GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Technische Universität Dortmund

Fakultät Kulturwissenschaften
Institut für deutsche Sprache und Literatur
Lehrstuhl für Linguistik der deutschen
Sprache und Sprachdidaktik

Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Technischer Bericht

Nr. 2013/1 (Meilenstein 1)

Automatische Klassifikation von Stützverbgefügen mithilfe von Data-Mining

BMBF-Verbundprojekt:

Korpus-basierte linguistische Recherche und Analyse mithilfe von Data-Mining (KobRA)

Förderkennzeichen: 01UG1245A
Projektlaufzeit: 01.09.2012 bis 31.08.2015
Projektkoordination: Prof. Dr. Angelika Storrer
Bearbeiter: Thomas Bartz, Christian Pölitz, Nadja Radtke

Dortmund, den 31.8.2013

Das diesem Bericht zugrunde liegende Vorhaben wird mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter den Förderkennzeichen 01UG1245A-D gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Automatische Klassifikation von Stützverbgefügen mithilfe von Data-Mining

1. Problemstellung und Projektkontext

Das übergreifende Ziel des KobRA-Projekts besteht darin, durch den Einsatz innovativer Data-Mining-Verfahren (insbesondere Verfahren des maschinellen Lernens) die Möglichkeiten der empirischen linguistischen Arbeit mit strukturierten Sprachressourcen zu verbessern. Die Verfahren werden an linguistischen Fallstudien erprobt, die an konkrete Forschungsprojekte angebunden sind. Die in diesem Report dokumentierte Fallstudie bearbeitet einen Problemtyp, der in vielen korpusbasierten linguistischen Untersuchungen auftritt: Zu einem empirisch-quantitativ zu untersuchenden sprachlichen Phänomen lassen sich zwar umfangreiche Trefferlisten aus Korpora gewinnen. Diese Trefferlisten sind aber nicht unmittelbar nutzbar, weil sie viele falsch positive Treffer enthalten, die mit der vorhandenen Korpustechnologie auch nicht weiter ausgefiltert werden können. Gesucht werden deshalb Data-Mining-Verfahren, die den Linguisten dabei unterstützen, falsch positive Treffer aus großen Suchtrefferlisten auszusondern.

Die Fallstudie bezieht sich auf Forschungen zu einem Konstruktionstyp, der in diesem Report als Stützverbgefüge (SVG, engl. ‚support verb construction', franz. ‚construction à verbe support') bezeichnet wird.[1] SVG bestehen aus einem Verb (‚Stützverb') und einem meist abstrakten Nomen (‚prädikatives Nomen'), die gemeinsam einen Prädikatsausdruck bilden. Syntaktisch lassen sich zwei Typen unterscheiden: Stützverben mit dem prädikativen Nomen im Akkusativ (Beispiel 1.1) und Stützverben mit dem prädikativen Nomen in der Präpositionalphrase (Beispiel 1.2):

1.1 Die Methoden **fanden** keine **Anwendung**.
1.2 Klaus **bringt** seinen Wagen **ins Rollen**.

Die Beispiele 1.1 und 1.2 zeigen ein für unsere Studie relevantes Charakteristikum von Stützverben: Stützverben sind zwar aus Vollverben (hier: *finden* und *bringen*) entstanden; ihre Bedeutung ist aber im Zuge der Grammatikalisierung ‚verblasst'[2]. Die zentrale Funktion der Stützverben besteht darin, gemeinsam mit dem prädikativen Nomen ein komplexes Prädikat zu bilden; die Bedeutung dieses komplexen Prädikats wird hauptsächlich vom Nomen getragen. Wenn man die Stützverbgefüge in 1.1 und 1.2 mit Beispielen wie 1.3 und 1.4 vergleicht,

[1] In der deutschsprachigen Literatur findet man auch die Bezeichnungen ‚Funktionsverbgefüge', ‚Nominalisierungsverbgefüge' , ‚Streckverbgefüge' oder ‚verbonominale Konstruktionen'; einen Überblick über Merkmale und Terminologie geben u.a. van Pottelberge (2001), Langer (2005), Heid (2004), Storrer (2006/2007).

[2] In der englischen Literatur findet man deshalb auch den Ausdruck ‚light verb' statt ‚support verb'.

in denen dasselbe Verb als Vollverb auftritt, wird der Unterschied zwischen ‚verblasstem' Stützverb und Vollverb deutlich.

1.3 Die Touristen **fanden** keine Unterkunft.
1.4 Klaus **bringt** seinen Sohn ins Schwimmbad.

Das Problem, an dem die hier beschriebenen Experimente ansetzen, besteht darin, dass man Vollverbverwendungen wie 1.3 und 1.4 nicht zuverlässig anhand morphosyntaktischer Merkmale von Stützverbverwendungen wie in 1.1 und 1.2 unterscheiden kann. Für die Identifizierung von Stützverben ist vielmehr ein semantisches Merkmal (‚semantisch verblasst') relevant, das in den großen Referenzkorpora (z.B. in den Korpora der Projektpartner: DWDS, vgl. Geyken 2007; DeReKo, vgl. Kupietz et al. 2010, 2009; TüBa-D/Z, vgl. Telljohann et al. 2012) nicht annotiert bzw. nicht regelbasiert zu ermitteln ist. Wer Frequenzdaten zu Stützverben in Korpora erheben und vergleichen möchte, steht deshalb immer vor der Aufgabe, aus der Gesamtmenge der Treffer, die zu einem Verb wie *finden* oder *bringen* ausgegeben werden, die Teilmenge der Treffer zu bestimmen, in denen das Verb als Stützverb vorkommt. Da typische Stützverben wie *bringen*, *kommen*, *stehen*, *machen* zu den hochfrequenten Verben des Deutschen zählen, kann eine solche Teilmengenbildung nur mit großem Zeitaufwand manuell-intellektuell durchgeführt werden. Manuelle Klassifikationen von Korpusbelegen haben zudem ergeben, dass das Verhältnis zwischen Stützverbverwendungen und Vollverbverwendungen bei den verschiedenen Verben erheblich variiert (vgl. Kamber 2008, S. 461, Storrer 2013). Dies bedeutet, dass die Anteile für jedes Verb einzeln ermittelt werden müssen.

Zur Lösung des Problems wurden in der hier beschriebenen Fallstudie auf der Basis von manuell vorklassifizierten Daten verschiedene Experimente mit Data-Mining-Verfahren durchgeführt und evaluiert, die dabei helfen, aus einer Treffermenge zu einem Verb die Teilmenge der Stützverbverwendungen zu identifizieren (vgl. 3.2.2). Diese manuell vorklassifizierten Daten stammen aus einem Projekt, in dem die folgenden Teilfragen auf der Grundlage umfangreicher Korpusdaten untersucht wurden:

- **Zeitliche Entwicklung:** Verändern sich die Frequenz und der Bestand der Verben und der Gefüge über das 20. Jahrhundert hinweg?
- **Textsortenspezifik**: Wie verteilen sich die Vorkommen über verschiedene Textsortenbereiche?

Es handelte sich um ein Teilprojekt im Vorhaben ‚Bericht zur Lage der deutschen Sprache', das gemeinsam von der Union der deutschen Akademien der Wissenschaften und der Deutschen Akademie für Sprache und Dichtung durchgeführt und 2009-2011 von der Thyssen-Stiftung finanziell gefördert wurde (vgl. Sprachbericht 2013). Die Fragestellungen, das methodische Vorgehen und die Ergebnisse des Projekts sind ausführlich in Storrer (2013) beschrieben. Wir nehmen im Folgenden auf dieses Projekt mit dem Ausdruck ‚Projekt Sprachbericht' Bezug.

Der Report ist folgendermaßen aufgebaut: Im folgenden Abschnitt 2 beschreiben wir zunächst die verwendete Datengrundlage und die linguistischen Vorarbeiten, die in die Data-Mining-Experimente eingeflossen sind. Abschnitt 3 erläutert das Vorgehen bei den Experimenten und die eingesetzten Data-Mining-Methoden; in Abschnitt 4 werden die Ergebnisse der Evaluation dargestellt. Die Verfahren und ihre Weiterentwicklung werden u.a. in ein Dissertationsprojekt einfließen, bei dem das System und der Bestand deutscher Stützverbgefüge für die Lehre

im Bereich Deutsch als Fremdsprache aufbereitet und in einem wiki-basierten Wörterbuch dargestellt werden[3]. Abschnitt 5 gibt einen Ausblick auf die geplanten Erweiterungen.

