Die dynamische Verknüpfung von Kollokationen mit Korpusbelegen und deren Repräsentation im DWDS-Wörterbuch

Alexander Geyken geyken@bbaw.de, Tel.: +49 (0)30 20370-390

1. Einführung

Die Beschreibung von Kollokationen im DWDS-Wörterbuch (DWDS = Digitales Wörterbuch der deutschen Sprache) nutzt die Möglichkeiten des digitalen Mediums in mehrfacher Weise: erstens dadurch, dass die Extraktion der Kollokationen mittels statistischer Methoden nahezu vollständig korpusbasiert erfolgt. Dadurch ist stets eine transparente und nachprüfbare Rückbindung von der im Wörterbuch aufgeführten Zitierform der Kollokation zu allen Korpusbelegen für diese Kollokation möglich. Zweitens wird es durch die Trennung der Kollokation im Wörterbuch von ihren Korpusbeispielen möglich, die Verbindung der Kollokation im Wörterbuch zu den Kollokationsbelegen dynamisch zu halten. Angesichts wachsender Korpora ist dies eine wichtige Eigenschaft, da damit die Aktualität des Wörterbuchs auf Belegebene stets gewährleistet ist. Eine dritte Charakteristik der Kollokationsbeschreibung im DWDS ist die Entkopplung von Kollokation und Belegen im Redaktionssystem. Im Redaktionssystem wird nur die Zitierform der Kollokation angegeben, die Rückbindung auf die Korpusbelege erfolgt dann über automatisch berechnete Verweise auf die Korpusbelege. Schließlich ergeben sich über die dynamische Verknüpfung von Zitierform und Korpusbelegen auch keine Platzprobleme: nur die Zitierform wird im Wörterbuch festgehalten, die Kollokationsbelege liegen in einer eigenen lexikalischen Datenbank und können nach Bedarf vom Nutzer des Systems angefordert werden.

In diesem Beitrag soll zunächst der Hintergrund des DWDS-Wörterbuchs dargestellt werden. Im zweiten Abschnitt erfolgt eine kurze Charakterisierung des im DWDS-Wörterbuch verwendeten Kollokationsbegriffs. Dessen Einbettung in die Wörterbuchstruktur des DWDS-Wörterbuchs wird im dritten Abschnitt beschrieben. Das eigentliche digitale Herzstück der Kollokationsbeschreibung im DWDS-Wörterbuch ist das DWDS-Wortprofil, eine auf syntaktischer Analyse und statistischer Auswertung basierende automatische Kollokationsextraktion, deren Grundlagen und Qualität in Abschnitt 4 dargestellt werden. In Abschnitt 5 soll anhand einiger Beispiele illustriert werden, wie die Arbeitsteilung der automatischen Kollokationen und der lexikographischen Intuition in der täglichen lexikographischen Arbeit aussieht. Schließlich geben wir im letzten Abschnitt einen Ausblick auf die künftige Arbeit.

2. Hintergrund: Das Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS)

Ziel des an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) beheimateten Projekts Digitales Wörterbuch der deutschen Sprache (DWDS) ist die Schaffung eines „Digitalen Lexikalischen Systems“ – eines umfassenden, jedem Benutzer über das Internet zugänglichen Informationssystems, das Auskunft über den deutschen Wortschatz in Vergangenheit und Gegenwart gibt. Das Projekt ist in drei Phasen von jeweils 6 Jahren zwischen 2007 und 2024 geplant (Klein/Geyken 2010). Während der ersten Phase werden drei an der BBAW und ihren Vorgängerinstitutionen erarbeitete Wörterbücher digital aufbereitet und in das Informationssystem des DWDS eingebunden: das historische Deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm, das Etymologische Wörterbuch des Deutschen (erarbeitet

