N-GRAMM ALS AUSGANGSVARIABLE FÜR DIE KOLLOKATIONSFORSCHUNG. EINE ANALYSE AM BEISPIEL DER WISSENSCHAFTSSPRACHE

Markéta Hotařová
Filozofická fakulta, Masarykova univerzita

HOTAŘOVÁ, Markéta: N-Gramm as an output variable for research into collocations. An analysis on the example of scientific language. *Philologia*, vol. XXX n° 2 (2020): 103–110.

Abstrakt: Ustálené viacslovné spojenia môžu niesť vo vedeckom úze nemeckého a českého jazyka spoločnú stopu. Tento príspevok sa zaoberá n-gramami ako signifikantnou veličinou pre výskum kolokácií a prezentuje jednotlivé výsledky a závery z vykonanej štúdie zaoberajúcou sa typickosťou nemeckých a českých vedeckých článkov z oblasti sociálnych vied a ich viditeľnými rozdielmi a spoločnými znakmi.
Kľúčové slová: n-gram, vedecký jazyk, kolokácie

Abstract: Multi-word units can have a common mark in the scientific language of German and Czech. This paper deals with n-grams as a significant variable for collocation research and presents partial results and conclusions from a study dealing with the typicality of German and Czech scientific articles in the field of social sciences and their visible differences and similarities.
Keywords: n-gram, scientific language, collocation

MEHRWORTVERBINDUNGEN UND AKTUELLE TENDENZEN IHRER ERFORSCHUNG IN DER KORPUSLINGUISTIK

Die Erforschung von mehr oder weniger festen Mehrwortverbindungen gehört in den letzten Jahrzehnten zum Hauptinteresse der Korpuslinguistik, die aus verschiedenen Perspektiven betrachtet werden kann. Dies wird an fol-

genden Arbeiten deutlich: aus fachsprachlicher Sicht (Kratochvílová-Zündorf 2015; Cedillo 2004), aus lexikographischer Sicht (Ďurčo 2014; Kratochvílová-Zündorf 2015), aus didaktischer Sicht (Ďurčo 2016) oder und aus der Sicht des Sprachgebrauchs (Bubenhofer 2009; Steyer 2013) und viele andere. Ihre textlinguistische Rolle ist nicht zu bestreiten. Die rekurrenten sprachlichen Mittel leisten u. a. einen wesentlichen Beitrag zum besseren Textverständnis. Dazu kommt es bereits bei der Textproduktion und das rekurrente Auftreten ist somit als „prädiktabel und modellierbar" zu betrachten (Heinemann – Heinemann 2002, 222) und gerade diese zwei Merkmale sind Untersuchungsgegenstand einer Stichprobenanalyse von Texten aus dem wissenschaftssprachlichen Bereich.

SPRACHLICHE MUSTER IN DER WISSENSCHAFTSSPRACHE

Bevor die Ergebnisse der Studie zu Mehrwortverbindungen im wissenschaftssprachlichen Sprachausschnitt der Sozialwissenschaften vorgestellt werden, werden an dieser Stelle Prämissen und Arbeitshypothesen für die Untersuchung vorgestellt.

Es wird davon ausgegangen, dass sich die erwähnte ‚Modellierbarkeit' vor allem in den Lexemen des alltäglichen Charakters und ihren Kombinationen manifestiert. Bei der kontrastiv angelegten Untersuchung werden Gemeinsamkeiten und Unterschiede im wissenschaftssprachlichen Gebrauch im Deutschen und Tschechischen anhand der sozialwissenschaftlichen Artikel thematisiert. Diesem Forschungsziel liegen zwei Thesen zugrunde: Einerseits handelt es sich um die ‚Universalitätsthese', andererseits um die ‚Relativitätsthese' (vgl. Steinhoff 2009, 99). Während die erste These die gemeinsamen kulturübergreifenden Merkmale in mehreren Wissenschaftssprachen sieht, lehnt die zweite These etwaige Gemeinsamkeiten ab. Graefen (1997, 68) betrachtet die heutige und die im amerikanisch-europäischen Raum betriebene Wissenschaft „als einen gemeinsamen Typ wissenschaftlicher Aktivität".