2. Datengrundlage und linguistische Vorarbeiten

2.1 Datenerhebung

Die in den Experimenten genutzten Daten wurden im Zuge des Projekts Sprachbericht erhoben. Aus den insgesamt in diesem Projekt erhobenen Datenbeständen haben wir für die Experimente die Verben *bringen, kommen* und *finden* ausgewählt, weil zu diesen Daten umfangreiche manuelle Annotationen vorhanden waren. Die Daten stammen aus zwei Korpusbeständen, die im Folgenden kurz skizziert werden:

Das **Kernkorpus des Projekts ‚Digitales Wörterbuch der deutschen Sprache (DWDS)'**, das im Folgenden **‚DWDS-KK'** abgekürzt wird, ist ein Referenzkorpus zur deutschen Sprache des 20. Jahrhunderts, das an der Berlin-Brandenburgischen Akademie der Wissenschaften aufgebaut wurde. Es umfasst 100.600.993 Textwörter, die in ausgewogenem Verhältnis über die Dekaden des 20. Jahrhunderts verteilt sind. Da jede Dekade auch eine vergleichbare Zahl von Textwörtern aus vier verschiedenen Textsortenbereiche (Belletristik, Gebrauchstexte, Wissenschaft, Zeitung) enthält, eignet sich das Korpus nicht nur für die Untersuchung der Frequenzentwicklung über das 20. Jahrhundert hinweg, sondern auch für den Vergleich der Vorkommensfrequenzen in den unterschiedlichen Textsortenbereichen. Die Daten sind teilweise urheberrechtlich geschützt, standen aber für die Auswertungen im Projekt vollständig zur Verfügung.

Die Daten wurden von uns am 09.02.2012 erhoben, dabei unterteilten wir die Datensätze nach den vier Textsortenbereichen. Tabelle 1 zeigt die Vorkommensfrequenzen zu den Verben *bringen, finden* und *kommen* sowie ihre Verteilung auf die vier Textsortenbereiche.

Verb	Gesamt	Belletristik		Gebrauchsliteratur		Wissenschaft		Zeitung	
bringen	64.629	18.006	27,86%	14.301	22,13%	12.653	19,58%	19.669	30,43%
finden	82.162	21.704	26,42%	17.215	20,95%	21.345	25,98%	21.898	26,65%
kommen	165.094	71.399	43,25%	36.068	21,85%	23.924	14,49%	33.703	20,41%

Tabelle 1: Vorkommensfrequenzen im DWDS-KK

Das Wikipedia-Korpus/Artikelseiten (‚**Wiko-A'**) und das Wikipedia-Korpus/ Diskussionsseiten (‚**Wiko-D'**) spiegeln die Version der Deutschen Wikipedia vom 13.08.2010 wider, die linguistisch am UKP (Ubiquitous Knowledge Processing Lab) der TU Darmstadt aufbereitet und für das Projekt zur Verfügung gestellt wurden (vgl. Zesch et al. 2007). Wiko-A umfasst 558.882.506 Textwörter; Wiko-D umfasst 234.770.301 Textwörter.

Tabelle 2 zeigt die die Vorkommensfrequenzen der Verben *bringen, finden* und *kommen* in den beiden Teilkorpora.

Verb	Wiko-A	Wiko-D
bringen	124.675	69.582
finden	333.262	380.315
kommen	433.125	232.653

Tabelle 2: Vorkommensfrequenzen in Wiko-A und Wiko-D

[3] Radtke, Nadja (in Vorbereitung): Konzeption und korpusbasierter Aufbau einer Wiki-Ressource zu deutschen Stützverbgefügen. Dissertation, TU Dortmund.

2.2. Datenaufbereitung

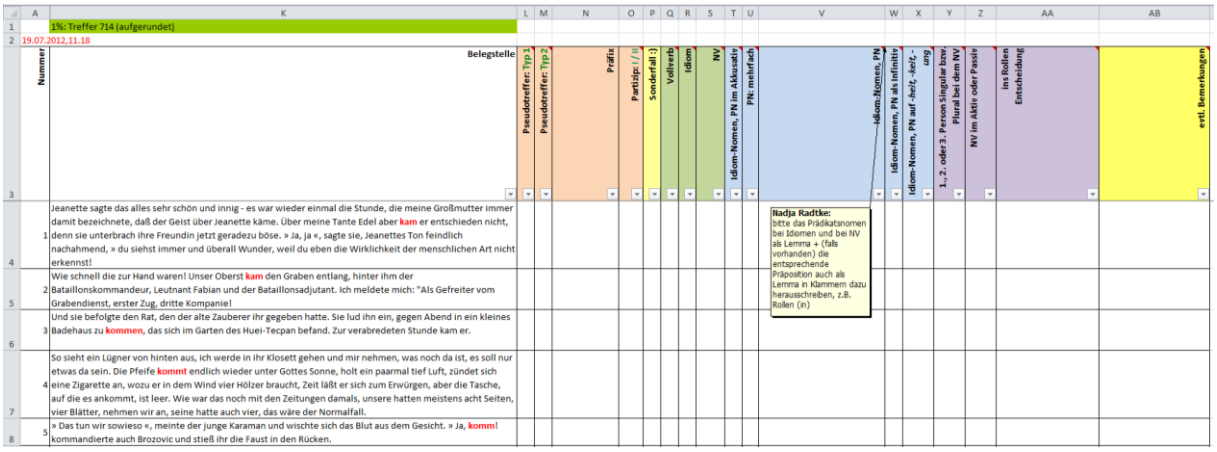
Wie bereits in Abschnitt 1 erläutert, lassen sich Vollverbverwendungen anhand der Form oder morphosyntaktischer Merkmale nicht zuverlässig von Stützverbverwendungen unterscheiden. Im Projekt Sprachbericht konnten wir deshalb bei den Untersuchungen zur Frequenzentwicklung und zur Textsortenspezifik nur mit Stichproben arbeiten, die wir im Hinblick auf verschiedene Merkmale manuell vorklassifiziert haben (vgl. im Detail Storrer 2013).

Die vom jeweiligen Korpusrecherchesystem ausgegebenen Textsegmente, die wir im Folgenden als ‚Treffer-Snippets' bezeichnen, wurden allesamt in Excel-Dateien bearbeitet. Wie der Ausschnitt in Abbildung 1 zeigt, belegt jedes Treffer-Snippet eine Tabellenzeile. Im Snippet ist das gesuchte Verb farbig bzw. durch festgelegte Sonderzeichen hervorgehoben (z.B.: „Sein Mut &&*findet*&& überall die Anerkennung der Anwesenden"). Bei den Korpora Wiko-A und Wiko-D wurde nur ein Satzkontext ausgegeben; die Snippets des DWDS-KK umfassen drei Sätze.

Die Metadaten zu den Snippets (Erscheinungsdatum, Textsorte etc.) sind in jeweils separaten Spalten vermerkt. Auch die manuelle Annotation linguistischer Merkmale wird in separaten Spalten festgehalten. Annotiert wurde, ob das Verb im Snippet als Stützverb verwendet wird; diese Information war für die im Folgenden beschriebenen Experimente relevant. Die Annotation für das Projekt Sprachbericht berücksichtigte aber noch weitere linguistische Merkmale (vgl. Abbildung 2), die für künftige Experimente genutzt werden können.

