

Collo-Phenomena, Lexicography and AI – Leveraging Computational Advances for more Efficient Collocation Candidate Extraction

Phrasalex III, Innsbruck, 12 July 2024

Peter Uhrig, Friedrich-Alexander-Universität Erlangen-Nürnberg

Collo-Phenomena

Collocation, Collostruction, and the company they keep

technical implementation

- usually as a window-based approach: “The usual measure of proximity is a maximum of four words intervening” (Sinclair 1991: 170)
- *...that I will not let another year go by without a carnivorous plant in my office, happily eating away on all these bugs...* (jenniferheller.com)

Collocation Candidate Extraction

length of list

Various techniques to reduce the length of the list of collocation candidates:

- smaller window size
- Part-of-Speech patterns (“Sketch Grammar”)
- syntactic parsing, e.g. dependency annotation

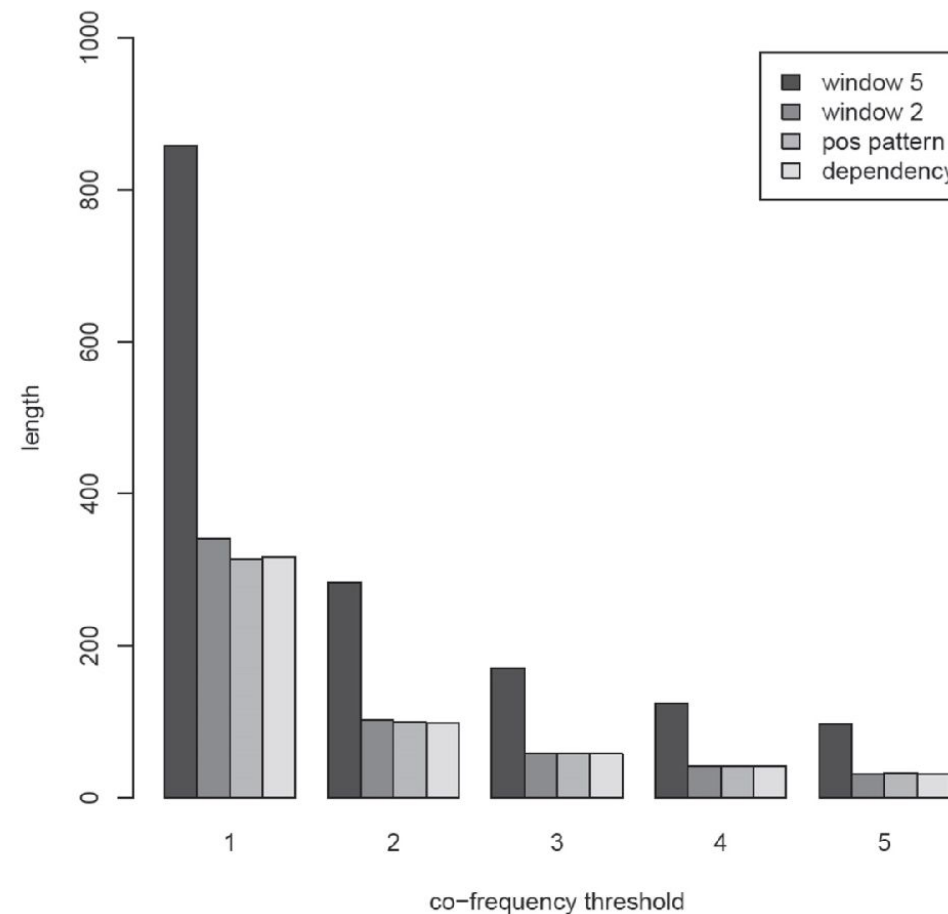


Figure 15: V + N: average length of collocation candidate lists

Parsed Corpora for Collocation Candidate Extraction

Not a new idea

Kenneth Ward Church and Patrick Hanks.
1990. [Word Association Norms, Mutual Information, and Lexicography](#). *Computational Linguistics*, 16(1):22–29.



Patrick Hanks, Erlangen, December 2009

Parsed Corpora for Collocation Candidate Extraction

Not a new idea

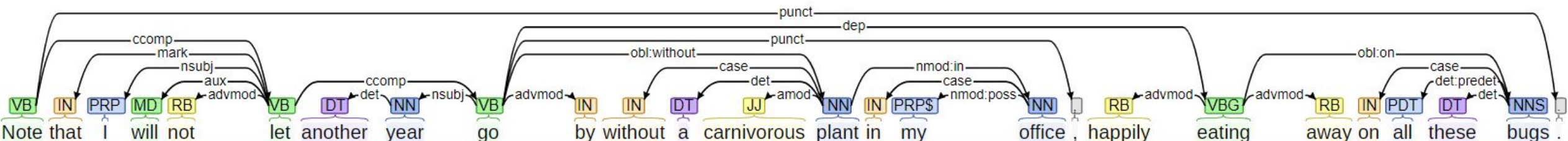
What can you drink?				What can you do to a telephone?			
Verb	Object	Mutual Info	Joint Freq	Verb	Object	Mutual Info	Joint Freq
<i>drink/V</i>	<i>martinis/O</i>	12.6	3	<i>sit_by/V</i>	<i>telephone/O</i>	11.78	7
<i>drink/V</i>	<i>cup_water/O</i>	11.6	3	<i>disconnect/V</i>	<i>telephone/O</i>	9.48	7
<i>drink/V</i>	<i>champagne/O</i>	10.9	3	<i>answer/V</i>	<i>telephone/O</i>	8.80	98
<i>drink/V</i>	<i>beverage/O</i>	10.8	8	<i>hang_up/V</i>	<i>telephone/O</i>	7.87	3
<i>drink/V</i>	<i>cup_coffee/O</i>	10.6	2	<i>tap/V</i>	<i>telephone/O</i>	7.69	15
<i>drink/V</i>	<i>cognac/O</i>	10.6	2	<i>pick_up/V</i>	<i>telephone/O</i>	5.63	11
<i>drink/V</i>	<i>beer/O</i>	9.9	29	<i>return/V</i>	<i>telephone/O</i>	5.01	19
<i>drink/V</i>	<i>cup/O</i>	9.7	6	<i>be_by/V</i>	<i>telephone/O</i>	4.93	2
<i>drink/V</i>	<i>coffee/O</i>	9.7	12	<i>spot/V</i>	<i>telephone/O</i>	4.43	2
<i>drink/V</i>	<i>toast/O</i>	9.6	4	<i>repeat/V</i>	<i>telephone/O</i>	4.39	3
<i>drink/V</i>	<i>alcohol/O</i>	9.4	20	<i>place/V</i>	<i>telephone/O</i>	4.23	7
<i>drink/V</i>	<i>wine/O</i>	9.3	10	<i>receive/V</i>	<i>telephone/O</i>	4.22	28
<i>drink/V</i>	<i>fluid/O</i>	9.0	5	<i>install/V</i>	<i>telephone/O</i>	4.20	2
<i>drink/V</i>	<i>liquor/O</i>	8.9	4	<i>be_on/V</i>	<i>telephone/O</i>	4.05	15
<i>drink/V</i>	<i>tea/O</i>	8.9	5	<i>come_to/V</i>	<i>telephone/O</i>	3.63	6
<i>drink/V</i>	<i>milk/O</i>	8.7	8	<i>use/V</i>	<i>telephone/O</i>	3.59	29
<i>drink/V</i>	<i>juice/O</i>	8.3	4	<i>operate/V</i>	<i>telephone/O</i>	3.16	4
<i>drink/V</i>	<i>water/O</i>	7.2	43				
<i>drink/V</i>	<i>quantity/O</i>	7.1	4				

Church & Hanks (1990)

technical implementation

- usually as a window-based approach: “The usual measure of proximity is a maximum of four words intervening” (Sinclair 1991: 170)
- ...*that I will not let another year go by without a carnivorous plant in my office, happily eating away on all these bugs...* (jenniferheller.com)

Enhanced++ Dependencies:



