und Aufspaltens, weil sie das Ansehen und die Integration der Sowjetunion in die Weltgemeinschaft durch den Abbau der ideologischen Feindbilder fördern will.

8 1988-09-30 Zeitung: ZEIT Mit dem Abbau der alten Feindbilder tut sich die DDR offenbar schwer.

9 1982-05-15 Zeitung: ADG Er fordere deshalb die Demonstranten auf, auf beiden Seiten für einen Abbau der Feindbilder einzutreten, die Resignation gegenüber der weltweiten Aufrüstung abzuwehren und eine aktive gewaltfreie persönliche Haltung anzustreben.

Tabelle 1: Konkordanzansicht für Relation GMOD: Abbau, Feindbild

Die Stärke des syntaktischen Ansatzes zeigt sich bei der Extraktion von syntaktischen Funktionen. Hierfür reicht in der Regel nicht nur ein lokaler Kontext, sondern der gesamte Satz muss ausgewertet werden, da Komplemente von Verben im allgemeinen Fall nicht mit lokalen Funktionen extrahiert werden können. Ein Beispiel hierfür liefert die Relation Verb-Präpositionalphrase *auskommen ohne Feindbild*. Diese Relation enthält alle Sätze im Korpus, bei denen das Verb *auskommen* – mit und ohne abtrennbarem Verbpartikel – in Relation mit der Präpositionalphrase *ohne Feindbild* steht. Durch Klick auf *auskommen ohne* in Abbildung 4 gelangt man zu den Satzkontexten in Tabelle 2, bei denen die Konstruktion sowohl in Verbzweit- wie auch Verbletztstellung vorkommt.

auskommen_ohne:

1 1993-10-22 Zeitung:ZEIT Hamburg 1993; 206 S., 26, -DM fast ganz ohne dieses Feindbild auskommt:

2 1991-09-26 Zeitung:ZEIT Hysterie und Haß. Doch die Unfähigkeit des einstigen Freiheitshelden gegen Kreml und Kommunismus, der jetzt ohne Feindbilder nicht mehr auskommt, der über den Weltmarkt nichts, aber über die Weltverschwörung gegen Georgien alles weiß, hat Hysterie und Haß gesät.

3 1989-10-20 Zeitung:ZEIT „Das Land kam vorübergehend ohne Feindbilder aus „, heißt es im Begleitbuch.

4 1972-07-28 Zeitung:ZEIT Das kritische Denken des Autorenteams kommt ohne Feindbild nicht aus.

Tabelle 2: Konkordanzansicht für Relation V-PP: auskommen, ohne Feindbild

4.2 grau

Im vorangegangen Abschnitt wurden die verschiedenen Nutzungsmöglichkeiten des Wortprofils auf der Webplattform www.dwds.de erläutert. Noch nicht angeschnitten wurde die lexikographische Qualität des Wortprofils. Aufgrund des vom Korpus abgedeckten Zeitraums ist hierfür der Vergleich mit einem großen einsprachigen deutschen Gegenwartswörterbuch naheliegend. Für den Vergleich haben wir zwei große Wörterbücher herangezogen: das große Wörterbuch der deutschen Sprache in 10 Bänden des Dudenverlags (GWDS) und das Wörterbuch der deutschen Gegenwartssprache (WDG). Da der Vergleich relativ ausführlich dargestellt werden soll, um die Bandbreite der möglichen Vergleichsparameter zu illustrieren, soll er an dieser Stelle nur für ein Wort demonstriert werden, nämlich für das Adjektiv *grau*. Dieses Adjektiv haben wir ausgewählt, weil es häufig genug für ein ausgeprägtes Wortprofil ist, weil es mehrere Lesarten hat und weil es keine grundsätzlichen Bedeutungsveränderungen in den letzten Jahren erfahren hat. Aus urheberrechtlichen Gründen beziehen wir uns nur auf den Artikel des WDG (siehe Anhang), weisen aber an den entsprechenden Stellen auf die Unterschiede zum GWDS hin. In der Tat zeigt sind die Artikelbeschreibungen des Stichworts *grau* in beiden Wörterbüchern in ihrer Bedeutungsstruktur im großen und ganzen vergleichbar, wenn auch nicht gleich. Einige wichtige Unterschiede zwischen beiden Wörterbüchern lassen sich feststellen. Das GWDS hat die Lesarten 1 und 2 des WDG, bei der der konkrete und der bildliche Gebrauch beschrieben wird, zu einer zusammengefasst. Das GWDS führt im Gegenzug eine neue Lesart ein, bei der *grau* im Sinne von „sich an der Grenze der Legalität bewegen“ (umgangssprachlich) verwendet wird, wie beispielsweise in *graue Händler*. Darüber hinaus verzeichnet das GWDS einige geläufige syntaktische Relationen, die das WDG nicht aufführt, beispielsweise *graue Schläfen*, *graues Brot* oder *grau meliert*. Umgekehrt fallen auch einige Relationen weg, wie beispielsweise *graue Theorie*, *graue Maus* oder *in Ehren ergraut*. Vom Umfang her halten sich beide Wörterbücher die Waage.