Sprachliche Muster bzw. Mehrwortverbindungen in der Wissenschaftssprache sind bereits zum Thema einiger Arbeiten geworden, als Beispiele sind Goldhahn (2017), Wallner (2014) oder Dönninghaus (2005) zu nennen. Im Folgenden wird somit versucht, anhand der ermittelten n-Gramme die Routine des wissenschaftlichen Handelns im Deutschen und Tschechischen zu manifestieren.

DAS KONZEPT DER N-GRAMME ALS UNTERSUCHUNGSGEGENSTAND UND DIE DARAUS FOLGENDEN FRAGEN

Mit dem Begriff ‚Muster' wird auf Bubenhofers (vgl. 2008, 409) Konzept zurückgegriffen, nach dem sich im ‚Muster' häufig auftretende Phänomene des typischen Usus widerspiegeln.

Auch ‚n-Gramm' geht auf Mehrwortverbindungen zurück, jedoch als operationalisierte und für die korpuslinguistische Analyse bestimmte Größe, die auf die Anzahl der aufeinanderfolgenden Wörter in einer Wortverbindung hinweist (vgl. Manning – Schütze 2002, 193).

Die Flexibilität der weniger festen Mehrworteinheiten verbirgt sich vor allem in der Anzahl an Kombinationsmöglichkeiten und auch in der Möglichkeit, die Spannweite der in den Mehrworteinheiten vorkommenden Lexeme zu definieren (vgl. Bubenhofer 2015, 496). Im Unterschied zur Kollokation hat ein n-Gramm keine feste lexikalische oder semantische Funktion und bietet somit einen größeren Forschungs- bzw. Interpretationsraum an.

Deshalb wurden gerade n-Gramme als Untersuchungseinheiten zur folgenden korpusgeleiteten Forschung mit dem beabsichtigten Ziel, aus den ermittelten Sprachdaten „abstraktere Einheiten" (Bubenhofer 2009, 122) zu extrahieren, die aufgrund ihrer Häufigkeit „abstraktere semantische oder pragmatische Funktionen" (Hein – Bubenhofer 2015, 180) haben können, herangezogen.

KORPUS UND -METHODE

Das zur Untersuchung bestimmte Korpus wurde aus rezensierten, wissenschaftlichen Zeitschriftenartikeln aus dem Themenbereich der Sozialwissenschaften aus den letzten zwei Jahrzehnten zusammengestellt. Die deutschen und tschechischen Artikel, die jeweils in den drei deutschsprachigen Ländern (Deutschland, Österreich und der Schweiz) und in der Tschechischen Republik publiziert wurden, bilden das Korpus mit einem Gesamtumfang von 871 041 Wörtern. Dieses Korpus, das aus zwei einzelsprachigen Teilkorpora gebildet wurde, reiht sich zu den sog. Vergleichskorpora, für die Texte aus mehreren Sprachen und ein gleiches Ziel, z. B. der gemeinsame Themenbereich, charakteristisch sind. Somit ist klar, dass es sich in diesem Fall um keine Übersetzungen handelt (vgl. Lemnitzer – Zinsmeister 2010, 104).

Die Gemeinsamkeiten und Unterschiede im wissenschaftssprachlichen Gebrauch lassen sich in den beiden Teilkorpora vergleichen, dies entspricht dem Schema ‚tertium comparationis', in dem die ermittelten Ergebnisse jeweils für

die Einzelsprache in Bezug auf die Fragestellung beschrieben werden (vgl. Gansel – Nefedov 2018, 53–54).

Zur Ermittlung der n-Gramme wurde das korpuslinguistische Softwaretool N-Gram Statistics Package (NSP) herangezogen. Zu seinen Vorteilen gehört der Einsatz der selbst gewählten Assoziationstests (vgl. Banerjee/Pedersen 2003, 370), seine kostenfreie Nutzung und nicht zuletzt auch die Möglichkeit, korpusgeleitet Mehrworteinheiten zu berechnen und größere Datenmengen zu analysieren (vgl. Bubenhofer 2009, 182). Als Nachteil, der jedoch von SprachwissenschaftlerInnen zu bewältigen ist, werden bestimmte Programmierkenntnisse des ‚Perl'-Programms betrachtet.

Der eigentliche Analyseprozess erfolgt dann in drei Schritten:

1) n-Gramme werden berechnet, ihre Verteilungen werden anhand des Assoziationstests ‚Log-Likelihood' geprüft, und extrahiert.
2) In den n-Grammen wird induktiv nach Substantivbasen gesucht, die dann zu abstrakteren funktional-semantischen Klassen gefasst werden.
3) Die einzelsprachigen Ergebnisse aus den Teilkorpora werden miteinander verglichen und interpretiert.