Collocation Candidate Extraction

Adj + N

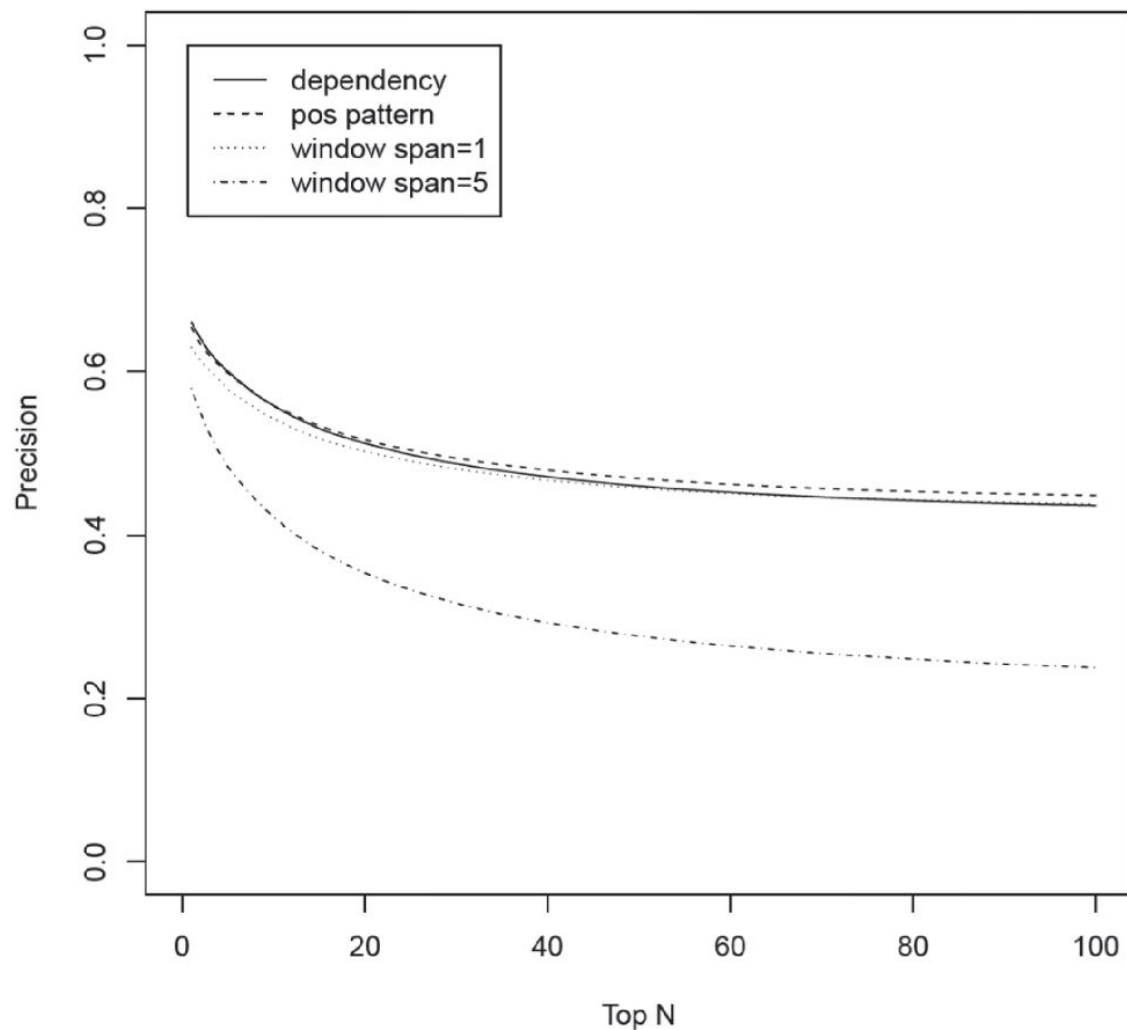


Figure 16: ADJ + N: Precision with t-score; frequency threshold = 5

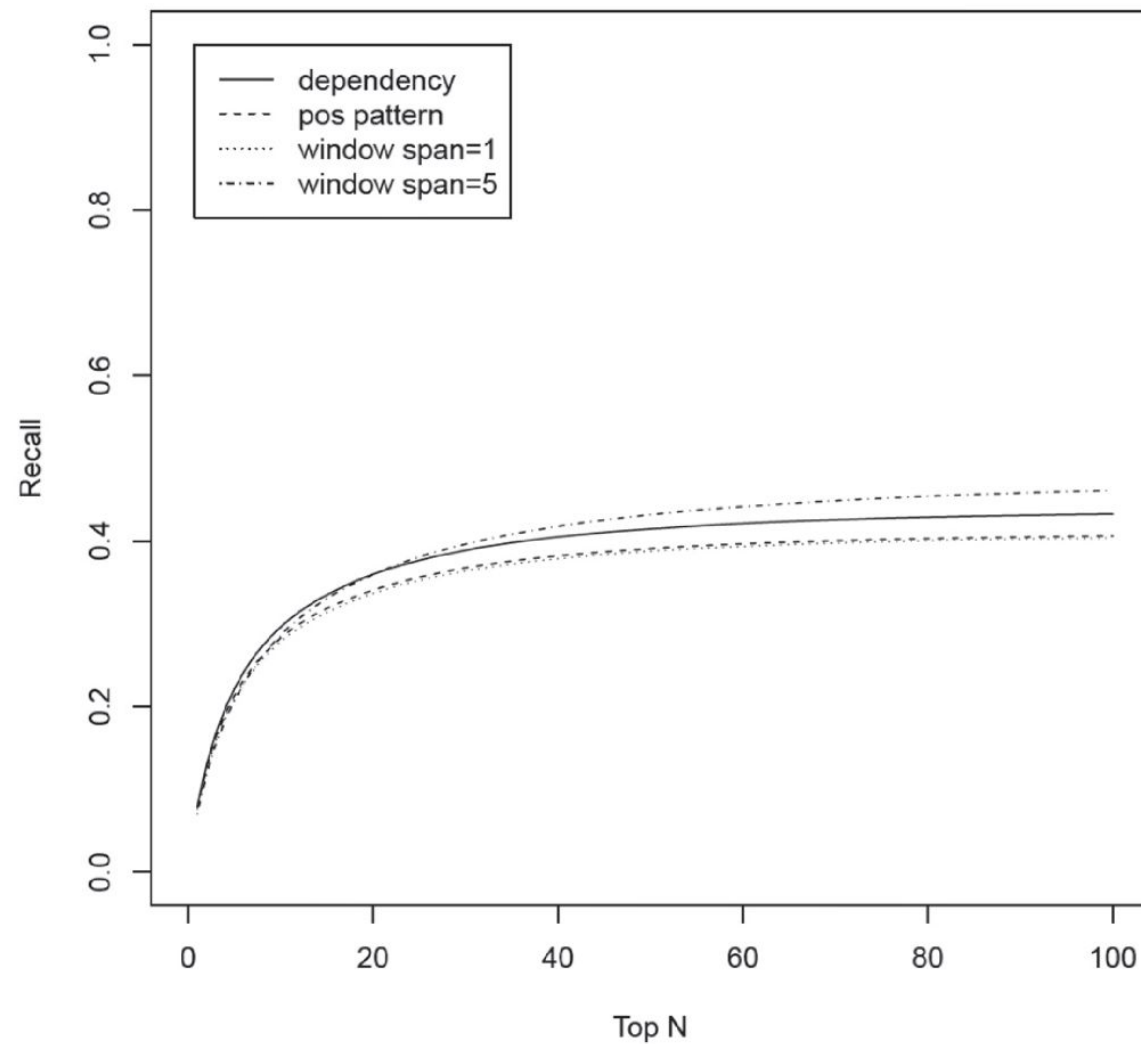


Figure 17: ADJ + N: Recall with t-score; frequency threshold = 5

Collocation Candidate Extraction

N + V

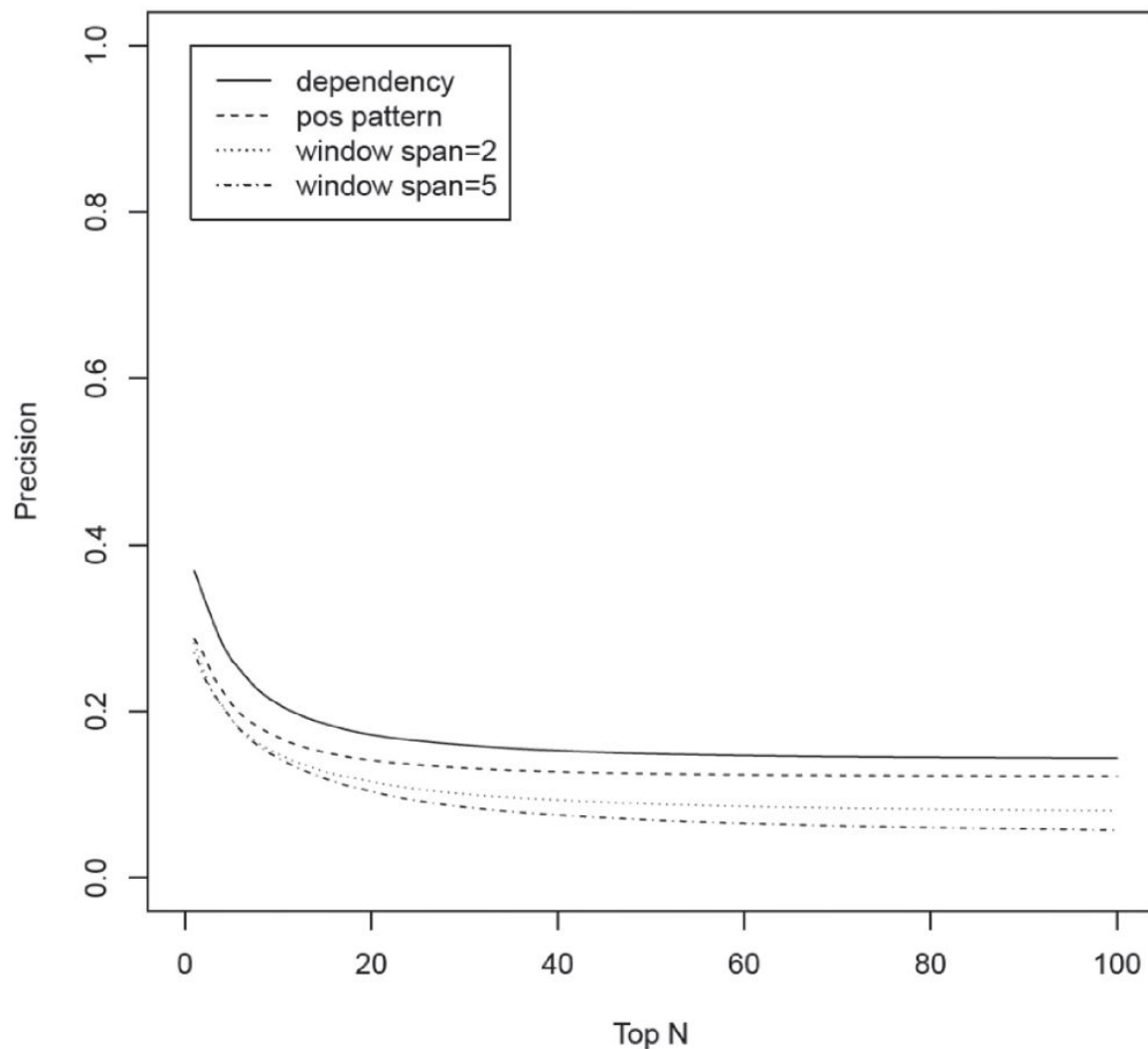


Figure 18: N + V: Precision with t-score; frequency threshold = 5

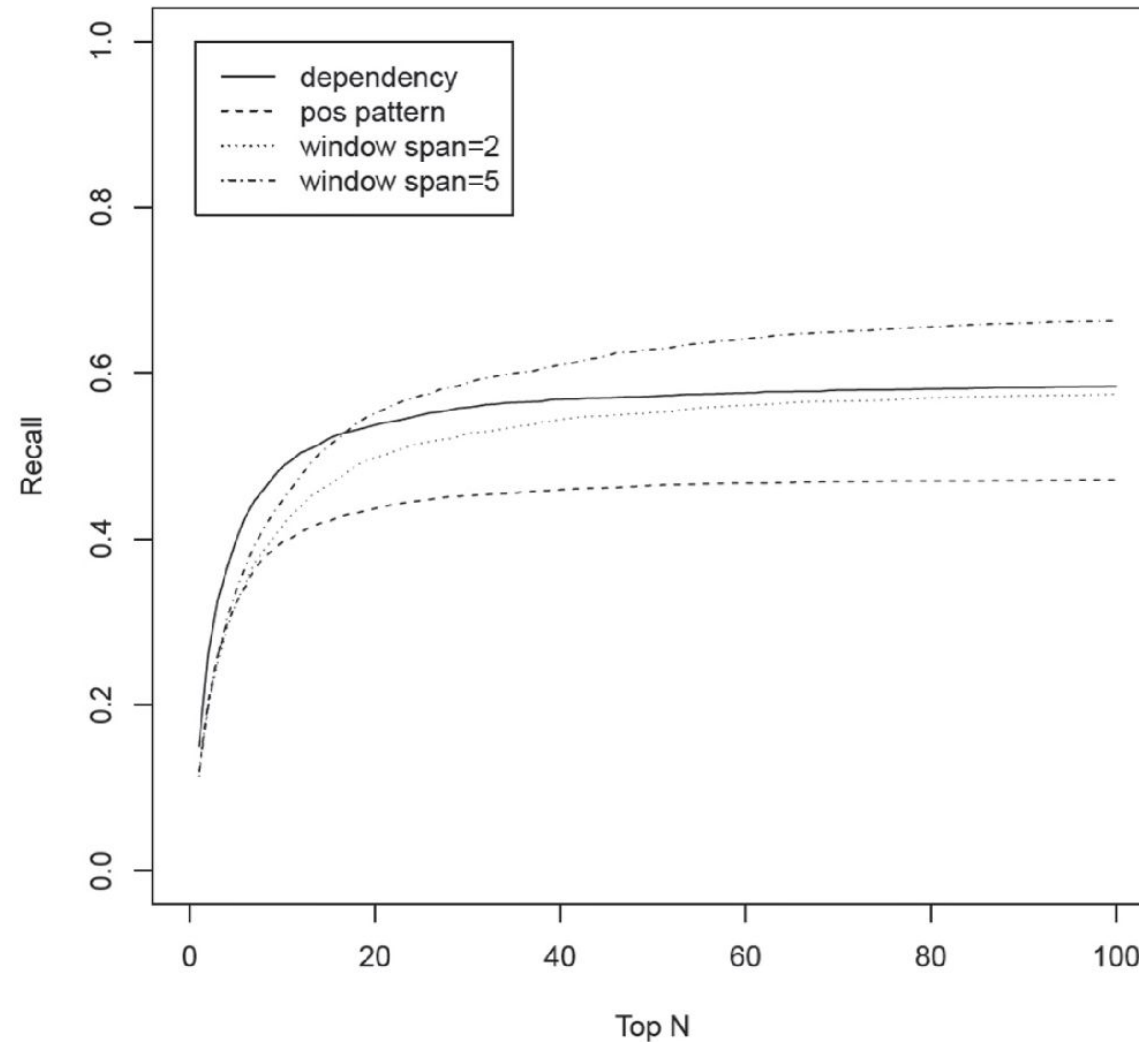


Figure 19: N + V: Recall with t-score; frequency threshold = 5

- substantial differences between parsers & parsing schemes
- spaCy is robust and shows good results for all types of relations
- log-likelihood works well on all relations, but on individual relations, specific association measures are better (e.g. t-score for ADJ-N, Mi_{conf} for V-ADV, ...)

(Uhrig et al. 2018)

E-View-Alation

Visualizing the interaction of gold standards and extraction parameters



- *gold standard*: selected entries from the BBI and OCD2 collocation dictionaries (Bartsch224 set)
 - *parameters*: 20 association measures, 13 corpora, 8 context sizes, 4 frequency thresholds
- (Evert et al. 2017)

BBI and OCD2

measure for *measure*

measure I *n.* 1. a cubic; dry; liquid; metric ~ 2. a tape ~ 3. in a certain ~ (in large ~) 4. (misc.) for good ~ ('as smt. extra'); made to ~ ('custom-made'); to take smb.'s ~ ('to evaluate smb.') (see also **measures**)

measure II *v.* 1. (d; tr.) to ~ against (to ~ one's accomplishments against smb. else's) 2. (P; intr.) the room ~s twenty feet by ten

measures *n.* 1. to carry out, take ~ 2. coercive; compulsory; draconian; drastic, harsh, stern, stringent; emergency; extreme, radical; preventive, prophylactic; safety, security; stopgap, temporary ~ 3. ~ to + inf. (we took ~ to insure their safety) 4. ~ against (to take ~ against smuggling)

BBI (Benson et al. 1986)

measure *noun*

1 official action to deal with a problem

ADJ. appropriate, effective, necessary, practical ◊ We urge you to adopt all necessary ~s to guarantee people's safety. | key | extraordinary, special ◊ We had to resort to extraordinary ~s to find employees. | simple | desperate, draconian, drastic, extreme, harsh, radical, repressive, strong, tough | emergency, urgent (esp. BrE) | interim, short-term, stopgap, temporary | additional | defensive, precautionary, preventative, preventive | disciplinary, punitive | corrective, remedial | protective, safety, security ◊ New security ~s were implemented to prevent further violence. | conservation, control ◊ Development of new water sources needs to be combined with conservation ~s. | austerity, cost-cutting, economy (esp. BrE), efficiency | economic, policy | government

MEASURE + VERB employ, implement, impose, institute, introduce, take, use ◊ Special ~s are being taken to protect the local water supplies. ◊ The authorities are using increasingly repressive ~s. | adopt, approve, enact, pass ◊ The Committee unanimously approved the ~. | defeat, oppose, veto ◊ The mayor threatened to veto a ~ passed by the city council. | propose, suggest

MEASURE + VERB be aimed at sth, be designed to, be intended to

PREP. ~ against ◊ tougher ~s against racism | ~ for ◊ ~s for reducing delays

2 amount/quantity of sth

ADJ. broad, considerable, fair, generous, great, significant, substantial, wide

PREP. ~ of ◊ He poured me a generous ~ of gin.

PHRASES in considerable, large, some, etc. ~ ◊ His success was due in large ~ to your help. | in equal ~ ◊ He's been praised and condemned in equal ~. | in no small ~

3 unit of size/quantity

ADJ. accurate, direct, fair, objective, precise ◊ This figure alone is not a fair ~ of our success. | broad ◊ GDP is considered the broadest ~ of a country's economic activity. | indirect | important ◊ The price of housing relative to income is an important ~ of real income. | quantitative, statistical ◊ They tried to formulate a quantitative ~ of well-being. | standard, standardized ◊ Higher scores on this standardized ~ indicate greater creativity. | imperial, metric | behavioural/behavioral (esp. AmE), performance ◊ Companies can use their stock price as a performance ~.

VERB + MEASURE calculate, derive, obtain ◊ This ~ is obtained by dividing corporate profits by corporate bond yields. | provide ◊ This figure provides an objective ~ of risk.

PREP. ~ of ◊ an accurate ~ of length

PHRASES weights and ~s

4 sign of sth

ADJ. crude, reliable, simple, true, useful, valid ◊ A reliable ~ of progress is whether your children can do something they couldn't do before. | important ◊ Accepting the lower salary was seen as an important ~ of commitment.

PREP. ~ of ◊ Landed income was the true ~ of the gentry.

