HUMBOLDT-UNIVERSITÄT ZU BERLIN

Problemkind der Korpuslinguistik: Das Lexikon in Struktur, Gebrauch und Analyse

Anna Shadrova

Humboldt-Universität zu Berlin,
Forschungsgruppe "Emerging Grammars in Language Contact Situations – A Comparative Approach" (RUEG), DFG, Referenznr. 313607803

Berlin, 23/05/2022

Overview

Usage-based turn

Seit den 1980er/90er Jahren „gebrauchsbasierte Wendung" mit Entwicklung verschiedener Grammatikansätze (z.B. Konstruktionsgrammatik Pattern Grammar, Word Grammar etc., Goldberg, 2013; Hoey, 2005; Hunston, 2012) u.a. in der Spracherwerbsforschung (L1 und L2, Ortega, 2013, 2015; Tomasello, 2009), Überschneidungen mit Funktionalismus (z.B. Bybee, 2013).

Gängige Postulate:

- Idiom Principle (Sinclair, 1991)
- Untrennbarkeit von Lexikon und Syntax = Zeichenhaftigkeit von Strukturen
- Phraseologisches Kontinuum
- Sprachgebrauch (belegt in Korpora) als eigentlicher Untersuchungsgegenstand

Idiom Principle Sinclair, 1991, p. 110

"(..) a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments."

Idiom Principle Sinclair, 1991, p. 110

"(..) a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments."

Open-choice principle:
"This is a way of seeing language text as the result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness. This is probably the normal way of seeing and describing language. It is often called a 'slot-and-filler' model, envisaging texts as a series of slots which have to be filled from a lexicon which satisifies local restraints. At each slot, virtually any word can occur. Since language is believed to operate simultaneously on several levels, there is a very complex pattern of choices in progress at any moment, but the underlying principle is simple enough", (ebd. 109).

Schätzungen zum Anteil usueller Wortverbindungen

- Conklin and Schmitt, 2012: 1/3 bis 1/2 “of discourse”
- Wray, 2002, p. 119: “formulaic processing is the default”; “construction out of, and reduction into, smaller units by rule occurs only as necessary”
- Erman and Warren, 2000 schätzen zwischen 50 und 90%

Schätzungen zur Menge usueller Wortverbindungen

"An analysis of the authentic data processed in preparation for the Oxford Dictionary of Current Idiomatic English (Cowie, Mackin & McCaig, 1975/1983), for example, yielded literally thousands of such stable multi-word units (see Cowie, 1988). Similarly, the Oxford Dictionary of Phrasal Verbs (Cowie & Mackin, 1993) and the Oxford Dictionary of English Idioms (Ayto, 2010) between them contain some 15,000 multi-word expressions." Singleton and Leśniewska, 2021, p. 49

Ursachen und/oder Funktionen von usuellen Wortverbindungen

- Ursache: Entrenchment? ("repetition of linguistic material is indicative of structure: repeatable structures are evidence for the units of linguistic cognition." Reitter, 2008, p. 14)
  - Ergebnis des Lernprozesses? Bates et al., 1988; Bates and Goodman, 1999; Dąbrowska and Lieven, 2005; MacWhinney, 2014; Tomasello, 2000, 2009
  - Frequenz/Konvention?

Ursachen und/oder Funktionen von usuellen Wortverbindungen

- Ursache: Entrenchment? ("repetition of linguistic material is indicative of structure: repeatable structures are evidence for the units of linguistic cognition." Reitter, 2008, p. 14)
  - Ergebnis des Lernprozesses? Bates et al., 1988; Bates and Goodman, 1999; Dąbrowska and Lieven, 2005; MacWhinney, 2014; Tomasello, 2000, 2009
  - Frequenz/Konvention? (Achtung, zirkulär)

Ursachen und Funktionen usueller Wortverbindungen

- Funktion (Zweck): Weniger Arbeit?
- Funktion (Zweck): Bessere Verarbeitbarkeit? (z.B. schnellere Verarbeitung, bessere Abrufbarkeit, Conklin and Schmitt, 2008; Ellis et al., 2014; Siyanova-Chanturia et al., 2011)

Ursachen und Funktionen usueller Wortverbindungen

- Funktion (Zweck): Weniger Arbeit?
- Funktion (Zweck): Bessere Verarbeitbarkeit? (z.B. schnellere Verarbeitung, bessere Abrufbarkeit, Conklin and Schmitt, 2008; Ellis et al., 2014; Siyanova-Chanturia et al., 2011)

  "The existence of collocational patterning has to do with the fact that human beings prefer to save effort whenever possible. Also, it is certainly connected with the huge demands made on us by the extreme rapidity of speech production, which are such that we have to exploit every opportunity to make savings on processing time." Singleton and Leśniewska, 2021, p. 50
- Funktion (Zweck): Funktionelle Routinenhaftigkeit? (z.B. Kommunikationsvorteil durch Eindeutigkeit komplexer Zeichen, Wray, 2002, u.a.)

Lexikosyntax

Anziehungskräfte zwischen Lexikon und Syntax, z.B.

- idiosynkratische Argumentstrukturpräferenzen von semantisch ähnlichen Verben Dux, 2020; Faulhaber, 2011
- unterschiedliche Produktivität von Argumentstrukturslots Herbst, 2014; Zeldes, 2012
- lexikalische Präferenzen von syntaktischen Strukturen, z.B. Ditransitivkonstruktionen

Lexikosyntax

Anziehungskräfte zwischen Lexikon und Syntax, z.B.

- idiosynkratische Argumentstrukturpräferenzen von semantisch ähnlichen Verben Dux, 2020; Faulhaber, 2011
- unterschiedliche Produktivität von Argumentstrukturslots Herbst, 2014; Zeldes, 2012
- lexikalische Präferenzen von syntaktischen Strukturen, z.B. Ditransitivkonstruktionen

Schlussfolgerung der gebrauchsbasierten Linguistik

Lexikalische Elemente sind nicht frei kombinierbar, sondern in hohem Umfang koselektionsbeschränkt.

Schlussfolgerung der gebrauchsbasierten Linguistik

Lexikalische Elemente sind nicht frei kombinierbar, sondern in hohem Umfang koselektionsbeschränkt.

...und das hat irgendeinen Grund oder eine Funktion.

Overview

Beispiel

Forschungsfrage: Wie ähnlich sind sich Muttersprachler:innen und Lerner:innen auf verschiedenen Erwerbsstufen im Gebrauch solcher koselektionsbeschränkter Einheiten?

Beispiel

Forschungsfrage: Wie ähnlich sind sich Muttersprachler:innen und Lerner:innen auf verschiedenen Erwerbsstufen im Gebrauch solcher koselektionsbeschränkter Einheiten?

Hintergrund: Es ist bekannt, dass Lerner:innen Schwierigkeiten mit dem Erwerb zu haben scheinen (z.B. Granger, 2011; Nesselhauf, 2005; Paquot, 2019; Paquot and Granger, 2012

Kobalt (Shadrova, 2021a; Zinsmeister et al., 2012)

- Essays von Germanistikstudierenden an weißrussischen und chinesischen Universitäten + L1-Vergleichstexte
- unterschiedliche Erwerbsstände (ca. B1 – C2)
- 171 Texte, davon 20 L1, 62 CH, 89 BEL
- Thema: Geht es der Jugend heute besser als früheren Generationen?
- 90 min Erhebungszeit, OnDaF (evaluierter C-Test) zur Sprachstandsmessung, zusätzlich Metadaten zur Sprachbiographie
- verfügbar über Annis https://korpling.org/annis3/ und https://doi.org/10.5281/zenodo.5730224

Beispiel aus Kobalt: DEU_010

“Ich werde mich jetzt auf Deutschland beschränken, da ich selbst hier lebe. Die Vorteile der heutigen Jugend in Deutschland sind offensichtlich. Wir haben das Glück, sehr viel Freiheit auch durch Gesetze genießen zu können. Wir können uns kleiden, wie wir wollen, ohne dass es jemanden interessiert. Es ist Platz für Individualität gegeben. Zudem gibt bzw. sollte es keine Unterschiede zwischen den Geschlechtern geben. Ein Teilschritt ist bereits erreicht, jeder hat das Recht auf Bildung. Man hat mehr Rechte als Jugendlicher als die Generationen vor uns.”

