Erschienen in:
Stefan Engelberg, Meike Meliss, Kristel Proost und Edeltraud Winkler (Hg.): Argumentstruktur zwischen Valenz und Konstruktion. Tübingen: Narr Francke Attempto (Studien zur deutschen Sprache - 68), S. 451–467

ARNE ZESCHEL

SEMIAUTOMATISCHE IDENTIFIKATION VON ARGUMENTSTRUKTURKONSTRUKTIONEN IN GROSSEN KORPORA[1]

1. Einleitung

Wie lassen sich Argumentstrukturkonstruktionen in nicht eigens dafür ausgezeichneten Korpora identifizieren? Hinter diesem Problem verbergen sich *zwei* grundlegende Fragen, die sich für empirische Studien im Bereich der verbalen Argumentstruktur stellen:

– Wie viele eigenständige Konstruktionen des relevanten Typs gibt es überhaupt (ggf. innerhalb eines bestimmten eingeschränkten Untersuchungsbereichs) und welche sind es?
– Wie lassen sich die Instanzen einer bereits identifizierten/postulierten Konstruktion in einem Korpus exhaustiv ermitteln?

Vor diesen zwei Fragen stellt sich natürlich die noch einmal grundlegendere, gemäß welchen Kriterien eine Struktur überhaupt als eigenständige (d.h. unabhängig beschreibungsrelevante) Konstruktion bzw. als konkrete Instanz einer solchen Konstruktion zu werten ist. Dieser Punkt wird hier nicht behandelt. Stattdessen soll es im Folgenden darum gehen, wie Argumentstrukturkonstruktionen und ihre Instanzen nach erfolgter Operationalisierung des Konstruktionsbegriffs möglichst vorannahmenfrei in großen Korpora zu identifizieren sind.

Dazu werden in Abschnitt 2 zunächst mögliche Strategien der Datenerhebung miteinander verglichen. Abschnitt 3 stellt eine datengeleitete Identifikationsprozedur vor, die im Rahmen des IDS-Projekts „Verben und Argumentstrukturen“ zur Identifikation präpositionaler Argumentstrukturmuster des Deutschen verwendet wird. Aus Platzgründen kann der Analysegang dabei nur knapp und ohne technische Details beschrieben werden. Schließlich werden die Ergebnisse einer Pilotstudie berichtet (Abschnitt 4) und eine kurze Zusammenfassung gegeben (Abschnitt 5).

[1] Ich danke Maike Huth und Agata Sokolowski für ihre Unterstützung bei der Datenanalyse.

Die Darstellung des Verfahrens erfolgt am Beispiel präpositionaler Argumentstrukturkonstruktionen mit der Präposition *vor*. Obwohl die vorgestellte Methode auch für die erste der eingangs genannten Fragen einschlägig ist (Wie viele verschiedene und welche präpositionalen Argumentstrukturkonstruktionen mit *vor* gibt es im Deutschen?), liegt der Fokus hier auf dem zweiten Problem: Wie lassen sich die Instanzen einer gegebenen Konstruktion möglichst vorannahmenfrei und exhaustiv in großen Korpora ermitteln? Die Diskussion erfolgt dabei am Beispiel des deutschen Pendants der so genannten *swarm*-Konstruktion des Englischen (Levin 1993; Dowty 2000). In Zeschel (2011) wird dafür argumentiert, drei Subtypen dieser Konstruktion zu unterscheiden (alle Belege in diesem Beitrag aus DeReKo, Institut für Deutsche Sprache 2013):

(1) a. Ich schrie vor Schmerz und gestikulierte wild. (Nürnberger Nachrichten, 2.6.2007)
b. Diesmal war sie es, die fast platzte vor Stolz und Glück. (taz, 8.7.2002)
c. Obwohl die österreichische Journalistengruppe die Expo an einem normalen Wochentag besucht, schwirrt das Gelände vor Schulklassen, deren Lehrer größte Mühe haben, die von einer vermeintlichen Sensation zur anderen hastenden Kinder beieinander zu halten. (Die Presse, 19.6.1998)

Im Folgenden wird der Subjektreferent dieses Musters als 'X' bezeichnet, der Referent des eingebetteten Nominals in der *vor*-PP als 'Y' sowie die drei in (1) illustrierten Subtypen als Varianten 'A', 'B' und 'C' der Konstruktion. Variante A ist eine kausale Adjunktkonstruktion. (Intuitiv) typische Instanzen setzen ein belebtes Subjekt mit einer starken körperlichen Empfindung oder Emotion in Beziehung, die X zu einer bestimmten unwillkürlichen Reaktion auf das Erleben von Y veranlasst. Das Verb kann sowohl intransitiv als auch transitiv sein (*etwas vor Schreck fallen lassen*). Die Funktion von Variante B lässt sich als 'intensivierte Eigenschaftszuschreibung' kennzeichnen. Eine Kausalrelation liegt nicht (bzw. nurmehr semantisch gebleicht) vor: Es wird nicht assertiert, dass sich das Verbereignis faktisch vollzieht. Typischerweise sind es wiederum belebte Subjekte, denen bestimmte abstrakte Eigenschaften und Emotionen zugeschrieben werden. Das Verb ist typischerweise intransitiv. Aus Gründen semantischer Gradienz (Ist die

