Brünner Beiträge zur **Germanistik und Nordistik** 32 / 2018 / **1**
https://doi.org/10.5817/BBGN2018-1-7

Möglichkeiten der korpusbasierten sprachwissenschaftlichen Analyse. Am Beispiel der Extraktion von Kollokationen im Intercorp und Sketch Engine

Possibilities of Corpus-based Linguistic Analysis. Example of Extraction of Collocations in InterCorp and Sketch Engine

Markéta Valíčková

Abstract

Collocations are for translators a challenging aspect of a language, since their translations in the target language have to fulfill the same function as in the source language. The paper depicts the methodology of research into Czech collocations and their German equivalents in the corpus manager InterCorp, and discusses the potential of corpus-linguistic instruments as useful assistants in the translation process.

Keywords

collocations, InterCorp, Sketch Engine, extraction, Ad Translationem

Dieser Beitrag diskutiert die praktische Bedeutung der Korpuslinguistik und ihrer Methoden für die Sprachwissenschaft. Das Potential der korpus-basierten Übersetzungswissenschaft soll gerade am Beispiel der Extraktion von Kollokationen vorgestellt werden, die für die Nicht-MuttersprachlerInnen eine der größten fremdsprachlichen Herausforderungen darstellen.

Der authentische Sprachgebrauch lässt sich grundsätzlich anhand vier Quellen untersuchen. Mithilfe von verschiedenen Korpusmanagern wird das Sprachmaterial empirisch durchsucht. Die erworbenen Ergebnisse können für verschiedene Zwecke eingesetzt werden, wie z.B. im Fremdsprachenunterricht, in der Übersetzungswissenschaft, bei der Erstellung von Wörterbüchern und Grammatiken oder in der sprachwissenschaftlichen Forschung. Vor diesem Hintergrund lässt sich also feststellen, dass das Korpus als „Sammlung von Texten oder Textteilen [bezeichnet wird], die bewusst nach bestimmten sprachwissenschaftlichen Kriterien ausgewählt und geordnet werden." (SCHERER 2006: 3). Für jede sprachwissenschaftliche Analyse stehen zwei Verfahren zur Verfügung. Bei der korpus-basierten Methode steht am Anfang der Untersuchung eine Theorie, die durch die Korpusdaten erklärt, veranschaulicht und überprüft wird (BUBENHOFER 2009: 100). Während die korpus-basierte Methode eine deduktive Vorgehensweise benutzt, und so die Theorie durch die Korpusdaten zu erklären bzw. zu widerlegen oder zu bestätigen versucht, stehen die Korpusdaten in der korpus-geleiteten Methode am Anfang der Untersuchung. (Vgl. BUBENHOFER 2009: 321)

Sind neue Kollokationen automatisch auffindbar? Mithilfe von korpuslinguistischen Instrumenten werden die Wortkombinationen aufgrund ihrer Frequenz berechnet und extrahiert. „Es ist nicht überraschend, wenn zwei Wörter in einem Satz im gesamten Korpus auftreten. Bemerkenswert ist jedoch ihre Kombination, die frequenter vorkommt als die zufällige Verteilung der Wörter im Korpus wäre." (BUBENHOFER 2009: S. 2) Solche statistischen Auswertungen der größeren Menge von Ergebnissen führen dann nicht nur zu den Informationen über den Sprachgebrauch, sondern auch zur Veranschaulichung der semantischen Verbindungen (z.B. Tisch, Stühle, Betten gehören zum selben semantischen Feld – Möbel) und der sprachlichen Konventionen (z.B. Auf einen Tisch legt man Geld oder Karten). (BUBENHOFER 2010: 1–2)

Kollokationen als fixierte polylexikalische Einheiten stellen für die NichtmuttersprachlerInnen formale und inhaltliche Probleme dar, da sie meist als komplexe Einheiten nicht erkannt werden. (BUBENHOFER 2010: 197) Viele Forscher einigen sich nicht auf einen eindeutigen Begriff, grundsätzlich wird aber „die Tendenz zur Ausweitung des Phraseologie-Begriffes sowie das Bedürfnis für die Begriffsbestimmung von Mehrworteinheiten neue, nicht traditionelle Kriterien zu wählen" (BUBENHOFER/PTASHNYK 2010: 12) beobachtet. Colson (2003: 45) fordert „eine statistisch operationalisierbare Definition von Kollokationen", um möglichst objektive Belege zu erhalten. Als Nachteil bezeichnet er die semantisch oder kognitiv begründbaren Kriterien (z.B. Idiomatizität). Die Kollokationen lassen sich nach zwei Ansätzen unterscheiden. Der lexikographische Ansatz (F. J. Hausmann) besteht in „komplexen Versprachlichungen zur Benennung des Sachverhalts." (KRATOCHVÍLOVÁ 2011: 74–81) Der statistisch orientierte Ansatz in der Kollokationsforschung betrachtet die Kollokationen als „statistisch relevantes Mitvor-