	A	B	C	D	E	F	G	H	I	J	L
1											1%: Treffer 217 (aufgerundet)
2	19.07.2012,11.13										
3	**Nummer**	**Nummer (ursprünglich)**	**Random**	Bibliographie (komplett)	Randomisierungsnummer	Erstveröffentlichung	Rechtsstatus	Pfad	Bibliographie (einzeln)	Bibliographie (einzeln)	**Belegstelle mit Brezeln**
4	1	77450	92	Gütersloh, Albert Paris, Sonne und Mond, München: Piper 1962	455	31.12.1962	OR0W		Belletristik	Sonne und Mond	Der Schöpfer ist auch der Schöpfer der heuristischen Prinzipien. Woraus denn sonst - weil niemand sich von vorneherein, aus dem bloßen Grunde unverdienten Existierens, lächerlich &&findet&& -, wenn nicht aus der erwerbbaren tiefen Einsicht, das einzige, überall und jederzeit von der Gottheit düpierte Wesen zu sein, sollte jene heilsame Komik entspringen, kraft welcher wir mit der absoluten Abhängigkeit und Unfähigkeit der eigenen Person ironischen Ernst machen? Herr Andree sprang übertrieben viel hin und her; aber übertrieben viel nur für den verehrten Leser, der die Ruhe selber ist, wenn er die ihm gemäße Kamera auf die ihr ebenfalls gemäße Umwelt richtet.
5	2	1461	160	Baum, Vicky, Menschen im Hotel, Berlin: Ullstein 1929	291	31.12.1929	OR1S		Belletristik	Menschen im Hotel	Er hat sie vier Tage vergraben und versteckt gehalten, was beinahe schon so gut wie gestohlen ist. Und jetzt ist er so weit und hat sich durchgebissen, will sich von ihr trennen und will sie - als &&gefunden&& - zurückgeben. Da stand er nun mit seinem Herzklopfen vor Nr.
6	3	57926	218	Nadolny, Sten, Selim oder Die Gabe der Rede, München: Piper 1990	80	31.12.1990	OR0W		Belletristik:Roman	Selim oder Die Gabe der Rede	Er bewegte die Arme in der Luft - offenbar ein Geschichtenerzähler, der von einer abenteuerlichen Segelfahrt um die Welt erzählte, und alles war erfunden. Dieses Denkmal, &&fand&& Selim, wirkte sympathisch. Am Kiosk sah er, was er suchte.

Abbildung 1: Excel-Tabelle mit importierten Treffer-Snippets aus dem DWDS-KK für das Verb *finden*, Hervorhebung durch festgelegte Sonderzeichen („&&"); Metadaten in separaten Spalten.



	Nummer	Belegstelle	Pseudotreffer: Typ 1	Pseudotreffer: Typ 2	Präfix	Partizip: I / II	Sonderfall :)	Valenzb.	Idiom	NV	Idiom-Nomen, PN im Akkusativ	PN: mehrfach	Idiom-Nomen, PN	Idiom-Nomen, PN als Infinitiv	Idiom-Nomen, PN auf -heit, -keit, -ung	1., 2. oder 3. Person Singular bzw. Plural bei dem NV	NV im Aktiv oder Passiv	ins Rollen Entscheidung	ertl. Bemerkungen
1		1%: Treffer 714 (aufgerundet)																	
2		19.07.2012,11.18																	
4	1	Jeanette sagte das alles sehr schön und innig - es war wieder einmal die Stunde, die meine Großmutter immer damit bezeichnete, daß der Geist über Jeanette käme. Über meine Tante Edel aber kam er entschieden nicht, denn sie unterbrach ihre Freundin jetzt geradezu böse. » Ja, ja «, sagte sie, Jeanettes Ton feindlich nachahmend, » du siehst immer und überall Wunder, weil du eben die Wirklichkeit der menschlichen Art nicht erkennst!																	
5	2	Wie schnell die zur Hand waren! Unser Oberst kam den Graben entlang, hinter ihm der Bataillonskommandeur, Leutnant Fabian und der Bataillonsadjutant. Ich meldete mich: "Als Gefreiter vom Grabendienst, erster Zug, dritte Kompanie!																	
6	3	Und sie befolgte den Rat, den der alte Zauberer ihr gegeben hatte. Sie lud ihn ein, gegen Abend in ein kleines Badehaus zu kommen, das sich im Garten des Huei-Tecpan befand. Zur verabredeten Stunde kam er.																	
7	4	So sieht ein Lügner von hinten aus, ich werde in ihr Klosett gehen und mir nehmen, was noch da ist, es soll nur etwas da sein. Die Pfeife kommt endlich wieder unter Gottes Sonne, holt ein paarmal tief Luft, zündet sich eine Zigarette an, wozu er in dem Wind vier Hölzer braucht, Zeit läßt er sich zum Erwürgen, aber die Tasche, auf die es ankommt, ist leer. Wie war das noch mit den Zeitungen damals, unsere hatten meistens acht Seiten, vier Blätter, nehmen wir an, seine hatte auch vier, das wäre der Normalfall.																	
8	5	» Das tun wir sowieso «, meinte der junge Karaman und wischte sich das Blut aus dem Gesicht. » Ja, komm! kommandierte auch Brozovic und stieß ihr die Faust in den Rücken.																	

Nadja Radtke:
bitte das Prädikatsnomen bei Idiomen und bei NV als Lemma + (falls vorhanden) die entsprechende Präposition auch als Lemma in Klammern dazu herausschreiben, z.B. Rollen (in)

Abbildung 2: Ergebnisse der manuellen Analysen mit Hinweisen zur Annotation als Kommentare (Ausschnitt aus der Datei zu *kommen*, DWDS-KK).

2.3 Spezifikation linguistischer Merkmale für die Klassifikationsverfahren

Aus den Forschungsarbeiten zu Stützverbgefügen sind Erkenntnisse zu morphosyntaktischen und distributionellen Merkmalen von Stützverbgefügen hervorgegangen, die sich für die automatischen Klassifikationsverfahren nutzen lassen. Als eine Vorarbeit für die in Abschnitt 3 beschriebenen Experimente wurden Merkmale zusammengestellt, die sich als Indizien für das Vorliegen von SVG werten lassen. Für die Experimente wurden zwei Merkmalslisten erstellt: Die in 2.3.1 dargestellte Liste bezieht sich auf typische Merkmale von prädikativen Nomina; die in 2.3.2 dargestellte Liste bezieht sich auf typische Merkmale von Stützverben.

Zur Erläuterung der Merkmale wird in beiden Tabellen auf die folgende Beispielsätze verwiesen:

1.1 Sein Mut **findet** überall **Anerkennung**.
1.1.1 Sein Mut **wird** überall **Anerkennung finden**.
1.1.2 Sein Mut **hat** überall **Anerkennung gefunden**.
1.1.3 Klaus versucht, überall **Anerkennung zu finden**.
1.1.4 Klaus hofft, dass sein Mut überall **Anerkennung findet**.
1.1.5 Klaus hofft, dass sein Mut überall **Anerkennung finden wird**.
1.1.6 Sein Mut **hat** überall die **Anerkennung** der Anwesenden **gefunden**.
1.1.7 Sein Mut **hat** überall die **Anerkennung**, nach der er fortwährend sucht, **gefunden**.
1.1.8 Sein Mut ist der Grund für die **Anerkennung**, die er überall **findet**.
1.1.9 **Anerkennung hat** sein Mut überall **gefunden**.

1.2 Klaus **bringt** den Wagen **zum Stehen**.
1.2.1 Klaus **wird** den Wagen **zum Stehen bringen**.
1.2.2 Klaus **hat** den Wagen **zum Stehen gebracht**.
1.2.3 Klaus versucht, den Wagen **zum Stehen zu bringen**.
1.2.4 Klaus hofft, dass er den Wagen **zum Stehen bringt**.
1.2.5 Klaus hofft, dass er den Wagen **zum Stehen bringen wird**.
1.2.6 Klaus bringt den Wagen, dessen Bremsen defekt sind, **zum Stehen**.
1.2.7 **Zum Stehen** lässt sich der Wagen bei diesem Gefälle niemals **bringen**.

2.3.1 Das prädikative Nomen

Stärkstes Indiz für das Vorliegen eines Stützverbgefüges in einem Satz sind zunächst (A) formale/distributionelle und (B) morphosyntaktische Merkmale, die sich auf die (z.T. präpositional angebundene) nominale Komponente des Gefüges beziehen. Die Reihenfolge der aufgeführten Merkmale gibt jeweils deren Priorisierung an (oben=höchste).