Der quantitative Vergleich der Wörterbücher mit dem Wortprofil zeigt zunächst einmal folgendes: das WDG verzeichnet 43 syntaktische Relationen, das GWDS nur 25 Relationen. Viele der Relationen im WDG sind jedoch „erwartbar“, wie beispielsweise *Bart* oder *Haarsträhne*, wenn man ohnehin *Haar* aufführt. Insofern bietet das WDG hier keinen Bedeutungsüberschuss gegenüber dem GWDS. Im DWDS-Wortprofil werden 7727 Relationen extrahiert, darunter 388 verschiedene Relationen (f>=4, sal>0). Dies ist mehr als das Zehnfache der im Wörterbuch aufgeführten Wortverbindungen. Es stellen sich somit die Fragen nach der gemeinsamen Schnittmenge von Wortprofil und WDG (markiert als (+WP;+WDG)), nach denjenigen Kookur-

renzen, die im Wortprofil, aber nicht im WDG sind (+WP;-WDG), und umgekehrt nach der Qualität derjenigen Wortverbindungen, die nicht im Wortprofil, dafür aber im WDG sind (-WP;+WDG).

(+WP;+WDG)

Zunächst einmal ergibt der Abgleich von Wortprofil und WDG, dass sich in der Schnittmenge folgende 29 Kookurrenzen mit *grau* befinden:

Vorzeit, Maus, Haare, Alltag, Bart, Himmel, Star, Wolke, Mauer, Theorie, Altertum, Haarsträhnen, Uniform, Augen, Haupt, Kostüm, Gestein, alt, färben, Scheitel, Elend, Öde, Regenwolken, Wölkchen, Morgen, Meer, Gesicht, Haut, Strand

Nimmt man noch Synonyme oder nahe Hyperonyme hinzu, so findet man auch weitere zwei WDG-Einträge im WP: Frühlicht (WP: *Morgenlicht*), Ferne (WP: *Unendlichkeit*)

(+WP;-WDG)

Über 90% aller salienten syntaktischen Relationen, genauer 369 der 398 Relationen des Wortprofils, sind nicht im WDG. Diese Zahl sagt zunächst nichts über die lexikalische Relevanz dieser Relationen aus. Bei unserem Vergleich haben wir daher die Einträge auf ihre lexikalische Relevanz angesehen, wobei wir einen Unterscheidung zwischen hoch salient (sal>10), salient (sal>5) und schwach salient (sal<5) gemacht haben.