kommen der lexikalischen Einheiten bzw. als Wortassoziationen" (KRATOCHVÍLOVÁ 2011: 74–81). Anhand des mathematischen Modellierens der assoziativen Cluster wird das potentielle Miteinandervorkommen der Lexeme (Kookkurrenzen) berechnet. Dabei sollen aber verschiedene Bezugsgrößen, wie z.B. Korpusgröße, Korrelationen etc. in Betracht gezogen werden.

Bei der Extraktion von Kollokationen stehen mehrere methodologische Wege zur Verfügung. Nach der Durchführung bleibt aber die Frage offen, ob es sich wirklich um Kollokationen bzw. um Phraseologismen handelt und ob sie als lexikalisiert bezeichnet werden können. Deshalb werden die so extrahierten lexikalischen Einheiten als Kandidaten an Kollokationen bezeichnet. Für das Auffinden können verschiedene Kriterien festgesetzt werden, wie z.B. geringe Variabilität des Phraseologismus u.a. Somit ist die automatisierte Suche nach unbekannten Phraseologismen erst in der Zukunft durchführbar. (Vgl. BUBENHOFER 2010: 51)

Wie kann eine Liste von Kandidaten an Kollokationen vom InterCorp und Sketch Engine erstellt werden? Beim ČNK bzw. beim tschechischen Korpus der Reihe SYN kommen drei Möglichkeiten in Frage. Beim InterCorp ist es gewissermaßen eingeschränkt und die genauen Vorgehensweisen müssen noch klar und eindeutig definiert werden, aber man kann grundsätzlich zwei Möglichkeiten anwenden. Im Parallelkorpus InterCorp sind 232 Millionen Wörter enthalten (Version 9). Sie ist über eine spezielle Schnittstelle Park zugänglich und jede einsprachige Version aller Korpusparallelen fungiert als ein vollwertiges einsprachiges Korpus mit allen gängigen Instrumenten. Im Unterschied zu Referenzkorpora zeichnet sich das InterCorp durch einen stetigen Zuwachs im Umfang der Paralleltexte und auch in der Anzahl der Sprachen aus. InterCorp kann aber nicht als Referenzkorpus verwendet werden, weil die Texte stilistisch, regional und originalsprachlich nicht ausgewogen sind. InterCorp enthält belletristische und journalistische Texte, des Weiteren auch Aufzeichnungen aus dem Europäischen Parlament sowie Untertitel in 39 Sprachen. Neben den deutschen und anderen europäischen Sprachen sind hier auch das Malaiische, Hebräische, Romani, Vietnamesische oder Japanische zu finden.[1]

Mithilfe der Funktion MEET können die Konkordanzen einer bestimmten Kollokation gefunden werden, hingegen wird aber nicht ermittelt, wie sich eine Kollokation mit einem bestimmten Schlüsselwort verbindet. Zur Veranschaulichung wurde die Suchabfrage (meet [lemma="prát"][lemma="prádlo"] -3 3) verwendet.

1 Český národní korpus (2018): InterCorp In: Ústav Českého národního korpusu FF UK, Praha. https://ucnk.ff.cuni.cz/intercorp/?req=page:info (2. 4. 2018).

Abbildung 1
Funktion MEET

Die Funktion UNION dient zur Vereinigung der Abfragen des Typs MEET (union (meet...) (meet...)). In den Ergebnissen werden die Konkordanzen mit dem Lemma „prát“ im Kontext mit dem Lemma „prádlo“ und auch die Konkordanzen mit dem Lemma „mýt“ in Verbindung mit „nádobí“ gezeigt. Anhand dieser Funktion können die Frequenzen der Kollokationen und auch ihrer Variabilitätsmaße veranschaulicht werden. Die Suchabfrage lautet (union (meet [lemma=“prát“][lemma=“prádlo“] -3 3) (meet [lemma=“mýt“][lemma=“nádobí“] -3 3)).