Merkmal-Kategorie	**Merkmale: Das prädikative Nomen ...**	**Beispiele**
A1 (Großschreibung)	ist ein Ausdruck mit Großschreibung des Anfangsbuchstabens	
A2 (Position)	hat einen Abstand von höchstens 3 Wörtern zum Satzschlusszeichen am rechten Ende des Satzes	1.1-1.1.5 1.2-1.2.6
	oder hat einen Abstand von höchstens 3 Wörtern zum linken Ende des Satzes	1.1.9
	oder steht unmittelbar links vor einem Komma	1.2.7
A3 (Kookkurrenzen)	folgt im Abstand von höchstens 2 Wörtern auf *in*/*ins*, *zu*/*zum*/*zur*	1.2
B1 (Wortart)	ist ein Nomen (*NN*)[4]	
B2 (Phrasenstruktur)	ist Kopf einer Nominalphrase (*NP*)	1.1.10
	oder ist Kopf einer Nominalphrase und Konstituente einer Präpositionalphrase (*PP*)	1.2.8
B3 (Satzfunktion)[5]	wird als Akkusativ-Objekt annotiert (*OA*)	1.1.10
	oder wird als Modifizierer (*MO*)/„Collocational Verb Construction“ (*CVC*) annotiert	1.2.8
A4 (Endung)	endet auf *ung*/*ungen*/*heit*/*keit*	1.1
	oder endet auf *en*/*ung*/*heit*/*keit*, wenn *in*/*ins*, *zu*/*zum*/*zur* in einem Abstand von höchstens 2 Wörtern vorausgehen.	1.2

Tabelle 3: Indizien für das Vorliegen eines prädikativen Nomens

2.3.2 Das Stützverb

Indizien für das Vorliegen eines Stützverbs lassen sich weiterhin aus folgenden Merkmalen des Stützverbs ableiten:

Merkmal-Kategorie	**Merkmale: Das Stützverb ...**	**Beispiele**
B1 (Wortart)	ist ein Vollverb (*VVFIN*)[6]	
	oder tritt in einem Satz als Partizip (*VVPP*) zusammen mit einem Hilfsverb (*VAFIN*) auf	1.1.2 1.2.2
	oder tritt in einem Satz als Infinitiv (*VVINF*) zusammen mit einem Hilfsverb (*VAFIN*) oder Modalverb (*VMFIN*) auf	1.1.5 1.2.5
	oder tritt in einem Satz als zu-Infinitiv (*VVIZU*) auf	1.1.3, 1.2.3
B2 (Morphologie)	Stützverb, Hilfs- oder Modalverben treten als finite Verben (*V*FIN*) bevorzugt in der der 3. Person Singular oder Plural auf (*person: 3; number: singular/Sg, plural/Pl*)	

[4] Part-of-Speech-Tags des Stuttgart-Tübingen-Tagsets STTS, vgl. Schiller et al. (1999).

[5] Bei B3 ist zu beachten, dass diese Kategorie von automatischen Parsern u.U. unzuverlässig annotiert wird. Überhaupt sind ja auch die o. angegebenen Merkmale *OA* und *MO* falsch, denn bei den SVG-Komponenten handelt es sich um Prädikatsbestandteile. Das korrekte Edge-Label *CVC* (‚collocational verb construction‘) wird jedoch nach unseren Erfahrungen bisher allenfalls von Dependenz-Parsern und ebenfalls nicht zuverlässig vergeben.

[6] Stützverben werden bislang von den automatischen linguistischen Verarbeitungswerkzeugen als ‚Vollverben‘ analysiert (*VVFIN* nach Stuttgart-Tübingen-Tagset STTS, vgl. Schiller et al. 1999). Das STTS enthält keine eigenen Tags für Stützverben.

A1 (Formen)	tritt bevorzugt in folgenden Formen auf: *findet*/*finden*, *fand*/*fanden*, *hat*/*haben gefunden*, *wird*/*werden finden*; *bringt*/*bringen*, *brachte*/*brachten*, *hat*/*haben gebracht*, *wird*/*werden bringen*; *kommt*/*kommen*, *kam*/*kamen*, *ist*/*sind gekommen*, *wird*/*werden kommen*	1.1.1-1.1.5

Tabelle 4: Indizien für das Vorliegen eines Stützverbs

3. Beschreibung der Data-Mining-Experimente

3.1 Vorüberlegungen und Aufbau der Experimente

Wie bereits erläutert, lassen sich Stützverbverwendungen von den Vollverbverwendungen, aus denen sie hervorgegangen sind, anhand morphosyntaktischer Merkmale nicht zuverlässig unterscheiden. Ausschlaggebend für die Klassifikation ist ein semantisches Merkmal (‚semantisch verblasst', s. 1.), das die Anwendbarkeit regelbasierter Verfahren einschränkt. Für den Einsatz von Data-Mining-Verfahren spricht hingegen die Fähigkeit dieser Verfahren, im Wort-, bzw. morphosyntaktischen Kontext oder in den Belegmetadaten gegebene latente Informationen zu nutzen, um die Gefüge von den Konstruktionen der Restgruppe zu unterscheiden. Weil das zu klassifizierende Phänomen theoretisch klar umrissen ist und mit den manuell klassifizierten Datenbeständen Trainingsdaten in hinreichendem Umfang zur Verfügung stehen, empfiehlt sich der Einsatz eines maschinellen Lernverfahrens, das systematische statistische Auffälligkeiten in einer begrenzten Menge manuell klassifizierter Daten auf ungesichtete Daten anwenden und für deren automatische Klassifizierung nutzen kann. Konkret wird bei einem solchen Verfahren die Klassifikation durch komplexe statistische Abbildungen von Suchtreffern (‚Treffer-Snippets') und darin enthaltenen Wörtern bzw. anderen Merkmalen auf Kategorien maschinell gelernt. Die Abbildungen, sogenannte ‚Classifier', können genutzt werden, um einem Suchtreffer oder einem Wort eine bestimmte Kategorie zuzuordnen.

Erste Ansätze automatischer Klassifikationsverfahren in der Informatik gehen in die frühen 60er Jahre zurück. Bereits Maron (1965) schlägt ein Verfahren zur automatischen Klassifikation von Dokumenten anhand dessen Inhalts vor. Ein entscheidender Fortschritt in diesem Bereich wurde mit der Anwendung sogenannter ‚Support-Vector-Maschinen' erzielt, die es erlauben, mit großen Mengen an Merkmalen eines Dokuments (auch irrelevante und redundante Merkmale sind möglich) effizient ein Klassifikationsmodell zu lernen. In diesem Zusammenhang ist besonders die Arbeit von Joachims (1998) hervorzuheben. Classifier werden inzwischen in vielen verschiedenen Anwendungsbereichen eingesetzt, Möglichkeiten und Grenzen der automatischen Klassifikation von Suchtreffern aus Korpora sind hingegen kaum erforscht. Die Computerlinguistik stellt Expertise in Bezug auf die linguistische Aufbereitung der Korpusdaten durch Wortarten- und Syntaxannotationen bereit. Metadaten ermöglichen zudem die Zuordnung von Belegen zu Textsorten und Zeiträumen (z.B. im DWDS-Kernkorpus). Welche Merkmale als sogenannte ‚Features' das maschinelle Lernen von Classifiern verbessern und wie Treffer-Snippets und Merkmale idealerweise für das Verfahren zu repräsentieren sind, sind interessante und unerforschte Fragen.

Die folgenden Abschnitte erläutern die Experimente, die zur Lösung der in Abschnitt 1 dargestellten Problemstellung durchgeführt wurden. Auf Grundlage der in Abschnitt 2 beschriebenen Daten wurden in drei unterschiedlichen Treatments jeweils Classifier gelernt sowie anschließend quantitativ und qualitativ evaluiert. Die in Abschnitt 3.2 beschriebenen Treatments unterscheiden sich in den Features, die für das maschinelle Lernen zur Anwendung gebracht wurden, wobei die Tiefe der Datenanreicherung schrittweise gesteigert wird:

1. **Bags-of-Words-Ansatz (BoW):** ohne linguistische Aufbereitung der Daten
2. **BoW + PoS-Tags:** mit Wortartenannotationen für jedes Textwort,
3. **Linguistische Expertise (Expert):** Berücksichtigung morphosyntaktischer und distributioneller Merkmale aus den Listen in 2.3.1 und 2.3.2.

Als quantitatives Maß für die Zuverlässigkeit der Verfahren dient das gewichtete harmonische Mittel aus Präzision (Precision) und Ausbeute (Recall), wobei Genauigkeit und Ausbeute gleich gewichtet werden. Der auf diese Weise ermittelte F_1-Wert stellt ein Standardmaß für die Beurteilung automatischer Klassifikationsverfahren dar (vgl. Manning et al. 2008). Zusätzlich werden die Ergebnisse manuell qualitativ evaluiert (s. 4.2).