5 (AmE) in MUSIC → See also BAR
ADJ. first, opening ◊ The band began playing the opening ~s.
VERB + MEASURE hear, play, sing ◊ He placed his fingers on the keys and played a few ~s.
PREP. ~ of ◊ the first few ~s of Mozart's third violin concerto | per ~, to a/the ~ ◊ a steady rhythm of four beats to a ~

NOTE

Weights and measures

... of ◊ I always drink gallons of water.
half a ..., a quarter of a ... ◊ half a pound of ham
a half/quarter hour, inch, mile, ounce, pint, pound ◊ a quarter pound of cheese
cubic ..., square ... ◊ a maximum flow of 3300 cubic feet a second
... square ◊ The room is about 25 feet square.
... broad, deep, high, long, tall, thick, wide ◊ The new dock was 230 m long and 92 m wide.
... bigger, cooler, faster, heavier, lighter, slower ◊ The climate was several degrees warmer than it is now.
about ..., approximately ..., around ... ◊ I kilogram = approx. 2.2 pounds
be ..., cover ..., measure ..., span ..., stretch (for) ... (used with measures of distance and area) ◊ The National Park covers 3000 acres. ◊ The sandy beach stretches for over four miles.
be ..., weigh ... (used with measures of weight) ◊ She weighed over 200 pounds.
be ..., last ..., take ... (used with measures of time) ◊ It takes approximately 365 and a quarter days for the earth to revolve around the sun.
in a ... ◊ How many inches are there in a foot?
in ... ◊ We were asked to estimate the temperature of the room in degrees centigrade.
to a/the ... ◊ My car does 25 miles to the gallon.
a ..., per ... ◊ They're \$4.50 a dozen. ◊ ten ounces of platinum per ton of ore
by the ... ◊ Apples are sold by the kilogram.
of ... ◊ The path will be built to a width of two and a half feet.
... in area, length, size, volume, weight ◊ Killer whales are up to 30 feet in length.
... by ... ◊ a huge room measuring 50 m by 18 m
to the nearest ... ◊ Give your answer to the nearest mile.

measure *verb*

1 find the size of sth

ADV. accurately, exactly, precisely, reliably | empirically, objectively, quantitatively ◊ Education policy places too much emphasis on things that can be quantitatively ~d. | carefully | directly ◊ Any type of data that could not be directly ~d was rejected. | indirectly | easily | experimentally | up ◊ We need to ~ up the room for a new carpet.

VERB + MEASURE be able to, can ◊ You can now ~ its length more accurately. | be easy to, be possible to | be difficult to, be hard to, be impossible to

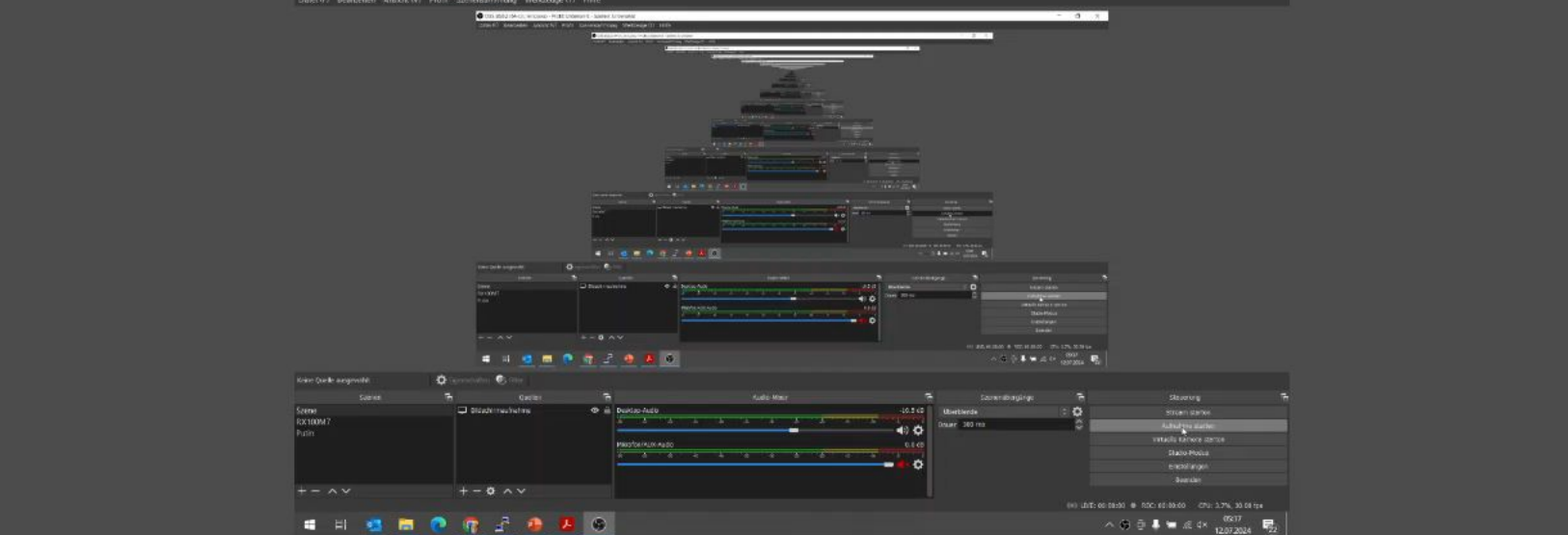
PREP. for ◊ She's being ~d for her wedding outfit. | in ◊ Cloth is ~d in yards.

2 judge the importance/value/effect of sth

ADV. easily ◊ The policy's impact cannot be easily ~d. | effectively | objectively

VERB + MEASURE can | be easy to, be possible to | be difficult to, be hard to, be impossible to ◊ It is hard to ~ the benefits to society of the system. | be used to ◊ the criteria that are used to ~ performance

PREP. according to ◊ It is really possible to ~ the skills of such jobs according to objective standards? | against ◊ The school's performance is ~d against a strict set of criteria. | by ◊ The policy's effectiveness cannot be ~d by numbers alone. | in terms of ◊ Success cannot be ~d merely in terms of the size of your salary.



Keine Quelle ausgewählt

Eigenschaften Filter

Szenen

Szene
RX100M7
Putin

Quellen

Bildschirmaufnahme

Audio-Mixer

Desktop-Audio -10.5 dB

Mikrofon/AUX-Audio 0.0 dB

Szenenübergänge

Überblende

Dauer 300 ms

Steuerung

Stream starten

Aufnahme starten

Virtuelle Kamera starten

Studio-Modus

Einstellungen

Beenden

Crossmodal Collostructions

Patterns in multimodal communication

Collostructional analysis always starts with a particular **construction** and investigates which **lexemes** are strongly attracted or repelled by a particular **slot** in the construction (i.e. occur more frequently or less frequently than expected); crucially, such 'slots' can exist at different levels of linguistic structure (for example, the ditransitive construction may be said to have four slots corresponding to the subject, the verb, and the indirect and direct objects, and the past-tense construction may be said to have a slot corresponding to the verb occurring in the past tense). Lexemes that are attracted to a particular construction are referred to as *collexemes* of this construction; conversely, a construction associated with a particular lexeme may be referred to as a *collostruct*; the combination of a collexeme and a collostruct will be referred to as a *collostruction*. (Stefanowitsch/Gries 2003: 214f)

Collostructional analysis always starts with a particular **construction** and investigates which **lexemes** are strongly attracted or repelled by a particular **slot** in the construction (i.e. occur more frequently or less frequently than expected).

The combination of two arbitrary constructions on arbitrary levels of representation that occur significantly more frequently together than expected.

be said to have a slot corresponding to the verb occurring in the past tense). Lexemes that are attracted to a particular construction are referred to as *collexemes* of this construction; conversely, a construction associated with a particular lexeme may be referred to as a *collostruct*; the combination of a collexeme and a collostruct will be referred to as a *collostruction*. (Stefanowitsch/Gries 2003: 214f)

Crossmodal Collostructions

Head moving horizontally only – Interjections

lemma	cofreq	p-value	collostructional strength	odds ratio	attracted/repelled
no	36254	= 0.001	3.120	1.022	attracted
uh-uh	43	= 0.292	0.535	1.139	attracted
eh	105	= 0.393	0.406	1.043	attracted
yes	28277	< 2.20E-16	64.797	0.881	repelled
oh	21758	< 2.20E-16	52.125	0.879	repelled
okay	20567	< 2.20E-16	48.695	0.879	repelled
hello	2436	< 2.20E-16	39.199	0.726	repelled
hi	4210	< 2.20E-16	36.901	0.787	repelled
yeah	43019	< 2.20E-16	36.512	0.925	repelled
wow	5415	< 2.20E-16	25.073	0.839	repelled
hey	7689	= 8.24E-13	12.084	0.903	repelled
ok	1223	= 1.23E-11	10.911	0.792	repelled
yep	932	= 5.80E-09	8.237	0.795	repelled
please	2288	= 1.77E-07	6.751	0.875	repelled
man	12695	= 2.69E-06	5.571	0.950	repelled
wham	28	= 5.55E-06	5.256	0.410	repelled
ah	609	= 2.32E-05	4.635	0.815	repelled
aw	118	= 0.000	3.626	0.680	repelled
nope	298	= 0.011	1.961	0.846	repelled
bam	139	= 0.019	1.726	0.801	repelled
uh	771	= 0.044	1.360	0.924	repelled
nah	86	= 0.080	1.097	0.822	repelled
amen	185	= 0.082	1.089	0.876	repelled
oops	116	= 0.089	1.049	0.850	repelled
damn	166	= 0.376	0.425	0.964	repelled