prädizierte Handlung tatsächlich faktisch assertiert oder liegt lediglich ein Fall von Hyperbole vor?) ist Variante B gemäß dem genannten Kriterium jedoch nicht trennscharf von der mitunter auch transitiven Variante A abzugrenzen. Variante C ist eine semantisch spezialisierte Lokativkonstruktion, die das gänzliche Erfüllt-/Bedeckt-/Saturiertsein von X mit Y impliziert. Wie in Variante B wird anstelle einer Kausalrelation zwischen zwei unabhängig prädizierten Sachverhalten nur eine einzige Proposition realisiert. Im Unterschied zu Variante B handelt es sich bei den Argumenten um Konkreta, und es wird das gehäufte Vorhandensein von Entität Y an Ort X prädiziert (*die Ablage quillt über vor Papieren*). Instanziierende Verben sind intransitiv. Allen drei Mustern ist die schematische Bedeutung 'X ist voller Y' gemein, ihre je unterschiedlichen Funktionen und semantischen Rollenkonfigurationen gehen jedoch auch mit unterschiedlichen Verwendungsrestriktionen bezüglich der jeweils verwendbaren Verben einher: Während Variante C lexikalisch stark restringiert ist, lassen sich für Variante B bestimmte semantische Restriktionen herausarbeiten, innerhalb derer die Konstruktion lexikalisch variabel instanziierbar ist, wohingegen Variante A gänzlich unbeschränkt ist, da sich eine prinzipiell unendliche Menge von Kausalrelationen im Rahmen dieser Konstruktion formulieren lässt (Zeschel 2011).

Auf Basis dieser knappen Charakterisierung ergeben sich eine Reihe weiterführender Fragen: Gibt es ggf. noch weitere Subtypen der Konstruktion als die in (1) postulierten? Wie lassen sich Produktivitätseigenschaften der gefundenen Varianten messen und „kreative", nichtantizipierte lexikalische Instanzen finden? Wie steht es um die relative Geläufigkeit der Varianten im Verhältnis zueinander – ist es z.B. plausibel, Variante B als bloßen figurativen „Ableger" einer nach wie vor primären Variante A zu betrachten? Und was ist zum Verhältnis der einzelnen Varianten zu formal und semantisch ähnlichen Konstruktionen mit *vor* in ihrer systemischen „Nachbarschaft" zu sagen? Die inhaltliche Beantwortung dieser Fragen würde den Rahmen des vorliegenden Beitrags sprengen. Im Folgenden soll es stattdessen um die Vorstellung eines Analyseverfahrens gehen, mit dem sich entsprechende Fragen zunächst überhaupt empirisch angehen lassen.

2. Zur Identifikation von Konstruktionen in Korpora

Das sicherlich geläufigste Erhebungsverfahren in konstruktionsorientierten Korpusstudien zur Argumentstruktur ist gerade nicht konstruktions-, sondern verbbasiert. Ausgehend von bestimmten Initialbeobachtungen werden vermutete semantische Restriktionen einer Konstruktion überprüft, indem Vorkommen der Konstruktion mit Verben gesucht werden, die bereits bekannten Verwendungen semantisch ähneln. Listen zu konkordierender Elemente werden aus Thesauri gewonnen und zusätzlich introspektiv ergänzt (vgl. etwa Proost 2009; Zeschel 2012). Das Vorgehen ermöglicht eine systematische Überprüfung gegebener semantischer Domänen bezüglich des Auftretens zugehöriger Verben in der Zielkonstruktion. Problematisch daran ist zum einen, dass Verben typischerweise polysem und semantische Verbklassen randbereichsunscharf sind. Aus diesem Grund ist es schwierig, beispielsweise ein exhaustives „Inventar deutscher Geräuschverben" zu erstellen (insbesondere, wenn auch rare und kreative Verbverwendungen und -bildungen mit Blick auf die Produktivität einer Konstruktion von Interesse sind). Dennoch wird man auf diese Weise aber zumindest die unkontroversen und geläufigen Vertreter einer Klasse problemlos ermitteln und überprüfen können. Zum anderen ist problematisch, dass nur Klassen überprüft werden können, deren mögliches Auftreten in der Konstruktion auch bereits antizipiert wurde. Stößt man im Rahmen der eingehenderen Beschäftigung mit einer Konstruktion dann auf eine neue, bislang nicht antizipierte Klasse instanziierender Verben, kann man diese natürlich ebenso nach dem geschilderten Verfahren systematisch nacherheben. Unklar bleibt dabei allerdings, ob man auf diese Weise am Ende auch tatsächlich alle relevanten Klassen gefunden haben wird. Gänzlich unmöglich ist es schließlich, die lexikalischen Füllungen von Konstruktionen zu ermitteln, deren Verbslot genuin offen ist, wie hier im Fall von Variante A.

Eine zweite, ebenfalls verbbasierte Strategie schränkt den Umfang der Datenerhebung von vorneherein auf erwartete typische Realisierungen der Zielkonstruktion ein, indem auch Merkmale für typisch befundener Kontexte der Zielverben im Rahmen des Musters in die Suchanfragen inkorporiert werden (Engelberg in diesem Band). Diese Strategie ist auf eine Maximierung der Präzision der Suchanfragen ausgerichtet, so dass der Nachbearbeitungsaufwand der Datenerhebung minimiert

wird. Damit einher geht allerdings, dass auch nur Instanzen der antizipierten Realisierungen gefunden werden und sich die Resultate ausschließlich auf diesen Ausschnitt der Daten beziehen.