3.2 Technische Beschreibung der Experimente

3.2.1 Vorverarbeitung

Die Treffer-Snippets liegen als Sequenzen von Wörtern vor, die zunächst vorverarbeitet werden müssen, um als Eingabe für maschinelle Lernverfahren dienen zu können. Für die Repräsentation der Snippets existieren unterschiedliche Ansätze. Anknüpfend an 3.1 wurden folgende Ansätze erprobt:

3.2.1.1 Bags-of-Words

Zum einen nutzten wir einen Bags-of-Words-Ansatz, wobei jedes Treffer-Snippet als großer Vektor mit Einträgen für jedes Wort der Gesamtmenge aller Wörter in einer Suchergebnisliste dargestellt wird (ein sogenannter ‚Wortvektor'). In einer Trefferliste mit *N* Wörtern ist der Vektor *N*-dimensional. Die Elemente der Wortvektoren können binär sein und das bloße Vorkommen eines Wortes in einem Treffer-Snippet oder Häufigkeiten des Wortes in einem Snippet und in allen Snippets der Trefferliste darstellen. Formal ist ein Wortvektor *v* für einen endlichen Text definiert als ein *N*-dimensionaler Vektor, d.h. alle möglichen Texte enthalten *N* unterschiedliche Wörter. Für *v* gilt, dass die *i*-te Komponente die Anzahl der Vorkommen oder (normalisierte) Frequenz von Wort *i* im Text ist. Ordnet man diese Wörter, so kann man jedes Wort über einen Index *i* identifizieren. Damit definieren wir eine Abbildung *Phi*, die die Treffer-Snippets (hier wie ‚Texte' behandelt) als Wortvektoren abbildet. Dies geschieht formal so:

$\varphi(d) = (f(\mathrm{w}1,d), f(\mathrm{w}2,d), \ldots, f(\mathrm{w}N,d))$, wobei $f(\mathrm{w}i,d)$ die Anzahl oder (normalisierte) Frequenz von Wort *i* in Text *d* (für ‚document') angibt.

3.2.1.2 Bags-of-Words + PoS-Tags

Zweitens wurden Part-of-Speech-Tags (PoS-Tags) verwendet, um die Wortartenzugehörigkeit der Wörter in jedem Beleg-Satz und damit mögliche überzufällige Kumulierungen bestimmter Wortarten zu repräsentieren. Die PoS-Tags wurden mithilfe des Open-NLP Taggers (Morton et al. 2005) und des Stanford Parsers (Klein & Manning 2003) für deutschsprachige Daten automatisch annotiert (z.B.: „Sein/PPOSAT Mut/NN findet/VVFIN überall/ADJD die/HD Anerkennung/NN der/ART Anwesenden/NN"). Diese Tags werden ähnlich zum Bags-of-Words-Ansatz auf einen Vektor abgebildet, wobei dann jede Komponente für das Vorhandensein eines bestimmten PoS-Tags in den Treffer-Snippets steht. Dieser Ansatz macht am meisten Sinn, wenn jedes Snippet nur aus einem einzigen Satz besteht. Deshalb haben wir für die Experimente dieses Treatments nur jeweils denjenigen Satz der Snippets berücksichtigt, der die potenzielle Stützverbform (gekennzeichnet durch „&&") enthält.

3.2.1.3 Linguistische Expertise

Drittens wurde eine Möglichkeit der Repräsentation bereits vorhandener linguistischer Expertise für den Classifier erprobt. Mithilfe eines regelbasierten Verfahrens wurde dazu für die Treffer-Snippets überprüft, welche vorgegebenen morphosyntaktischen und distributionellen Merkmale der in 2.3 beschriebenen Listen zutreffen und das Zu- bzw. Nichtzutreffen jeweils auf einen binären Vektor abgebildet. Nach dieser Vorgehensweise wird jedes Snippet also durch einen Vektor repräsentiert, bei dem jede Komponente für das Vorhanden- oder Nicht-Vorhandensein eines bestimmten morphosyntaktischen bzw. distributionellen Merkmals steht. Die Merkmale beziehen sich z.T. auf PoS- und syntaktische Annotationen und sind auf die Satzebene beschränkt. Daher wurde als Basis erneut nur jeweils derjenige Satz der Snippets berücksichtigt, der die potenzielle Stützverbform (gekennzeichnet durch „&&“) enthält.

3.2.2 Klassifikationsaufgabe

Wie in Abschnitt 1 erläutert, besteht die zu lösende Aufgabe in der Bereitstellung eines Verfahrens zur Klassifikation von Treffer-Snippets auf das Vorkommen bestimmter Verben in der Verwendung als Stützverben bzw. in anderen Verwendungen. Formal wollen wir einen Classifier $c(d)$ lernen, der für einen gegebenen Satz ein bestimmtes Verb als Stützverb oder Verb einer Restmenge (Vollverben plus weitere Verwendungen) klassifiziert.

Ein für diese Aufgabe geeignetes Verfahren ist die ‚Stützvektormethode' (kurz SVM), deren Überlegenheit auch für Aufgaben der Dokumentklassifikation in der Dortmunder Informatik bereits Joachims (1998) gezeigt hat. Neu ist jedoch die Anwendung der SVM auf Treffer-Snippets aus Korpora. Formal wird dabei eine lineare Hyperebene für den Raum gesucht, der durch die bei der Vorverarbeitung (s. 3.2.1) erzeugten Wortvektoren aufgespannt ist. Die manuell klassifizierten Trainingsdaten bestimmen die Lage dieser Hyperebene, die so definiert ist, dass sie den Raum der Treffer-Snippets mit Stützverben vom Raum der Treffer-Snippets ohne Stützverben trennt und möglichst weit von den jeweils am nächsten liegenden Wortvektoren entfernt ist. Dies hat verschiedene Vorteile: Für die exakte Lagebestimmung der Hyperebene werden nicht alle Wortvektoren (d.h. Snippets) benötigt, sondern nur die am nächsten liegenden sogenannten ‚Stützvektoren'. Außerdem garantiert der möglichst breite Rand um die Hyperebene, dass auch solche Treffer-Snippets noch zutreffend klassifiziert werden können, die von den Trainingsdaten geringfügig abweichen.

Wir verwenden einen binären Classifier, der definiert ist auf Basis einer linearen Funktion $g(d) = \langle w,\varphi(d)(d)\rangle + b$, wobei w ein Vektor in Raum RN ist, b ein Bias-Term und $\langle .,. \rangle$ das Skalarprodukt in R. Der Classifier ist weiterhin definiert durch $c(d) = 1$, falls $g(d) >= 0$ und $c(d) = -1$, falls $g(d) < 0$. Dabei steht 1 für das Vorhandensein eines Stützverbs und -1 für dessen Nicht-Vorhandensein. Die Aufgabe ist nun, den optimalen Vektor w zu bestimmen. Dieser soll so gewählt werden, dass $g(d) >= 0$ ist für alle Sätze d, die ein Stützverb enthalten, und $g(d)<0$ ist für alle Sätze, die kein Stützverb enthalten. Dazu werden die manuell klassifizierten Trainingsdaten benötigt. Der Vektor w wird so gewählt, dass die Hyperebene $g(d)$ die Menge der Trainingsdaten wie oben verlangt trennt. Weiterhin muss w so gewählt werden, dass die Klassifikation neuer, ungesichteter Treffer-Snippets mit hoher Wahrscheinlichkeit richtig vorhergesagt wird. Dies kann man gewährleisten, wenn die Trainingsdaten im Raum der Wortvektoren, also $\{\varphi(d)\}$, einen maximalen Abstand zu $g(d)$ haben. Details zum Verfahren siehe Cristianini & Shawe-Tayler (2004).

3.3 Verwendete Tools

Für sämtliche Experimente wurde das Data-Mining-Tool ‚RapidMiner' (früher: ‚YALE', Mierswa et al. 2006) verwendetet, das eine Vielzahl an Data-Mining-Verfahren sowie Metho-

den zum Einlesen von Daten und zur Evaluierung von Lernverfahren beinhaltet. Weiterhin verfügt RapidMiner über eine Erweiterung, die das Einlesen und Transformieren von Texten in die verschiedenen Repräsentationsformen ermöglicht.

Für die Vorverarbeitung wurde des Text-Mining-Plugin des RapidMiner verwendet, das das zeilenweise Einlesen von Excel-Tabellen ermöglicht. Mithilfe des Plugins konnten die in den Experimenten verwendeten Datensätze zusammen mit den Informationen aus den manuellen Analysen (vgl. Abschnitt 2.2) eingelesen und weiterverarbeitet werden. Anschließend erfolgte durch geeignete Operatoren die Abbildung in Wortvektoren und das Training sowie die Evaluation optimaler Support-Vektor-Maschinen (SVM).