Crossmodal Collostructions

Uh-uh



Crossmodal Collostructions

Head moving vertically only – Interjections

lemma	cofreq	p-value	collostructional strength	odds ratio	attracted/repelled
yes	2881	< 2.20E-16	53.555	1.361	attracted
okay	1931	< 2.20E-16	19.985	1.249	attracted
yeah	3430	= 4.88E-07	6.312	1.090	attracted
oh	1829	= 4.73E-06	5.325	1.113	attracted
hi	416	= 5.22E-05	4.283	1.222	attracted
yep	102	= 0.002	2.770	1.369	attracted
ok	128	= 0.003	2.578	1.303	attracted
bam	21	= 0.005	2.310	1.928	attracted
wow	464	= 0.023	1.643	1.101	attracted
eh	13	= 0.030	1.530	1.859	attracted
please	198	= 0.036	1.444	1.144	attracted
hello	232	= 0.059	1.229	1.114	attracted
aw	16	= 0.069	1.158	1.533	attracted
wham	7	= 0.075	1.127	1.969	attracted
man	952	= 0.119	0.925	1.040	attracted
hey	598	= 0.144	0.843	1.046	attracted
nah	10	= 0.147	0.833	1.491	attracted
amen	17	= 0.246	0.610	1.218	attracted
no	2537	= 0.249	0.603	1.014	attracted
nope	26	= 0.300	0.523	1.128	attracted
oops	10	= 0.409	0.388	1.118	attracted
damn	13	= 0.409	0.388	1.098	attracted
ah	50	= 0.428	0.369	1.034	attracted
uh-uh	1	= 0.229	0.640	0.352	repelled
uh	54	= 0.396	0.402	0.953	repelled

Crossmodal Collostructions

Correct recognition – possible nod before *No*



Crossmodal Collostructions

baldonado?!



Leveraging LLMs for Collocation Dictionaries

Warning: work in progress, very early stages

Prompt: List collocates of the noun "takeover".

Model Response:

Sure, here are some common phrases that use the word "takeover":

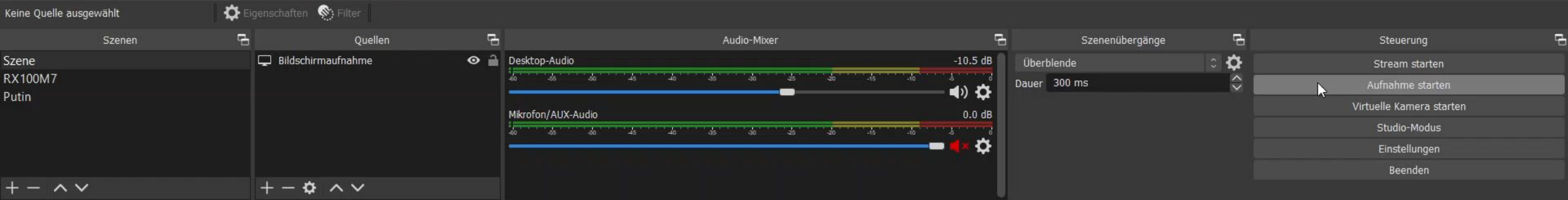
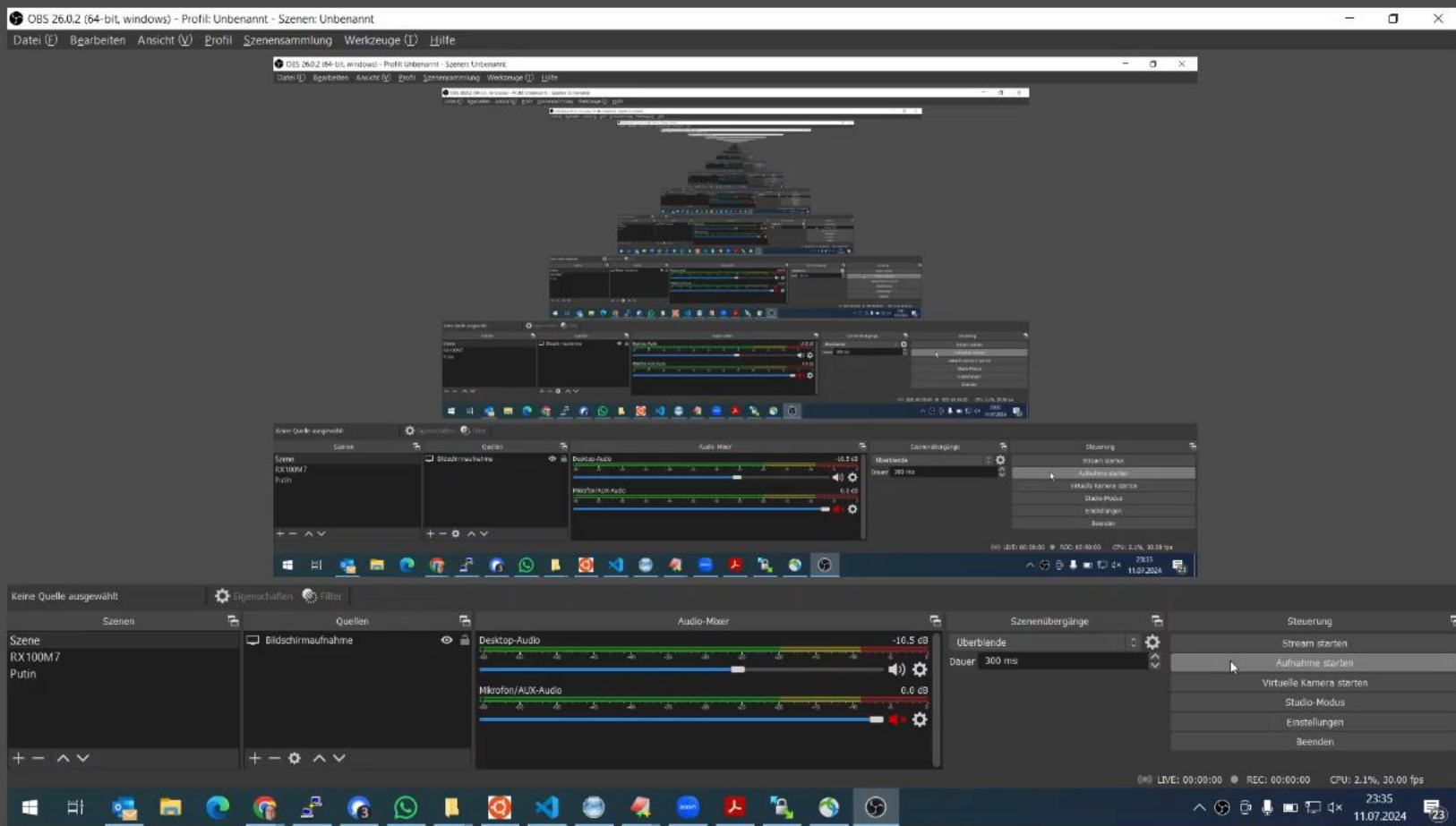
1. **Hostile takeover:** This refers to a type of corporate acquisition where the target company's management resists the buyout attempt.
2. **Mergers and acquisitions (M&A):** Takeover is often used as a synonym for acquisition, which involves one company buying another.
3. **Corporate takeover:** A company gains control over another by acquiring its stock or assets.
4. **Asset takeover:** One company takes over another company's assets, but not necessarily its liabilities.
5. **Leveraged buyout (LBO):** A financial transaction in which a company is purchased with a combination of equity and significant amounts of borrowed money, structured in such a way that the company's cash flow or assets are used to secure and repay the borrowed money.
6. **Management buyout (MBO):** When a company's existing managers acquire all or part of the business from their current owners.
7. **Friendly takeover:** In this scenario, both parties involved agree on the terms of the sale, and there is no resistance from the target company's management.
8. **Acquisition bid:** An offer made by one company to purchase another.
9. **Bidder:** The company making an acquisition bid.
10. **Target:** The company being acquired.

Prompt: List the collocations of the word "leg" and group the collocates semantically in the same way as it is done in the Oxford Collocations Dictionary for Students of English.

Model Response:

Certainly! Here are some common collocates of the word "leg" organized into semantic groups based on the Oxford Collocations Dictionary for Students of English:

1. **Parts of the leg:** *ankle, calf, foot, hip, knee, shin, thigh*.
 - Examples: She twisted her ankle while running. He has strong calves from years of cycling.
2. Length/height of legs: *long, short, slender, spindly, knobbly*.
 - Examples: She inherited her father's long legs. His short legs made him look even smaller.
3. Leg movements: *cross, fold, stretch, swing, kick, jump, limp, march, run, straddle, walk*.
 - Examples: She crossed her legs at the knee. He folded his arms and legs defensively.
4. Supporting oneself using legs: *balance, brace, climb, perch, rest, stand, support*.
 - Examples: He balanced precariously on one leg. She braced herself against the wall with her hands and feet.
5. Clothing for legs: *boot, jeans, pants, skirt, stockings, sock, trouser*.
 - Examples: She wore black jeans and boots. He put on socks before slipping on his shoes.
6. Conditions related to legs: *aching, cramp, injured, lame, numb, paralyzed, painful, restless, sprained, tired*.
 - Examples: Her legs were aching after the hike. He had a cramp in his calf muscle.
7. Other collocates: *break, bruise, cut, limb, stub, twist, wound*.
 - Examples: She broke her leg skiing last winter. He stubbed his toe on the coffee table.