Des Weiteren sind auch Verfahren möglich, die umgekehrt auf maximalen Rücklauf abzielen. Im Fall gänzlich schematischer Konstruktionen ist dazu ein geparstes Korpus erforderlich, in dem direkt nach bestimmten syntaktischen Strukturen gesucht werden kann. Während für die Korrektheit von POS-Tagging bereits sehr gute Werte erzielt werden (Giesbrecht/Evert 2009), ist die Verlässlichkeit syntaktischer Auszeichnungen allerdings deutlich geringer. Obwohl die Suchanfrage selbst den Kreis gefundener lexikalischer Instanziierungen einer gegebenen Position also nicht bereits unweigerlich einschränkt, ist hierbei nicht nur mit einem bestimmten Schnitt falscher Positive, sondern auch mit einer (in diesem Zusammenhang problematischeren) Dunkelziffer falscher Negative zu rechnen, die der folgenden Analyse von vorneherein verborgen bleiben. Ist die untersuchte Konstruktion hingegen nur teilschematisch, d.h. ein oder mehr Elemente des Musters lexikalisch fix, ist ein genuin exhaustiver Rücklauf möglich – dies allerdings nur auf Kosten einer minimalen Präzision. Wird beispielsweise nach potenziell relevanten Konstruktionen mit *vor* gesucht, ist letztlich nur eine Suche nach *vor* selbst offen genug, um nicht-antizipierte lexikalische Füllungen bestimmter Positionen im Kontext der vollen Bandbreite möglicher Linearisierungen, Modifikationen und weiterer denkbarer Komplikationen zu ermitteln. Andererseits ist klar, dass eine überwältigende Mehrzahl der damit erzielten Treffer aus für den gegebenen Zweck irrelevanten Verwendungen von *vor* bestehen wird. Da gerade für vergleichsweise seltene Argumentstrukturmuster auch keine zu kleinen Stichproben gezogen werden dürfen, um noch genügend Instanzen der eigentlichen Zielkonstruktion zu ermitteln, stellt sich also die Frage, ob ein derartiges Verfahren mit einem angesichts des zu erwartenden Nutzens vertretbaren Aufwand einhergeht. Sicherlich scheidet die Option, die erzielten Treffer sämtlich einzeln durchzulesen, bei einer entsprechend großen Belegmenge aus. Die Frage, die sich stellt, ist also, ob die – letztendlich unumgängliche – händische Klassifikation der Daten geeignet (semi)automatisch vorentlastet werden kann, so dass große Mengen irrelevanter Belege jeweils en bloc eliminiert und am Ende eine mit zumutbarem Aufwand bearbeitbare Restmenge an Daten übrig bleibt. Der folgende Abschnitt

skizziert eine Variante eines solchen Verfahrens, die im Rahmen des IDS-Projekts „Verben und Argumentstrukturen“ zum Einsatz kommt.

3. Datengeleitete Musteridentifikation

Ausgangspunkt der hier vorgestellten Pilotstudie ist eine Datenbasis von einer Million Belegsätzen mit *vor*, die aus einer geschichteten Stichprobe deutscher, österreichischer und Schweizer Pressetexte aus DeReKo gezogen wurden.[2] Der Ausgangsdatensatz wurde mit TreeTagger (Schmid 1994) wortartenannotiert. Diese Datenbasis wurde dann in einem zweistufigen Verfahren semiautomatisch um bestimmte Klassen hier irrelevanter Belege reduziert. Die beiden Stufen gliedern sich jeweils in mehrere Teilschritte, die im Folgenden kurz erläutert werden.

3.1 Elimination von Ausschlussmustern

Mit dem Begriff „Ausschlussmuster“ werden hier Gruppen häufiger Verwendungen bezeichnet, die aus strukturellen oder semantischen Gründen für die gegebene Analyse irrelevant sind und möglichst exhaustiv aus den Daten entfernt werden sollten. Im vorliegenden Fall sind dies:

- Belege ohne lexikalisches Verb (bzw. „VV.*“-Verbtag, z.B. in Dachzeilen von Zeitungsartikeln oder im Fall von Kopulakonstruktionen)
- Belege mit Tokenisierungsfehlern (z.B. bei Wortresten wie *Vor- und Nachteile*)
- Belege mit Verbpartikel- statt Präpositionsverwendung (z.B. *es kommt vor*)
- Belege, in denen die Zielform Bestandteil einer festen Mehrworteinheit ist (z.B. *nach wie vor*)
- Belege, in denen die Zielform Bestandteil einer Temporalangabe ist (z.B. *vor der Halbzeit*)

[2] Die Stratifikation orientierte sich zunächst an den Proportionen der durchschnittlichen Bevölkerungszahlen der drei Staaten im berücksichtigten Zeitraum (1991-2012) und innerhalb der so gewichteten Textmengen dann an den durchschnittlichen Auflagenstärken/Reichweiten der berücksichtigten Titel im relevanten Zeitraum (laut Auflagenkontrolleinrichtungen bzw. Verlagsauskunft).

– Belege, in denen die Zielform Bestandteil einer Lokalangabe ist (z.B. *vor dem Haus*)
– Belege, in denen die PP als Attribut eines Adjektivs oder Nomens fungiert (z.B. *sicher vor Blitzen*, *der Schutz vor Konsequenzen*)
– Belege, in denen das in der PP eingebettete Nomen determiniert oder quantifiziert ist, da dies als inkompatibel mit der Zielkonstruktion betrachtet wurde (vgl. **strotzen vor der/einer Kraft*, **wimmeln vor fünfzig Agenten*)

Bei den ersten vier Schritten steht fest, dass alle gefundenen Instanzen des jeweiligen Ausschlussmusters für die gegebene Analyse irrelevant sind und ohne weitere Überprüfungen aus der Datenbasis entfernt werden können. Andere Schritte wie etwa die Entfernung mutmaßlicher Temporal- und Lokalangaben erfordern stichprobenhafte Validierungen, indem eine bestimmte Anzahl der ausgesonderten Belege händisch darauf überprüft wird, ob das Zielvorkommen darin tatsächlich das jeweilige Ausschlussmuster instanziiert oder nicht.