In Abbildung 3 sind die einzelnen Schritte visualisiert:

1. Einlesen der Texte,
2. Erzeugen der Wortvektoren,
3. Kreuzvalidierung mit SVM.

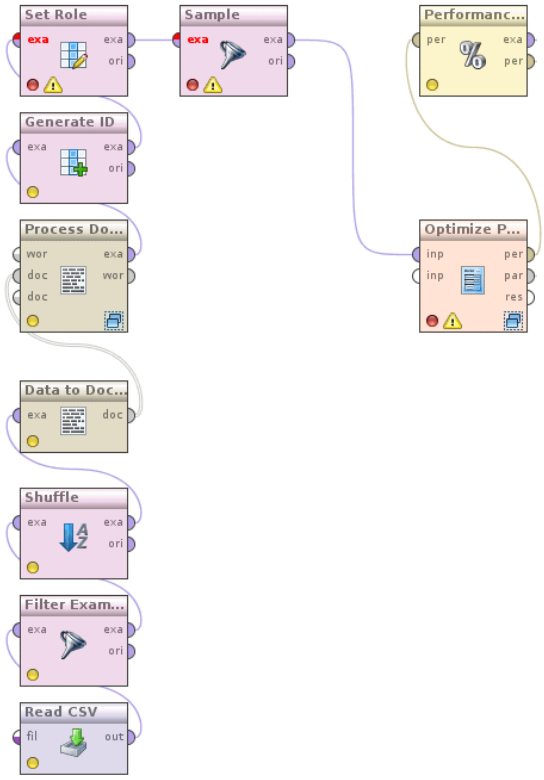


Abbildung 3: Data-Mining-Prozess

Durch den Operator ‚Read CSV' werden die Treffer-Snippets zeilenweise eingelesen. In jeder Zeile steht ein Treffer-Snippet und die manuelle Klassifikation: Stützverb oder Rest. Der nächste Operator ‚Filter Examples' filtert alle Snippets, die keine Informationen darüber enthalten, ob ein Stützverb vorliegt, oder nicht. Danach werde alle Snippets gemischt und mittels des Operators ‚Data to Document' in eine interne Datenstruktur kopiert, aus welcher im nächsten Schritt durch den Operator ‚Process Documents' die Wortvektoren erzeugt werden. Im Weiteren wird durch ‚Set Role' das Label als Zielvariable für einen Classifier deklariert und mittels ‚Sample' gleichviele Snippets mit und ohne Stützverb gezogen. Im Anschluss werden die Daten an einen komplexen Operator weitergeleitet, der die optimalen SVM-Parameter und die Güte des Verfahrens mittels Fünffach-Kreuzvalidierung bestimmt. Im letzten Schritt wird das Gütekriterium ausgegeben. Dies wurde für alle Datensätze durchgeführt.

4. Evaluation

4.1 Quantitative Evaluation

Für die quantitative Evaluation des Verfahrens wurde eine Fünffach-Kreuzvalidierung durchgeführt. Dabei werden die manuell klassifizierten Daten fünfmal in fünf gleichgroße zufällige Mengen von Treffer-Snippets geteilt, wobei jeweils auf vier Mengen der Classifier gelernt, d.h. der optimale Vektor *w* und der Bias *b* bestimmt und auf der fünften Menge getestet wird. Als Gütekriterium für den Classifier dient der F_1-Score, das gewichtete harmonische Mittel aus Präzision (Precision) und Ausbeute (Recall), wobei Genauigkeit und Ausbeute gleich gewichtet werden; formal: F_1 = 2*(Präzision*Ausbeute)/(Präzision+Ausbeute). Der Durchschnitt der bei den Einzelläufen ermittelten Güte des Classifiers ergibt die jeweiligen F_1-Scores, die für die verwendeten Daten bislang zwischen 60 und 80% liegen, für die Wiko-A- und Wiko-D-Daten sogar tendenziell besser als für die Daten aus dem DWDS-KK. Grund dafür dürfte in erster Linie die wesentlich größere Menge an Trainingsdaten in den Wikipedia-Korpora sein, möglicherweise verbessern aber auch bestimmte textsortenspezifische Merkmale das Resultat. Die für die einzelnen Textsortenbereiche des DWDS-KK ermittelten F_1-Scores variieren ebenfalls textsortenspezifisch. Die folgenden Tabellen 5-7 zeigen die F_1-Scores für die einzelnen Textsortenbereiche des DWDS-KK in den drei Treatments:

bringen	**Belletristik**	**Gebrauchsliteratur**	**Wissenschaft**	**Zeitung**
BoW	69,7	67,6	72,6	70,3
BoW+Tags	67,1	67,2	76,8	71,7
Expert	65,4	66,9	63,5	65,5

Tabelle 5: F_1-Scores für die Güte der automatischen Klassifikation von Treffer-Snippets zu *bringen* in den einzelnen Textsortenbereichen des DWDS-KK und in den drei Treatments

kommen	**Belletristik**	**Gebrauchsliteratur**	**Wissenschaft**	**Zeitung**
BoW	72,4	65,4	74,9	67,5
BoW+Tags	74,5	71,1	76,9	67,7
Expert	71,6	67,5	69,4	63,7

Tabelle 6: F_1-Scores für die Güte der automatischen Klassifikation von Treffer-Snippets zu *kommen* in den einzelnen Textsortenbereichen des DWDS-KK und in den drei Treatments

finden	**Belletristik**	**Gebrauchsliteratur**	**Wissenschaft**	**Zeitung**
BoW	-[7]	71,7	68,3	67,6
BoW+Tags	-[4]	58,5	68,8	73,6
Expert	-[4]	69,6	68,2	67,6

Tabelle 7: F_1-Scores für die Güte der automatischen Klassifikation von Treffer-Snippets zu *finden* in den einzelnen Textsortenbereichen des DWDS-KK und in den drei Treatments

Überraschend ist außerdem die Überlegenheit bereits des BoW-Ansatzes gegenüber dem Lernen auf Basis festgelegter linguistischer Merkmale (‚Expert'). Offensichtlich liefert die im

[7] *finden* wird in den untersuchten Daten aus dem Textsortenbereich Belletristik des DWDS-KK sehr selten als Stützverb gebraucht. Sehr häufig sind hingegen Verwendungen wie *etw. schön* oder *schlecht finden*. Für diesen Datensatz konnte der F_1-Score nicht ermittelt werden, weil keines der Snippets als SVG-Beleg klassifiziert wurde (Recall 0%).

Kontext von SVG auftretende Wortwahl entscheidende Hinweise für die Klassifizierung. Möglicherweise lassen sich die erzielten Ergebnisse aber verbessern, wenn BoW-Ansatz und linguistische Merkmale kombiniert werden. Dies wird gegenwärtig erprobt.

Zusammen mit der Vorhersage, ob ein gegebenes Treffer-Snippet ein Stützverb enthält, wird für jedes Snippet auch ein Konfidenzwert *p* für die statistische Sicherheit der Vorhersage geliefert. Dieser Wert gibt die Irrtumswahrscheinlichkeit für die Vorhersage an: je geringer der Betrag der Konfidenz, desto unsicherer die Klassifikation im jeweiligen Fall (Berechnung und Details s. Platt 1999, Rüping 2006). Für die getesteten Daten liegen die Konfidenzwerte zwischen -4 und +4. Dieser Konfidenzwert kann für verschiedene weitergehende Analyseschritte genutzt werden:

1. Vielversprechend scheint nach ersten Experimenten beispielsweise die Berücksichtigung des Konfidenzwertes bei der Ausweitung der manuell klassifizierten Trainingsdaten. Konkret könnten dazu aktiv solche Treffer-Snippets ausgewählt werden, die nur mit geringer Sicherheit als Stützverb bzw. Rest klassifiziert wurden. Auf diese Weise ließe sich sicherstellen, dass v.a. solche Snippets manuell klassifiziert werden, die die Güte des Classifiers möglichst wirksam steigern können. Aktuell werden Datensätze mit geringer Konfidenz aus dem DWDS-KK aktiv manuell nachanalysiert, um die Trainingsdatenmenge zu verbessern.

2. Konfidenzwerte lassen sich außerdem für anschließende qualitative Analysen oder Visualisierungen nutzen. Korpusnutzer könnten beispielsweise für bestimmte Fragestellungen nur die Menge der am sichersten klassifizierten Snippets berücksichtigen, die in den meisten Fällen einen erheblichen Anteil der Gesamttreffermenge ausmacht, der manuell mit vergleichbarem zeitlichem Aufwand nicht erreichbar wäre (s. 4.2). In Visualisierungen könnten die Konfidenzwerte genutzt werden, um potenziell unsichere Treffer (die ‚Grauzone') zu markieren.