Dieser Analyseprozess wird als Methode der kontrastiven n-Gramm-Analyse bezeichnet, die eingesetzt wird, um die Tendenzen im aktuellen wissenschaftlichen Sprachgebrauch anhand einer Stichprobe im Kontext Sozialwissenschaften zu skizzieren. Teil dieses Forschungsvorhabens sind u. a. zwei Analysen, deren Teilergebnisse an dieser Stelle zu präsentieren sind. Einerseits handelt es sich um die dominanten abstrakten Substantivbasen und andererseits um die Heckenausdrücke bzw. „vorsichtige Formulierungen" (Graefen – Thielmann 2007, 91). Gleichzeitig soll überprüft werden, ob und in welchem Maß sich diese Methode für eine solche Fragestellung eignet.

Der Fokus der Untersuchung reichte von Bigrammen bis Hexagrammen in der jeweiligen Sprache, die entsprechend anhand des Softwaretools NSP extrahiert und nach dem ‚Log-Likelihood-Test' sortiert wurden. Zunächst wurden die Grenzwerte (bei Bi- und Trigrammen 4, bei Tetragrammen 3 und bei Penta- und Hexagrammen 2), ab welcher Mindestanzahl die erforschten n-Gramme in die Ergebnisse einbezogen werden, ermittelt. Somit wurden als Analysebasis zu den Analysen 19 080 n-Gramme herangezogen.

ERGEBNISSE UND INTERPRETATION

Die erste Analyse stützt sich auf Steyer (vgl. 2013, 337–338), die in den induktiv gewonnenen Wortverbindungen Abstraktionen sieht, die im Sprach-

gebrauch rekurrent auftreten und gleichzeitig stützt sie sich auch auf Graefen (vgl. 2001, 192), die auf aus der Alltagssprache entlehnten Lexeme fokussiert und dabei ihre Aufmerksamkeit darauf lenkt, dass gerade die Kombination von solchen Lexemen ihre „wissenschaftsspezifische Bedeutung" generiert (vgl. Graefen 2001, 192). Somit wurden unter fünf funktional-semantischen Klassen – Delimitationsperspektive, Extensionsperspektive, Vergleichsdimension, Bestimmung des Maßes und Hervorhebungen – folgende drei dominante abstrakte Substantivbasen ermittelt:

Delimitationsperspektive	*RÁMEC – RAHMEN* *SMYSL – SINN, ART und WEISE* *ZÁKLAD – GRUNDLAGE, BASIS*
Extensionsperspektive	*ZÁVISLOST, SOUVISLOST – ZUSAMMENHANG* *SOULAD – EINKLANG* *VZTAH – BEZIEHUNG*
Vergleichsdimension	*SROVNÁNÍ, POROVNÁNÍ – VERGLEICH, POLARISIERUNG* *ROZDÍL, ROZPOR – UNTERSCHIED, GEGENSATZ, DIFFERENZ* *STRANA*
Bestimmung des Maßes	*MÍRA* *POČET, PODÍL – (AN)TEIL, ANZAHL*
Hervorhebungen	*SKUTEČNOST – TATSACHE* *DŮRAZ* *ROLE*

Laut der berechneten Ergebnisse handelt es sich sowohl im Deutschen als auch im Tschechischen in der Mehrheit um Präpositionalphrasen nach dem Schema [Präposition] + ([Artikel]) + [Substantiv]. Anhand der induktiv ausgewerteten Datenmenge lässt sich feststellen, dass die deutsche und tschechische Wissenschaftssprache über eine ähnliche Gedankenführung in der Wahl der kombinierten Lexeme alltagssprachlichen Charakters verfügt. Somit wird die ‚Universalitätsthese' empirisch gestützt (vgl. Steinhoff 2009, 99–100). Darüber hinaus ist bei den AutorInnen von wissenschaftlichen Artikeln die Tendenz festzustellen Sachverhalte vornehmlich zu vergleichen, zu zählen und hervorzuheben. Sie bringen auch oft Fremdwörter als ‚einheimische' Wörter zum Ausdruck (wie z. B. *Kontext, Perspektive, Relation, Polarisierung*). Dass dem Einsatz von Fremdwörtern in wissenschaftlichen Artikeln eine Funktion zukommt, ist nicht zu bestreiten. Aber ihr nicht ganz transparenter semantischer

