

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 unbekanntem 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).

InterCorp v8 - Czech		InterCorp v8 - German	
<input type="checkbox"/> konsalk-hypnotizer	Alespoř si tam ještě mnozí perou své prádlo ... *	<input type="checkbox"/> konsalk-hypnotizer	Zumindest waschen dort noch viele ihre Wäsche ... *
<input type="checkbox"/> Kanders-Heserist_letko	Tereza prala prádlo a vedle vany měla položenou knihu .	<input type="checkbox"/> Kanders-Heserist_letko	Teresa wusch die Wäsche, und neben der Wanne hatte sie ein Buch liegen .
<input type="checkbox"/> _SUBTITLES	- Jenže u nás se prádlo nepere !	<input type="checkbox"/> _SUBTITLES	- Dachte ich auch, ist aber nicht so !
<input type="checkbox"/> konsalk-byta_jich_des	Můj otec je železničář v Kazani, moje matka pracuje v prádelně, ve které se pere především špinavé prádlo z pracovních táborů .	<input type="checkbox"/> konsalk-byta_jich_des	Mein Vater ist Gleisarbeiter in Kasan, meine Mutter schuftet in einer Wäscherei, die vor allem das Dreckzeug aus den Arbeitslagern säubert .
<input type="checkbox"/> Adla-KlemparsVityare	co mě se všichni zbláznili , jakpak světlé šedý sako ? to se má už jeden ptal na světlé šedý sako , peru prádlo tedy za tím náspem , vše , je tam poko a tam já dycky peru prádlo , a jak tak peru - jen si to poslechněte , mladé pane , kampak byste kváloral, tedy jak tak peru , přišel mi podívat se takhle pod traf , tam co sou olivový kaše a koukám jak tam řákej člověk , takovej přelástej , hraje v těch kačích , něco hledá , za chvíli mi zase přišel podívat se pod traf a to byste nevítil , mladé pane , ten člověk zase hraje v těch kačích , ale v těch , co sou dáí , a zas tam něco hledá a pak mi zase přišel podívat se takhle pod traf - tak počkejte , mladé pane , af sám to dopovím - a ten člověk se tam zase hraje v kačích a něco hledá , ne že by má to zajímalo , já měla svůj prádlo , svou práci , ale jen dyt mi přišlo podívat se pod traf , tak sem tam toho člověka , vše , toho plekátého dycky všlela , jo , jo , byt zrovna takovej , jak říkáte , mladé pane , a pak vám nagejdou ten člověk -jo Marcelovi to bude trvat etě dlouho ,	<input type="checkbox"/> Adla-KlemparsVityare	said he doch alle verrückt , was für ein halbgraues Jackett ? da hat schon einer nach einem halbgrauen gefragt , ich wasche da hinter dem Damm , da fließt ein Bach , verstanden , dort wasche ich meine Wäsche , und wie ich da wasche - hören Sie nur ruhig zu , junger Mann , Ele mit Wiele , das sag ich immer , also wie ich da wasche , da fällt mir ein , dort unter den Bahndamm zu gucken , wo die Olivenbäume wachsen , und sieh da , ein Kerl mit einer Glotze , der willt in den Büschen , sucht was , nach einer Wiele kommt es mir wieder , unter den Bahndamm zu guk-ken , nicht zu glauben , junger Herr , der willt immer noch in den Büschen , der Kerl , aber ein Stückchen weiter , und wieder sucht er etwas , und dann ist es mir zumute , nochmals unter den Bahndamm zu gucken - warten Sie doch , junger Herr , ich bin noch nicht fertig , und der Kerl willt da wieder und sucht in den Büschen , nicht daß es mich interessiert , ich hab zu tun , genug zu tun , nur wenne es mir so kommt , unter den Bahndamm zu gucken , da seh ich dort immer auch den Kerl , den Glatzkopf , jaja , genauso sah er aus , wie Sie es sagen , junger Herr , und dann auf einmal , stellen Sie sich vor , der Mensch - jawohl , mit Marcello kann es noch eine Wiele dauern ,
<input type="checkbox"/> alenece-dorastesteny	Prala cizí prádlo , což byla stejné těžká práce , jako když se její manžel oháněl kládivem nad kovadlinou .	<input type="checkbox"/> alenece-dorastesteny	Sie wusch fremde Wäsche , eine ebenso mühsame Arbeit wie die ihres Mannes mit Hammer und Ambö .
<input type="checkbox"/> Woodoro-Porcarat_micla	Jo , a prádlo radši ani neperte , všechno se musí rovnou spátit , obléčení , příkravky , zkrátka všechno ... *	<input type="checkbox"/> Woodoro-Porcarat_micla	Und waschen Sie auf keinen Fall seine Leine . Es muss alles verbrannt werden , die Kleidung , die Bettwäsche , alles ... *

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)).