unter Leitung von Wolfgang Pfeifer) und das zwischen 1961 und 1977 publizierte Wörterbuch der deutschen Gegenwartssprache ([WDG]). Ferner werden die in den Projekten DWDS und Deutsches Textarchiv erstellten Korpora in das DWDS-Informationssystem integriert und gemeinsam mit den Wörterbüchern abfragbar gemacht. Insbesondere sind dies das zeitlich und nach Textsorten ausgewogene DWDS-Kernkorpus (Geyken 2007, 100 Millionen Textwörter), das historische Referenzkorpus des Deutschen Textarchivs (ca. 50 Millionen Textwörter, www.deutschestextarchiv.de) und die vorwiegend aus elektronischen Zeitungsquellen stammenden DWDS-Ergänzungskorpora mit einer Größe von derzeit etwa 2,5 Milliarden laufenden Textwörtern. Alle Korpora werden fortlaufend erweitert. Ein Kernstück der ersten Projektphase besteht in der Formalisierung und Überführung der elektronischen Fassung des WDG in eine lexikalische Datenbank: das DWDS-Wörterbuch (Herold/Geyken 2008, Herold 2011).

Das DWDS-Wörterbuch unterscheidet sich vom WDG nicht nur durch die strengere Schemabeschreibung, sondern stellt auch eine moderate lexikographische Überarbeitung des WDG dar, insbesondere hinsichtlich Änderungen von ideologisch belasteten Wortartikeln und der Anpassung an die neue Rechtschreibung (Klein/Geyken 2010). Ziel des DWDS-Wörterbuchs ist die Erstellung eines auf wissenschaftlichen Prinzipien beruhenden synchronen Wörterbuchs mit gewissen historischen Anteilen (im Wesentlichen: Belegchronologie und etymologische Angaben bzw. Wortgeschichten). Das DWDS-Wörterbuch wird während der zweiten und dritten Projektphase in den Jahren 2013 bis 2024 bearbeitet werden. Mit Beginn der zweiten Phase wird sich die Anzahl der Projektmitarbeiter auf 10 Lexikographen erhöhen, die eine Wortstrecke von etwa 25.000 hochfrequenten Lemmata bearbeiten werden, die entweder zu neu waren, um in das WDG aufgenommen zu werden oder aus anderen Gründen keine Aufnahme in das WDG fanden. In der dritten Phase soll der Grundbestand des DWDS-Wörterbuchs auf allen Ebenen der Mikrostruktur überarbeitet werden. Im Folgenden wird nur ein kleiner Ausschnitt daraus beschrieben, derjenige der Kollokationen.

3. Kollokationen im DWDS-Wörterbuch

Als Kollokationen werden im DWDS-Wörterbuch Mehrwortverbindungen codiert, deren Gesamtbedeutung sich zwar aus der Bedeutung der Einzelwörter erschließen lässt, die aber dennoch keine beliebige Kombinatorik aufweisen, sondern in gewissem Maße als Muster im Sprachgedächtnis abgespeichert sind. Eine pragmatische, weil für die lexikographische Arbeit praktisch verwendbare Charakterisierung der Kollokationen basiert auf Hausmann. Kollokationen im hausmannschen Sinne sind „typische, spezifische Zweier/Dreier Beziehungen zwischen Wörtern“, die aus Basis und Kollokator bestehen (Hausmann 1984). Diese Kollokationen werden im DWDS-Wörterbuch als Kollokationen vom Typ1 bezeichnet. Beispiele hierfür sind (die Basis ist hier fettgedruckt, der Kollokator kursiv) *schütteres* ***Haar***, *heikles* ***Thema***, ***Unfall*** *bauen* (*bauen* in dieser Bedeutung ist nicht ohne *Unfall* denkbar…). Die Terminologie im DWDS weicht hier von Hausmann ab, der unter Kollokationstypen die möglichen syntaktischen Relationen zwischen Basis und Kollokator fasst. Auf diese wird in Abschnitt 5.1 näher eingegangen.