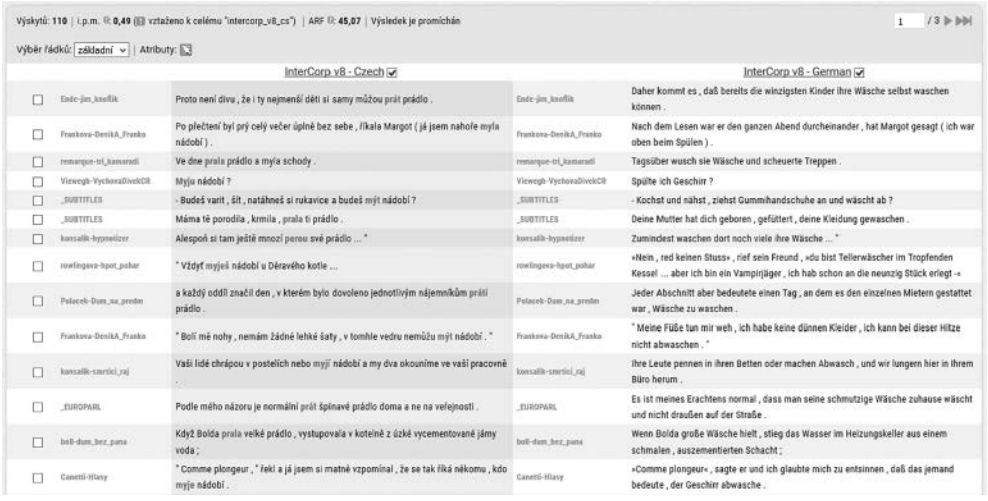


Abbildung 2
Funktion UNION

Wichtig sind die Positionen rechts und links von KWIC, die für potentielle Kollokationen gehalten werden. Empfohlen werden Positionen mit dem Intervall -3 und +3 von KWIC. Dank der Kollokationsprofile kann man z.B. zwischen den verschiedenen Bedeutungen von Polysemen unterscheiden.

Kollokationen konnten früher nur mithilfe der Introspektion identifiziert werden. Heute werden dazu die statistischen Assoziationsmaße benutzt, die in meisten Fällen die Frequenz der Kollokatoren und die Frequenz der ganzen Wendung in Beziehung setzen, und ev. auch die Größe des Korpus. Zu den beliebtesten gehören MI-score, t-score, log-likelihood und logDice. Keine von ihnen kann aber als universal für das Suchen von allen Kollokationen bezeichnet werden. Jede automatische Suche von Kollokationen mithilfe von Assoziationsmaßen fordert noch eine linguistische Interpretation.

Sketch Engine gehört zu jenen Methoden, die die reinen Signifikanzmaße mit dem Wissen über syntaktische Strukturen oder Semantik kombinieren. Dieser Korpusmanager ist mit den von der Masaryk-Universität gestellten Zugangsdaten frei verfügbar.

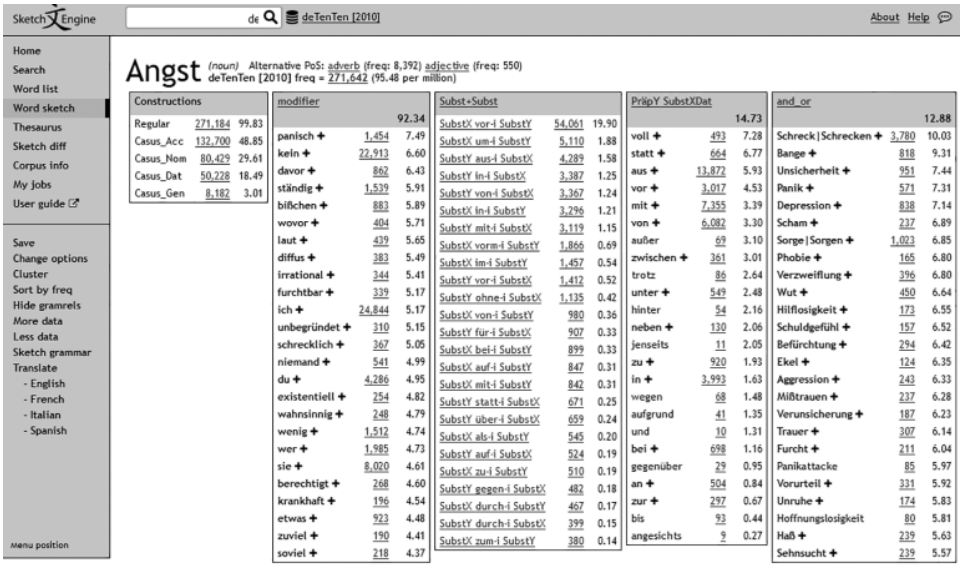
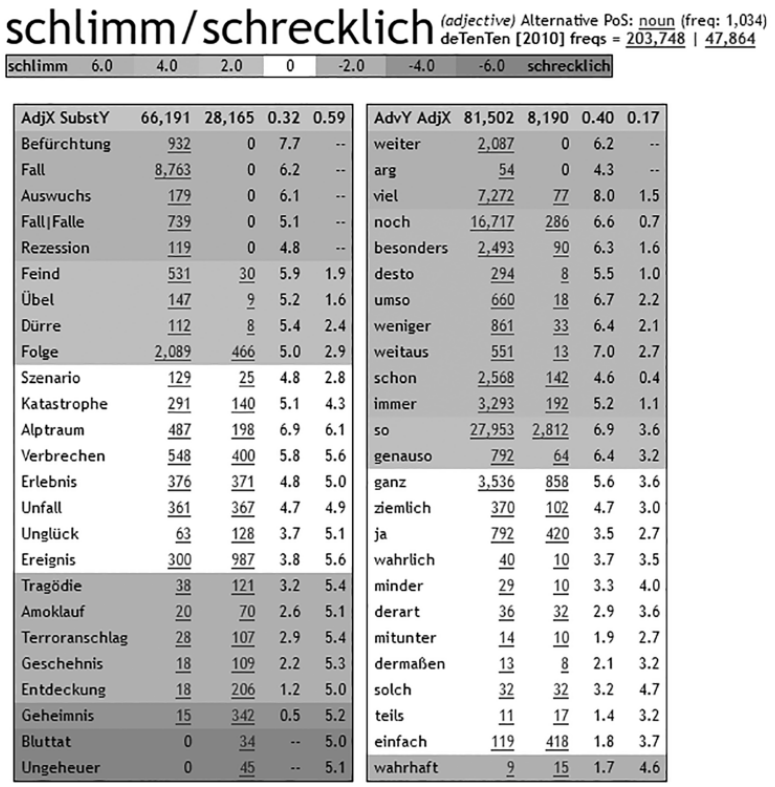