Hoch salient sind insgesamt 54 syntaktische Relationen, 10 davon im WDG; unter den verbleibenden finden sich folgende lexikographisch relevante Wortverbindungen (geordnet nach absteigender Salienz):

- *graue Eminenz*
- *graue Maus* (fehlt im WDG im übertragenen Sinne)
- *grauer Kapitalmarkt*
- *graue Schläfen*
- *grauer Pfandbriefmarkt* (im Sinne von: geschlossener Fonds)
- *Graue Ackerschnecke* (limax agrestis)
- *graue Zellen*

Ferner fehlen folgende Eigennamen-Kontexte:

- *Graue Panther* (politische Partei)
- *Graue Wölfe* (extremistische politische Partei)
- *Graue Kloster* (Gymnasium in Berlin)

Salient (Sal>5) sind weitere 143 Wortverbindungen im Wortprofil; davon sind 12 im WDG. Unter den verbleibenden 132 Relationen sind die lexikographisch relevanten Wortverbindungen (geordnet nach absteigender Salienz):

- *graues Einerlei*
- *graues Umweltpapier*
- *graue Asche* (von Toten oder von verbranntem Material)
- *graue Habenzinsen* (auf dem nicht freien Kapitalmarkt)
- *grauer Markt*
- *grauer Haarkranz*
- *grauer Nadelstreifen*
- ferner alle Adjektiv-Modifizierer: beispielsweise 'schütter' wie in *schütteres graues Haar*, ‚voll' wie in *voll grauer Haare*

Schwach salient: unter den Wortverbindungen mit einer Salienz < 5 sind nur noch wenige lexikographisch interessante Verbindungen zu finden; u.a. *graue Erbsen*, *graue Liste* oder der *Graue Burgunder*. Die meisten sind transparent und erwartbar, wie beispielsweise *graue Krawatte*, *graue Socke* oder *graue Matte*.

(-WP;+WDG)

Umgekehrt finden sich 14 Wortverbindungen im WDG, die nicht im Wortprofil auftauchen:

Frühlicht, Stoff, wie eine Maus, bei Nacht sind alle Katzen grau, keine grauen Haare wachsen lassen, in Ehren grau geworden, Gesichtsfarbe, Woche, Tag, Monat, Mittelalter, vor grauen Jahren, Ferne, Zukunft

Diese Zahl scheint auf den ersten Blick recht hoch, relativiert sich aber schnell, wenn man sie anhand ihrer Gebräuchlichkeit analysiert und mit den Alternativen vergleicht, die das Wortprofil für einige dieser Wortverbindungen bereitstellt. In der Folge soll dies anhand der Lesarten des WDG detaillierter gezeigt werden.

WP und WDG im Lesartenvergleich

Das WDG unterscheidet insgesamt fünf Lesarten von *grau*.

Lesart 1: /Mischfarbe aus Schwarz und Weiß/

(+WP;+WDG): *Auge, Gestein, Haut, Himmel, Kostüm, Mauer, Maus, Meer, Regenwolke, Strand, Uniform, Wolke*

(+WP;-WDG): unter anderem *Granit, Beton, Anzug, Kittel, Kutte*

(-WP;+WDG): *Stoff*. Dafür sind im Gegenzug im WP einige Synonyme, wie *Flanell*, *Samt* oder *Wolle*. Die Vergleichskonjunktion *grau wie eine Maus* findet sich nur im WDG (auch nicht im GWDS); im Gegenzug ist aber im WP das wesentlich häufiger gebrauchte *graue Maus* enthalten, mit beispielsweise folgendem Beleg: *Arminia*

Bielefeld hat schon seit ein paar Jahren den Ruf als „graue Maus" der Liga weg. Die ZEIT, 26.12.2008

Lesart 2a: /farblos, bleich, durch zunehmendes Alter/

(+WP;+WDG): *Bart, Haar, Haarsträhne, Haupt, Scheitel, Star, alt*

(+WP;-WDG): u.a. *Schläfen, Strähne, Haarkranz*

(-WP;+WDG):

sich keine grauen Haare wachsen lassen (bildlich, umgangssprachlich im Sinne von sich keine Sorgen machen); wird als Konstruktionsmuster vom Wortprofil derzeit nicht extrahiert, ist aber unter den Beispielen von „grau Haar" mehrfach zu finden.

in Ehren grau geworden: taucht in den Korpora insgesamt 4 Mal auf, ist aber im Wortprofil nicht salient; aber dafür gibt es über 100 Treffer im Korpus für "in Ehren ergraut". Diese sind unter dem Wortprofil von „Ehre" zu finden.