Kollokationen vom Typ1 stehen im Gegensatz zu den unspezifischen Wortverbindungen wie *Haus bauen* oder *Buch kaufen*. Zwischen diesen beiden Polen stehen Kollokationen vom Typ2. Dies sind Mehrwortverbindungen, bei denen es mehrere Möglichkeiten für Kollokate gibt, die jedoch nicht beliebig, sondern semantisch oder pragmatisch motiviert sind. Beispiele hierfür sind: *Ball abspielen*, *Ball zuwerfen*, *Recht anwenden*, *Recht brechen*, *Recht auf Mitbestimmung*, *Recht des Stärkeren*. Beiden Typen (Typ1 und Typ2) von Kollokationen ist ge-

mein, dass mit ihnen die „nicht erwartbaren“ Mitspieler (Kollokatoren) beschrieben werden sollen. Grenzfälle zwischen Typ1 und Typ2 können dann auftreten, wenn es zwei oder mehr Kollokatoren gibt, von denen eines besonders häufig auftritt.

Kollokationen werden im DWDS-Wörterbuch auch abgegrenzt von den Verbindungen, deren Gesamtbedeutung gar nicht oder nur teilweise aus den Einzelbedeutungen erschließbar ist. Diese werden in der Werkstattsprache des DWDS-Wörterbuchs als Phrasem codiert. Wir verwenden die Bezeichnung „Phrasem“ hier in einem weiteren Sinn als in der linguistischen Literatur üblich, wo es nur idiomatische Wendungen bezeichnet. Darüber hinaus werden Phraseme im DWDS-Wörterbuch auch zur Codierung von Grußformeln, Sprichwörtern, rhetorischen Fragen etc. verwendet.

Eine weitere wichtige Charakteristik der Kollokationsbeschreibung im DWDS-Wörterbuch besteht darin, dass die Kandidaten für Kollokationen grundsätzlich aus dem statistischen Wortprofil des DWDS bezogen werden sollen (s. Abschnitt 5). Die Mehrzahl der Kollokationen, die über die statistische korpusbasierte Wortprofilanalyse für die Neueinträge extrahiert werden, ist vom Typ2, also distributionell nicht spezifisch, aber usualisiert. Banale bzw. gänzlich unspezifische, aber dennoch statistisch signifikante Kookurrenzen muss der Lexikograph aussortieren. Mit der Verwendung des statistischen Wortprofils als Basis für die Kollokationsbeschreibung entfällt für den Lexikographen auch die Auswahl geeigneter Belege für eine Kollokation. Diese werden nicht explizit in das Wörterbuch geschrieben, sondern im Nachhinein über die automatische Verknüpfung mit dem Wortprofil hinzugefügt.

4. Modellierung der Kollokationen im DWDS-Wörterbuch und Arbeit mit dem Schema

4.1 Schemabeschreibung im DWDS-Wörterbuch

Die Grundidee der Schemabeschreibung des DWDS-Wörterbuchs besteht in der konsistenten Auszeichnung auf allen Strukturebenen des Wörterbuchs, angefangen bei Formangaben und grammatischen Angaben über Bedeutungsangaben, pragmatische Markierungen, Beispiele und Zusätze sowie Belege bis zu den Verweisstrukturen. Das DWDS-Wörterbuchschema besteht aus ca. 50 verschiedenen Beschreibungselementen auf Element-Ebene sowie festen Wertelisten (beispielsweise bei grammatischen Angaben wie Genus oder Angaben zu Sach- oder Fachgebieten). Die Schema-Beschreibung des DWDS-Wörterbuchs liegt in Form von RELAX-NG und Schematron-Regeln vor und ist mit den Richtlinien der TEI P5 kompatibel. Die elektronische Fassung des DWDS-Wörterbuchs ist in diesem Schema validierbar (Herold 2011).

Für die anstehende Arbeit der Lexikographen wurde ein reduziertes Schema, eine sogenannte Werkstattsprache entworfen, die aber automatisch in das DWDS-Wörterbuchschema konvertierbar ist. Die Lexikographen erarbeiten die Artikel innerhalb eines lexikographischen Redaktionssystems.[1]Dieses besteht im Wesentlichen aus der lexikographischen Rechercheumgebung der DWDS-Website und einem DWDS-Framework, welches für die Autorenumgebung des oXygen-Editors entwickelt wurde. Die von dem Lexikographenteam erstellten Artikel werden zentral über das in eine graphische Benutzeroberfläche von oXygen integrierteVersionierungssystem Apache Subversion (SVN) verwaltet.