Beispiel aus Kobalt: BEL_008

“Was junge Leute im Alter ab 16 anbetrifft, so ist es für sie notwendig, mit der Zeit gleich zu schreiten, um möglichst eine bessere Perspektive zu bekommen: Dabei geht es um das Studium, um die Karriere, die zukünftige Familie. Jeder strebt danach, mehr Geld zu verdienen, um sein Leben zu erleichtern. Um sich selbst die Schönheiten modernen Lebens leisten zu können, braucht man Geld. Deshalb versuchen immer mehr Studenten, das Studium mit einem Job zu verbinden, was üblicherweise schwer ist. Aber solche frühere Selbstständigkeit und Unabhängigkeit vom Elternhaus hat zum Verhältniswechsel in der Gesellschaft geführt. Solche Erscheinungen wie “wilde Ehen” trifft man ziemlich oft, was auf Missverständnis von der Seite der älteren Generation stößt.” (BEL_008, onDaF=125, B2/fast C1)

Beispiel aus Kobalt: CH_034

“Nachdem wir den Universitätsabschluss machten, können wir keine Arbeit haben. Und dann können wir nicht leben. Das ist die Realität. Es gibt also viele Leute, die finden, dass die Jugend heute unhöflicher ist als die frühere Generation. Die Jugenden heute grüßen die ältere Leute nicht. Sie haben ihren eigenen Charakter. Sie machen die Haare komisch und die komische Farbe gefällt den älteren Leuten nicht. Die Jugenden hören immer Rockmusik. Das geht nur den älteren Leuten auf die Nerven. Sie haben immer ein komisches Interesse. Das können und möchten die ältere Leute nicht verstehen.” (CH_034, onDaF=72, nicht ganz B1)

Assoziationsmaße

Wenn Wort a 10 mal vorkommt und Wort b 100 mal, und sie zusammen 5 mal vorkommen, ist das dann viel oder wenig?

Häufig berechnet als

- Varianten bedingter Wahrscheinlichkeit
- Entropie/Surprisal/Mutual Information
- ΔP

Bouma, 2009; Gries, 2013, 2015; Gries and Ellis, 2015; Krenn, Evert, et al., 2001; Orliac and Dillinger, 2003; Pecina, 2010, u.v.a

Problem: Aussagekraft von Assoziationsmaßen

Schon lexikographisch schwierig:

- z.B. die relative Häufigkeit eines Worts im Korpus hängt für die meisten Wörter stärker von der Korpusgröße ab als von seiner eigentlichen Frequenz. Das spiegelt sich in widersprüchlichen Assoziationsmaßen wieder.
- vielfältige andere Probleme (Shadrova, 2020, v.a. Kap. 3 und 4, Shadrova, 2021b, 2022)

Problem: Lexikalische Vielfalt

	subcorpus	docs	verb_lex	noun_lex
1	ch_095	10	153	378
2	ch_115	24	318	697
3	ch_130	17	304	655
4	ch_160	11	231	481
5	l1	20	419	749
6	rus_075	11	107	214
7	rus_095	27	290	644
8	rus_115	21	318	674
9	rus_130	20	380	694
10	rus_160	10	298	506

Problem: Lexikalische Vielfalt

	lemma	ch_095	ch_115	ch_130	ch_160	l1	rus_075	rus_095	rus_115	rus_130
1	Auge	1	2	1	2	1	1	1	2	4
2	Beispiel	6	7	6	8	15	5	21	3	8
3	Computer	8	16	6	9	6	11	38	23	19
4	Familie	11	26	20	12	17	3	36	17	16
5	Freund	3	7	2	1	7	5	31	14	15
6	Geld	5	10	3	1	7	2	25	17	23
7	Generation	67	205	105	60	166	49	125	117	91
8	Internet	16	25	13	9	14	13	40	39	23
9	Jugend	115	195	139	68	140	57	159	106	88
10	Jugendliche	17	94	101	42	94	5	60	38	55
11	Kind	6	37	35	36	88	10	60	37	38
12	Krieg	1	9	5	5	16	7	16	16	21
13	Leben	21	66	60	28	33	24	99	106	83
14	Möglichkeit	3	8	5	4	18	14	78	54	44
15	Mutter	2	14	2	2	8	3	12	9	6
16	Schule	10	13	10	9	13	2	13	15	19
17	Seite	6	14	12	7	7	2	11	7	17
18	Situation	11	13	5	2	12	2	19	13	21
19	Tag	3	9	7	3	2	3	9	7	19
20	Welt	10	41	28	18	12	4	38	37	45
21	Zeit	8	49	30	8	16	19	106	64	77
22	Arbeit	12	15	4	3	4	3	11	14	19
23	Eltern	18	43	42	25	23	18	55	43	71
24	Freiheit	3	2	6	4	8	5	1	6	7
25	Großeltern	1	3	4	5	2	3	14	12	8
26	Land	2	19	19	5	16	4	24	17	23
27	Meinung	17	42	15	17	19	5	28	17	16
28	Fernseher	1	2	2	2	3	2	7	2	1
29	Mensch	8	17	10	11	17	17	84	69	71
30	Problem	4	26	23	14	38	12	43	40	54
31	Technik	4	22	14	3	8	6	6	4	7
32	Zukunft	2	13	10	1	5	1	26	15	10

Problem: Lexikalische Vielfalt

- Lexeme, die in jedem Subkorpus in mindestens fünf Dokumenten vorkommen: Internet, Jugend, Leben, Zeit, Eltern
- Lexeme, die in acht oder neun (von zehn) Subkorpora in mindestens fünf Dokumenten vorkommen: Jugendliche, Kind, Schule, Welt, Meinung, Mensch, Computer, Möglichkeit, Situation, Problem, Frage, Jahr, Leute

Problem: Lexikalische Vielfalt

Obwohl die Texte einander ähnlich erscheinen, und obwohl alle Sprecher:innengruppen viele verschiedene Wörter verwenden, gibt es sehr wenig lexikalische Übereinstimmung über Sprecher:innen hinweg.

Problem: Lexikalische Vielfalt

Obwohl die Texte einander ähnlich erscheinen, und obwohl alle Sprecher:innengruppen viele verschiedene Wörter verwenden, gibt es sehr wenig lexikalische Übereinstimmung über Sprecher:innen hinweg.
Das ist schlecht, wenn man dieselben Kombinationen vergleichen will.
Wenn sie *prefabricated* sind, und Sprecher:innen sie als komplexe Zeichen zur Eindeutigkeit verwenden, sollten sie aber dieselben sein.

Problem: Kombinatorik

	subcorpus	docs	verb_lex	noun_lex	potential
1	ch_095	10	153	378	57834
2	ch_115	24	318	697	221646
3	ch_130	17	304	655	199120
4	ch_160	11	231	481	111111
5	l1	20	419	749	313831
6	rus_075	11	107	214	22898
7	rus_095	27	290	644	186760
8	rus_115	21	318	674	214332
9	rus_130	20	380	694	263720
10	rus_160	10	298	506	150788

Problem: Kombinatorik

Beispiel:

- BEL-115: 304 Lexeme als Akkusativobjekt, 148 Verblexeme, die mit Akkusativobjekten auftreten = 44 992 mögliche V-OBJA-Kombinationen
- Zahl der im Subkorpus vorkommenden Verben mit Akkusativobjekt (Token): 726
- Wenn alle gleich wahrscheinlich wären: p= $\frac{1}{44992}$, p(2xdieselbe) = $\frac{1}{44992^2}$

Problem: Kombinatorik

Beispiel:

- BEL-115: 304 Lexeme als Akkusativobjekt, 148 Verblexeme, die mit Akkusativobjekten auftreten = 44 992 mögliche V-OBJA-Kombinationen
- Zahl der im Subkorpus vorkommenden Verben mit Akkusativobjekt (Token): 726
- Wenn alle gleich wahrscheinlich wären: p= $\frac{1}{44992}$, p(2xdieselbe) = $\frac{1}{44992^2}$
- Ziehen mit Zurücklegen ohne Reihenfolge: $\frac{(44992+726-1)!}{726!(44992-1)!} = 1.3527 \cdot 10^{1617}$ Möglichkeiten
- Angenommen, nur 1% sind semantisch überhaupt möglich: $2.276 \cdot 10^{337}$ Möglichkeiten und angenommen, davon werden nur 10% überhaupt frei kombiniert: 72 mal ziehen aus 449: *immer noch* $3.352 \cdot 10^{89}$.
  - Zahl der Atome im Universum wird zwischen 10^{78} und 10^{82}

Problem: Kombinatorik

Beispiel:

- BEL-115: 304 Lexeme als Akkusativobjekt, 148 Verblexeme, die mit Akkusativobjekten auftreten = 44 992 mögliche V-OBJA-Kombinationen
- Zahl der im Subkorpus vorkommenden Verben mit Akkusativobjekt (Token): 726
- Wenn alle gleich wahrscheinlich wären: p= $\frac{1}{44992}$, p(2xdieselbe) = $\frac{1}{44992^2}$
- Ziehen mit Zurücklegen ohne Reihenfolge: $\frac{(44992+726-1)!}{726!(44992-1)!} = 1.3527 \cdot 10^{1617}$ Möglichkeiten
- Angenommen, nur 1% sind semantisch überhaupt möglich: $2.276 \cdot 10^{337}$ Möglichkeiten und angenommen, davon werden nur 10% überhaupt frei kombiniert: 72 mal ziehen aus 449: *immer noch* $3.352 \cdot 10^{89}$.
  - Zahl der Atome im Universum wird zwischen 10^{78} und 10^{82}
- Selbst wenn man nur 7 mal aus 449 zieht, immer noch $7.647 \cdot 10^{14}$ Möglichkeiten

Problem: Kombinatorik

Jede Kombination von Koselektionen ist extrem unwahrscheinlich. Das kann man nicht seriös gradieren.
Shadrova, 2020, 2022

Overview

RUEG

Research Group Emerging Grammars: A Comparative Approach

- Kontrolliertes Korpus mit Daten von Herkunftssprecher:innen des Deutschen in den USA, des Griechischen, Russischen, Türkischen in Deutschland, mit Parallelldaten von Monolingualen aus denselben Ländern und Daten derselben Sprecher:innen in der Majoritätssprache
- Beispiel: Thorsten wohnt in den USA und wächst mit Deutsch als Herkunftssprache auf. Wir nehmen Thorsten auf Deutsch und Englisch auf. Zusätzlich nehmen wir monolingual John aus den USA auf Englisch und den monolingualen Timo aus Deutschland aus Deutschland auf Deutsch auf.
- Aufgabe: Videobeschreibung in vier Settings (formal/informell, mündlich/schriftlich)
- Video: https://osf.io/szfhd/

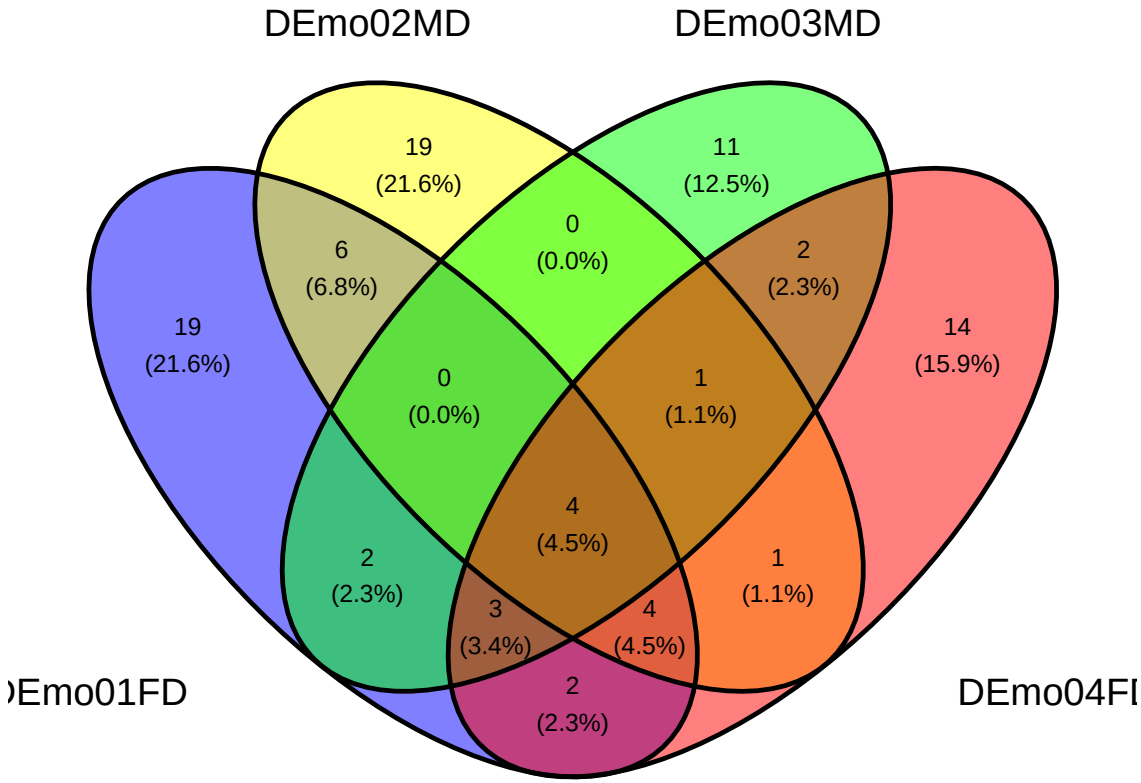
Lexikalische Diversität: RUEG

In 267 Texten (alle Sprecher:innengruppen), formal schriftlich, tritt *Einkauf* mit diesen Verben auf (Häufigkeit in Klammern):

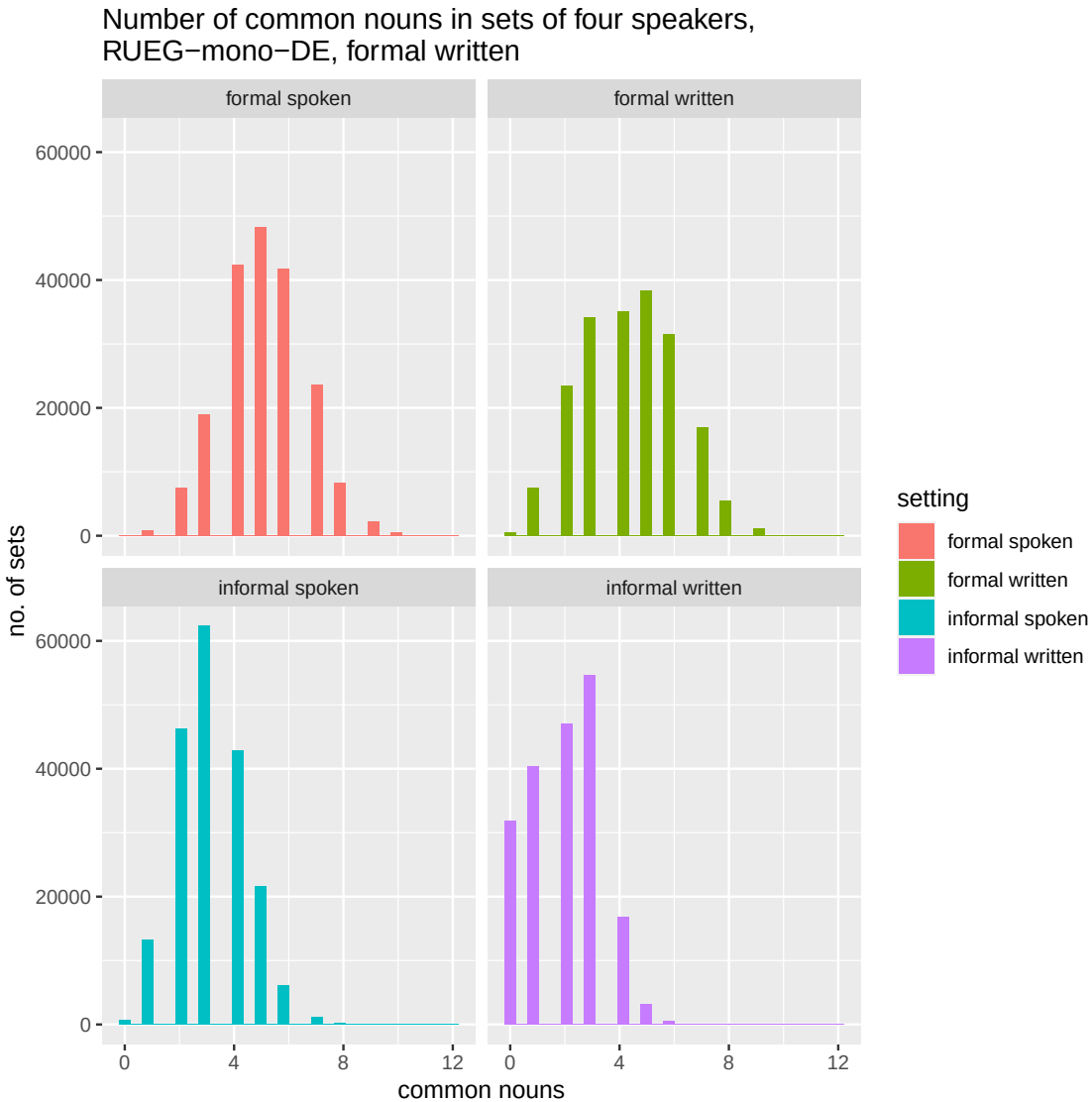
- packen (27), fallen (24), räumen (19), einpacken (17), helfen (14), einräumen (14), laden (12), aufsammeln (12), verstauen (12), einsammeln (13), fallen lassen (10), einladen (7), auspacken (7), legen (7), sein (6), holen (6), ausladen (6), entladen (5), aufheben (5), ausräumen (5), verladen (5), rollen (3), einsortieren (3), runterfallen (3), herunterfallen (3), reißen (3), kümmern (3), fliegen (2), verlieren (2), beladen (2), sortieren (2), bepacken (2), reinpacken (2)
- *keins davon* in mehr als 10% der Texte (max=27)
- nur einmal kommen vor: belagern, verteilen, plumsen, reintun, hineinlegen, herausholen, reinlegen, landen, mitreißen, leeren, zurückräumen, tun, ausparken, verschütten, sammeln, zurückzupacken, zusammenräumen, hinunterfallen, umrennen, raustransportieren, miteinräumen, zurückpacken (zusammen 22)
-