Das eigentliche Ausschlussverfahren erfolgt über eine schrittweise Subkorpusbildung. Zu Beginn werden alle Belege ohne lexikalisches Verbtag von der Ausgangsdatenmenge abgezogen. Der zweite Reduktionsschritt wird dann innerhalb des resultierenden Teilkorpus vollzogen, der dritte dann nur noch innerhalb des wiederum aus Schritt 2 gewonnenen Subsubkorpus und so weiter. Dabei basieren natürlich nicht alle Schritte auf so selbsterklärenden und problemlos zu überprüfenden Kriterien wie etwa dem Vorhandensein eines „VV.*"-Tags im relevanten Satz. Im Folgenden deshalb einige Erläuterungen, wie die je relevanten Muster identifiziert und definiert wurden: Als lexikalisierte Mehrworteinheit wurden alle Zwei- und Dreiwortsequenzen mit der Komponente *vor* gewertet, die in der Online-Version des Duden-Wörterbuchs einen eigenständigen Eintrag verzeichnen. Zur Ermittlung wurden aus der vor diesem Schritt verbleibenden Datenmenge alle lexikalischen Bi- und Trigramme, in denen das Wort *vor* auftrat und die eine Häufigkeit von mindestens 100 hatten (entsprechend 0,1 Vorkommen in 1.000 Treffern für *vor*), auf ihr Gelistetsein im Duden überprüft und die Instanzen aller gefundenen Typen entfernt.

Für die Identifikation mutmaßlicher Temporal- und Lokalangaben wurde zunächst eine Frequenzliste aller Nomina mit einer Häufigkeit von mindestens 100 erstellt, die mit einem Wortabstand von 0-2 (ohne

dazwischen stehendes weiteres Nomen) innerhalb desselben Satzes rechts neben der Zielpräposition auftraten. Jedes Element auf dieser Liste wurde dann daraufhin bewertet, ob es sich mutmaßlich um eine Temporal- bzw. Lokalangabe handelt. Dazu wurden verschiedene relevante Unterkategorien aus den Daten gewonnen, die hier am Beispiel der Temporalangaben veranschaulicht werden: Bei der Durchsicht der Liste ergaben sich neben genuin temporalen Intervall- (*vor fünf Minuten*) und Phasenbezeichnungen (*vor Beginn*) auch mehrere Gruppen „domänenspezifisch salienter Zeitmarken", deren Prominenz in den Daten sich daraus erklärt, dass in den zugrundeliegenden Pressetexten immer wieder über die immer gleichen Inhalte berichtet wird, innerhalb derer bestimmte Wörter eben jene „salienten Zeitmarken" bezeichnen, die das berichtete Geschehen strukturieren. So handelt es sich etwa bei *vor dem Anpfiff* aufgrund der großen Prominenz des Wortes *Anpfiff* als zeitlicher Referenzpunkt innerhalb der Sportberichterstattung mit hoher Wahrscheinlichkeit um eine Temporalangabe. Zu den identifizierten rekurrenten Domänen in den Daten zählten neben Inhalten aus den klassischen journalistischen Ressorts Politik, Wirtschaft, Sport und Kultur auch die Kategorien „Saliente Zeitmarken allgemein" (*vor Weihnachten*), „Biografische Etappen und Zäsuren" (*vor dem Abitur*) und „Historische Zäsuren" (*vor dem Krieg*). Ob es sich bei den instanziierenden Ausdrücken dann aber tatsächlich immer um Temporalangaben handelt, ist natürlich nicht gesagt (*die Angst vor einem neuen Krieg*). Aus diesem Grund ist bei diesen Klassifikationen jeweils eine stichprobenhafte Validierung nötig, wie hoch der Anteil fälschlicherweise aussortierter Kombinationen von *vor* mit dem jeweiligen Signalwort ist. Generell wurden aussortierte Daten auch nicht „weggeworfen", da es am Ende der Prozedur noch eine abschließende Gesamtvalidierung gab (siehe Abschnitt 4). Hier wie bei allen anderen Klassifikationsentscheidungen im Rahmen des Verfahrens wurden die Listen von zwei verschiedenen Beurteilern getrennt bearbeitet, strittige Fälle nach Diskussion aufgelöst und abschließend nur übereinstimmend bewertete Kandidaten einer Kategorie auf die jeweilige Ausschlussliste geschrieben und aus der Datenbasis entfernt.

Präpositionalattribute zu nominalen und adjektivischen Köpfen (*die Warnung vor etwas, sicher vor etwas*) wurden mithilfe des „Wörterbuchs deutscher Präpositionen" (Müller 2013) identifiziert. Aus der elektronischen Version des Textes wurden all jene Adjektive und Nomen ex-

trahiert und auf eine Ausschlussliste geschrieben, für die konventionelle Anschlüsse mit *vor* gelistet waren. Entfernt wurden dann alle Belege, in denen ein Element dieser Liste der Präposition entweder direkt voranging oder noch ein optionales, über intervenierende POS-Kategorien identifizierbares Attribut dazwischenstand (*der Schutz der Bevölkerung vor Konsequenzen*). Determinierte und quantifizierte NPn innerhalb der *vor*-PP waren dann wieder recht einfach über das POS-Tagging zu identifizieren.