[1] http://www.dwds.de/projekt/lexarbeitsplatz/.

4.2 Modellierung der Kollokationen im XML-Wörterbuchschema

Die Kollokationen sind im DWDS-Wörterbuchschema in der Werkstattsprache innerhalb der Lesartenbeschreibung codiert (vgl. Abb. 1). Der Vorzug der Werkstattsprache besteht für die praktische lexikographische Arbeit darin, dass Mikrostrukturen, die im TEI-Schema eingebettet codiert sind, beispielsweise durch ein Attribut-Wert-Paar, in der Werkstattsprache explizit als Element codiert und somit im XML-Editor einfacher handhabbar sind. Ein Beispiel hierfür ist das Element Konstruktionsmuster in Abbildung 1, welches in einem TEI-P5-Schema als *cit type="pattern"* codiert werden müsste. Wenn man dann zusätzlich noch weitere Typisierungen, beispielsweise nach syntaktischer Klasse hinzufügen möchte, wird der Redaktionsprozess im XML-Editor schnell unübersichtlich.

Das Schema-Fragment der Lesartenbeschreibung ist rekursiv definiert und ermöglicht beliebig viele Einbettungen von Unterlesarten. Auf jeder Lesartenebene können Formangaben, die Syntagmatik, die Diasystematik, Frequenzangaben und Verweise beschrieben werden. Diese Informationen werden der Definition der Lesart vorangestellt. Auf die Definition wiederum folgen die Verwendungen, in denen Phraseme, Kollokationen (vom Typ1 oder Typ2, notiert als <Kollokation1> bzw. <Kollokation2>), Kompetenzbeispiele oder Korpusbelege beschrieben werden.

```
Lesart= element Lesart {
        Formangabe *
        , Syntagmatik ?
        , Diasystematik *
        , Frequenzangabe ?
        , Verweise *
        , Definition ?
        , Verwendungen *
        , Lesart *
}

Verwendungen = element Verwendungen {
        Phrasem *
        &Kollokation1 *
        & Kollokation2 *
        & Beleg *
        & Kompetenzbeispiel *
}

Kollokation1 = element Kollokation1 {
        , attribute type { 'ATTR' | 'CJ' | 'OBJA' ... }
        , Zitierform
        , Paraphrase ?
        , Diasystematik?
        , Verweise
}
```

Abb. 1: Lesartenbeschreibung im DWDS-Werkstattschema

Eine Kollokation wird in ein <Kollokation1>- oder ein <Kollokation2>-Element eingeschlossen. Die Kollokation umfasst die Basis (das Stichwort, unter dem die Kollokation aufgeführt wird) und den Kollokator. Diese werden nicht näher in der XML-Struktur unterschieden, sondern als Text in das Element <Zitierform> eingeschlossen (vgl. die Beispiele in Abbildung 2).

```
<Kollokation1type= ″OBJA″>
        <Zitierform> Termin einhalten</Zitierform>
</Kollokation1>
<Kollokation1type= ″MOD″>
        <Zitierform> sündhaft teuer</Zitierform>
</Kollokation1>
<Kollokation2type= ″OBJA″>
        <Zitierform> Ball abspielen</Zitierform>
</Kollokation2>
```

Abb. 2: Beispiele für die XML-Struktur der Kollokationen

Falls es für ein Stichwort mehrere Kollokationen gibt, werden diese zunächst nach dem Kollokationstyp (Typ 1 und Typ 2), zweitens nach syntaktischen, drittens nach semantischen und schließlich nach pragmatischen Kriterien gruppiert. Die syntaktischen Gruppierungen werden im Werkstattschema über den Relationsnamen referenziert, der auf dem DWDS-Wortprofil (vgl. Abschnitt 5) basiert und damit die automatische Identifizierung der Korpusbelege über den Schlüssel im DWDS-Wortprofil ermöglicht. Die folgenden drei Beispiele illustrieren das Zusammenspiel dieser Kriterien.