Bezug lässt den übermittelten Inhalt mehrdeutig werden. Damit hängt auch die zweite durchgeführte Analyse zusammen, die das oft vorkommende Phänomen des ‚vorsichtigen Sprachgebrauchs' in den geistes- und sozialwissenschaftlichen Artikeln (vgl. Dönninghaus 2005, 349) sowohl im Deutschen als auch im Tschechischen anhand der untersuchten Stichprobe reflektiert. Somit wurde im untersuchten Sprachmaterial der Frage nachgegangen, welche Tendenzen der Hedging-Strategie in beiden Einzelsprachen zu beobachten sind, wobei unter Heckenausdrücken bzw. Hedges „vorsichtige[…] Formulierungen" zu verstehen sind (vgl. Graefen – Thielmann 2007, 91). Insgesamt ist das häufige Vorkommen des Modalverbs *können* sowohl im Deutschen als auch im Tschechischen zu verzeichnen, das in Form von Konjunktiv II (*als bislang nachgezeichnet werden können; analysieren zu können; auf diese Weise können; können wir / by mohly; bychom mohli*) oder als Teil verschiedener verbaler Konstruktionen realisiert wurde. Häufig lassen sich auch andere Modalverben verzeichnen, wie z. B. *sollen*, wieder im Konjunktiv II (*měla být; by to mělo být*) oder *müssen* in negierter Form (*nemusí nutně*). Des Weiteren ist auch eine bedeutende Menge tschechischer verbaler Konstruktionen mit dem Kopulaverb *sein* oder der Phrase *es lässt sich* zu finden (*je možné; by bylo možné očekávat / lze předpokládat)*.

Anhand der deutschen Sprachdaten wurde eine große Vielfalt von verschiedenen lexikalischen Mitteln, wie Adjektive (*die relative Menge; eine gewisse*) oder Pronomina (*den meisten Fällen*) festgestellt. Zur Hedging-Strategie tragen auch deutsche Infinitivkonstruktionen (*ist anzunehmen; ist zu vermuten; zu sein scheint*) bei.

SCHLUSSFOLGERUNGEN

Der Einsatz des korpuslinguistischen Softwaretools NSP hat sich auch aufgrund des diesem Tool inhärenten Signifikanztests ‚Log-Likelihood' im induktiven, korpusgeleiteten Modus bewährt, und zwar im Hinblick auf das Auffinden dominanter Substantivbasen als auch in der Identifizierung von Heckenausdrücken, denen noch eine weitere korpusbasierte Arbeit gewidmet wird. Aufgrund der unterschiedlichen Realisierungen des Subjekts *wir* vor allem im Tschechischen, z. B. in Form von Ellipsen, ist diese Frage weiter zu erforschen. Des Weiteren wurde ermittelt, dass die deutsche und tschechische Wissenschaftssprache im Kontext der Sozialwissenschaften oft auf gemeinsamen Konzepten hinsichtlich der verwendeten Kombinationen von Lexemen beruhen, die in Texten als Mehrwortverbindungen vorkommen. Die Teilergebnisse der durchgeführten kontrastiven n-Gramm-Analyse stellen einen Beitrag zur

Erforschung der Typizität des wissenschaftlichen Sprachgebrauchs in beiden Vergleichssprachen dar und bilden zugleich eine Grundlage für weitere Untersuchungen im Rahmen des Forschungsprojekts ‚LaCon - Language configurations in humanities' in Brno.

Literatuverzeichnis

BANERJEE, Satanjeev – Ted PEDERSEN. 2003. "The Design, Implementation, and Use of the Ngram Statistics Package." In *Computational Linguistics and Intelligent Text Processing,* edited by Gelbukh, Alexander, CICLing 2003. Lecture Notes in Computer Science, vol 2588, 370–381. Berlin/Heidelberg: Springer.

BUBENHOFER, Noah. 2008. „Diskurse berechnen? Wege zu einer korpuslinguistischen Diskursanalyse." In *Methoden der Diskurslinguistik. Sprachwissenschaftliche Zugänge zur transtextuallen Ebene*, edited by Spitzmüller, Jürgen, and Warnke, Ingo H., Linguistik – Impulse & Tendenzen 31, 407–434 Berlin: De Gruyter.

BUBENHOFER, Noah. 2009. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse.* Berlin/New York: De Gruyter.