4.2 Qualitative Evaluation

Die unter 4.1 erläuterten, durch das Klassifikationsverfahren für jedes Treffer-Snippet ausgegebenen Konfidenzwerte wurden für weitergehende qualitative Analysen genutzt. Konkret sollte überprüft werden, inwiefern sich die Konfidenzwerte als Maß für die Zuverlässigkeit der automatischen Klassifikation mit der Einordnung der Snippets durch Experten decken.

Dazu wurde eine Zufallsstichprobe von automatisch klassifizierten Snippets der Wikipedia-Korpora Wiko-A und Wiko-D aus folgenden acht Wertebereichen gezogen und manuell überprüft:

1. $p >= -4$ und < -3,
2. $p >= -3$ und < -2,
3. $p >= -2$ und < -1,
4. $p >= -1$ und < 0,
5. $p >= 0$ und < 1,
6. $p >= 1$ und < 2,
7. $p >= 2$ und < 3,
8. $p >= 3$ und < 4.

Für jeden Wertebereich wurden 250 bzw. – für den Fall, dass die Trefferzahl unter 250 liegt – die Gesamtmenge der Treffer manuell gesichtet.

Die Ergebnisse bestätigen die Gültigkeit der durch das automatische Verfahren berechneten Konfidenz auch im Hinblick auf die intellektuelle Beurteilung der Treffer durch linguistische Experten. Die folgenden Tabellen 8-13 zeigen durchgängig, dass in den Bereichen $p >= 0$, in denen die durch das Verfahren automatisch als Stützverb-Belege klassifizierten Treffer liegen, auch aus Sicht der Experten der Anteil der Belege für Stützverben an der gesichteten Treffermenge überwiegt. Für die Bereiche $p < 0$ gilt entsprechend der umgekehrte Befund, in diesen Bereichen überwiegen bei automatischem Verfahren wie Experten die Treffer ohne Stützverben. In den Bereichen hoher Konfidenz beträgt die Übereinstimmung sogar annähernd 100%.

Interessant ist der Umgang des automatischen Verfahrens mit unvollständigen Snippets bzw. falsch positiven Treffern. Falsch Positive werden zutreffend überwiegend der Restgruppe zugeordnet. Unvollständige Snippets sind differenzierter zu betrachten: Sie können durchaus Belege für Stützverb-Vorkommen darstellen, wurden bei der manuellen Klassifikation aber wegen des fehlenden Kontexts und dadurch bedingten hohen subjektiven Beurteilungsniveaus grundsätzlich nicht gezählt. Das automatische Verfahren kann diese Snippets auf Basis des gelernten Classifier-Modells jedoch teils mit hoher Konfidenz klassifizieren (s. z.B. Tabelle 10).

bringen (Wiko-A)

		Automatisches Verfahren (Konfidenz für Vorhersage):							
		Stützverb				**kein Stützverb**			
		4 > *p* >= 3	3 > *p* >= 2	2 > *p* >= 1	1 > *p* >= 0	0 > *p* >= -1	-1 > *p* >= -2	-2 > *p* >= -3	-3 > *p* >= -4
Experten	**Stützverb**	0	44	241	172	46	7	0	0
	kein Stützverb	0	0	8	73	172	154	116	1
	Snippet unvollst.	0	0	0	2	3	8	3	0
	falsch positiv	0	0	1	3	29	81	131	6
	gesichtet gesamt	0	44	250	250	250	250	250	7
	Treffer gesamt	0	44	3.131	27.379	61.257	29.855	1.809	7

Tabelle 8: Vorkommen von *bringen* als Stützverb (manuell erhoben) in durch das automatische Verfahren ermittelten Konfidenzbereichen; Daten: Treffer-Snippets Wiko-A

bringen (Wiko-D)

		Automatisches Verfahren (Konfidenz für Vorhersage):							
		Stützverb				**kein Stützverb**			
		4 > *p* >= 3	3 > *p* >= 2	2 > *p* >= 1	1 > *p* >= 0	0 > *p* >= -1	-1 > *p* >= -2	-2 > *p* >= -3	-3 > *p* >= -4
Experten	**Stützverb**	0	42	224	150	43	11	2	0
	kein Stützverb	0	0	25	92	176	179	141	2
	Snippet unvollst.	0	0	0	2	2	2	1	0
	falsch positiv	0	0	1	6	29	58	106	1
	gesichtet gesamt	0	42	250	250	250	250	250	3
	Treffer gesamt	0	42	2.790	23.080	34.799	7.946	283	3

Tabelle 9: Vorkommen von *bringen* als Stützverb (manuell erhoben) in durch das automatische Verfahren ermittelten Konfidenzbereichen; Daten: Treffer-Snippets Wiko-D

finden (Wiko-A)

		Automatisches Verfahren (Konfidenz für Vorhersage):							
		Stützverb				**kein Stützverb**			
		4 > *p* >= 3	3 > *p* >= 2	2 > *p* >= 1	1 > *p* >= 0	0 > *p* >= -1	-1 > *p* >= -2	-2 > *p* >= -3	-3 > *p* >= -4
Experten	**Stützverb**	1	208	238	207	38	1	0	0
	kein Stützverb	0	0	2	35	152	152	78	22
	Snippet unvollst.	4	42	10	5	22	26	45	40
	falsch positiv	0	0	0	3	38	71	127	188
	gesichtet gesamt	5	250	250	250	250	250	250	250
	Treffer gesamt	5	303	6.169	27.243	80.853	158.448	54.466	2.427

Tabelle 10: Vorkommen von *finden* als Stützverb (manuell erhoben) in durch das automatische Verfahren ermittelten Konfidenzbereichen; Daten: Treffer-Snippets Wiko-A

finden (Wiko-D)

		Automatisches Verfahren (Konfidenz für Vorhersage):							
		Stützverb				**kein Stützverb**			
		4 > *p* >= 3	3 > *p* >= 2	2 > *p* >= 1	1 > *p* >= 0	0 > *p* >= -1	-1 > *p* >= -2	-2 > *p* >= -3	-3 > *p* >= -4
Experten	**Stützverb**	0	0	234	218	31	1	0	0
	kein Stützverb	0	0	0	30	203	234	244	246
	Snippet unvollst.	0	0	16	2	2	3	1	3
	falsch positiv	0	0	0	0	14	12	5	1
	gesichtet gesamt	0	0	250	250	250	250	250	250
	Treffer gesamt	0	0	301	4.772	63.933	242.345	63.833	1.329

Tabelle 11: Vorkommen von *finden* als Stützverb (manuell erhoben) in durch das automatische Verfahren ermittelten Konfidenzbereichen; Daten: Treffer-Snippets Wiko-D

kommen (Wiko-A)

		Automatisches Verfahren (Konfidenz für Vorhersage):							
		Stützverb				**kein Stützverb**			
		4 > *p* >= 3	3 > *p* >= 2	2 > *p* >= 1	1 > *p* >= 0	0 > *p* >= -1	-1 > *p* >= -2	-2 > *p* >= -3	-3 > *p* >= -4
Experten	**Stützverb**	129	250	244	192	36	5	0	0
	kein Stützverb	0	0	5	51	179	189	139	90
	Snippet unvollst.	1	0	1	0	9	22	20	93
	falsch positiv	0	0	0	6	25	34	91	67
	gesichtet gesamt	130	250	250	250	250	250	250	250
	Treffer gesamt	130	2.886	16.408	38.705	129.977	202.041	37.377	1.372

Tabelle 12: Vorkommen von *kommen* als Stützverb (manuell erhoben) in durch das automatische Verfahren ermittelten Konfidenzbereichen; Daten: Treffer-Snippets Wiko-A

kommen (Wiko-D)

		Automatisches Verfahren (Konfidenz für Vorhersage):							
		Stützverb				**kein Stützverb**			
		4 > *p* >= 3	3 > *p* >= 2	2 > *p* >= 1	1 > *p* >= 0	0 > *p* >= -1	-1 > *p* >= -2	-2 > *p* >= -3	-3 > *p* >= -4
Experten	**Stützverb**	0	3	242	191	33	6	3	0
	kein Stützverb	0	0	3	51	189	170	135	20
	Snippet unvollst.	0	2	5	2	2	3	3	0
	falsch positiv	0	0	0	6	26	71	109	214
	gesichtet gesamt	0	5	250	250	250	250	250	234
	Treffer gesamt	0	5	729	9.376	71.296	131.017	17.717	234