Im ersten Beispiel (*Jeans*, vgl. Abbildung 3) findet eine Mischung von syntaktischen und semantischen Gruppierungen statt. Die syntaktischen Gruppierungen werden explizit typisiert. Dies geschieht durch die syntaktische Relation ATTR, die im Wortprofil die Adjektiv-Nomen-Relation bezeichnet. Die semantische Gruppierung wird hingegen nur über Kommentare (die in oXygen als „processing instructions“ codiert werden) festgehalten, da die semantischen Klassen zu offen sind, um sie in das enge Korsett einer geschlossenen Wertemenge zu pressen. Im vorliegenden Fall bezieht sich die semantische Gruppierung auf die Kriterien der Farblichkeit in der ersten Gruppe und die stoffliche Qualität in der zweiten Gruppe:

```
<Kollokation2 type= ″ATTR″>
        <Zitierform>helle Jeans</Zitierform>
        <Zitierform> dunkle Jeans</Zitierform>
        …
        <Zitierform> weiße Jeans</Zitierform>
</Kollokation2>
<Kollokation2 type= ″ATTR″>
        <Zitierform> abgetragene Jeans</Zitierform>
        <Zitierform>ausgebeulte Jeans</Zitierform>
        …
        <Zitierform> zerschlissene Jeans</Zitierform>
</Kollokation2>
```

Abb. 3: Kollokationen von *Jeans*

Im zweiten Beispiel (*Allergie*, vgl. Abbildung 4) werden drei Kollokationsgruppen gebildet, die sowohl syntaktisch als auch semantisch unterschiedlich sind: Gruppe 1 bezeichnet eine Substantiv-Koordination (Relation CJ), die die semantische Relation von Begriff/Oberbegriff zum Ausdruck bringt, Gruppe zwei die Eigenschaft (ATTR) und die dritte Gruppe bezeichnet schließlich eine inchoative Handlung in Form einer Nomen-Verb-Relation (OBJA).Codiert werden diese Kollokationen im DWDS-Wörterbuch wie folgt:

```
<Kollokation1 type= ″CJ″>
        <Zitierform>Allergien und Unverträglichkeiten</Zitierform>
</Kollokation1>
<Kollokation2 type= ″ATTR″>
        <Zitierform>eine heftige Allergie</Zitierform>
          <Zitierform>eine schwere Allergie</Zitierform>
          <Zitierform>eine starke Allergie</Zitierform>
</Kollokation2>
<Kollokation2 type= ″OBJA″>>
        <Zitierform>eine Allergie auslösen</Zitierform>
</Kollokation2>
```

Abb. 4: Kollokationen von *Allergie*

Im dritten Beispiel (*Handy*, vgl. Abbildung 5) werden syntaktisch gleiche Relationen (SUBJ) nach ihrer stilistischen Färbung gruppiert. Der Skopus des Elements Stilebene umfasst dabei alle Zitierformen des übergeordneten Kollokationsblocks. Diese Art der Codierung stellt ein Zugeständnis an die Anforderungen der täglichen Arbeit im Artikelredaktionssystem dar, bei der die Diasystematik einfach und einheitlich über alle Elemente zugreifbar sein muss. Bei der Konvertierung der Werkstattsprache in das DWDS-Wörterbuchschema werden diese „seriellen“ Abhängigkeiten wieder in hierarchische umgewandelt.

```
<Kollokation2 type= ″SUBJ″>
        <Zitierform>das Handy piepst</Zitierform>
        <Zitierform>das Handy bimmelt</Zitierform>
        <Diasystematik>
                <Stilebene>umgangssprachlich</Stilebene>
        <Diasystematik>
</Kollokation2>
<Kollokation2 type= ″SUBJ″>
        <Zitierform>das Handy klingelt</Zitierform>
        <Zitierform>das Handy vibriert</Zitierform>
</Kollokation2>
```

Abb. 5: Kollokationen von *Handy*

Die Gruppierungen lassen sich für die Präsentationsebene entsprechend ihren Typisierungen anzeigen oder nach alphabetischer Reihenfolge oder nach statistischer Signifikanz sortieren. Letztere wird über den Abgleich mit der Datenbank des DWDS-Wortprofils realisiert, die im folgenden Abschnitt beschrieben wird.