Abbildung 3
Kollokationen im Sketch Engine



schlimm/schrecklich *(adjective)* Alternative PoS: noun (freq: 1,034)
deTenTen [2010] freqs = 203,748 | 47,864

schlimm	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	schrecklich

AdjX SubstY	66,191	28,165	0.32	0.59
Befürchtung	932	0	7.7	--
Fall	8,763	0	6.2	--
Auswuchs	179	0	6.1	--
Fall|Falle	739	0	5.1	--
Rezession	119	0	4.8	--
Feind	531	30	5.9	1.9
Übel	147	9	5.2	1.6
Dürre	112	8	5.4	2.4
Folge	2,089	466	5.0	2.9
Szenario	129	25	4.8	2.8
Katastrophe	291	140	5.1	4.3
Alptraum	487	198	6.9	6.1
Verbrechen	548	400	5.8	5.6
Erlebnis	376	371	4.8	5.0
Unfall	361	367	4.7	4.9
Unglück	63	128	3.7	5.1
Ereignis	300	987	3.8	5.6
Tragödie	38	121	3.2	5.4
Amoklauf	20	70	2.6	5.1
Terroranschlag	28	107	2.9	5.4
Geschehnis	18	109	2.2	5.3
Entdeckung	18	206	1.2	5.0
Geheimnis	15	342	0.5	5.2
Bluttat	0	34	--	5.0
Ungeheuer	0	45	--	5.1

AdvY AdjX	81,502	8,190	0.40	0.17
weiter	2,087	0	6.2	--
arg	54	0	4.3	--
viel	7,272	77	8.0	1.5
noch	16,717	286	6.6	0.7
besonders	2,493	90	6.3	1.6
desto	294	8	5.5	1.0
umso	660	18	6.7	2.2
weniger	861	33	6.4	2.1
weitaus	551	13	7.0	2.7
schon	2,568	142	4.6	0.4
immer	3,293	192	5.2	1.1
so	27,953	2,812	6.9	3.6
genauso	792	64	6.4	3.2
ganz	3,536	858	5.6	3.6
ziemlich	370	102	4.7	3.0
ja	792	420	3.5	2.7
wahrlich	40	10	3.7	3.5
minder	29	10	3.3	4.0
derart	36	32	2.9	3.6
mitunter	14	10	1.9	2.7
dermaßen	13	8	2.1	3.2
solch	32	32	3.2	4.7
teils	11	17	1.4	3.2
einfach	119	418	1.8	3.7
wahrhaft	9	15	1.7	4.6

Abbildung 4
Kombinierbarkeit der Kollokationen

Die Lemmata sind mit roter und grüner Farbe gekennzeichnet. Die roten Konkordanzen können mit den roten Lemmata kombiniert werden, die grünen Konkordanzen haben dann die Tendenz zur Kombination mit dem grünen Lemma. Die weißen Konkordanzen können mit beiden Lemmata vorkommen. Volle Farbtöne stehen dann für stärkere Kollokationen.