Oxford Collocations Dictionary for Students of English, 2nd Edition, CD-ROM



```
1 <?xml:version="1.0"?>
2 <entry:gen="true".id="a_#gen_e".guid="00000000"><h:gen="1">A</h><p:p="n"/><z_p>.noun</z_p><note-g><note.type="adj_n"><eb>A</eb>.is used before these nouns: .<Ref.topic="grade_n_e"><xr:
href="grade_n_e"><xh>grade</xh></xr></Ref><z>,</z><Ref.topic="string_n_e"><xr:href="string_n_e"><xh>string</xh></xr></Ref><z>,</z><Ref.topic="team_n_e"><xr:
href="team_n_e"><xh>team</xh></xr></Ref><z>,</z><Ref.topic="vitamin_n_e"><xr:href="vitamin_n_e"><xh>vitamin</xh></xr></Ref></note></note-g></entry>
3
4 <?xml:version="1.0"?>
5 <entry:gen="true".id="a3_#gen1_e".guid="00000001"><h:gen="2">A3</h><p:p="n"/><z_p>.noun</z_p><note-g><note.type="adj_n"><eb>A3</eb>.is used before these nouns: .<Ref.topic="paper_n_e"><xr:
href="paper_n_e"><xh>paper</xh></xr></Ref><z>,</z><Ref.topic="sheet_n_e"><xr:href="sheet_n_e"><xh>sheet</xh></xr></Ref></note></note-g></entry>
6
7 <?xml:version="1.0"?>
8 <entry:gen="true".id="a4_#gen1_e".guid="00000002"><h:gen="2">A4</h><p:p="n"/><z_p>.noun</z_p><note-g><note.type="adj_n"><eb>A4</eb>.is used before these nouns: .<Ref.topic="paper_n_e"><xr:
href="paper_n_e"><xh>paper</xh></xr></Ref><z>,</z><Ref.topic="sheet_n_e"><xr:href="sheet_n_e"><xh>sheet</xh></xr></Ref></note></note-g></entry>
9
10 <?xml:version="1.0"?>
11 <entry:gen="true".id="aa_#gen_e".guid="00000003"><h:gen="1">AA</h><p:p="n"/><z_p>.noun</z_p><note-g><note.type="adj_n"><eb>AA</eb>.is used before these nouns: .<Ref.topic="battery_n_e"><xr:
href="battery_n_e"><xh>battery</xh></xr></Ref><z>,</z><Ref.topic="meeting_n_e"><xr:href="meeting_n_e"><xh>meeting</xh></xr></Ref></note></note-g></entry>
12
13 <?xml:version="1.0"?>
14 <entry:gen="true".id="aaa_#gen_e".guid="00000004"><h:gen="1">AAA</h><p:p="n"/><z_p>.noun</z_p><note-g><note.type="adj_n"><eb>AAA</eb>.is used before these nouns: .<Ref.topic="battery_n_e"><xr:
href="battery_n_e"><xh>battery</xh></xr></Ref></note></note-g></entry>
15
16 <?xml:version="1.0"?>
17 <entry:id="abandon_v_e".guid="00000005"><h:
id="abandon_v_1">abandon</h><infl><inflection>abandon</inflection><inflection>abandons</inflection><inflection>abandoned</inflection><inflection>abandoning</inflection></infl><p:p="v"/><z_p>.
verb</z_p><n-g:n="1"><z_n>1</z_n><z>.</z><d>leave.sbsth</d><sl-g:sl="adv"><z_title>adverb.</z_title><sb-g:sb="a"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">hastily</cl><x>The town had
been hastily .</x></sb-g></sl-g><sl-g:sl="phrases"><z_title>phrases.</z_title><sb-g><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:idx="na".parent="phrases">~.sb.to their fate</cl></sb-g><sb-g:
sb="a"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:idx="na".parent="phrases">be left .</cl></sb-g><sb-g:sb="a"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:idx="na".parent="phrases">be found
.</cl><x>The car was found .</x></sb-g></sl-g><sl-g><n-g:n="2"><z_n>2</z_n><z>.</z><d>stop.<lookup.lookup="doing">doing</lookup></lookup.lookup="supporting">supporting</lookup>.
sth</d><sl-g:sl="adv"><z_title>adverb.</z_title><sb-g:sb="b"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">altogether</cl><z_cl>,</z_cl><z>.</z><cl:
parent="adv">completely</cl><z_cl>,</z_cl><z>.</z><cl:parent="adv">entirely</cl><z_cl>,</z_cl><z>.</z><cl:parent="adv">totally</cl><x>The government does not propose to ~ the project
altogether.</x></sb-g><sb-g:sb="b"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">effectively</cl><z_cl>,</z_cl><z>.</z><cl:parent="adv">largely</cl><z_cl>,</z_cl><z>.</z><cl:
parent="adv">virtually</cl><x>This principle has now been effectively .</x></sb-g><sb-g:sb="f"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">eventually</cl><z_cl>,</z_cl><z>.</z><cl:
parent="adv">finally</cl></sb-g><sb-g:sb="e"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">simply</cl><x>Traditional policies were simply .</x></sb-g><sb-g:sb="c"><z_sb-g><img:
src="bullet.png"/></z_sb-g><cl:parent="adv">formally</cl></sb-g><sb-g:sb="d"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">abruptly</cl><z>.</z><g="espa"/><z_g><it>esp.
AmE</it></z_g><z>.</z><z_cl>,</z_cl><z>.</z><cl:parent="adv">immediately</cl><z_cl>,</z_cl><z>.</z><cl:parent="adv">quickly</cl></sb-g><sb-g:sb="d"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:
parent="adv">gradually</cl></sb-g><sb-g:sb="d"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">quietly</cl><x>The plans for reform were quietly .</x></sb-g><sb-g:sb="f"><z_sb-g><img:
src="bullet.png"/></z_sb-g><cl:parent="adv">temporarily</cl></sb-g><sb-g:sb="f"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:parent="adv">voluntarily</cl></sb-g></sl-g><sl-g:
sl="verbandhwd"><z_title>verb.+ abandon.</z_title><sb-g:sb="a"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:pos="pvpart".parent="vah".idx="n">be forced to</cl><z_cl>,</z_cl><z>.</z><cl:pos="pvpart".
parent="vah".idx="n"><lookup.lookup="have_to">have to</lookup></cl></sb-g><sb-g:sb="a"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:pos="pvpart".parent="vah".idx="n">decide to</cl></sb-g></sl-g><sl-g:
sl="prep"><z_title>preposition.</z_title><sb-g:sb="a"><z_sb-g><img:src="bullet.png"/></z_sb-g><cl:idx="na".parent="prep">for</cl><x>He .</x></sb-g><sb-g><z_sb-g><img:
src="bullet.png"/></z_sb-g><cl:idx="na".parent="prep">in.</lookup.lookup="favour">favour</lookup></lookup.lookup="favor">favor</lookup>.of</cl><x>She .</x></sb-g></sl-g></n-g>
18 </entry>
19 </xml>
```

```
<entry id="abandon_v_e" guid="000000005">
  <h id="abandon_v_1">abandon</h>
  <infl>
    <inflection>abandon</inflection>
    <inflection>abandons</inflection>
    <inflection>abandoned</inflection>
    <inflection>abandoning</inflection>
  </infl>
  <p p="v"/>
  <z_p> verb</z_p>
  <n-g n="1">
    <z_n>1</z_n>
    <z> </z>
    <d>leave sb/sth</d>
  <sl-g sl="adv">
    <z_title>adverb </z_title>
    <sb-g sb="a">
      <z_sb-g>
        
      </z_sb-g>
      <cl parent="adv">hastily</cl>
      <x>The town had been hastily ~ed.</x>
    </sb-g>
  </sl-g>
  <sl-g sl="phrases">
    <z_title>phrases </z_title>
    <sb-g>
      <z_sb-g>
        
      </z_sb-g>
      <cl idx="na" parent="phrases">~ sb to their fate</cl>
    </sb-g>
    <sb-g sb="a">
      <z_sb-g>
        
      </z_sb-g>
      <cl idx="na" parent="phrases">be left ~ed</cl>
    </sb-g>
    <sb-g sb="a">
      <z_sb-g>
        
      </z_sb-g>
      <cl idx="na" parent="phrases">be found ~ed</cl>
      <x>The car was found ~ed in a nearby town.</x>
    </sb-g>
  </sl-g>
</n-g>
```

```
<n-g n="2">
  <z_n>2</z_n>
  <z> </z>
  <d>
    stop
    <lookup lookup="doing">doing</lookup>
    /
    <lookup lookup="supporting">supporting</lookup>
    sth
  </d>
  <sl-g sl="adv">
    <z_title>adverb </z_title>
    <sb-g sb="b">
      <z_sb-g>
        
      </z_sb-g>
      <cl parent="adv">altogether</cl>
      <z_cl>,</z_cl>
      <z> </z>
      <cl parent="adv">completely</cl>
      <z_cl>,</z_cl>
      <z> </z>
      <cl parent="adv">entirely</cl>
      <z_cl>,</z_cl>
      <z> </z>
      <cl parent="adv">totally</cl>
      <x>The government does not propose to ~ the project altogether.</x>
    </sb-g>
    <sb-g sb="b">
      <z_sb-g>
        
      </z_sb-g>
      <cl parent="adv">effectively</cl>
      <z_cl>,</z_cl>
      <z> </z>
      <cl parent="adv">largely</cl>
      <z_cl>,</z_cl>
      <z> </z>
      <cl parent="adv">virtually</cl>
      <x>This principle has now been effectively ~ed.</x>
    </sb-g>
    <sb-g sb="f">
      <z_sb-g>
        
      </z_sb-g>
      <cl parent="adv">eventually</cl>
      <z_cl>,</z_cl>
      <z> </z>
```

```
<cl parent="adv">finally</cl>
</sb-g>
<sb-g sb="e">
  <z_sb-g>
    
  </z_sb-g>
  <cl parent="adv">simply</cl>
  <x>Traditional policies were simply ~ed</x>
</sb-g>
<sb-g sb="c">
  <z_sb-g>
    
  </z_sb-g>
  <cl parent="adv">formally</cl>
</sb-g>
<sb-g sb="d">
  <z_sb-g>
    
  </z_sb-g>
  <cl parent="adv">abruptly</cl>
  <z> (</z>
  <g g="espan">
    <z_g>
      <it>esp. AmE</it>
    </z_g>
    <z></z>
    <z_cl>,</z_cl>
    <z> </z>
    <cl parent="adv">immediately</cl>
    <z_cl>,</z_cl>
    <z> </z>
    <cl parent="adv">quickly</cl>
  </sb-g>
  <sb-g sb="d">
    <z_sb-g>
      
    </z_sb-g>
    <cl parent="adv">gradually</cl>
  </sb-g>
  <sb-g sb="d">
    <z_sb-g>
      
    </z_sb-g>
    <cl parent="adv">quietly</cl>
    <x>The plans for reform were quietly ~ed</x>
  </sb-g>
  <sb-g sb="f">
    <z_sb-g>
```