5. „Assistierte“ Kollokationsextraktion

5.1 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. 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. Das DWDS-Wortprofil beruht auf einer syntaktischen Voranalyse der Korpusdaten durch den Shallow Parser Syncop (SYNtactic COnstraint Parsing, vgl. Didakowski 2007). Die Berechnung des DWDS-Wort-

profils selbst 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. Die Methodik des Wortprofils ist anderweitig ausführlich beschrieben (Geyken et al. 2009). An dieser Stelle beschränken wir uns aus Platzgründen auf die praktischen Ergebnisse des Wortprofils.

Der derzeitige Prototyp des DWDS-Wortprofils (Wortprofil_2010) ist unter www.dwds.de abfragbar. Er beruht auf einer Mischung eines Referenz- und eines Zeitungskorpus, des DWDS-Kernkorpus und des ZEIT-Archivs (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. Anhand eines Beispiels sollen die verschiedenen, vom DWDS-Wortprofil extrahierten Informationen verdeutlicht werden. Beispielsweise werden für das Stichwort *Feindbild* im DWDS-Wortprofil 32 verschiedene syntaktische Relationen mit insgesamt 384 Vorkommen extrahiert. Diese werden in Form einer Schlagwortwolke dargestellt (vgl. Abbildung 6). Die Voraussetzung für die Aufnahme einer Relation in das Wortprofil ist, dass dafür wenigstens vier Belege im Korpus vorkommen. Damit soll verhindert werden, dass okkasionelle Verbindungen fälschlicherweise in das Wortprofil aufgenommen werden.

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

- Adjektiv-Nomen (Etikett: ATTR): altes, antibolschewistisches, äußeres, gemeinsames, gepflegtes, ideales, ideologisches, intaktes, klares, klassisches, linkes, neues, primitives, richtiges, schlichtes, überkommenes, verblasstes, westliches
- 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

Abb. 6: DWDS-Wortprofil für *Feindbild*

Die Relationstypen lassen sich über das Wortprofil-Fenster ansteuern, indem man den Relationenfilter anklickt (vgl. Abbildungen 6 und 7). Es werden dann die syntaktischen Relationstypen aufgeklappt (vgl. Abbildung 7). Klickt man auf einen 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 bereits weiter oben erwähnte 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. 7: DWDS-Wortprofil für *Feindbild* – 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 6 auf die Verb-Verbindung *auskommen_ohne*, gelangt man zu den in Abbildung 8 gezeigten insgesamt vier Satzkontexten, die vom Analysesystem aus dem Korpus extrahiert wurden.

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.

Abb. 8: Syntaktische Relation (auskommen_ohne, Feindbild) mit dem Relationstyp: V_PP

5.2 Zur Qualität des Wortprofils

Im vorangegangenen Abschnitt wurden die verschiedenen Nutzungsmöglichkeiten des Wortprofils 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). Aus Platzgründen soll der Vergleich 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. Eine ausführliche Darstellung aller Vergleichsparameter findet sich in Geyken (2011). Zusammengefasst hier 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 auch im direkten Vergleich nieder: Mit einer hohen statistischen Signifikanz (Salienz s>5) enthält das Wortprofil mehr als 20 lexikographisch relevante Beispiele, die nicht im WDG verzeichnet sind. Bei einer Salienz von unter fünf (s<5) im Wortprofil nimmt die Dichte der lexikographisch relevanten Relationen stark ab: von den knapp 200 syntaktischen Relationen (s<5) sind lediglich fünf 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 (WP) hat aber hier die gebräuchlicheren Beispiele: *grauer Kapitalmarkt* oder *grauer Markt* (WP) statt *graue Händler* oder *graues Material* (GWDS). Schließlich, 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. In Abschnitt 6 wird gezeigt, dass sich diese positive Beurteilung des Wortprofils für die Probeartikel zum DWDS-Wörterbuch übertragen lässt.