Lesart 2b: /farblos, bleich, durch Blutleere/

(+WP;+WDG): *Gesicht*

(+WP;-WDG): keine Einträge

(-WP;+WDG):

graue Gesichtsfarbe: ist im WP nicht enthalten, da es bereits im zugrundeliegenden Korpus mit insgesamt nur drei Belegen unter dem derzeit gewählten Salienzschwellwert von 4 liegt.

vor Haß (kommt im WDG nur als Literaturbeleg vor (Marchwitza)). In der Korpora gibt es diesen Beleg gar nicht. Auch im Web findet man hierfür nur einen Treffer[2] – und dieser stammt noch dazu aus dem WDG: „ Sie starrte mich grau vor Haß an" (Marchwitza Jugend 248)

Lesart 3: /übertragen/, trostlos und öde

(+WP;+WDG): *Alltag, Elend, Elend, Morgen, Theorie, öde*

Letztere Konstruktion ist im WDG nur durch das Goethezitat belegt: *grau, treuer Freund, ist alle Theorie* . Im Wortprofil finden sich dazu 140 Belege mit anderen Konstruktionen, die illustrieren, dass „graue Theorie" nicht nur im Zusammenhang mit dem Goethe-Zitat gebraucht wird, z.B.

- Das Ethos ist kein Istzustand, es ist ein Sollzustand, trotzdem nicht graue Theorie, sondern eine sehr praktische Sache. (ZEIT, 04.01.2010)
- Und jede Eröffnungsvorbereitung graue Theorie sein lässt. (ZEIT, 28.8.2009)
- Dass dies keine graue Theorie ist, zeigt ein Fall in den USA. (ZEIT, 18.5.2009)

2 Suchanfrage unter www.google.de vom 6.12.2010

- Alles graue Theorie, die uns Winzer in die Illegalität treibt. (ZEIT, 21.11.2008)
- Doch wie sich zeigte, blieb das graue Theorie. (ZEIT, 27.8.2008)
- Die Betriebshandelsspanne erweist sich, nämlich, als graue Theorie. (ZEIT, 24.7.2008)
- …

(+WP;-WDG): keine Einträge

(-WP;+WDG): *Monat*, *Tag* und *Woche*. Alle drei Begriffe entstammen einem Literaturbeleg von Feuchtwanger (s. Anhang). Diese Kookurrenzen sind nicht salient, dafür sind aber nahe Synonyme im Wortprofil: *Herbsttag*, *Nachmittag*, *Novembertag* oder *Vormittag*.

das graue Elend kriegen (sich tief unglücklich fühlen, zeigen) ist nicht im Wortprofil, dafür aber die einfachere Attribut-Nomen-Relation *graues Elend*.

Lesart 4: /weit zurückliegend längst vergangen/

(+WP;+WDG): *Altertum, Vorzeit*

(-WP;+WDG): *Mittelalter*

Lesart 5: /unbestimmt, ungewiss/

(+WP;+WDG): keine Relationen sind im Schnitt

(-WP;+WDG): *graue Ferne, Zukunft*; diese sind beide nicht im Wortprofil salient.

