Empirical methods in linguistics [05]

Corpus linguistics 1

Prof. Dr. Martin Schäfer
Department of English Studies, Universität Leipzig

What is a corpus?

"A corpus is a collection of (1) machine-readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety." Brezina and Gablasova (2018)

Corpus linguistics and data collection methods

- observation, survey, experiment: the researcher has to collect the data first
- the construction (=compilation) of a corpus can be seen as one variant of observation

Why corpora?

Why corpora?

- authentic observational data on actual language use
- sampled to be representative
- versatile and easy to use because it is machine-readable

Why corpora? A simple example

- What is the past tense of *sneak*?

Why corpora? A simple example

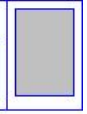
- What is the past tense of *sneak*?
- Dictionaries: yes, both forms exist

Why corpora? A simple example

- What is the past tense of *sneak*?
- Dictionaries: yes, both forms exist
- Corpora allow us to go further:
  - Equally common? Any sociolinguistic signature (varieties, age groups, socioeconomic status, modes, registers, etc.)?
  - Any long-term development going on here? One form gradually replacing the other perhaps?

A sneak peak

The changing usage frequency of one of the two forms during the past 200 years of written American English:

SECTION	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
FREQ	572	0	0	0	0	0	0	1	0	1	4	1	6	10	8	27	45	44	94	130	201
WORDS (M)	405	7.0	13.7	15.8	16.5	16.9	18.8	20.1	20.4	22.0	23.1	25.7	27.7	27.4	28.7	29.1	28.8	29.9	33.1	34.8	35.5
PER MIL	1.41	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.05	0.17	0.04	0.22	0.36	0.28	0.93	1.56	1.47	2.84	3.73	5.67
SEE ALL YEARS AT ONCE																					

Corpus linguistics as a method

"Corpus linguistics is the investigation of *linguistic research questions* based on the complete and systematic analysis of the distribution of linguistic phenomena in a linguistic corpus."

Source: Stefanowitsch (2020, p. 55), the fourth attempt at a definition, this time giving the *linguistic interpretation*

Which corpus?

Your research question determines the choice of corpus!

- language: monolingual, multilingual ...
- modality: written/spoken/signed language
- purpose: general vs. specific registers
- annotations: POS, syntax, semantics ...
- size/persistence
- temporal dimension: synchronic, diachronic

The corpus collection for this class

https://www.english-corpora.org/

Working with a corpus

"complete and systematic analysis"

Working with a corpus

"complete and systematic analysis"

- recall: get everything that we want
- precision: get only what we want

Working with a corpus

"complete and systematic analysis"

- recall: get everything that we want
- precision: get only what we want

Example: the verb *to play*



Working with a corpus

"complete and systematic analysis"

- recall: get everything that we want
- precision: get only what we want

Example: the verb *to play*



- only the exact string *play* (other inflectional forms are missing)
- disregardless of whether it is a noun or a verb or whatever (leads to false positives)

Working with a corpus

- create clever search queries (lemma queries: `PLAY`/`[play]`)
- use all available annotation (POS queries `play_vv`)
- do manual editing (almost always needed :))

Working with a corpus

- create clever search queries (lemma queries: `PLAY`/`[play]`)
- use all available annotation (POS queries `play_vv`)
- do manual editing (almost always needed :))
- Use the HELP!

Working with a corpus

"the distribution of linguistic phenomena"

Working with a corpus

"the distribution of linguistic phenomena"

- quantitative: frequency of occurrence (LIST/CHART)
- qualitative: display in context/concordance (LIST/KWIC)

Working with a corpus

quantitative information: *delicious* in the COCA

- LIST: click on OPTIONS, change DISPLAY to PER MIL
- LIST: tick SECTIONS
- LIST: try `=delicious`
- Try CHART!