5.3 DWDS-Wortprofil und lexikographische Intuition

Die Beschreibung der Kollokationen im DWDS-Wörterbuch basiert derzeit auf einer Mischung von aus dem Wortprofil extrahierten Relationen und Kompetenzkollokationen. Dass Kompetenzkollokationen auf absehbare Zeit das Wortprofil ergänzen müssen, ergibt sich aus der Tatsache, dass Kollokation und statistische Kookkurrenz zwar korrelieren, aber nicht identisch sein müssen. So kann es durchaus Kollokationen geben, die sich in den überwiegend schriftlichen Korpora des DWDS kaum niederschlagen und somit auch nicht im Wortprofil extrahiert werden können. Ein Beispiel für solch eine niedrigfrequente Kollokation liefert beispielsweise die Kollokationsbasis *Äußeres* mit den von Hausmann erwähnten Kollokatoren *angenehm*, *gepflegt*, *attraktiv*, *ansprechend* (alle im Wortprofil) und *einnehmend* (nicht im Wortprofil). Damit zusammenhängend ist die Tatsache, dass die im Wortprofil aufgelisteten Kollokationskandidaten Ergebnisse der hiermit dokumentierten Diskurse, weniger jedoch

Ergebnis sprachlicher Möglichkeiten oder Präferenzen darstellen. Es gibt im Wortprofil beispielsweise *konservative* Blogger, aber keine *progressiven*, weil der *konservative Blogger* der im Diskurs markierte Fall ist. Es gibt *erfolgreiche* Blogs, aber keine *langweiligen* oder *erfolglosen*, weil es sich über Letztere nicht zu reden lohnt.

6. Erprobung der Kollokationsanalyse

In einer Erprobungsphase wurden 136 Stichwörter lexikographisch für das DWDS-Wörterbuch ausgearbeitet. Die Ausarbeitung orientierte sich formal und inhaltlich an der strukturierten Version des elektronischen WDG, da die Probeeinträge zusammen mit der Grundsubstanz in das DWDS-Wörterbuch einfließen sollen. Kriterium für die Auswahl dieser Stichwörter war, dass sie im WDG nicht belegt sind, in heutigen Korpora jedoch hochfrequent sind und somit Kandidaten für Volleinträge im DWDS-Wörterbuch darstellen. Bei der Auswertung dieser Arbeit soll im Folgenden nur auf die Kollokationen Bezug genommen werden. Bei den ausgearbeiteten 136 Einträgen wurden 33„hausmannnsche“ Kollokationen (Typ1) und 402 Kollokationen vom Typ2 in den Einträgen vermerkt. Nur ein relativ geringer Anteil, nämlich insgesamt 27 Kollokationen, konnte nicht aus dem Wortprofil extrahiert werden und wurde als Kompetenzkollokation vermerkt. Beispiele hierfür lassen sich aus allen Relationstypen finden:

- Adjektiv-Nomen (4 Kollokationen): eloquenter Schreibstil, inflatorische Verwendung, tägliche Charterflüge
- Objekt-Verb (14): Übertragungsrechte makeln, Gebäude observieren, Countdown abbrechen
- Verb-Adverb (4): dauerhaft archivieren, eloquent vertreten
- Präpositionalobjekt-Verb (1): zum Administrator ernennen
- Subjekt-Verb (3): der Jet startet, setzt auf; das Kraftwerk emittiert

7. Ausblick

Die Beschreibung der Kollokationen im DWDS-Wörterbuch basiert derzeit – und wohl auch noch auf absehbare Zeit – auf einer Mischung von aus dem Wortprofil extrahierten Relationen und Kompetenzkollokationen.