Lexikalische Überlappung: RUEG

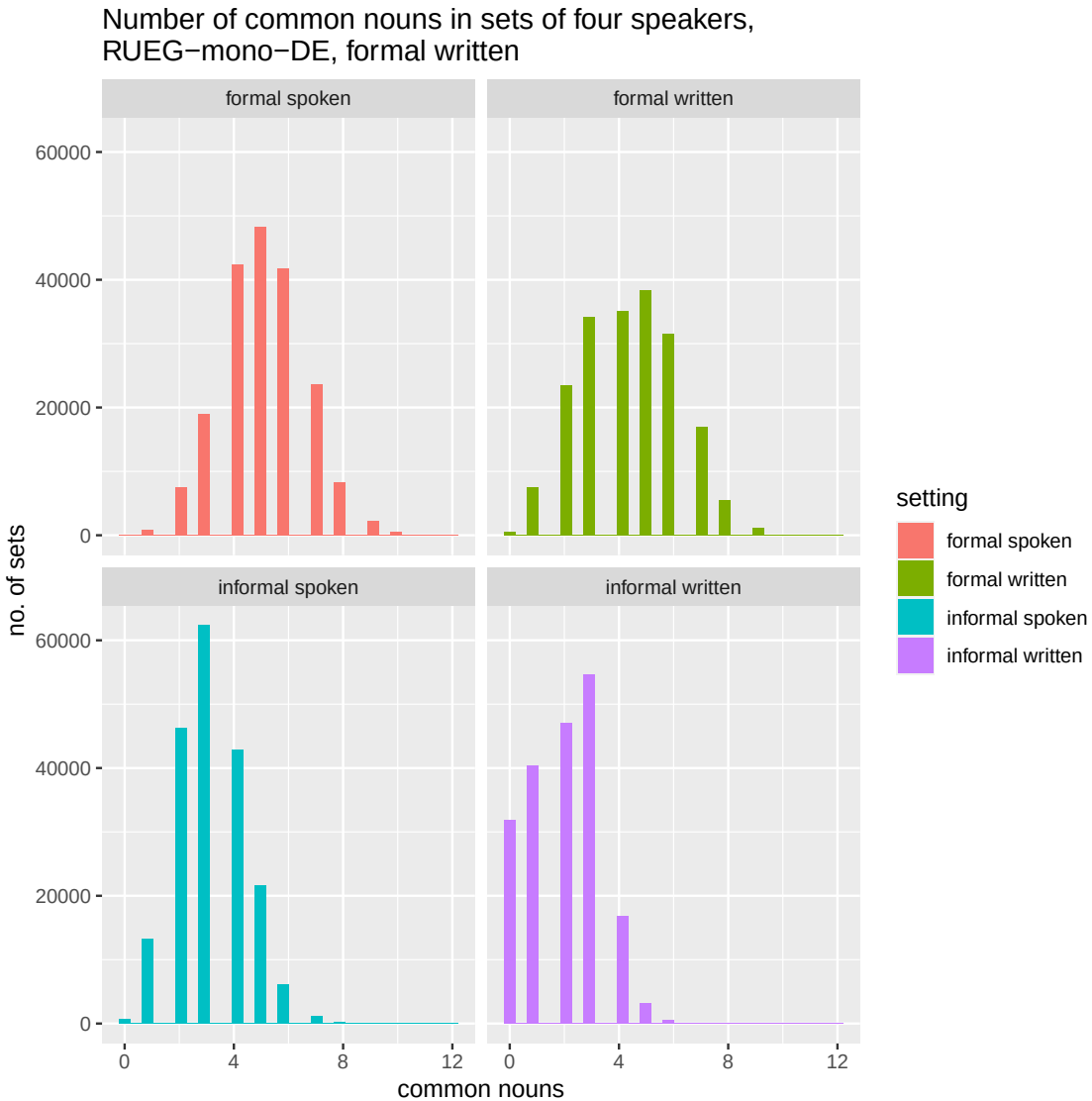
RUEG informell gesprochen, monolingual Deutsch: Vier beliebige Texte, alle Nomen-, Verb-, Adjektiv- und Adverblexeme



Lexikalische Überlappung: RUEG

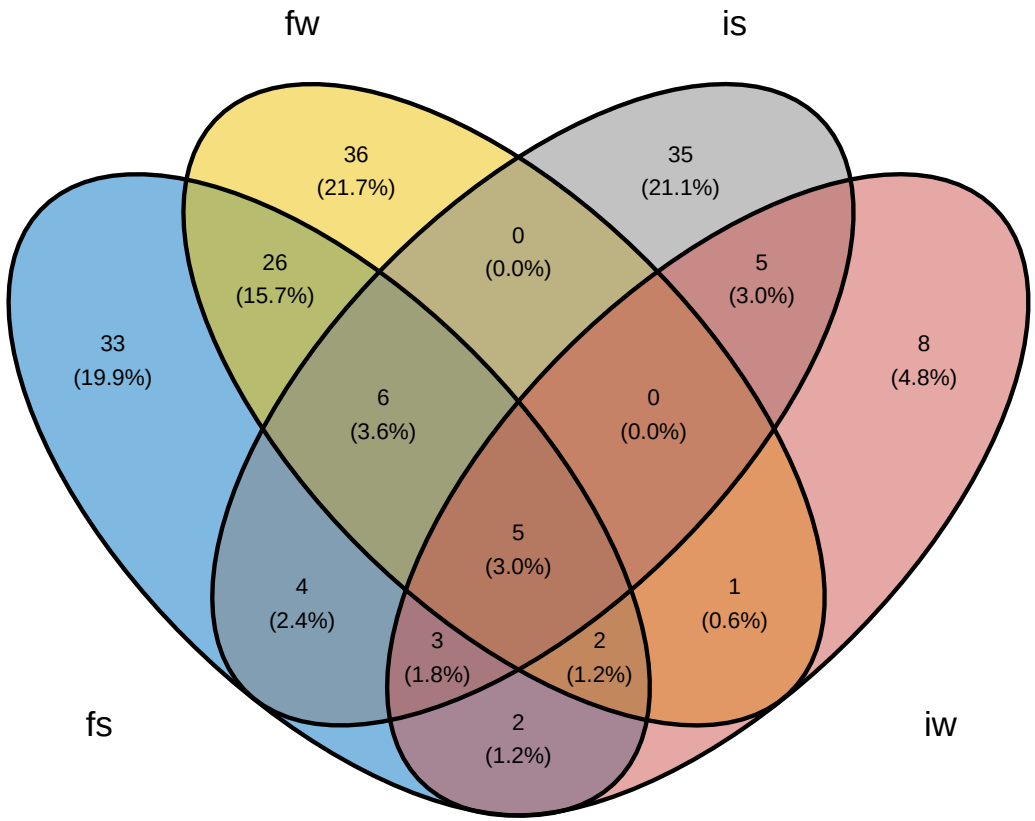


Lexikalische Überlappung: RUEG



Lexikalische Diversität Einzelsprecher: RUEG

Text DEmo50MD_fwD (formal, schriftlich, monolingual Deutsch, männlich, jugendlich):



Lexikalische Diversität Einzelsprecher: RUEG

Text DEmo50MD_fwD (formal, schriftlich, monolingual Deutsch, männlich, jugendlich):

- Fünf Wörter kommen in allen vier Settings vor: *Hund, Auto, Ball, bremsen, kommen*
- insgesamt 166 Wörter (hier gezählt: Adjektive, Nomen, Verben ohne haben, sein, werden)
- In jedem Setting ist die Zahl der Lexeme, die nur in diesem Setting vorkommen, 1/3 bis über 2/3