Zusammengefasst noch einmal die wichtigsten Schlüsse, die sich aus dem Beispiel *grau* ableiten lassen. Wortprofile können im Vergleich zu großen einsprachigen Wörterbüchern ein Vielfaches der syntaktischen Relationen zu einem Wort enthalten. So verzeichnet das Wortprofil zu *grau* knapp 400 verschiedene Relationen, wohingegen das WDG 43, das GWDS nur 25 Wortverbindungen aufführen. Dies schlägt sich im auch direkten Vergleich nieder: mit einer hohen Salienz (sal>5) enthält das Wortprofil mehr als 20 lexikographisch relevante Beispiele, die nicht im WDG verzeichnet sind. Bei einer Salienz von unter 5 im Wortprofil nimmt die Dichte der lexikographisch relevanten Relationen stark ab. Von den knapp 200 syntaktischen Relationen (s<5) sind nicht mehr als 5 als lexikographisch relevant einzustufen. Erstaunlich ist zunächst, dass das Wortprofil nur etwa 70% der im Wörterbuch verzeichneten Wortverbindungen als syntaktische Relation enthält. Zu den meisten dieser fehlenden Verbindungen führt das Wortprofil jedoch gebräuchlichere Alternativen auf; in anderen Fällen existiert die Wortverbindung nur als Literaturzitat. Man findet im Wortprofil auch eine zusätzliche Lesart, die zwar nicht im WDG, jedoch im GWDS verzeichnet ist. Das Wortprofil hat aber hier die gebräuchlicheren Beispiele: *grauer Kapitalmarkt* oder *grauer Markt* (WP) statt *graue Händler* oder *graues Material* (GWDS). Schließ-

lich, das zeigt das Beispiel *graue Theorie*, findet man mit dem Wortprofil zahlreiche authentische Kontexte und Verwendungen, die von dem einzigen dazu im WDG aufgeführten Goethezitat (*grau, mein Freund, ist alle Theorie*) abweichen. Eine Konstruktion übrigens, die in das GWDS nicht mehr aufgenommen wurde.

5. Potentiale des Wortprofils

Das Beispiel *grau* in Abschnitt 4.2 macht deutlich, dass in dem gegenwärtigen DWDS-Wortprofil mit einem zugrundeliegenden Korpus der Größe von 500 Millionen Textwörtern eine große Reichhaltigkeit von Wortverbindungen steckt, die es auch im Vergleich mit großen einsprachigen Wörterbüchern interessant macht. Bevor wir zum Abschluss dieses Beitrags kurz auf die Potenziale dieses Ansatzes eingehen, sollen zuvor noch ein paar einschränkende Anmerkungen zur Reichweite der in 4.2 gemachten Beobachtungen gemacht werden.

Erstens wird es notwendig sein, diese Befunde auf der Basis aller offenen Wortarten anhand einer hinreichend großen Anzahl von Wörtern zu replizieren. Eine entsprechende Studie, die einen Vergleich von 50 Wortprofilen mit entsprechenden Einträgen in WDG-Artikeln ähnlich dem in Abschnitt 4.2 durchführt, ist derzeit im Gange. Dabei zeigt sich, und das ist die zweite Relativierung, dass die Abdeckung der Relationsextraktion mit dem Parser Syncop für den Verbbereich im Allgemeinen weitaus geringer ist als für den nominalen und den adjektivischen Bereich. Dies ist deshalb unmittelbar einleuchtend, da viele syntaktische Relationstypen von Nomina und Adjektiven mit lokalen Abhängigkeiten zu erfassen sind, wohingegen bei den Verben syntaktische Funktionen, also phrasenübergreifende Abhängigkeiten zu extrahieren sind.