InterCorp v8 - Czech		InterCorp v8 - German	
<input type="checkbox"/> Ende-jin_kavalka	Proto není divu , že i ty negmanší děti si samy můžou práť prádlo .	<input type="checkbox"/> Ende-jin_kavalka	Daher kommt es , daß bereits die winzigsten Kinder ihre Wäsche selbst waschen können .
<input type="checkbox"/> Frankova-Denkla_Franko	Po přelčení byl přj celý večer úplně bez sebe , říkala Margot (já jsem nahole myla nádobí) .	<input type="checkbox"/> Frankova-Denkla_Franko	Nach dem Lesen war er den ganzen Abend durcheinander , hat Margot gesagt (ich war oben beim Spülen) .
<input type="checkbox"/> remaque-tri_kamaradi	Ve dne prala prádlo a myla schody .	<input type="checkbox"/> remaque-tri_kamaradi	Tagsüber wusch sie Wäsche und schaeuerte Treppen .
<input type="checkbox"/> Viewegh-VychodivickCB	Myje nádobí ?	<input type="checkbox"/> Viewegh-VychodivickCB	Spülte ich Geschir ?
<input type="checkbox"/> _SUBTITLES	- Budeš varit , šit , natáhneš si rukavice a budeš mýt nádobí ?	<input type="checkbox"/> _SUBTITLES	- Kochst und nähst , ziehst Gummihandschuhe an und wäscht ab ?
<input type="checkbox"/> _SUBTITLES	Máma tě porodila , kmlia , prala ti prádlo .	<input type="checkbox"/> _SUBTITLES	Deine Mutter hat dich geboren , gefüttert , deine Kleidung gewaschen .
<input type="checkbox"/> konsalk-hypnotizer	Alespoř si tam ještě mnozí perou své prádlo ... *	<input type="checkbox"/> konsalk-hypnotizer	Zumindest waschen dort noch viele ihre Wäsche ... *
<input type="checkbox"/> rowlingove-hpat_pohar	* Vždyť myje nádobí u Džavého kotle ...	<input type="checkbox"/> rowlingove-hpat_pohar	*Nein , red keinen Stuss , rief sein Freund , »du bist Tellerwäscher im Tropfenden Kessel ... aber ich bin ein Vampirjäger , ich hab schon an die neunzig Stück erliegt «
<input type="checkbox"/> Polaceb-Dum_na_predni	a každý oddíl značil den , v kterém bylo dovoleno jednotlivým nájemníkům práti prádlo .	<input type="checkbox"/> Polaceb-Dum_na_predni	Jeder Abschnitt aber bedeutete einen Tag , an dem es den einzelnen Mietern gestattet war , Wäsche zu waschen .
<input type="checkbox"/> Frankova-Denkla_Franko	* Boží mě nohy , nemám žádné lehké šaty , v tomhle vedru nemůžu mýt nádobí . *	<input type="checkbox"/> Frankova-Denkla_Franko	* Meine Füße tun mir weh , ich habe keine dünnen Kleider , ich kann bei dieser Hitze nicht abwaschen ... *
<input type="checkbox"/> konsalk-saricki_raj	Vaši lidé chrapou v postelích nebo myjí nádobí a my dva okouníme ve vaší pracovně .	<input type="checkbox"/> konsalk-saricki_raj	ihre Leute pennen in ihren Betten oder machen Abwasch , und wir lungern hier in ihrem Büro herum .
<input type="checkbox"/> _EUROPABL	Podle mého názoru je normální práť špinavé prádlo doma a ne na veřejnosti .	<input type="checkbox"/> _EUROPABL	Es ist meines Erachtens normal , dass man seine schmutzige Wäsche zuhause wäscht und nicht draußen auf der Straße .
<input type="checkbox"/> bol-dim_žerz_pama	Když Boida prala velké prádlo , vystupovala v kotelně z úzké vycentované jámy voda ;	<input type="checkbox"/> bol-dim_žerz_pama	Wenn Boida große Wäsche hielt , stieg das Wasser im Heizungskeller aus einem schmalen , auszentrierten Schacht ;
<input type="checkbox"/> Caeneti-Hlavy	* Comme plongeur , feki a já jsem si matně vzpomínal , že se tak říká někomu , kdo myje nádobí .	<input type="checkbox"/> Caeneti-Hlavy	*Comme plongeur , sagte er und ich glaubte mich zu entsinnen , daß das jemand bedeute , der Geschir abwasche .