3.2 Elimination sonstiger Fehltreffer

Nicht alle irrelevanten Verwendungen in der Datenbasis lassen sich einer der allgemeinen Klassen zuschlagen, die im letzten Abschnitt diskutiert wurden. Um zusätzlich auch partikularere, ggf. aber individuell dennoch sehr häufige Fehltreffer zu entfernen, schloss sich deshalb noch eine zweite Filterstufe an. Für diesen Schritt wurde aus den verbleibenden Daten aus jedem Satz die Präposition, das nächste rechts stehende Nomen/Pronomen (mutmaßlich der Kopf der von *vor* eingebetteten Phrase) sowie das der Präposition innerhalb des Satzes nächststehende lexikalische Verb extrahiert (mutmaßlich dasjenige Verb, das die *vor*-PP einbettet). So ergab sich eine Liste in den Daten diskontinuierlicher und abfolgevariabler Einheiten wie etwa *sagen+vor+Journalisten*, *stellen+vor+Probleme, erzielen+vor+Steuern* etc., die hier im Folgenden als „Kopftripel" bezeichnet werden, da es sich mutmaßlich um den Kopf der VP, der PP und der darin eingebetteten NP handelt. Neben der Extraktion des jeweiligen Tripels für jeden verbleibenden Beleg erfolgte in diesem Schritt noch eine Reihe weiterer Aufbereitungen wie etwa eine Partikelverberkennung (Ist die nächststehende lexikalische Verbform tatsächlich ein alleinstehendes Verb oder ggf. Bestandteil eines Partikelverbs in Distanzstellung?) und die Extraktion weiterer lexikalischer Kookkurrenzmerkmale für jedes Kopftripel (konkret die Erstellung von Frequenzlisten für jedes Nomen, Pronomen, Adjektiv und Adverb in einem Fenster von ±5 Wörtern innerhalb desselben Satzes rund um die Präposition). Die Aufbereitung erfolgte mithilfe der Software rcqp (Desgraupes/Loiseau 2012), einem Interface zwischen der statistischen Programmiersprache R (R Core Team 2014) und der Korpusanalyseplattform IMS Open Corpus Workbench (Evert/Hardie 2011).

Alle Tripel, die mindestens dreimal in den Daten auftraten, wurden in eine Liste geschrieben und wiederum durchklassifiziert. Unterschieden wurden dabei Tripel, hinter denen sich mit großer Sicherheit Treffer der Zielkonstruktion verbargen (*strotzen+vor+Kraft*), Tripel, hinter denen mit großer Sicherheit Fehltreffer standen (*führen+vor+Augen*), Tripel, die Belege aus der potenziell interessanten „Peripherie" der Zielkonstruktion subsumieren (*fliehen+vor+Gefahr*), sowie unklare (und häufig auch kryptisch anmutende) Kombinationen, hinter denen sich oft „Fehlzuordnungen" des Skripts verbargen (in dem Sinne, dass etwa das in der linearen Abfolge des Satzes nächststehende lexikalische Verb gar nicht dasjenige ist, das die PP syntaktisch einbettet). Ziel der Klassifikation war dabei nicht nur die Voridentifikation vermutlicher Zielkonstruktionen und potenziell interessanter Muster in ihrer semantischen Umgebung, sondern vor allem auch die weitere Entlastung der händischen Durchsicht der verbleibenden Daten um häufige Fehltreffer. So trat z.B. das erwähnte Tripel *führen+vor+Augen* allein 1.657-mal in den Daten auf. Für hinreichend sicher klassifizierbare Tripel wie dieses war es daher deutlich ökonomischer, sie auf der Typebene einmal als irrelevant zu markieren und alle Instanzen abzuwählen, als auf der Tokenebene 1.657 verschiedene Sätze mit der Wendung *jemandem etwas vor Augen führen* einzeln durchzulesen und manuell aus der Datenbasis zu entfernen.

Natürlich waren nicht alle Tripel per se problemlos als klar relevant oder klar irrelevant erkennbar. Zum einen kann in solchen Fällen die Zuschaltung der weiteren Kookkurrenzinformation hilfreich sein (für das Tripel *sehen+vor+Augen* mit seinen 57 Treffern z.B. werden als geläufige Kookkurrenzen u.a. *man*, *Hand, kaum* und *noch* ausgegeben, wodurch man sich auch ohne Ansicht der 57 Einzelbelege sehr schnell ein gutes Bild davon machen kann, worum es in den zugrundeliegenden Sätzen wohl geht). Zum anderen zielt dieser Filterschritt auch nur auf die Entfernung von Elementen ab, die tatsächlich mit großer Sicherheit als semantisch inkompatibel mit der Zielkonstruktion betrachtet werden können. Alle unsicheren oder gänzlich unklaren Fälle verblieben in der Datenbasis und wurden erst in der abschließenden manuellen Durchsicht geklärt.

Nach Abschluss dieser Klassifikation erfolgte ein letzter Reduktionsschritt, in dem die zugehörigen Belege der als klar inkompatibel ausgezeichneten Tripel entfernt wurden. Von diesem Punkt an wurde die

verbleibende Datenmenge manuell weiterbearbeitet. Durch den letzten Klassifikationsschritt gab es dafür nun vier getrennte Mengen verbleibender Daten: Zum einen waren dies die vermutlichen Treffer der Zielkonstruktion, die Treffer für Muster in ihrer ebenfalls zu betrachtenden systemischen „Umgebung" sowie Belege mit nicht eindeutig oder gar unmöglich zu klassifizierenden Tripeln. Zum anderen gab es noch jene Belege, die im letzten Schritt gar nicht mitklassifiziert worden waren, da das jeweilige Tripel nur ein- oder zweimal in den Daten auftrat. Mit Blick auf kreative Extensionen einer Konstruktion sind allerdings besonders diese raren Tripel von besonderem Interesse. Für die Pilotstudie, deren Ergebnisse im nächsten Abschnitt knapp berichtet werden, wurden von den verbleibenden Daten, die auf der Tripel-Ebene nicht eindeutig klassifizierbar waren oder für diesen Schritt gar nicht erst berücksichtigt wurden, 25% der Belege manuell ausgewertet. Nach Identifikation der tatsächlichen Treffer wurde dann noch einmal ein Abgleich aller in der Konstruktion gefundenen Verben gegen alle im Laufe des Verfahrens aussortierten Belege gemacht, um so ggf. zu Unrecht ausgeschlossene Belege zu ermitteln.