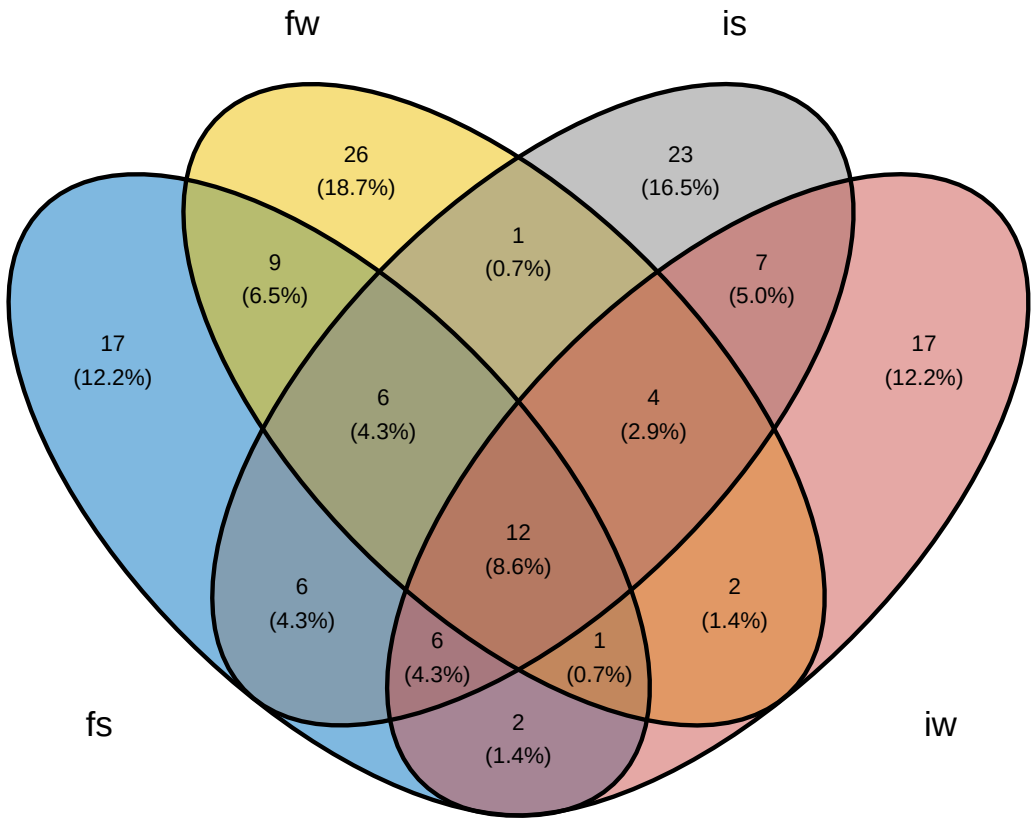
Lexikalische Diversität Einzelsprecher: RUEG

Text DEmo50MD_fwD (formal, schriftlich, monolingual Deutsch, männlich, jugendlich):

- Parkplatz, Parkplatzfläche, Parkfläche, Parkplatzrand, Bürgersteig
- Kleinwagen, Fahrzeug, Auto
- in Richtung Ball, in Richtung der Frau
- Frau mit dem Hund, Frau mit Hund

Lexikalische Diversität Einzelsprecherin: RUEG

Text *DEmo56FD_fwD* (formal, schriftlich, monolingual Deutsch, männlich, jugendlich):



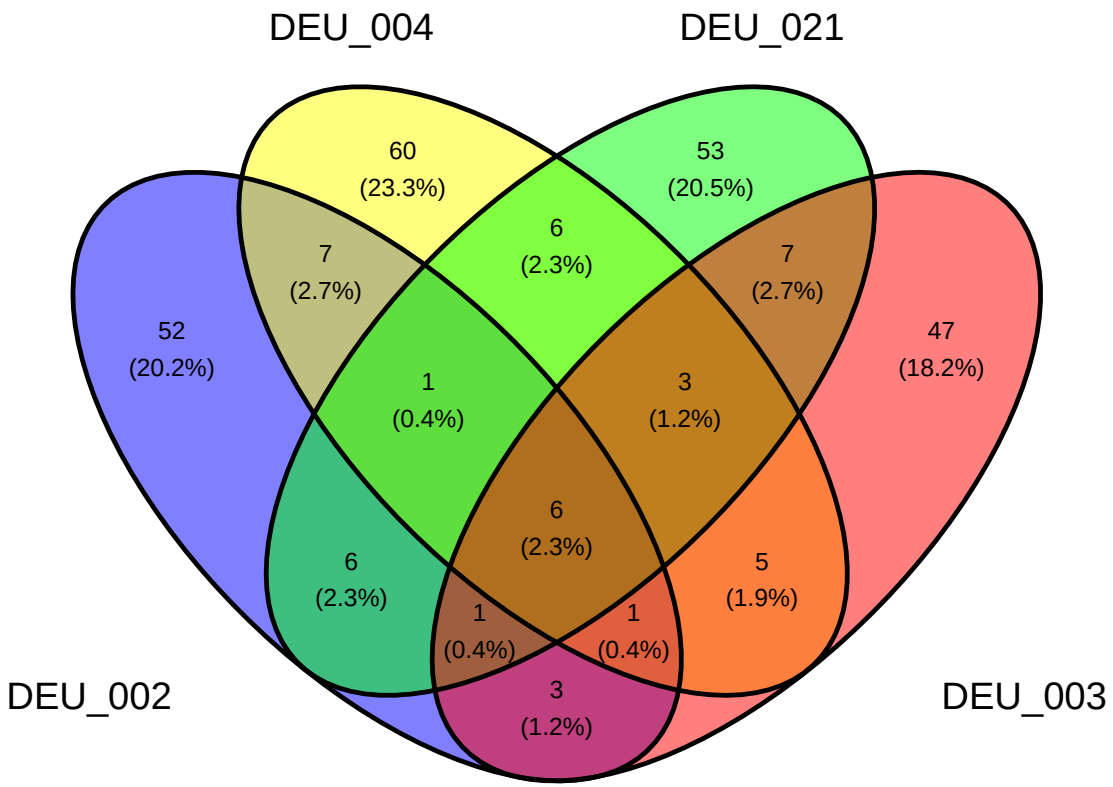
Lexikalische Diversität Einzelsprecherin: RUEG

Text DEmo56FD_fwD (formal, schriftlich, monolingual Deutsch, weiblich, jugendlich)

- Zwölf Wörter kommen in allen vier Settings vor: *Auto, bremsen, Familie, gerade, hinten, Hund, Kleinwagen, kläffen, kommen, Mann, Seite, Straße*
- insgesamt 139 Wörter (hier gezählt: Adjektive, Nomen, Verben ohne haben, sein, werden)
- In jedem Setting ist die Zahl der Lexeme, die nur in diesem Setting vorkommen, 1/4 bis 1/3