Die dritte Relativierung schließlich betrifft die Evaluierung der Qualität der extrahierten Relationen in Bezug auf deren Status als Kollokation. Hierzu lässt sich folgendes feststellen: Syntaktische Relationen weisen, wenn sie statistisch signifikant sind, eine Ähnlichkeit zu dem weit verbreiteten Kollokationsbegriff von Hausmann (1984) auf, weichen aber in zwei Aspekten entscheidend davon ab. Kollokationen im Hausmannschen Sinne bestehen aus einer Kollokationsbasis und einem Kollokator. Beispiele hierfür sind *schütteres* **Haar, Termin** *einhalten* oder *hoch* **erfreut** (die Basis ist hier fettgedruckt, der Kollokator kursiv). Syntaktische Relationstypen hingegen sind symmetrisch, d.h. wenn das Wort A in Relation mit dem Wort B steht, so steht auch umgekehrt B in Relation mit A. Der Grund hierfür ist ein pragmatischer, da bei automatisch extrahierten syntaktischen Relationen Basis und Kollokator nicht sicher bestimmt werden können und man somit die Abdeckung des Wortprofil-Ansatzes für den Nutzer erhöht, wenn man die Umkehrbarkeit der Relation und somit den Zugriff auf die Kollokation sowohl über die Basis als auch den Kollokator zulässt. Ganz abgesehen davon ist es bei einer Datenbankrepräsentation immer möglich, die Kollokation sowohl unter der Basis als auch unter dem Kollokator unterzubringen. Darüber hinaus sind bei Kollokationen im Hausmannschen Sinne die Menge der Kollo-

katoren für eine gegebene Basis sehr eingeschränkt. Vor allem geht es darum, „nicht erwartbare" Kollokatoren zu beschreiben. In einem völlig automatisierten System ist dies nicht möglich. Hier lassen sich für ein gegebenes Wort die syntaktisch relevanten Kookurrenzpartner nur über die statistische Signifikanz ermitteln. Diese aber ist im allgemeinen nicht gleichbedeutend mit einer Erwartbarkeit. Beispielsweise führt das DWDS-Wortprofil für das o.g. Nomen *Haar* 223 Adjektive auf, die allesamt statistisch signifikant in Bezug auf das Suchwort sind. Darunter gibt es eine ganze Reihe von „echten", weil nicht erwartbaren Kollokationen, wie beispielsweise *schütter*, *gegelt* oder *gescheitelt*, aber auch andere erwartbare signifikante Kollokate wie *schwarz*, *blond*, *lang* oder *nass*. Obzwar diese Kookurrenzen keine Kollokationen im Hausmannschen Sinne sind, so sind viele davon auch nicht beliebig, sondern semantisch oder pragmatisch motiviert. Beispiele hierfür wären für *Ball* beispielsweise *Ball abspielen*, *Ball zuwerfen*, oder *Recht anwenden*, *Recht brechen*, *Recht auf Mitbestimmung* für *Recht*.

Jenseits dieser drei Einschränkungen sind jedoch die Potenziale künftiger Entwicklungen von statistischen Wortprofilen enorm. Die Erweiterungen liegen zunächst in der Verbesserung der Abdeckung der Parser zur Extraktion der syntaktischen Relationen, vor allem aber in der Erweiterung und Vergrößerung der Korpusbasis. Wie bereits weiter oben erwähnt, lassen sich mit größeren, nach Textsorten breit gestreuten Korpora sinnvolle Wortprofile für weitaus mehr Stichwörter bilden und die Konfidenz der Salienz syntaktischer Relationen verbessern.

Darüber hinaus sind Wortprofile als Ressource für die Textproduktion von großen Wert. Mit statistischen Wortprofilen lassen sich sowohl typische Wortverbindungen für ein Wort auffinden als auch deren semantische und pragmatische Kontexte in authentischen Textzusammenhängen mit wenigen Mausklicks nachschlagen. Hierzu noch einmal ein Beispiel: Im GWDS beispielsweise findet man, wenn man unter dem Stichwort *grau* die Relation *grau meliert* nachschlägt, einen authentischen Textbeleg:

grau melierte Schläfen; sein Haar war inzwischen g. meliert, was seine Erscheinung noch würdevoller machte (Danella, Hotel 30) (GWDS, Eintrag grau, Lesart 1)