4. Ergebnisse und Diskussion

Tabelle 1 zeigt die Ergebnisse der einzelnen Datenreduktionsschritte für die Analyse von Belegen mit der Präposition *vor*:

#	**Eliminiert**	**Vorher**	**Nachher**	**Abzug**	**Agreement**
1	Tokenisierungsfehler	1.000.000	997.177	-0,3%	–
2	Belege ohne Vollverb	997.177	926.994	-7,0%	–
3	Verbpartikeln	926.994	783.248	-14,4%	–
4	Mehrworteinheiten	783.248	581.055	-20,2%	–
5	Temporalangaben	581.055	376.041	-20,5%	84,6%
6	Lokalangaben	367.041	304.096	-6,3%	89,3%
7	Regierte Präpositionalattribute	304.096	266.969	-3,7%	–
8	Determinierte/quantifizierte NPen	266.969	67.839	-19,9%	–
9	Semantische inkompatible Tripel	67.839	55.907	-1,2%	94,6%
	Gesamt	**1.000.000**	**55.907**	**-94,4%**	

Tab. 1: Reduktion der Ausgangsdatenbasis für *vor*

Die Gesamtreduktion der Datenmenge um mehr als 94% der Ausgangstreffer im Rahmen dieser Pilotstudie ist ein erstaunliches Ergebnis. Zu bedenken ist dabei allerdings, dass allein beinahe 20% des Ertrags durch den in diesem Fall konstruktionsspezifisch begründeten Ausschluss von determinierten und quantifizierten NPn innerhalb der *vor*-PP zustande kamen. Ob sich ähnlich drastische Resultate auch mit anderen Präpositionen und anderen Konstruktionen ergeben werden, ist offen. So oder so sind aber auch die hier erzielten über 55.000 Restbelege noch eine sehr große Datenmenge für eine händische Durchsicht. Für folgende Analysen wird daher zu erwägen sein, wie eine geeignet geschichtete Stichprobe aus verschiedenen Regionen der (Tripel-)Frequenzverteilung gezogen werden kann, um den letzten Schritt noch einmal handhabbarer zu machen. In der aktuellen Studie wurde wie erwähnt eine ungeschichtete Zufallsstichprobe von 25% der Daten zu den unklassifizierten bzw. unklassifizierbaren Tripeln manuell durchgesehen.

Damit nun zu den inhaltlichen Resultaten und der Frage, welche neuen Einblicke die vorgestellte Erhebungsstrategie erbringt. Beginnen wir dazu am äußeren Rand der Strukturen, die sich zu diesem Zeitpunkt noch in der Datenbasis befanden und als potenziell relevant/interessant betrachtet wurden. Die erste Station dabei sind Konstruktionen mit Verben wie *sich fürchten, kapitulieren* und *fliehen*, bei denen die *vor*-PP üblicherweise als Präpositionalobjekt mit semantisch intransparenter Präposition gewertet wird. Ganz so intransparent ist die Konstellation allerdings nicht, da auf die Konfrontation einer Entität mit einer zweiten Entität oder einem Sachverhalt Bezug genommen wird, d.h. eine zumindest abstrakte Gegenüberschaft vorliegt: Es ist die (konkrete oder abstrakte) Präsenz von etwas, die X zur bezeichneten Handlung veranlasst bzw. den bezeichneten Prozess in X auslöst. Diese Paraphrase zeigt an, dass hier bereits der Boden für eine Interpretation der *vor*-PP als Angabe einer Ursache oder Begründung für die Verbhandlung bereitet ist. In entsprechender Abwandlung gilt dies auch für transitive Verben wie *schützen, warnen* und *retten,* bei denen die Verbhandlung darauf ausgerichtet ist, das Zustandekommen einer bestimmten Auswirkung des mit *vor* angeschlossenen Gegenstands, Prozesses oder Sachverhalts auf den Referenten des direkten Objekts zu verhindern. In einigen spezialisierten Verwendungen dieser Verben gibt es auch durchaus enge Berührungspunkte mit Variante B der Zielkonstruktion:

(2) a. „Wir können uns vor Anfragen kaum retten“, berichtet Bauer. (Nürnberger Nachrichten, 12.9.2007)

b. Das Jugendcafé Ole im Alten Forsthaus kann sich derzeit vor Besuchern kaum retten, berichtet Jugendpfleger Reine Greve. (Frankfurter Rundschau, 16.9.1998)

(2a) besagt nichts anderes, als dass X über sehr viel Y verfügt, und (2b) erinnert zudem an die lokative Variante C, da es sich bei X um eine konkrete Örtlichkeit handelt.