Zusammenfassend lässt sich sagen, dass InterCorp und Sketch Engine zu jenen statistischen Methoden gehören, die Kookkurrenzen von Wortketten entweder als Nachbarschafts-, Satz- oder Fensterkombinationen darstellen. Dabei stellt sich die Frage, inwieweit es möglich ist, zuverlässige und lexikographisch verwertbare Informationen über das distributionelle Verhalten der Textwörter in Korpora durch die Kookkurrenzermittlungen zu gewinnen. Die grundlegenden Mängel der sprachstatistischen Verfahren liegen darin, dass verschiedene Maße immer andere, überlappende Teilmengen von Wortlisten mit unterschiedlichen Präferenzen liefern. Des Weiteren scheinen die generierten Wortlisten in dem Sinne problematisch zu sein, da nichts über den Typ der Beziehung zwischen Wortformen, die ein Kookkurrenzpaar bilden, gesagt wird. Die Interpretation der LinguistInnen ist somit unverzichtbar. Problematisch scheint auch die Ungenauigkeit der

morphologischen Markierung und der Lemmatisierung der Korpusdaten, die teilweise durch das „Verfeinern" des Suchalgorithmus (durch CQL) verringert werden kann – wenn das die These erlaubt. Der Vorteil besteht in der Tatsache, dass spezielle Abfragen und Stichproben möglich sind und dass der Kontext gezeigt wird. Grundsätzlich sollen dann die Repräsentativität und die Größe des Korpus im Vordergrund stehen und die Korrelation zwischen der Vorkommenshäufigkeit von Kollokationen und ihrer tatsächlichen Geläufigkeit bei den Sprechern einer Sprache in Betracht gezogen werden. In letzter Zeit tauchen neue kritische methodologische Ansätze auf. Dräger/Juska-Bacher oder Quasthoff/Schmidt/Hallsteindóttir sind der Ansicht, dass die Korpusstudien bzw. ihre Ergebnisse durch Online-Befragungen ergänzt werden sollen oder dass die fehlende Korrelation vor allem an der zu geringen Datengrundlage liegt (BUBENHOFER 2010: 14, 37). Für die Erstellung von Kollokationsprofilen und für die automatische Sortierung der Wortkombinationen können aber beide Werkzeuge als effektiv bezeichnet werden.

Quellen und Literatur

BUBENHOFER, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin, New York: de Gruyter. ISBN 978-3-11-021584-7.

BUBENHOFER, Noah, et al. (2010): Korpora, Web und Datenbanken: computergestützte Methoden in der modernen Phraseologie und Lexikographie Baltmannsweiler: Schneider Verlag Hohengehren.

BUBENHOFER, Noah/PTASHNYK, Stefaniya (2010): Korpora, Datenbanken und das Web: State of the Art computergestützter Forschung in der Phraseologie und Lexikographie. In: Korpora, Web und Datenbanken: computergestützte Methoden in der modernen Phraseologie und Lexikographie Baltmannsweiler: Schneider Verlag Hohengehren.

Český národní korpus (2018): InterCorpIn: Ústav Českého národního korpusu FF UK, Praha. https://ucnk.ff.cuni.cz/intercorp/?req=page:info (2. 4. 2018).

COLSON, J.-P. (2003): "Corpus Linguistics and Phraseological Statistics: a few Hypotheses and Examples". In: BURGER, H./GRÉCIANO, G./HÄCKI BUHOFER, A. (Hrsg.): Flut von Texten – Vielfalt der Kulturen, Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie. Baltmannsweiler: Schneider Verlag Hohengehren, 45–59.

KRATOCHVÍLOVÁ, Iva/ WOLF, Norbert Richard (2010): Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. Heidelberg: Winter.

KRATOCHVÍLOVÁ, Iva (2011): Kollokationen im Lexikon und im Text. Mehrwortverbindungen im Deutschen und Tschechischen. Studien und Quellen zur Sprachwissenschaft. Berlin: LIT Verlag.

SCHLOBINSKI, Peter (1996): Empirische Sprachwissenschaften. Wiesbaden: Springer.

Mgr. et Mgr. Markéta Valíčková / 399426@mail.muni.cz

Masarykova univerzita, Filozofická fakulta, Ústav germanistiky, nordistiky a nederlandistiky
Arna Nováka 1, 602 00 Brno, CZ