BUBENHOFER, Noah. 2015. „21. Muster aus korpuslinguistischer Sicht." In *Handbuch Satz, Äußerung, Schema*, edited by Dürscheid, Christa, and Schneider, Jan Georg, 485–502Berlin/New York: De Gruyter.

CEDILLO, Ana Caro. 2004. *Fachsprachliche Kollokationen. Ein übersetzungsorientiertes Datenbankmodell Deutsch-Spanisch*. Tübingen: Gunter Narr.

DÖNNINGHAUS, Sabine. 2005. *Die Vagheit der Sprache*. *Begriffsgeschichte und Funktionsbeschreibung anhand der tschechischen Wissenschaftssprache.* Wiesbaden: Harrassowitz.

ĎURČO, Peter. 2014. „Feste Wortverbindungen mit Adjektiven: Korpuslinguistischer Ansatz als Grundlage für bilinguale Lexikographie." In *Valenz und Kookkurrenz. Grammatische und lexikologische Ansätze*, edited by Ďurčo, Peter et.al, 147–161. Wien: Lit.

ĎURČO, Peter. 2016. „Zum Konzept der Kollokationsdidaktik und des Kollokationslernens bei Germanistikstudenten." In *Kollokationsforschung und Kollokationsdidaktik*, edited by Ďurčo, Peter, 147–172. Wien: Lit.

GANSEL, Christina – Sergej NEFEDOV. 2018. *Wissenschaftliches Schreiben. Ein Handbuch*. Wolgast: Steffen Media Usedom.

GOLDHAHN, Agnes. 2017. *Tschechische und deutsche Wissenschaftssprache im Vergleich*. Berlin: Frank & Timme GmbH.

GRAEFEN, Gabriele – Winfried THIELMANN. 2007. „Der Wissenschaftliche Artikel." In *Reden und Schreiben in der Wissenschaft*, edited by Auer, Peter, and Baßler, Harald, 67–98. Frankfurt am Main: Campus.

HEIN, Katrin – Noah BUBENHOFER. 2015. „Korpuslinguistik konstruktionsgrammatisch. Diskursspezifische N-Gramme zwischen statistischer Signifikanz und semantisch-pragmatischem Mehrwert.“ In *Konstruktionsgrammatik IV: Konstruktionen als soziale Konventionen und kognitive Routinen*, edited by Lasch, Alexander, and Ziem, Alexander, 179–206. Tübingen: Stauffenburg.

KRATOCHVÍLOVÁ (Zündorf), Iva. 2011. *Kollokationen im Lexikon und im Text. Mehrwortverbindungen im Deutschen und Tschechischen*. Berlin: LIT.

KRATOCHVÍLOVÁ (Zündorf), Iva. 2015. „Formulierungsroutinen und Konfigurationen der fachinternen Wirtschaftskommunikation als Spezialgebiet der fachsprachlichen Textlinguistik und Phraseologie.“ In *Fachkommunikation im Wandel*, edited by Satzger, Axel – Vaňková, Lenka – Wolf, Norbert Richard. Ostrava: Ostravská univerzita v Ostravě, 65–78. Filozofická fakulta.

KRETZENBACHER, Heinz Leonard. 1994. „Wie durchsichtig ist die Sprache der Wissenschaften?“ In *Linguistik der Wissenschaftssprache,* edited by Kretzenbacher – Heinz Leonard – Weinrich, Harald, 15–40. Berlin: De Gruyter.

LEMNITZER, Lothar – Heike ZINSMEISTER. 2010. *Korpuslinguistik*. *Eine Einführung*. Tübingen: Narr Francke Attempto.

MANNING, Christopher D. – Hinrich SCHÜTZE. 2002. *Foundations of Statistical. Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

STEYER, Kathrin. 2013. *Usuelle Wortverbindungen: Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

WALLNER, Franziska. 2014. *Kollokationen in Wissenschaftssprachen. Zur lernerlexikographischen Relevanz ihrer wissenschaftssprachlichen Gebrauchsspezifika*. Deutsch als Fremd- und Zweitsprache. Tübingen: Stauffenburg.

Mgr. et Mgr. Markéta Hotařová, Ph.D.
Ústav germanistiky, nordistiky
a nederlandistiky
Filozofická fakulta
Masarykova univerzita
Arne Nováka 1
602 00 Brno-střed
marketa.hotarova@mail.muni.cz