Das Beispiel deutet an, dass sich die Relation *grau meliert* in der Regel auf eine Person bzw. dessen Haar bezieht. Die Schlussfolgerung darüber, ob sich dies nur auf eine männliche Person bezieht oder ob Träger desselben positiv konnotiert wird, bleibt jedoch der Intuition des Nutzers überlassen. Mit dem Wortprofil lässt sich diese Vermutung mit einem Klick an nicht nur einem, sondern vielen weiteren authentischen Korpusbeispielen überprüfen. So findet man im DWDS-Wortprofil die Verbindung *grau meliert* mit einer Salienz von 38 und einer Frequenz von 51, die sich ausschließlich auf männliche Personen beziehen. Darüber hinaus belegen zahlreiche dieser Kontexte die positiven Konnotationen, die entweder auf den sozialen Status des Trägers oder andere pragmatische Kontexte zurückzuführen sind. Jeweils ein Beispiel hierzu:

Eduardo Montes ist ein stolzer Mann. *Grau meliertes* Haar, hellblaues Maßhemd, fein gepunktete Krawatte – fast zu elegant für sein profanes Chefzimmer in der Hofmannstraße. (ZEIT, 04.10.2006)

Oder dass man gute Kunden zum Essen in einen der sechs Speisesäle einlädt, wo

grau melierte Kellner auf Kosten des Hauses schon mal einen Grand Cru entkorken. (ZEIT, 31.10.2001)

Künftige Experimente mit anderen und größeren Korpora werden zeigen, inwieweit Werkzeuge dieser Art ebenso natürlich für das Verfassen von Texten sein werden, wie heutzutage bereits die Rechtschreib- und Grammatikkontrolle oder die Synonymvorschläge in Textverarbeitungen.

Literatur

Braun, S.; Kohn, K. und J. Mukherjee (2006). *Corpus Technology and Language Pedagogy*. Peter Lang. Frankfurt.

Church, K. W., und Hanks, P. (1989). Word association, norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83.

Didakowski, J. (2007): SynCoP - Combining syntactic tagging with chunking using WFSTs. Linguistik in Potsdam, In: Proceedings of FSMNLP 2007. Universitätsverlag Potsdam.

Fritzinger, F., Kisselew, M., Heid, U., Madsack, A., Schmid, H. (2009). Werkzeuge zur Extraktion von signifikantenWortpaaren als Web Service. Vortrag: GSCL Symposium Sprachtechnologie und eHumanities, Duisburg, 26-27 Februar 2009

Foth, K. (2005). Eine umfassende Dependenz-Grammatik des Deutschen. Universität Hamburg.

Hausmann, F.-J. (1984). Wortschatzlernen ist Kollokationslernen. In: Praxis des neusprachlichen Unterrichts. 31. Jg. (1984), S. 385-406.

Geyken, A.; Hanneforth, Th. (2005). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In: Proceedings of FSMNLP 2005, Lecture Notes in Artificial Intelligence. Springer.

Geyken, A. (2007): *The DWDS corpus: A reference corpus for the German language of the 20th century*. In: Fellbaum, Christiane (ed.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum Press, S. 23-41.

Geyken A., Didakowski, J. und Siebert, A. (2009). Generation of word profiles for large German corpora. In Yuji Kawaguchi, Makoto Minegishi and Jacques Durand (ed.). Corpus Analysis and Variation in Linguistics, p. 141–157.

Ivanova, K., Heid, U., Schulte im Walde, S., Kilgarriff, A, Pomikálek, J. (2008). Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. Proceedings of the 6th Conference on Language Resources and Evaluation. Marrakech, Morocco.

Jurish, B., (2003). A Hybrid Approach to Part-of-Speech Tagging. Final report, Project Kollokationen im Wörterbuch, BBAW, Berlin.

Kilgarriff, A.; Rychly, P., Smrz, P., and D.Tugwell (2004). The Sketch Engine. In Proceedings Euralex 2004. Lorient, France, July: 105-116.