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.

Angst (noun). Alternative PoS: *adverb* (freq: 8.392) *adjective* (freq: 550)
deTenTen [2010] freq = 271,642 (95.48 per million)

Constructions	modifier	Subst+Subst	PräpY SubstXDat	and_or
Regular 271,184 99.83	panisch + 1,454 7.49	SubstX vor-i SubstY 54,061 19.90	voll + 493 7.28	Schreck Schrecken + 3,780 10.03
Casus_Acc 132,700 48.85	kein + 22,913 6.60	SubstX um-i SubstY 5,110 1.88	statt + 664 6.77	Bange + 818 9.31
Casus_Nom 80,429 29.61	davor + 862 6.43	SubstY aus-i SubstX 4,289 1.58	aus + 13,872 5.93	Unsicherheit + 951 7.44
Casus_Dat 50,228 18.49	ständig + 1,539 5.91	SubstY in-i SubstX 3,387 1.25	vor + 3,017 4.53	Panik + 571 7.31
Casus_Gen 8,182 3.01	bilchen + 883 5.89	SubstX von-i SubstY 3,367 1.24	mit + 7,355 3.39	Depression + 838 7.14
	wovor + 404 5.71	SubstX in-i SubstY 3,276 1.21	von + 6,082 3.30	Scham + 237 6.89
	laut + 429 5.65	SubstY mit-i SubstX 3,119 1.15	außer 69 3.10	Sorge Sorgen + 1,021 6.85
	diffus + 383 5.49	SubstX vorm-i SubstY 1,866 0.69	zwischen + 361 3.01	Phobie + 165 6.80
	irrational + 344 5.41	SubstX im-i SubstY 1,457 0.54	trotz 86 2.64	Verzweiflung + 356 6.80
	furchtbar + 339 5.17	SubstY vor-i SubstX 1,412 0.52	unter + 549 2.48	Wut + 450 6.64
	ich + 24,844 5.17	SubstY ohne-i SubstX 1,135 0.42	hinter 54 2.16	Hilflosigkeit + 173 6.55
	unbegründet + 310 5.15	SubstX von-i SubstY 980 0.36	neben + 130 2.06	Schuldgefühl + 157 6.52
	schrecklich + 367 5.05	SubstY für-i SubstX 907 0.33	jenseits 11 2.05	Befürchtung + 294 6.42
	niemand + 541 4.99	SubstX bei-i SubstY 899 0.33	zu + 920 1.93	Ekel + 124 6.35
	du + 4,286 4.95	SubstX auf-i SubstY 847 0.31	in + 3,993 1.63	Aggression + 243 6.33
	existentiell + 254 4.82	SubstX mit-i SubstY 842 0.31	wegen 68 1.48	Mißtrauen + 237 6.28
	wahrscheinlich + 248 4.79	SubstY statt-i SubstX 671 0.25	aufgrund 41 1.35	Verunsicherung + 187 6.23
	wenig + 1,512 4.74	SubstY über-i SubstX 659 0.24	und 10 1.31	Trauer + 307 6.14
	wer + 1,785 4.73	SubstX als-i SubstY 545 0.20	bei + 698 1.16	Furcht + 211 6.04
	sie + 8,020 4.61	SubstY auf-i SubstX 524 0.19	gegenüber 29 0.95	Panikattacke 85 5.97
	berechtigt + 268 4.60	SubstX zu-i SubstY 510 0.19	an + 504 0.84	Vorurteil + 331 5.92
	krankhaft + 176 4.54	SubstY gegen-i SubstX 482 0.18	zur + 297 0.67	Unruhe + 174 5.83
	etwas + 923 4.48	SubstX durch-i SubstY 467 0.17	bis 93 0.44	Hoffnungslosigkeit 80 5.81
	zuviel + 190 4.41	SubstY durch-i SubstX 399 0.15	angesichts 9 0.27	Haß + 239 5.63
	soviel + 218 4.37	SubstX zum-i SubstY 380 0.14		Sehnsucht + 239 5.57