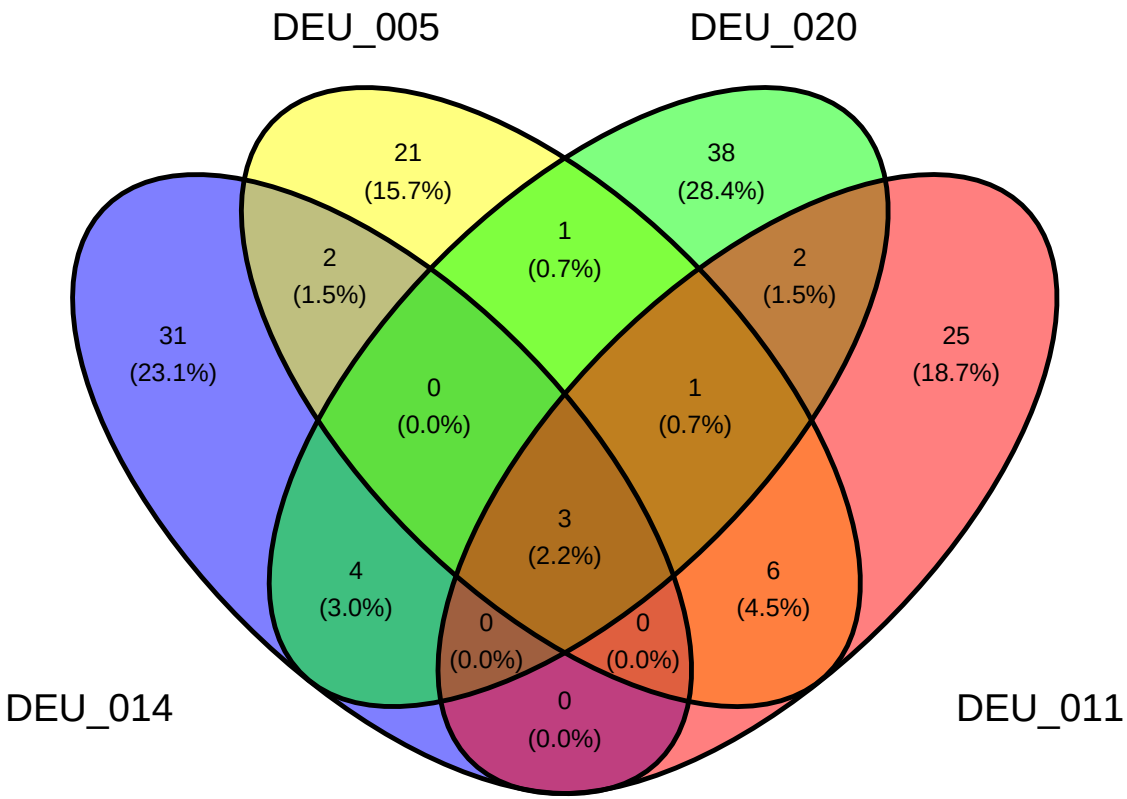
Kobalt L1

Nomen in einer Menge von vier Sprecher:innen



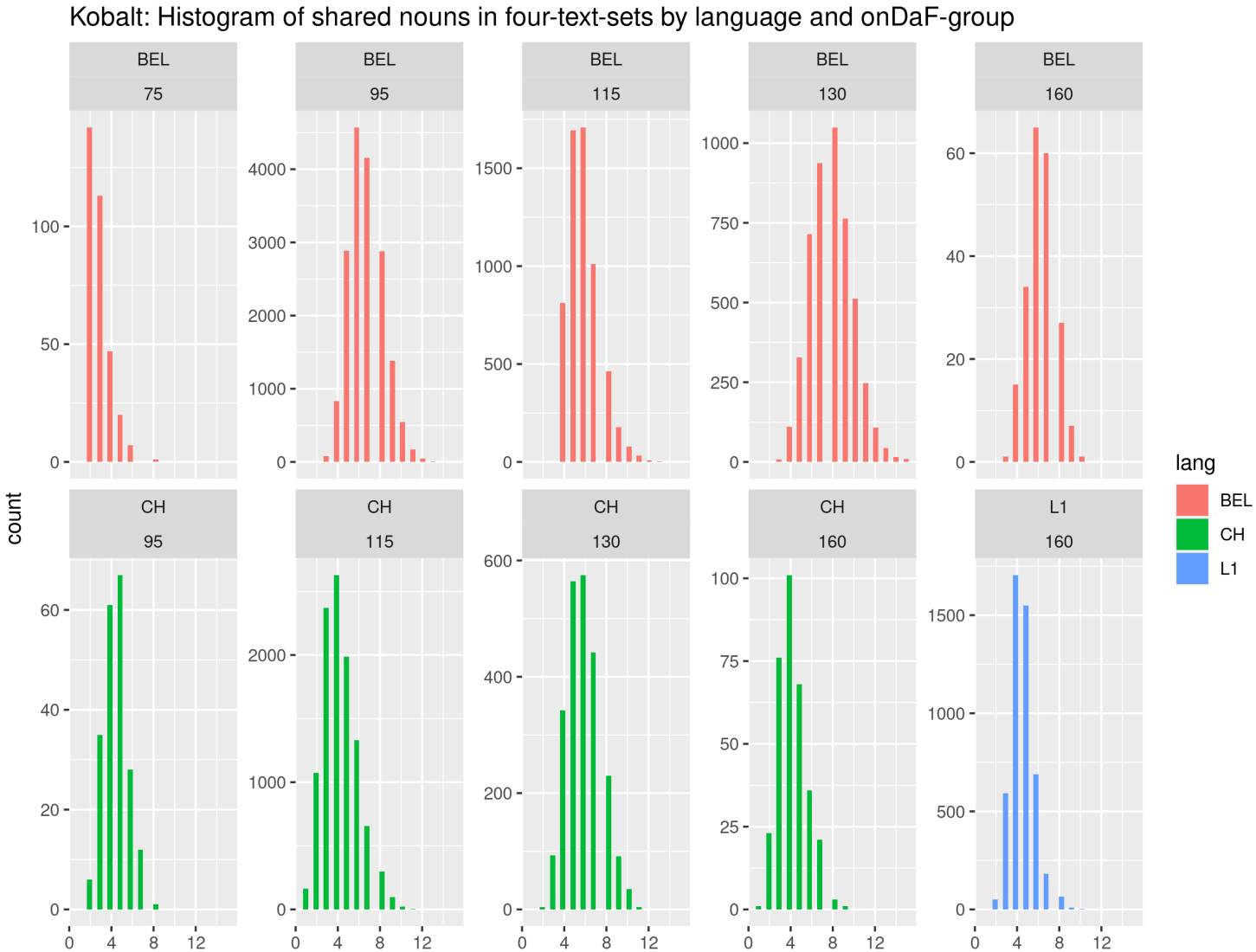
Kobalt L1

Verben in einer Menge von vier Sprecher:innen



Kobalt L1

Für alle Mengen von vier Sprecher:innen



Andere Korpora

- Ähnliches Bild für Falko, BeMaTaC, in Arbeit: BerDiaKo
- ähnlich gefunden schon in Chafe, 1980

Andere Korpora

- Ähnliches Bild für Falko, BeMaTaC, in Arbeit: BerDiaKo
- ähnlich gefunden schon in Chafe, 1980

Sprecher:innen verbalisieren dieselben Situationen im Wesentlichen nicht gleich. Das Lexikon im Gebrauch in aufgabenbasierten Korpora konvergiert im Wesentlichen nicht gegen eine Menge gleicher Lexeme.

Overview

Explanandum

- Wie passt die Beobachtung, dass Gruppen von Sprecher:innen im Wesentlichen *nicht* dieselben lexikalischen Elemente in derselben Situation verwenden zur Annahme, dass Sprache “by default” formelhaft sei?
- Wie schaffen es Sprecher:innen, so unterschiedliche Gruppen von Lexemen zu verwenden, und trotzdem alle zum selben Thema zu schreiben?
- Wie passt die beobachtete Instabilität von Frequenzen (Nicht-Stationarität) in Korpora zur Beobachtung, dass Häufigkeiten und Verteilungen im Lernprozess eine große Rolle spielen?

Explanandum

- Wie passt die Beobachtung, dass Gruppen von Sprecher:innen im Wesentlichen *nicht* dieselben lexikalischen Elemente in derselben Situation verwenden zur Annahme, dass Sprache “by default” formelhaft sei?
- Wie schaffen es Sprecher:innen, so unterschiedliche Gruppen von Lexemen zu verwenden, und trotzdem alle zum selben Thema zu schreiben?
- Wie passt die beobachtete Instabilität von Frequenzen (Nicht-Stationarität) in Korpora zur Beobachtung, dass Häufigkeiten und Verteilungen im Lernprozess eine große Rolle spielen?
- Was ist denn jetzt mit dem Idiom Principle?

Wo liegt der Fehler?

- Methode? (Eine der Beobachtungen war falsch)
- Interpretation? (Die Beobachtungen an sich waren richtig, sind aber falsch interpretiert)
- Theoretische Unschärfe? (Die Beobachtungen und Interpretationen waren richtig, aber sie beschreiben unterschiedliche Teile eines Phänomens)

Wo liegt der Fehler?

- Methode? (Eine der Beobachtungen war falsch)
  - z.B. Zählung bei Erman and Warren

Wo liegt der Fehler?

- Methode? (Eine der Beobachtungen war falsch)
  - z.B. Zählung bei Erman and Warren
- Interpretation? (Die Beobachtungen an sich waren richtig, sind aber falsch interpretiert)
  - Lexikographie vs. Gebrauch (Zoo vs. Farm)

Wo liegt der Fehler?

- Methode? (Eine der Beobachtungen war falsch)
  - z.B. Zählung bei Erman and Warren
- Interpretation? (Die Beobachtungen an sich waren richtig, sind aber falsch interpretiert)
  - Lexikographie vs. Gebrauch (Zoo vs. Farm)
  - Common sense vs. komplexe Auswahl (z.B. n-gram “ist der”, aber auch semantische vorhersehbare Dinge)

Wo liegt der Fehler?