Koskenniemi, K. (1990). Finite State Parsing and Disambiguation. In Proceedings of COLING, vol. 2, 229-232.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. COLING-ACL98. Montreal.

Pomikalek, J.; Rychly, P. and Kilgarriff, A (2009). *Scaling to Billion-plus Word Corpora*. Advances in Computational Linguistics. Special Issue of Research in Computing Science Vol 41, Mexico City.

Schiehlen, M. (*2003). A cascaded finite-state parser for. German*. In Proceedings of the 10th EACL, Budapest.

Wörterbücher

[GWDS] Duden - Das große Wörterbuch der deutschen Sprache in 10 Bänden (1999). Bibliographisches Institut, Mannheim. 3. Auflage.

[WDG] Klappenbach, R. und Steinitz, W. (Hg.) (1964-1977). *Wörterbuch der deutschen Gegenwartssprache (WDG)*. Berlin: Akademie-Verlag.

Anhang

Wörterbuchartikel *grau* im WDG

grau /Adj./ **1.** /*Mischfarbe aus Schwarz und Weiß*/ die g. Haut des Elefanten; g. wie eine Maus; ein g. Stoff, Kostüm; eine g. Uniform; er hat g. Augen; g. Gestein; g. Mauern; Am grauen Strand, am grauen Meer /Und seitab liegt die Stadt Storm 8,194; g. Rauch-, Regenwolken; der Himmel ist ganz g.; g. Frühlicht; Grau ist"s immer, wenn ein Morgen naht Brecht *Gedichte* 278; /*sprichw.*/ bei Nacht sind alle Katzen g. (*nachts erkennt man keine äußerlichen Unterschiede*) ein helles, dunkles, düsteres, bleiches, kaltes, fahles, farbloses Grau; das Grau der Regenwolken; das Blau des Himmels ging in Weiß und Grau über; eine Bluse in Grau; die Dame in Grau; sie kam in Grau, bevorzugt die Farbe Grau; *vgl.* Grauchen **2.** *farblos, bleich* **a)** *durch zunehmendes Alter*: g. Haar; eine g. Haarsträhne; Sie hatte bereits einen grauen Scheitel G. Hauptm. 4,552; ein g. Bart; Achtung vor einem g. Haupte (*einem alten Menschen*) haben sein g. Haar färben; alt und g. werden; er ist in Erhren g. geworden; der g. Star (*krankhafte Trübung der Linse im Auge*) /*bildl.*/ u m g . darüber, deshalb brauchst du dir keine g. Haare wachsen zu lassen (*darüber brauchst du dir keine Sorgen zu machen*) **b)** *durch Blutleere*: ein g. Gesicht; mein mageres graues Gesicht und die Trostlosigkeit meines Blickes Böll *Wort* 12; eine g. Gesichtsfarbe; Der alte Mann wurde ganz grau im Gesicht Brecht *Dreigroschenroman* 296; Sie starrte mich grau vor Haß an Marchwitza *Jugend* 248 **3.** /*übertr.*/ *trostlos, trübe, öde*: Es hieß warten, einen grauen Morgen und einen grauen Tag und eine graue Woche und einen grauen Monat Feuchtw. *Tag* 78; In Dresden ging der graue Alltag wieder los Renn *Kindheit* 23; Grau, teurer Freund, ist alle Theorie Goethe *Faust* I 2038; ihr erschien die Welt g. und öde; u m g . s c h e r z h . das g. Elend kriegen (*sich tief unglücklich fühlen, zeigen*)◎ ich kann das Grau in Grau unserer Nachkriegsepoche auf die Dauer nicht aushalten G. Hauptm. *Sonnenuntergang* I **4.** *weit zurückliegend, längst vergangen*: in g. Vorzeit; im g. Altertum, Mittelalter; vor g. Jahren, Zeiten; **5.** *unbestimmt, ungewiß*: das liegt noch in g. Ferne, Zukunft;

Quelle: WDG – elektronische Version, www.dwds.de