Extracting the Data

Instructions to ChatGPT, first round

I have a file that contains entries like the ones given below. Write a Python script that reads in the file, called 50lines.xml, and for each entry prints the headword (<h>), the inflected forms, the word class. Then have it print the relevant collocates, grouped as they are in the entry (e.g. adj, hwdandverb, verbandhwd, ...)

```
<?xml version="1.0"?>
```

```
<entry id="abbreviation_n_e" guid="000000013"> <h  
id="abbreviation_n_1">abbreviation</h> <infl>
```

Note that the file will not parse like this because the file is not an XML file but a collection of small XML files on their own lines. Please adapt the script to read this and then parse the lines containing XML data separately with `xml.etree.ElementTree`.

This is the error message I get:

Error parsing line:

no element found: line 1, column 0

Traceback (most recent call last):

```
File "/mnt/c/Daten/export_ocd2.py", line 42, in <module>  
    parse_xml_lines('50lines.xml')
```

```
File "/mnt/c/Daten/export_ocd2.py", line 37, in parse_xml_lines  
    parse_entry(entry)
```

```
File "/mnt/c/Daten/export_ocd2.py", line 5, in parse_entry  
    inflected_forms = [inflection.text for inflection in  
entry.find('infl').findall('inflection')]
```

AttributeError: 'NoneType' object has no attribute 'findall'

Some entries span more than one line, so that the parsing of these lines fails. Can you account for that?

The collocates that we want to display are inside the <cl> tag. Please print those instead of what the script prints currently.

Not extracting the Data

Fine-Tuning LLMs



Base model: Meta Llama3 8B

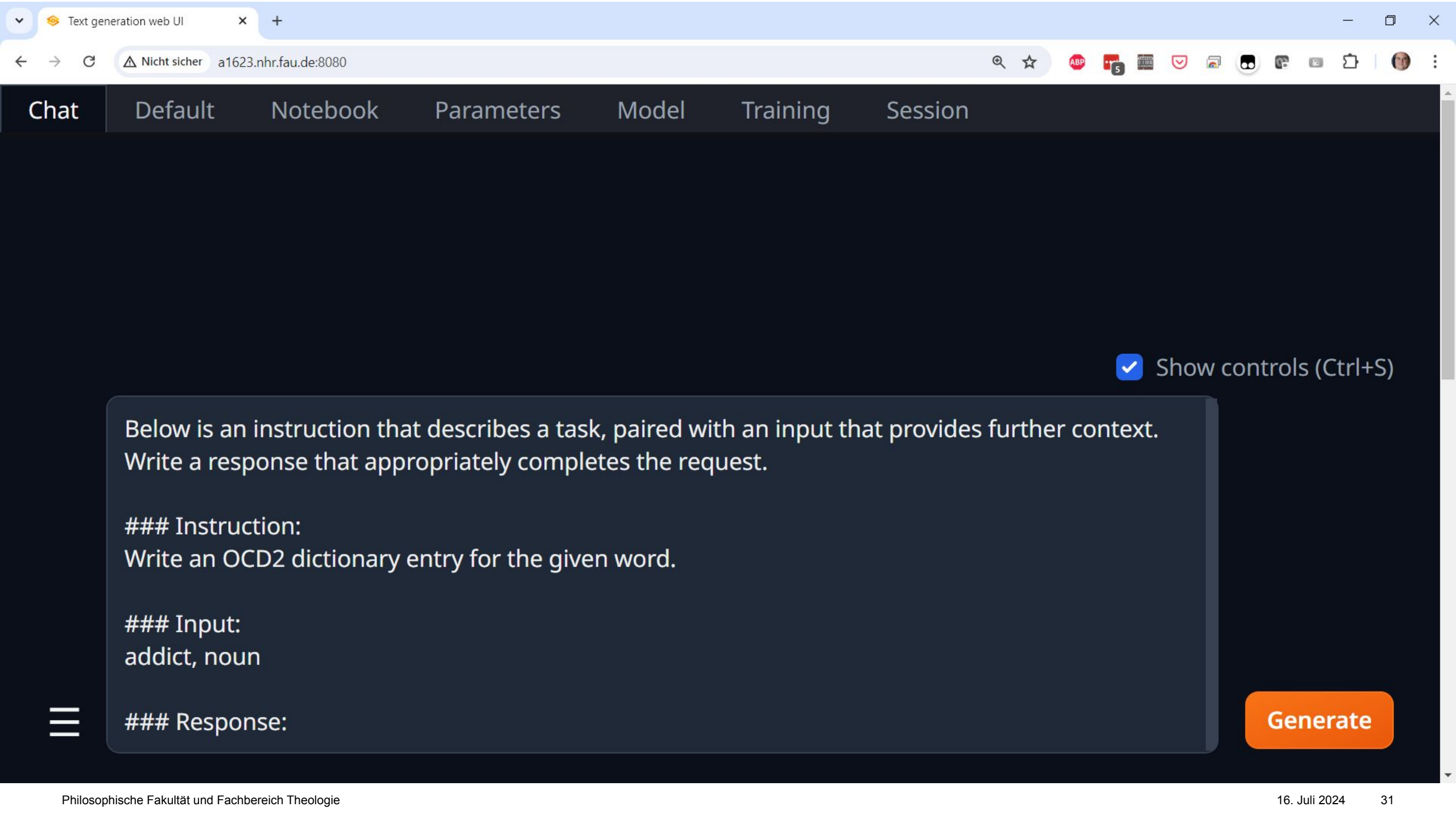
Three approaches:

- Low-Rank Adapter (LoRA)
- full fine tune with sample packing
- full fine tune without sample packing

Special thanks to Sebastian Wind, Thomas Zeiser, and the National High Performance Computing Centre at FAU, as well as the Competence Network for Scientific High Performance Computing in Bavaria.

```
[
  {
    "instruction": "Write an OCD2 dictionary entry for the given word.",
    "input": "abashed, adj.",
    "output": "<entry id=\"abashed_adj_e\" guid=\"000000008\"><h id=\"abashed_adj_1\">abashed</h><p
p=\"adj\"/><z_p> adj.</z_p><sl-g sl=\"verbs\"><z_title>verbs </z_title><sb-g sb=\"a\"><z_sb-g><img
src=\"bullet.png\"/></z_sb-g><cl idx=\"na\" parent=\"verbs\">be</cl><z_cl>,</z_cl><z> </z><cl idx=\"na\"
parent=\"verbs\">look</cl></sb-g></sl-g><sl-g sl=\"adv\"><z_title>adverb </z_title><sb-g sb=\"b\"><z_sb-g><img
src=\"bullet.png\"/></z_sb-g><cl parent=\"adv\" idx=\"n\">a little</cl><z_cl>,</z_cl><z> </z><cl
parent=\"adv\">slightly</cl><z_cl>,</z_cl><z> </z><cl>etc.</cl></sb-g><sb-g sb=\"c\"><z_sb-g><img
src=\"bullet.png\"/></z_sb-g><cl parent=\"adv\">suitably</cl><x>He glanced at Juliet accusingly and she looked suitably
~.</x></sb-g></sl-g></entry>"
  },

```



Chat

Default

Notebook

Parameters

Model

Training

Session

Show controls (Ctrl+S)

Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.

Instruction:

Write an OCD2 dictionary entry for the given word.