Eine weitere Verbesserung des Wortprofils ist auf drei Ebenen erreichbar: erstens auf der Ebene der Relationsextraktion mit Hilfe eines leistungsfähigeren Syntaxparsers. Diese Arbeiten, die insbesondere eine Verbesserung im Bereich der Verb-Relationen betreffen, sind bereits implementiert und Teil der nächsten Wortprofil-Version (DWDS-Wortprofil 2012). Auch durch die Variierung der statistischen Maße lassen sich Veränderungen und Verbesserungen der extrahierten Wortprofil-Relationen erzielen. So scheint die Ersetzung des bislang eingesetzten Salienzmaßes (Geyken et al. 2009) durch Dice-Koeffizienten (Rychly 2008) zu verbesserten Ergebnissen bei absolut gesehen hochfrequenten Wörtern zu führen, die mit dem Kollokator jedoch nicht häufiger als erwartbar auftauchen. Die dritte Ebene, auf der Verbesserungen der Qualität der extrahierten Wortprofile erzielt werden können, betrifft die Korpora. 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. Diese 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 allgemein-

sprachliches 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 1.000 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 nur etwa 5.000 Lemmata, die mehr als 1.000 Mal vorkommen. Bei dem 500 Millionen Textwörter großen Korpus, welches derzeit für das Wortprofil_2010 verwendet wird, gelangt man immerhin auf 15.000 Lemmata, die wenigstens 1.000 Mal im Korpus belegt sind. Wenn man die Schwellenwerte für die Mindestanzahl von Kontexten von vier auf drei senkt, verdoppelt sich in etwa die Anzahl der Lemmata. Dennoch ist auch diese Anzahl eine zu geringe Basis für ein umfangreiches einsprachiges Wörterbuch. Eine weitere Erhöhung der Korpusgrundlage um die weiteren Texte des DWDS-Korpus in einem Umfang von 2 Milliarden Textwörtern ist somit einer der nächsten geplanten Schritte.

8. Literatur

Didakowski, Jörg (2007): SynCoP – Combining syntactic tagging with chunking using WFSTs. In: Proceedings of FSMNLP 2007. Potsdam, S. 107-118.

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

Geyken, Alexander (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In:Abel, Andrea/Zanin, Renata (Hg.). Korpora in Lehre und Forschung. Bozen/Bolzano, S. 115-137.

Geyken, Alexander/Didakowski, Jörg/Siebert, Alexander (2009):Generation of word profiles for large German corpora. In: Kawaguchi, Yuji/Minegishi, Makoto/Durand, Jacques (Hg.):Corpus Analysis and Variation in Linguistics. Amsterdam, S.141-157.

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

Hausmann, Franz-Josef (1984): Wortschatzlernen ist Kollokationslernen. In: Praxis des neusprachlichen Unterrichts. 31. Jg. (1984), S. 395-406.

Herold, Axel (2011): Retrodigitalisierung und Modellierung des Wörterbuchs der deutschen Gegenwartssprache. In: Krafft, Andreas/Spiegel, Carmen (Hg.): Sprachliche Förderung und Weiterbildung transdisziplinär. Frankfurt/M. u.a., S. 197-213. (=Forum angewandte Linguistik 51)

Herold, Axel / Geyken, Alexander (2008): Adaptive word sense views for the dictionary database eWDG: The case of definition assignment. In: Storrer, Angelika/Geyken, Alexander/Siebert, Alexander/Würzner, Kay-Michael (Hg.): Text resources and lexical knowledge (TTCP 8). Berlin, S. 209-221.

Ivanova, Kremena/Heid, Ulrich/Schulte im Walde, Sabine/Kilgarriff, Adam/Pomikálek, Jan (2008): Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In: Proceedings of the 6th Conference on Language Resources and Evaluation. Marrakesch, Marokko, paper no. 537.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich/Schierholz, Stefan/Schweickard, Wolfgang/Wiegand, Herbert Ernst/Gouws, Rufus H./Wolski, Werner (Hg.):Lexikographica. Berlin/New York, S. 79-93.

Rychly, Pavel (2008): A Lexicographer-Friendly Association Score. In: Sojka, Petr /Horák, Aleš (Hg.): Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008. Brno, S. 6-9.

[WDG] Klappenbach, Ruth/Steinitz, Wolfgang (Hg.) (1964-1977): Wörterbuch der deutschen Gegenwartssprache. Berlin.