Tabelle 13: Vorkommen von *kommen* als Stützverb (manuell erhoben) in durch das automatische Verfahren ermittelten Konfidenzbereichen; Daten: Treffer-Snippets Wiko-D

5. Fazit und Anschlussarbeiten

Bereits die bislang getesteten Verfahren ermöglichen eine Analyse der Gesamttrefferlisten für beliebige Stützverben, durch die mit akzeptabler Genauigkeit bzw. zumindest mit bekannter ‚Grauzone' Aussagen über den Anteil der Stützverbverwendungen gemacht werden können. Die in Storrer (2013) noch exemplarisch an vier Verben und relativ kleinen Stichproben durchgeführten Studien zur Frequenzentwicklung und zur Textsortenspezifik der Gefüge können dadurch mit wesentlich geringerem manuellen Zeitaufwand mit weiteren Stützverben auf einer sehr viel breiteren Datengrundlage untersucht werden. Ein weiterer konkreter Einsatzbereich ist die lexikographische Erfassung der Stützverben in Internet-Wörterbüchern. Im Rahmen des Dissertationsprojekts von Nadja Radtke wird ein Wiki-Wörterbuch für DaF-Lerner aufgebaut werden, in dem die Verfahren genutzt werden, um die beschriebenen Verben nach

ihren Vorkommensfrequenzen zu ordnen und den Nutzern Hinweise auf die Textsortenspezifik der verschiedenen Verben und ihrer Gefüge zu geben.

In Bezug auf die Frage der Anwendbarkeit von Data-Mining-Verfahren – genauer: Klassifikationsverfahren auf Basis von Support-Vektor-Maschinen – auf Treffer-Snippets aus Korpussuchen können die Experimente ersten Aufschluss über die folgenden Teilfragestellungen gegeben:

- Wie viele Daten müssen sinnvollerweise manuell annotiert werden?
- Welche Zusatzinformationen sollten verwendet werden (PoS-Tags, Distributionsregeln etc.)?
- Sollte auf der kompletten Datenmenge oder auf Teilmengen (z.B. für unterschiedliche Textsorten) gelernt werden?
- Sind die an Daten zu ausgewählten Stützverben gelernten Verfahren auf andere Stützverben übertragbar? Welche weiteren Anpassungen/manuell annotierten Daten sind ggf. notwendig?

Aufbauend auf den in diesem Report dargestellten Erkenntnissen werden einzelne Fragestellungen in weiteren Experimenten vertieft. In den weiterführenden Arbeiten soll insbesondere erprobt werden, durch welche weiteren Merkmale und ggf. Merkmalskombinationen (z.B. N-Gramme, vollständige oder teilweise syntaktische Annotation der Treffer, Berücksichtigung weiterer manuell annotierter Merkmale wie typische prädikative Nomina/Suffixe etc., Textsorten-Metadaten) die Verfahren in ihrer Güte noch verbessert werden können.

Ergänzend wurde vom Tübinger Projektpartner ein Verfahren zur automatischen Erkennung von Präfixverben entwickelt, das als Filter vor das eigentliche Klassifikationsverfahren geschaltet werden kann, um einen sehr häufig vorkommenden Typ von falsch positiven Treffer-Snippets mit sehr guter Genauigkeit zu erkennen und vorab aus der Treffermenge auszufiltern. Es wird geprüft werden, wie sich der Filter auf die Güte der Klassifikationsverfahren auswirkt. Unabhängig von dem Nutzwert für die hier diskutierten Verfahren ist ein Werkzeug zur Erkennung von Präfixverben generell für korpusgestützte Untersuchungen zu deutschen Verben ein wichtiges Desiderat.

6. Zitierte Literatur

Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (Hg.): Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies. London u.a.: Continuum, 23–41.

Heid, Ulrich (2004): Spécifités morpho-syntaxiques des constructions à verbe support en allemand: Analyse de corpus. Linguisticae Investigationes 27(2), 309–325.

Heid, Ulrich/Fritzinger, Fabienne/Hauptmann, Susanne/Weidenkaff, Julia/Weller, Marion (2008): Providing corpus data for a dictionary for German juridical phraseology. In: Storrer, Angelika et al. (Hg.): Text Resources and Lexical Knowledge. Berlin u.a.: Mouton de Gruyter, 131–144.

Hinrichs, Erhard/Kübler, Sandra/Naumann, Karin/Heike Telljohann/Trushkina, Julia (2004): Recent Developments of Linguistic Annotations of the TüBa-D/Z Treebank. In: Proceedings of the Third Workshop on Treebanks and Linguistic Theories, Tübingen.

Joachims, Thorsten (1998): Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning, Berlin, Heidelberg: Springer.

Kamber, Alain (2008): Funktionsverbgefüge – empirisch. Eine korpusbasierte Untersuchung zu den nominalen Prädikaten des Deutschen. Tübingen: Max Niemeyer.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (Hg.): Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta: European Language Resources Association (ELRA), 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf

Klein, Dan & Manning, Christopher D. (2003): Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, 423–430.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich et al. (Hg.): Lexikographica. Berlin u.a.: de Gruyter, 79–93.

Kupietz, Marc & Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto & Kawaguchi, Yuji (Hg.): Working Papers in Corpus-based Linguistics and Language Education, No. 3. Tokyo: Tokyo University of Foreign Studies (TUFS), 53–59. http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf

Langer, Stefan (2005): A Formal Specification of Support Verb Constructions. In: Langer, Stefan & Schnorbusch, Daniel (Hg.): Semantik im Lexikon. Tübingen: Narr, 179–202.

Manning, Christopher D./Raghavan, Prabhakar/Schütze, Hinrich (2008): Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Mierswa, Ingo/Wurst, Michael/Klinkenberg, Ralf/Scholz, Martin/Euler, Timm (2006): YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), Philadelphia, USA.

Morton, Thomas/Kottmann, Joern/Baldridge, Jason/Bierner, Gann (2005): Opennlp: A java-based nlp toolkit. http://opennlp.sourceforge.net, 2005.

Nello Cristianini & John Shawe-Taylor (2004): Kernel Methods for Pattern Analysis. Cambridge: Cambridge University Press.

Platt, John (1999): Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, Alexander et al. (Hg.): Advances in Large Margin Classifiers. Cambridge: MIT Press.

Pottelberge, Jeroen van (2001): Verbonominale Konstruktionen, Funktionsverbgefüge. Vom Sinn und Unsinn eines Untersuchungsgegenstandes. Heidelberg: Winter.

Rüping, Stefan (2006): Robust Probabilistic Calibration. In: Proceedings of the European Conference on Machine Learning (ECML), Berlin, Heidelberg: Springer, 743–750.

Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht, Universitäten Stuttgart und Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

Sprachbericht 2013: Deutsche Akademie für Sprache und Dichtung & Union der deutschen Akademien der Wissenschaften (Hg.): Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache. Berlin/New York: de Gruyter.

Storrer, Angelika (2006): Funktionen von Nominalisierungsverbgefügen im Text. Eine korpusbasierte Fallstudie. In: Prost, Kristel & Winkler, Edeltraud (Hg.): Von der Intentionalität zur Bedeutung konventionalisierter Zeichen. Festschrift für Gisela Harras zum 65. Geburtstag. Tübingen: Narr, 147–178.

Storrer, Angelika (2007): Corpus-based Investigations on German Support Verb Constructions. In: Fellbaum, Christiane (Hg.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum Press.

Storrer, Angelika (2013): Variation im deutschen Wortschatz am Beispiel der Streckverbgefüge. In: Deutsche Akademie für Sprache und Dichtung & Union der deutschen Akademien der Wissenschaften (Hg.): Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache. Berlin/New York: de Gruyter. 171–209.

Telljohann, Heike/Hinrichs, Erhard/Kübler, Sandra/Zinsmeister, Heike/Beck, Kathrin (2012): Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technischer Bericht, Seminar für Sprachwissenschaft, Universität Tübingen.

Zesch, Torsten/Gurevych, Iryna/Mühlhäuser, Max (2007): Analysing and accessing Wikipedia as a lexical und semantic resource. In: Georg Rehm et al.: Data structures for Linguistic Resources and Applications. Tübingen, 197–205.