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	<u>932</u>	0	7.7	--					
Fall	<u>8,763</u>	0	6.2	--					
Auswuchs	<u>179</u>	0	6.1	--					
Fall Falle	<u>739</u>	0	5.1	--					
Rezession	<u>119</u>	0	4.8	--					
Feind	<u>531</u>	<u>30</u>	5.9	1.9					
Übel	<u>147</u>	<u>9</u>	5.2	1.6					
Dürre	<u>112</u>	<u>8</u>	5.4	2.4					
Folge	<u>2,089</u>	<u>466</u>	5.0	2.9					
Szenario	<u>129</u>	<u>25</u>	4.8	2.8					
Katastrophe	<u>291</u>	<u>140</u>	5.1	4.3					
Alptraum	<u>487</u>	<u>198</u>	6.9	6.1					
Verbrechen	<u>548</u>	<u>400</u>	5.8	5.6					
Erlebnis	<u>376</u>	<u>371</u>	4.8	5.0					
Unfall	<u>361</u>	<u>367</u>	4.7	4.9					
Unglück	<u>63</u>	<u>128</u>	3.7	5.1					
Ereignis	<u>300</u>	<u>987</u>	3.8	5.6					
Tragödie	<u>38</u>	<u>121</u>	3.2	5.4					
Amoklauf	<u>20</u>	<u>70</u>	2.6	5.1					
Terroranschlag	<u>28</u>	<u>107</u>	2.9	5.4					
Geschehnis	<u>18</u>	<u>109</u>	2.2	5.3					
Entdeckung	<u>18</u>	<u>206</u>	1.2	5.0					
Geheimnis	<u>15</u>	<u>342</u>	0.5	5.2					
Bluttat	0	<u>34</u>	--	5.0					
Ungeheuer	0	<u>45</u>	--	5.1					
AdvY AdjX	81,502	8,190	0.40	0.17					
weiter	<u>2,087</u>	0	6.2	--					
arg	<u>54</u>	0	4.3	--					
viel	<u>7,272</u>	<u>77</u>	8.0	1.5					
noch	<u>16,717</u>	<u>286</u>	6.6	0.7					
besonders	<u>2,493</u>	<u>90</u>	6.3	1.6					
desto	<u>294</u>	<u>8</u>	5.5	1.0					
umso	<u>660</u>	<u>18</u>	6.7	2.2					
weniger	<u>861</u>	<u>33</u>	6.4	2.1					
weitaus	<u>551</u>	<u>13</u>	7.0	2.7					
schon	<u>2,568</u>	<u>142</u>	4.6	0.4					
immer	<u>3,293</u>	<u>192</u>	5.2	1.1					
so	<u>27,953</u>	<u>2,812</u>	6.9	3.6					
genauso	<u>792</u>	<u>64</u>	6.4	3.2					
ganz	<u>3,536</u>	<u>858</u>	5.6	3.6					
ziemlich	<u>370</u>	<u>102</u>	4.7	3.0					
ja	<u>792</u>	<u>420</u>	3.5	2.7					
wahrlich	<u>40</u>	<u>10</u>	3.7	3.5					
minder	<u>29</u>	<u>10</u>	3.3	4.0					
derart	<u>36</u>	<u>32</u>	2.9	3.6					
mitunter	<u>14</u>	<u>10</u>	1.9	2.7					
dermaßen	<u>13</u>	<u>8</u>	2.1	3.2					
solch	<u>32</u>	<u>32</u>	3.2	4.7					
teils	<u>11</u>	<u>17</u>	1.4	3.2					
einfach	<u>119</u>	<u>418</u>	1.8	3.7					
wahrhaft	<u>9</u>	<u>15</u>	1.7	4.6					

Abbildung 4
 Kombinierbarkeit der Kollokationen

Die Lemmata sind mit roter und grüner Farbe gekennzeichnet. Die roten Konkordanzanzen können mit den roten Lemmata kombiniert werden, die grünen Konkordanzanzen haben dann die Tendenz zur Kombination mit dem grünen Lemma. Die weißen Konkordanzanzen 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 Kulturspezifität 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