Working with a corpus

qualitative information: the verb *hate* in the COCA

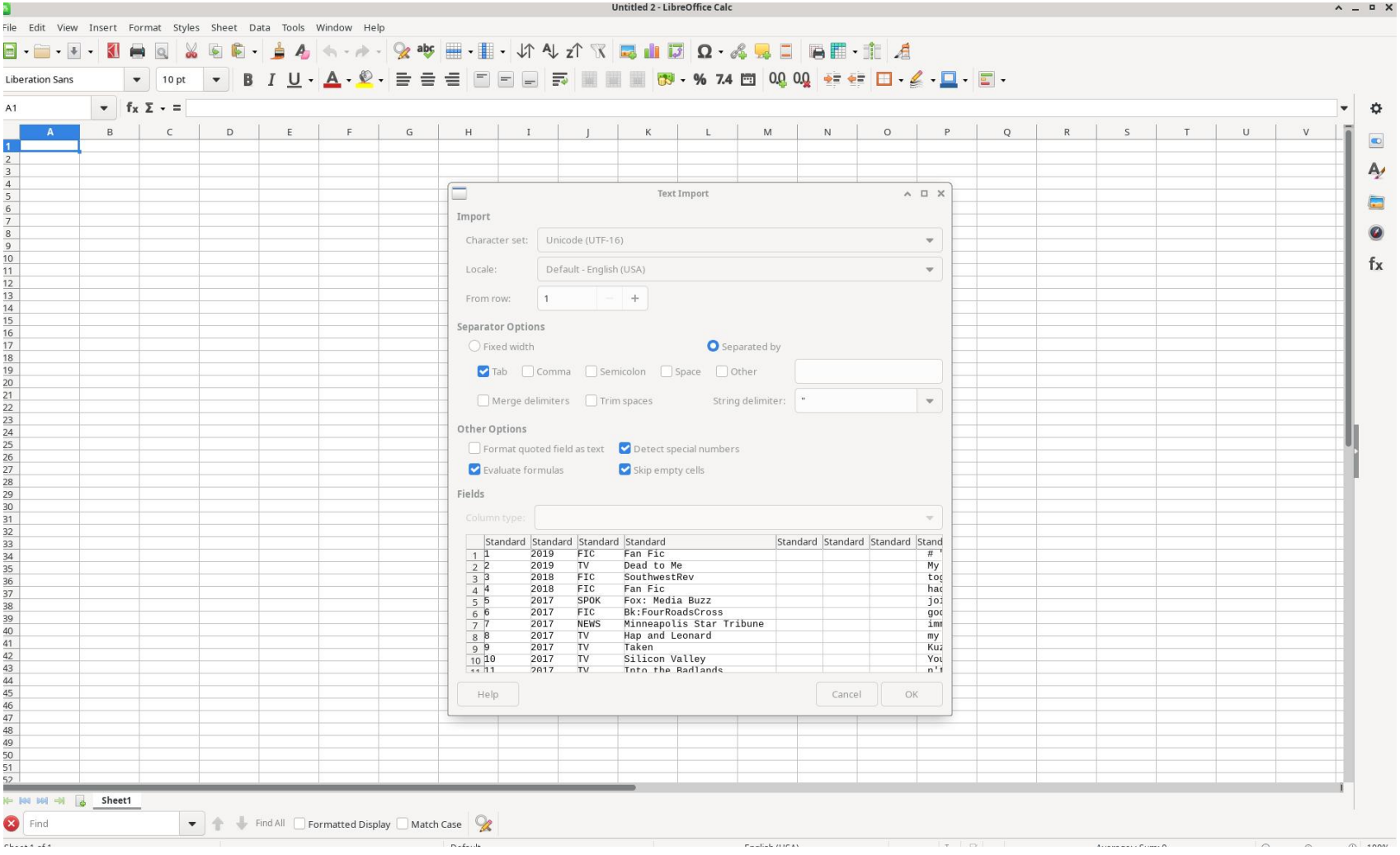
- search for all forms of the verb
- in the results list, click on **100** in the top blue line to select a random sample of 100 hits