- Methode? (Eine der Beobachtungen war falsch)
  - z.B. Zählung bei Erman and Warren
- Interpretation? (Die Beobachtungen an sich waren richtig, sind aber falsch interpretiert)
  - Lexikographie vs. Gebrauch (Zoo vs. Farm)
  - Common sense vs. komplexe Auswahl (z.B. n-gram “ist der”, aber auch semantische vorhersehbare Dinge)
- Theoretische Unschärfe? (Die Beobachtungen und Interpretationen waren richtig, aber sie beschreiben unterschiedliche Teile eines Phänomens)

Idiom principle als Ausdruck von Wahrnehmungsprototypisierung

z.T. sind Texte wahrscheinlich real lexikalisch nicht so ähnlich, wie man sie wahrnimmt

- Interessante Beobachtung zur Sprachverarbeitung und Kategorisierung von Semantik

Untersuchungsgröße

z.T. sind die berechneten Maße wahrscheinlich auf der falschen Grundgesamtheit berechnet:

- Maße, die das gesamte Lexikon in großen Korpora als Grundgesamtheit verstehen, sind sicher falsch
  - Hängen rechnerisch mehr davon ab, was sonst im Korpus passiert, als von den betreffenden Wortpaaren

Untersuchungsgröße

z.T. sind die berechneten Maße wahrscheinlich auf der falschen Grundgesamtheit berechnet:

- Maße, die das gesamte Lexikon in großen Korpora als Grundgesamtheit verstehen, sind sicher falsch
  - Hängen rechnerisch mehr davon ab, was sonst im Korpus passiert, als von den betreffenden Wortpaaren
  - Nicht linguistisch plausibel („Welches Wort nehme ich denn jetzt…Bundeskanzlerin oder Kaffeetasse? Nein, lieber Senkspreizfußeinlage. Oder grün?")

Untersuchungsgröße

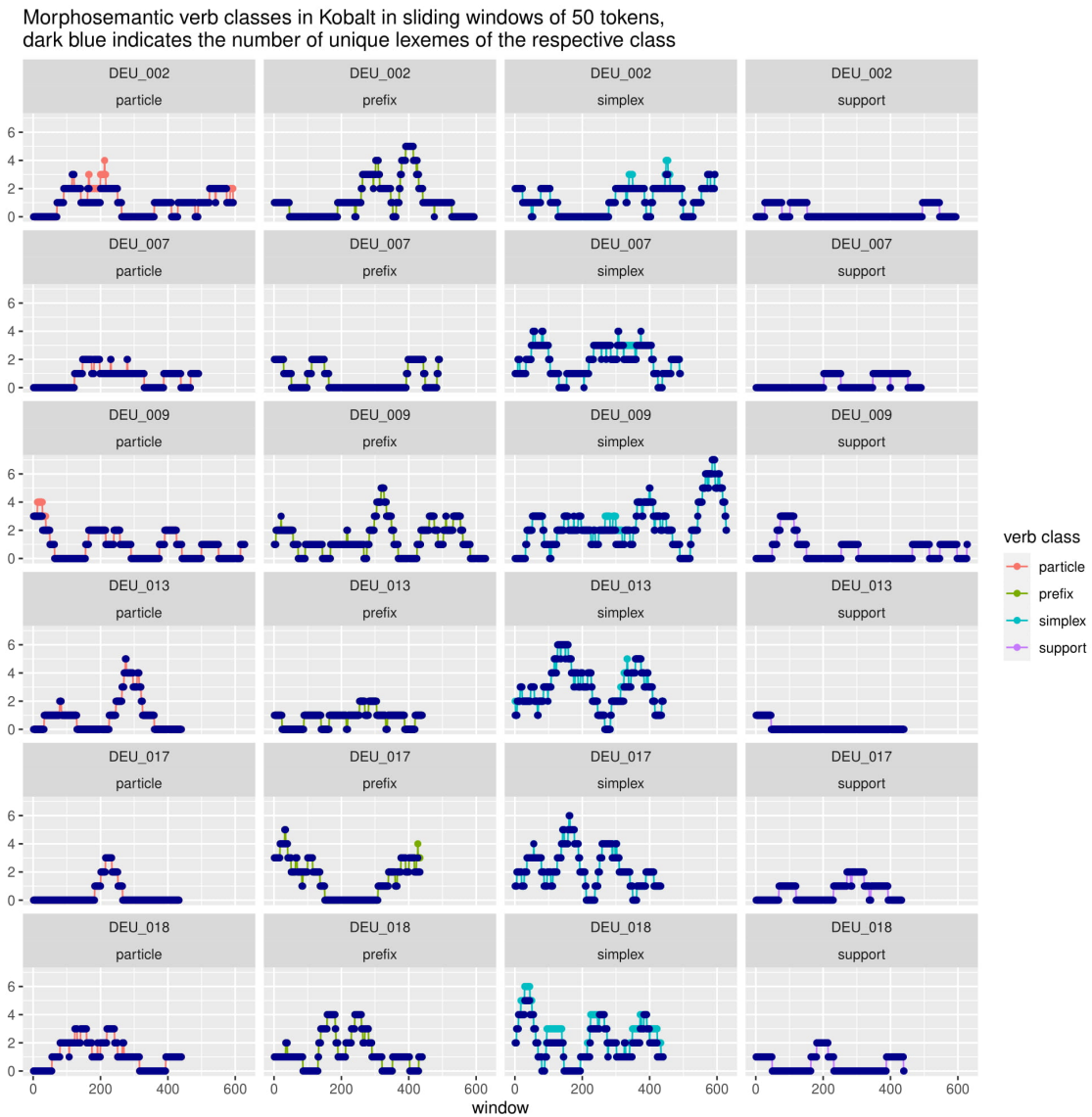
z.T. sind die berechneten Maße wahrscheinlich auf der falschen Grundgesamtheit berechnet:

- Maße, die das gesamte Lexikon in großen Korpora als Grundgesamtheit verstehen, sind sicher falsch
  - Hängen rechnerisch mehr davon ab, was sonst im Korpus passiert, als von den betreffenden Wortpaaren
  - Nicht linguistisch plausibel („Welches Wort nehme ich denn jetzt…Bundeskanzlerin oder Kaffeetasse? Nein, lieber Senkspreizfußeinlage. Oder grün?")
- Textübergreifende Maße sind vielleicht ungünstig (etwa als würde man alle Tiere im Zoo gleichzeitig beschreiben wollen)
- Hinweise dafür:
  - hohe inter- und intraindividuelle Varianz (Shadrova et al., 2021)
  - Pfadabhängigkeit

Zwischenfazit

Solange wir über das gesamte Korpus zählen, sitzen wir zwischen den Stühlen einer vermeintlich lexikographischen Beschreibung und einer Beschreibung des Lexikons im Gebrauch. Denn der Gebrauch ist notwendig individuell und (text)strukturiert.

Pfadabhängigkeit



Pfadabhängigkeit

1. thematisch bedingt
2. diskursbedingt (z.B. Register, semantic progression)
3. kognitiv bedingt (z.B. Priming)

Struktur

- Die Pfadabhängigkeit gibt uns Hinweise auf eine zu Grunde liegende Struktur
- Das bedingt oder interagiert anscheinend auch mit Prozessen der Wortbildung, verhält sich aber anders als in der Syntax (Problem für postulierte Untrennbarkeit von Syntax und Lexikon)

Ongoing & future Research

Process creates structure – structure mediates process:

- Vielleicht ist das Lexikon sowohl kategoriell (z.B. nach Abstraktionsgraden, Prototypizität(*Basic Level Category*, Emberson et al., 2019; Rosch, 1983; Rosch et al., 1976), vielleicht Spezifität) organisiert.
- Diese Struktur hätte wohl Einfluss auf den Schreibfluss (welche Wörter fallen mir zuerst ein)
- Und sie könnte bei unterschiedlichen Sprecher:innengruppen unterschiedlich sein – vielleicht könnte sie sogar erklären, warum L2-Sprecher:innen Probleme mit dem Erwerb von Koselektionsbeschränkungen haben

Overview

Fazit

Das Lexikon ist vor allem eins: extrem variabel.

Fazit

Das Lexikon ist vor allem eins: extrem variabel.