Input:

addict, noun

Response:

Generate

```
<entry id="addict_n_e" guid="000000161"><h
id="addict_n_1">addict</h><infl><inflection>addict</inflection><inflection>addicts</inflection></infl><p p="n"/><z_p>
noun</z_p><sl-g sl="adj"><z_title>adjective </z_title><sb-g sb="a"><z_sb-g></z_sb-g><cl
parent="adj">hard</cl><x>It is difficult to persuade hard drug ~s to seek help.</x></sb-g><sb-g sb="b"><z_sb-g></z_sb-g><cl parent="adj">drug</cl><z_cl>,</z_cl><z> </z><cl
parent="adj">heroin</cl><z_cl>,</z_cl><z> </z><cl parent="adj">opiate</cl><z_cl>,</z_cl><z> </z><cl
parent="adj">painkiller</cl><z_cl>,</z_cl><z> </z><cl parent="adj">smack</cl><z> (</z><g
g="espan"><z_g><it>esp. AmE</it></z_g><z>)</z><z_cl>,</z_cl><z> </z><cl
parent="adj">street</cl><z_cl>,</z_cl><z> </z><cl parent="adj">substance</cl></sb-g><sb-g sb="b"><z_sb-g></z_sb-g><cl parent="adj">alcohol</cl><z_cl>,</z_cl><z> </z><cl
parent="adj">cocaine</cl><z_cl>,</z_cl><z> </z><cl parent="adj">gambling</cl><z_cl>,</z_cl><z> </z><cl
parent="adj">marijuana</cl><z_cl>,</z_cl><z> </z><cl parent="adj">nicotine</cl><z_cl>,</z_cl><z> </z><cl
parent="adj">porn</cl><z> (</z><r>informal</r><z>)</z><z_cl>,</z_cl><z> </z><cl
parent="adj">pornography</cl><z_cl>,</z_cl><z> </z><cl>etc.</cl></sb-g
```


accusation *noun*

ADJECTIVE

- **serious**
- **baseless, false, unfounded, unsubstantiated**
- **ridiculous, wild**
- **bitter**
- **mutual**
They sank into mutual ~ and recrimination.
- **public**
- **veiled**
She made a lot of thinly veiled ~s.
- **renewed, repeated**

VERB + ACCUSATION

- **hurl, level at, make**
They have the nerve to level these ~s against one of our most resp
- **face**
Their father now faces an ~ of murder.
- **deny, dismiss, refute, reject**
- **prove, substantiate, support**
New evidence has emerged which supports the ~ against her.

ACCUSATION + VERB

- **fly, fly around**
There seem to be a lot of wild ~s flying around.

PREPOSITION

- **amid ~s**
He fled the country amid ~s of fraud.
- **~ against**
You made a public ~ of misconduct against Saunders.
- **~ of**
He was forced to defend himself against ~s of plagiarism.

PHRASES

- **bring an ~ against sb**
She rejected all the ~s brought against her.

accusation *noun*

ADJECTIVE

- **baseless** (*esp. AmE*), **false, groundless** (*esp. BrE*), **malicious, spurious** (*esp. BrE*), **unfounded, untrue** (*esp. BrE*)
- **damaging, defamatory, false, malicious, scandalous** (*esp. BrE*), **serious, slanderous** (*esp. BrE*)
- **alleged, unproven**
- **explicit, direct, specific**
There were no direct ~s of corruption.
- **vague**
- **criminal, felony** (*AmE*), **moral, seditious** (*esp. BrE*), **treasonable** (*BrE*)
- **political, racial** (*AmE*), **racialist** (*BrE*), **war**
He was arrested on charges of war ~s.
- **ritual, satanic**
ritual ~s of murder

VERB + ACCUSATION

- **bring, level, make**
*They brought ~s of murder against the two men.
She levelled the ~ at him that he had deliberately misled the committee.*
- **face**
He is facing ~s of fraud.
- **withdraw**
She withdrew her ~ that he had lied to the public.
- **confirm, support**
There is no evidence to support these ~s.
- **deny**
He denied ~s that he had accepted money from the company.

PREPOSITION

- **~ against**
~s against the president
- **~ of**
an ~ of corruption

PHRASES

- **an ~ of guilt**
The ~ of guilt is often made against people who have not committed a crime.

accusation *noun*

ADJECTIVE

- serious
- baseless, false, unfounded, unjustified, ridiculous, wild
- bitter
- mutual
They sank into mutual ~ and reciprocal ~.
- public
He made a public ~.
- veiled
She made a lot of thinly veiled ~s.
- renewed, repeated

VERB + ACCUSATION

- hurl, level at, make
They have the nerve to level these ~s at her face.
- face
Their father now faces an ~ of murder.
- deny, dismiss, refute, reject
- prove, substantiate, support
New evidence has emerged which ~s his ~.

ACCUSATION + VERB

- fly, fly around
There seem to be a lot of wild ~s flying around.

PREPOSITION

- amid
He flew ~ the ~s.
- ~ against
You made a public ~ of misconduct against Saunders.
- ~ of
He was forced to defend himself against ~s of plagiarism.

PHRASES

- bring an ~ against sb
She rejected all the ~s brought against her.

accusation *noun*



~ (nE), false, groundless (esp. BrE), malicious, spurious (esp. BrE), untrue (esp. BrE)
 ~atory, false, malicious, scandalous (esp. BrE), serious, unjustified (BrE)
 ~en
 ~specific
 ~ect ~s of corruption.

(AmE), moral, seditious (esp. BrE), treasonable (BrE)
 (AmE), racist (BrE), war
 ~ on charges of war ~s.

er

te

of murder against the two men.
 at him that he had deliberately misled the committee.

'fraud.

She withdrew her ~ that he had lied to the public.

confirm, support

There is no evidence to support these ~s.

deny

They were ~ed from the company.

~ against

~s against the president

~ of

an ~ of corruption

PHRASES

an ~ of guilt

The ~ of guilt is often made against people who have not committed a crime.

<https://bdlweb.phil.uni-erlangen.de/pgu/ocd2>

LLMs as Lexicographers?

some concluding remarks

LLMs as Lexicographers?

some concluding remarks



- Are we using LLMs for the wrong tasks?

Prompt: Give me the 10 most likely words in the slot masked with **** in the following sentence: "I **** forgot to take my pills." Just the words, no other commentary.

Response: Here are the 10 most likely words:

1. just
2. almost
3. nearly
4. totally
5. completely
6. partially
7. momentarily
8. briefly
9. simply
10. plain

LLMs as Lexicographers?

some concluding remarks

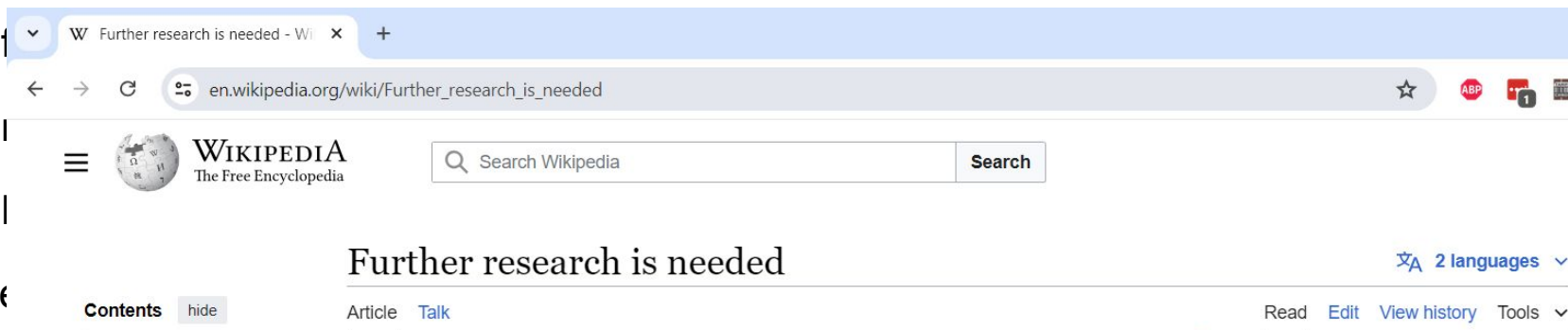


- Are we using LLMs for the wrong tasks?
- LLMs can be fine-tuned to learn the structure of specialized dictionary entries
- The fully fine-tuned LLMs were better at producing the right structure than the right content (but LoRA issues)
- The generated entries show remarkable idiosyncrasies (e.g. the markers of variety)
- Intuitively, the generated entries look worse than the lists extracted with the optimized methods discussed above, but...

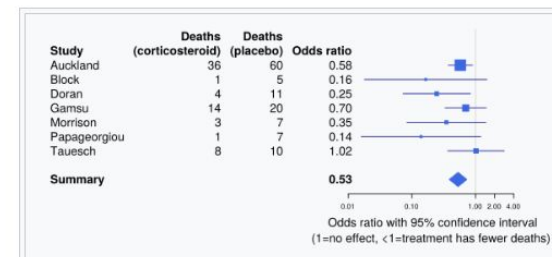
LLMs as Lexicographers?

some concluding remarks

- Are we using LLMs for...
- LLMs can be fine-tuned...
- The fully fine-tuned LLMs...
- The generated entries...
- Intuitively, the generated entries are often correct, but...



ove,



A **blobbogram** is designed to show whether further research is needed. Studies crossing the vertical line are inconclusive. Here the summary (bottom diamond) shows that the treatment prevented babies from dying. Further studies like these are not needed.

Meaning [\[edit\]](#)

Some [research journals](#) have banned the phrase "more research is needed" on the grounds that it is redundant;^[1] it is almost always true and fits almost any article, and so can be taken as understood.

A 2004 metareview by the [Cochrane collaboration](#) of their own [systematic medical reviews](#) found that 93% of the reviews studied made indiscriminate FRIN-like statements, reducing their ability to guide future research. The presence of FRIN had no correlation with the [strength of the evidence](#) against the medical intervention. Authors who thought a treatment was useless were just as likely to recommend researching it further.^[2]

Thank you!
peter.uhrig@fau.de