Einen Schritt näher an der Zielkonstruktion sind Varianten des Idioms *X sieht den Wald vor lauter Bäumen nicht.* Entsprechende Varianten gibt es in den Daten in derart großer Zahl, dass sich der Begriff der „festen“ Wendung dabei verbietet. Als Bedeutung dieses Musters wurde in Zeschel (2011) veranschlagt, dass das gehäufte Auftreten von Y es X erschwert, Handlung V (bzw. VP) erfolgreich auszuführen. Die gegenüber Ausdrücken des Typs *sich fürchten/kapitulieren/fliehen vor etwas* hinzukommende Implikation ist mithin die durch *lauter* markierte Häufung bzw. starke Ausprägung von Y (sowie die negative Polarität des Musters). Mischungen dieses Typs mit der Zielkonstruktion stellen Beispiele wie (3) dar, bei denen Y eine starke Empfindung von X bezeichnet:

(3) Kann sie vor lauter Liebe nicht mehr klar denken, oder sind es die Feste und schönen Kleider, die sie berauschen? (FAZ, 23.3.1999)

Daneben wurden jedoch auch Varianten des Musters gefunden, die eine geringere Ähnlichkeit zur Zielkonstruktion aufweisen:

(4) Vor lauter Finanzkrise geschieht zu wenig für Klimaschutz. (Nürnberger Nachrichten, 16.2.2009)

In diesen Fällen gibt es keinen belebten Subjektreferenten X, der sich angesichts der Konfrontation mit Y in einer bestimmten Weise verhält bzw. eine bestimmte Handlung ausführt/nicht ausführen kann. Stattdessen bleibt die von der Konstellation betroffene Entität, die zu einer bestimmten Handlung nicht in der Lage ist, formal implizit. Mit Blick auf die Zielkonstruktion hat das Muster mit *lauter* also sowohl semantisch ähnlichere als auch semantisch unähnlichere Varianten in den Daten als die erste diskutierte Gruppe.

Damit zum äußeren Rand der eigentlichen Zielkonstruktion, der Kausalangabe (Variante A). Es wurde bereits darauf hingewiesen, dass die Grenze zu der intensivierenden Attribuierungskonstruktion (Variante B) unscharf ist, da bei potenziell hyperbolischen Verwendungen wie (5) nicht klar ist, ob die Verbbedeutung faktisch assertiert ist oder nicht:

(5) An der Seitenlinie aber stand Schalke-Trainer Rangnick und zitterte vor Angst. (dpa, 27.4.2011)

Ein Punkt des Übergangs zwischen den beiden Konstruktionsvarianten sind also Instanzen mit Verben, die typisch für Variante A sind, jedoch im gegebenen Kontext allein für den expressiven Effekt der Konsekutivimplikation eingesetzt werden (*etwas ist so* <EIGENSCHAFT>, *dass* ...), ohne dass das Eintreten des Verbereignisses dabei tatsächlich behauptet würde. Das geschilderte Verfahren fördert daneben jedoch noch weitere, nicht-antizipierte Überlappungen zwischen den Varianten zutage:

(6) Dafür gibt es Miniaturen zu entdecken, die vor Schönheit, Exklusivität und sprachlichem Genie betören. (taz, 11.12.2004)

(6) lässt sich einerseits kausal paraphrasieren (die Miniaturen betören *deshalb, weil* sie so außerordentlich schön sind). Andererseits liegt funktional betrachtet eine intensivierte Eigenschaftszuschreibung vor, und die Kombination eines abstrakten Gegenstands – der (sprachlichen) *Miniatur* – mit der gleichfalls abstrakten Qualität *Schönheit* ist eine Partizipantenkonfiguration, die deutlich typischer für Variante B ist. Wie schon im Fall der Idiomvarianten mit *lauter* bringt die geänderte Datenerhebungsstrategie damit neue Querverbindungen und Berührungspunkte zwischen den postulierten Mustern ans Licht, die aus einer verbbasierten Perspektive verborgen bleiben.

Damit zum Fokus der Untersuchung im engsten Sinne, den figurativen Varianten der Zielkonstruktion. Hier wurden einige Instanzen des Musters gefunden, deren Verben keinem der in der verbbasierten Analyse in Zeschel (2011) veranschlagten semantischen Typen entsprechen:

(7) a. Die Ringkämpfer können vor Kraft nicht gehen. (FAZ, 17.3.2001)

b. David Garrett zeigt dampfend vor Selbstbewusstsein, dass er [...] (Hannoversche Allgemeine Zeitung, 9.11.2009)

c. „Das Haus lacht vor Silber" – Landesmuseum zeigt römisches Tafelsilber im Alten Haus. (Rhein-Zeitung, 8.10.1997)

Das erste Beispiel kommt auch in Abwandlungen mit *laufen* vor. Das zweite ließe sich zwar auf den postulierten semantischen Subtyp 'Y tritt aus X aus' zurückführen (mit instanziierenden Verben wie *überquellen, überschäumen* und *triefen*), zeigt damit aber an, dass die mutmaßlich relevante Verbklasse (Verben, die die Bewegung von Flüssigkeiten denotieren) zu eng veranschlagt wurde. Die Verwendung in (7c) hingegen steht allein und ist als kreativer Okkasionalismus zu werten. Insgesamt bleibt bezüglich der Konstruktionsvarianten B und C jedoch festzuhalten, dass die abweichenden Resultate nur marginale Details betreffen und beispielsweise keine gänzlich neue semantische *Klasse* von Instanzen gefunden wurde.