Working with a corpus

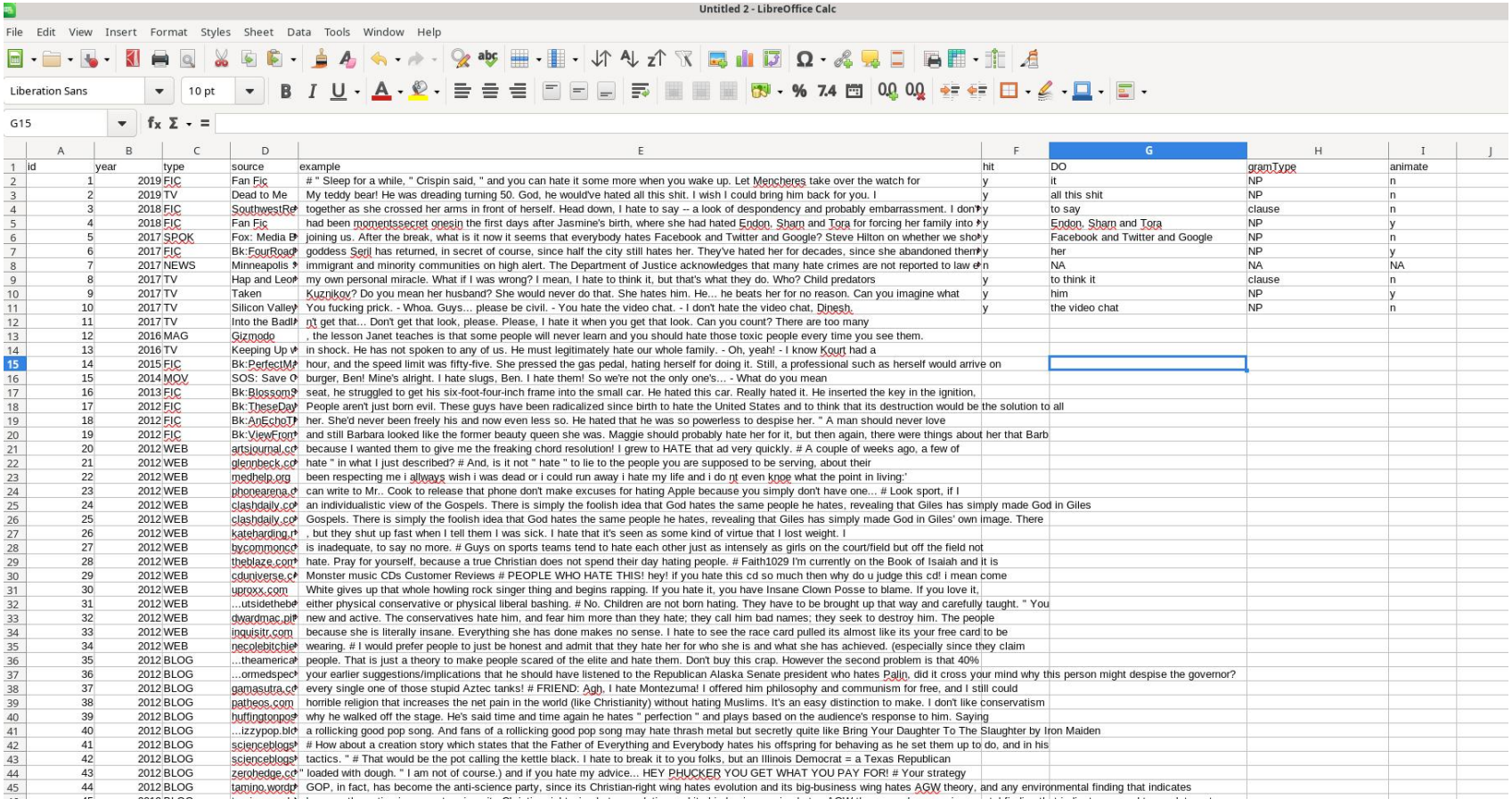
Next step: get the data into a spreadsheet (here: for CALC, for other spreadsheet software, a simple copy and paste might do

- use your mouse to mark all hits and copy them to your systems clipboard (Control/Steuerung c or right click and select **copy**)
- paste into CALC: use PASTE SPECIAL/TEXT IMPORT DIALOG