Das macht es zu einem methodologischen Alptraum:

- Datengröße muss enorm sein, um überhaupt gleiche Elemente zu kriegen (Größenordnungen mehr, als wir in kontrolliert erhobenen Korpora je haben werden)
- Vergleichbarkeit ist nicht anhand der eigentlichen Elemente gegeben, nur kategoriell möglich
- Aber: jedes lexikalische Element hat einen spezifischen Eigenwert
  $\rightarrow$ man kann auch nicht einfach so abstrahieren

  Kategorisierung ist extrem schwierig (Synonymie? Unscharfe Kategorien wie Abstraktheit, Spezifität)

Fazit

Das Lexikon ist vor allem eins: extrem variabel.
Das macht es zu einem methodologischen Alptraum:

- Datengröße muss enorm sein, um überhaupt gleiche Elemente zu kriegen (Größenordnungen mehr, als wir in kontrolliert erhobenen Korpora je haben werden)
- Vergleichbarkeit ist nicht anhand der eigentlichen Elemente gegeben, nur kategoriell möglich
- Aber: jedes lexikalische Element hat einen spezifischen Eigenwert $\rightarrow$ man kann auch nicht einfach so abstrahieren

  Kategorisierung ist extrem schwierig (Synonymie? Unscharfe Kategorien wie Abstraktheit, Spezifität)

Allerdings auch spannend, denn Sprecher:innen sind a) real ständig mit dieser Diversität konfrontiert und b) scheint es so etwas wie Koselektionsbeschränkungen ja auch zu geben. Aber wie?

Gebrauch vs. Lexikographie

Im jedem Korpus finden wir eine Auswahl des Lexikongebrauchs von einem oder mehreren Sprecher:innen in bestimmten Kontexten.

Diese Auswahl ist nicht repräsentativ. Beispiele:

1. Ein Zeitungskorpus ist nicht repräsentativ für den In- oder Output einzelner Sprecher:innen, selbst wenn sie Journalist:innen sind
2. Ein Dialogkorpus ist nicht repräsentativ für den In- oder Ouput derselben Sprecher:innen in anderen Dialogkontexten
3. Keins von beiden ist repräsentativ für "das Lexikon als solches"

   individuelles vs. "kollektives" Lexikon (Kollektivbegriff schwierig)

Gebrauch vs. Lexikographie

Im jedem Korpus finden wir eine Auswahl des Lexikongebrauchs von einem oder mehreren Sprecher:innen in bestimmten Kontexten.

Diese Auswahl ist nicht repräsentativ. Beispiele:

1. Ein Zeitungskorpus ist nicht repräsentativ für den In- oder Output einzelner Sprecher:innen, selbst wenn sie Journalist:innen sind
2. Ein Dialogkorpus ist nicht repräsentativ für den In- oder Ouput derselben Sprecher:innen in anderen Dialogkontexten
3. Keins von beiden ist repräsentativ für "das Lexikon als solches"

   individuelles vs. "kollektives" Lexikon (Kollektivbegriff schwierig)

Eine Sammlung verschiedener Korpora kann (vielleicht) *lexikographisch* repräsentativ sein, sie ist es aber nicht für den Gebrauch einzelner Sprecher:innen.

Ongoing & future research

Wir schauen uns stattdessen den Gebrauch kontextualisiert und möglichst vorbei an frequentistischen Methoden an:

- Häufen sich nominale Wortbildungsprozesse innerhalb bestimmter rhetorischer Strukturen in Kobalt? (mit Anke, Julia & Shujun, SFB)
- Wie unterscheidet sich der Gebrauch und die Produktivität von Partikelverben bei Herkunfts-, bilingualen Majoritäts- und monolingualen Sprecher:innen des Deutschen? (mit Anke, Mareike Keller & Gaja, RUEG)
- Gibt es auf unterschiedlichen L2-Erwerbsstufen Unterschiede in der Verteilung der lexikalischen Semantik (z.B. des Abstraktionsgrads und der Spezifität von Nomen) in Kobalt? Lassen sich daraus Unterschiede in der Struktur des Lexikons im Gebrauch ableiten?

Vielen Dank für die Aufmerksamkeit!

anna.shadrova@hu-berlin.de

https://hu.berlin/anna_shadrova

Referenzen I

Bates, E., Bretherton, I., & Snyder, L. (1988). From first words to grammar.

Bates, E., & Goodman, J. C. (1999). On the emergence of grammar from the lexicon. In B. MacWhinney (Ed.), *The emergence of language* (pp. 29–79). Lawrence Erlbaum Associates.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 31–40.

Bybee, J. L. (2013). Usage-based theory and exemplar representations of constructions. *The oxford handbook of construction grammar* (pp. 49–68). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0004

Chafe, W. L. (1980). The pear stories: Cognitive, cultural, and linguistic aspects of narrative production.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied linguistics*, *29*(1), 72–89.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61.

Dąbrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions.

Referenzen II

Dux, R. (2020). *Frame-constructional verb classes: Change and theft verbs in english and german* (Vol. 28). John Benjamins Publishing Company.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality.

Emberson, L. L., Loncar, N., Mazzei, C., Treves, I., & Goldberg, A. E. (2019). The blowfish effect: Children and adults use atypical exemplars to infer more narrow categories during word learning. *Journal of child language*, *46*(5), 938–954.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29–62.

Faulhaber, S. (2011). *Verb valency patterns: A challenge for semantics-based accounts* (Vol. 71). De Gruyter Mouton.

Goldberg, A. E. (2013). Constructionist approaches. In T. Hoffmann & G. Trousdale (Eds.), *The oxford handbook of construction grammar* (pp. 15–31). Oxford University Press.

Granger, S. (2011). From phraseology to pedagogy: Challenges and prospects. *Chunks in the Description of Language. A tribute to John Sinclair*, 123–146.

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, *10*(1), 95–125.

Referenzen III

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*(S1), 228–255. https://doi.org/10.1111/lang.12119

Herbst, T. (2014). The valency approach to argument structure constructions. *Constructions–collocations–patterns*, 167–216.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge. https://doi.org/10.4324/9780203327630

Hunston, S. (2012). *Pattern grammar*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0899

Krenn, B., Evert, S. et al. (2001). Can we do better than frequency? a case study on extracting pp-verb collocations. *Proceedings of the ACL Workshop on Collocations*, 39–46.

MacWhinney, B. (2014). Item-based patterns in early syntactic development. *Constructions, collocations, patterns*, *2562*, 33–69.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins Amsterdam.

Orliac, B., & Dillinger, M. (2003). Collocation extraction for machine translation. *Proceedings of Machine Translation Summit IX*, 292–298.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, *63*, 1–24. https://doi.org/10.1111/j.1467-9922.2012.00735.x

Referenzen IV

Ortega, L. (2015). Second Language Learning Explained? SLA across 10 Contemporary Theories. In B. Van Patten & J. Williams (Eds.), *Theories in second language acquisition - an introduction* (2nd ed., pp. 245–272). Routledge.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. https://doi.org/10.1017/S0267190512000098

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation*, *44*(1-2), 137–158.

Reitter, D. (2008). Context effects in language production: Models of syntactic priming in dialogue corpora.

Rosch, E. (1983). Prototype classification and logical classification: The two systems. *New trends in conceptual representation: Challenges to Piaget's theory*, 73–86.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Shadrova, A. (2020). *Measuring Coselectional Constraint in Learner Corpora: A Graph-based Approach* (PhD thesis). Humboldt-Universität zu Berlin.

Referenzen V

Shadrova, A. (2021a). *Kobalt: Extension Corpus and Annotation Guidelines for Verb Classification and Dependency Adjustments* (Version 1.0). Zenodo. https://doi.org/10.5281/zenodo.5730224

Shadrova, A. (2021b). Topic models do not model topics: Epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, *2021*.

Shadrova, A. (2022). It may be in the structure, not the combinations: Graph metrics as an alternative to statistical measures in corpus-linguistic research. In A. Kuczera & F. Diehr (Eds.), *Proceedings of graph technologies in the humanities 2020*.

Shadrova, A., Linscheid, P., Lukassek, J., Lüdeling, A., & Schneider, S. (2021). A Challenge for Contrastive L1/L2 Corpus Studies: Large Inter- and Intra-Individual Variation Across Morphological, but Not Global Syntactic Categories in Task-Based Corpus Data of a Homogeneous L1 German Group. *Frontiers in Psychology*, *12*, 5267. https://doi.org/10.3389/fpsyg.2021.716485

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Singleton, D., & Leśniewska, J. (2021). Phraseology: Where lexicon and syntax conjoin. *Research in Language and Education: An International Journal [RILE]*, *1*(1), 46–58.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*(2), 251–272.

Referenzen VI

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, *4*(4), 156–163. https://doi.org/10.1016/S1364-6613(00)01462-5

Tomasello, M. (2009). *Constructing a Language. A Usage-based Theory of Language Acquisition*. Harvard university press.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Zeldes, A. (2012). *Productivity in Argument Selection: From Morphology to Syntax* (Vol. 260). De Gruyter Mouton. https://doi.org/10.1515/9783110303919

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., & Skiba, D. (2012). Das wissenschaftliche Netzwerk "Kobalt-DaF". *Zeitschrift für germanistische Linguistik*, *40*(3), 457–458. https://doi.org/0.1515/zgl-2012-0030