Als letzter Punkt ist zu erwähnen, dass sich auch aus der abschließenden Validierung der Ausschlussmuster noch interessante Funde wie etwa der folgende ergaben:

(8) Er erinnert sich sehr genau an die Fragen der Journalisten, die alle „irgendwie trieften vor einer gewissen Schadenfreude, auch [...]" (FAZ, 10.9.2005)

Diese Beobachtung zeigt, dass es auch Ausnahmen zur oben vorausgesetzten Regel gibt, der zufolge determinierte und quantifizierte NPn innerhalb der *vor*-PP nicht auftreten können (vgl. Dowty 2000).

5. Zusammenfassung

Der vorliegende Beitrag hat ein Verfahren zur datengeleiteten Identifikation verbaler Argumentstrukturmuster in Korpusdaten vorgestellt. Die Prozedur eignet sich zum einen für die Entdeckung „neuer", d.h. nicht bereits vor Beginn der Untersuchung antizipierter Muster in den Daten (hier nicht illustriert). Zum anderen erlaubt sie auch die Ermittlung „neuer", nicht-antizipierter Instanzen und Mischungen einer bereits vorgegebenen Zielkonstruktion mit formal und semantisch ähnlichen Strukturen in ihrer systemischen Umgebung. Der vielleicht größte Vorteil des Verfahrens liegt darin, dass auch lexikalisch offene, nicht anhand introspektiv (mehr oder minder geeignet) zusammengestellter Verblisten konkordierbare Muster für eine korpuslinguistische Erkundung zugänglich gemacht werden.

Der Fokus der Herangehensweise liegt auf einem maximierten Rücklauf und eingangs minimaler Präzision, die im Rahmen einer schrittweisen Subkorpusbildung sukzessive verbessert wird. Das Verfahren eignet sich nur für teilschematische Konstruktion, bei denen mindestens eine Strukturposition lexikalisch invariant ist, so dass dieses Element exhaustiv konkordiert werden kann (z.B. eine Präposition im Fall von Argumentstrukturen mit präpositionaler Komponente). Besteht in einem gegebenen Slot nur stark eingeschränkte lexikalische Variabilität (z.B. eine Alternanz zwischen verschiedenen Präpositionen), ist auch denkbar, die einzelnen Varianten separat zu analysieren und die Resultate dann zusammenzuführen.

Die in Abschnitt 4 vorgestellte Pilotstudie illustriert in knapper Form einige der nicht-antizipierten Korpusfunde, die einer introspektiv-verbbasierten Analyseperspektive verborgen bleiben. Es bleibt abzuwarten, ob sich die Datenmenge in Folgestudien ebenso stark reduzieren lässt wie dies bei *vor* der Fall war. Unabhängig davon ist damit zu rechnen, dass das Verfahren jeweils den konkreten Erfordernissen des aktuellen Untersuchungsgegenstandes anzupassen ist (trivialerweise gibt es z.B. nicht für alle Präpositionen homographe Verbpartikeln, die aus den Daten auszuschließen sind) und sicher auch noch in verschiedener Hinsicht optimier- und ausbaubar ist (z.B. durch den Einbezug von Parsing). Genau wie eine breitere empirische Evaluation des Nutzen-Kosten-Verhältnisses des Ansatzes steht die Beantwortung dieser Fragen momentan jedoch noch aus.

Literatur

DeReKo = Institut für Deutsche Sprache (2013): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2013-I (Release vom 19.3.2013). Mannheim: Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo.

Desgraupes, Bernard/Loiseau, Sylvain (2012): rcqp: interface to the corpus query protocol. http://cran.r-project.org/web/packages/rcqp/index.html (Stand: 7.7.2015).

Dowty, David (2000): 'The garden swarms with bees' and the fallacy of 'argument alternation'. In: Ravin, Yael/Leacock, Claudia (Hg.): Polysemy: theoretical and computational approaches. Oxford, S. 111-128.

Evert, Stefan/Hardie, Andrew (2011): Twenty-first century corpus workbench: updating a query architecture for the new millennium. In: Proceedings of the Corpus Linguistics Conference, Birmingham, 20-22 Juli 2011. www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf (Stand: 7.7.2015).

Giesbrecht, Eugenie/Evert, Stefan (2009): Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus. In: Alegria, Iñaki/Leturia, Igor/Sharoff, Serge (Hg.): Proceedings of the Fifth Web as Corpus Workshop (WAC5), 7. September 2009, San Sebastian, Spanien, S. 27-35. https://www.sigwac.org.uk/attachment/wiki/WAC5/WAC5_proceedings.pdf (Stand: 7.7.2015).

Levin, Beth (1993): English verb classes and alternations: a preliminary investigation. Chicago.

Müller, Wolfgang (2013): Wörterbuch deutscher Präpositionen. 3 Bde. Berlin.

Proost, Kristel (2009): Warum man nach Schnäppchen jagen, aber nicht nach Klamotten bummeln kann. Die *nach*-Konstruktion zwischen Lexikon und Grammatik. In: Winkler, Edeltraud (Hg.): Konstruktionelle Varianz bei Verben. (= OPAL Sonderheft 4/2009). Mannheim, S. 10-41. http://pub.ids-mannheim.de/laufend/opal/pdf/opal09-4_proost.pdf (Stand: 7.7.2015).

R Core Team (2014): R: a language and environment for statistical computing. Wien. www.R-project.org (Stand: 7.7.2015).

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, S. 44-49.

Zeschel, Arne (2011): Den Wald vor lauter Bäumen sehen – und andersherum: Zum Verhältnis von 'Mustern' und 'Regeln'. In: Lasch, Alexander/Ziem, Alexander (Hg.): Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze. (= Stauffenburg Linguistik 58). Tübingen, S. 43-57.

Zeschel, Arne (2012): Incipient productivity: a construction-based approach to linguistic creativity. Berlin.