Working with a corpus



Working with a corpus: further annotation



Untitled 2 - LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window Help

Liberation Sans 10 pt

G15

	A	B	C	D	E	F	G	H	I	J
1	id	year	type	source	example	hit	DO	gramType	animate	
2	1	2019	FIC	Fan Fic	# " Sleep for a while, " Crispin said, " and you can hate it some more when you wake up. Let Mencheres take over the watch for	y	it	NP	n	
3	2	2019	TV	Dead to Me	My teddy bear! He was dreading turning 50. God, he would've hated all this shit. I wish I could bring him back for you. I	y	all this shit	NP	n	
4	3	2018	FIC	SouthwestRe	together as she crossed her arms in front of herself. Head down, I hate to say -- a look of despondency and probably embarrassment. I don't	y	to say	clause	n	
5	4	2018	FIC	Fan Fic	had been momentssecret onesin the first days after Jasmine's birth, where she had hated Endon, Sharn and Tora for forcing her family into	y	Endon, Sharn and Tora	NP	y	
6	5	2017	SPOK	Fox: Media B	joining us. After the break, what is it now it seems that everybody hates Facebook and Twitter and Google? Steve Hilton on whether we sho	y	Facebook and Twitter and Google	NP	n	
7	6	2017	FIC	Bk:FourRoad	goddess Seril has returned, in secret of course, since half the city still hates her. They've hated her for decades, since she abandoned them	y	her	NP	y	
8	7	2017	NEWS	Minneapolis S	immigrant and minority communities on high alert. The Department of Justice acknowledges that many hate crimes are not reported to law e	n	NA	NA	NA	
9	8	2017	TV	Hap and Leo	my own personal miracle. What if I was wrong? I mean, I hate to think it, but that's what they do. Who? Child predators	y	to think it	clause	n	
10	9	2017	TV	Taken	Kuznikov? Do you mean her husband? She would never do that. She hates him. He... he beats her for no reason. Can you imagine what	y	him	NP	y	
11	10	2017	TV	Silicon Valley	You fucking prick. - Whoa. Guys... please be civil. - You hate the video chat. - I don't hate the video chat, Dinesh.	y	the video chat	NP	n	
12	11	2017	TV	Into the Badl	n't get that... Don't get that look, please. Please, I hate it when you get that look. Can you count? There are too many					
13	12	2016	MAG	Gizmodo	, the lesson Janet teaches is that some people will never learn and you should hate those toxic people every time you see them.					
14	13	2016	TV	Keeping Up W	in shock. He has not spoken to any of us. He must legitimately hate our whole family. - Oh, yeah! - I know Kourt had a					
15	14	2015	FIC	Bk:PerfectM	hour, and the speed limit was fifty-five. She pressed the gas pedal, hating herself for doing it. Still, a professional such as herself would arrive on					
16	15	2014	MOV	SOS: Save O	burger, Ben! Mine's alright. I hate slugs, Ben. I hate them! So we're not the only one's… - What do you mean					
17	16	2013	FIC	Bk:Blossom	seat, he struggled to get his six-foot-four-inch frame into the small car. He hated this car. Really hated it. He inserted the key in the ignition,					
18	17	2012	FIC	Bk:TheseDay	People aren't just born evil. These guys have been radicalized since birth to hate the United States and to think that its destruction would be the solution to all					
19	18	2012	FIC	Bk:AnEchoT	her. She'd never been freely his and now even less so. He hated that he was so powerless to despise her. " A man should never love					
20	19	2012	FIC	Bk:ViewFron	and still Barbara looked like the former beauty queen she was. Maggie should probably hate her for it, but then again, there were things about her that Barb					
21	20	2012	WEB	artsjournal.co	because I wanted them to give me the freaking chord resolution! I grew to HATE that ad very quickly. # A couple of weeks ago, a few of					
22	21	2012	WEB	glennbeck.co	hate " in what I just described? # And, is it not " hate " to lie to the people you are supposed to be serving, about their					
23	22	2012	WEB	medhelp.org	been respecting me i allways wish i was dead or i could run away i hate my life and i do nt even knoe what the point in living:'					
24	23	2012	WEB	phonearena.c	can write to Mr.. Cook to release that phone don't make excuses for hating Apple because you simply don't have one... # Look sport, if I					
25	24	2012	WEB	clashdaily.co	an individualistic view of the Gospels. There is simply the foolish idea that God hates the same people he hates, revealing that Giles has simply made God in Giles					
26	25	2012	WEB	clashdaily.co	Gospels. There is simply the foolish idea that God hates the same people he hates, revealing that Giles has simply made God in Giles' own image. There					
27	26	2012	WEB	kateharding.n	, but they shut up fast when I tell them I was sick. I hate that it's seen as some kind of virtue that I lost weight. I					
28	27	2012	WEB	bycommonco	is inadequate, to say no more. # Guys on sports teams tend to hate each other just as intensely as girls on the court/field but off the field not					
29	28	2012	WEB	theblaze.com	hate. Pray for yourself, because a true Christian does not spend their day hating people. # Faith1029 I'm currently on the Book of Isaiah and it is					
30	29	2012	WEB	cduniverse.c	Monster music CDs Customer Reviews # PEOPLE WHO HATE THIS! hey! if you hate this cd so much then why do u judge this cd! i mean come					
31	30	2012	WEB	uproxx.com	White gives up that whole howling rock singer thing and begins rapping. If you hate it, you have Insane Clown Posse to blame. If you love it,					
32	31	2012	WEB	...utsidethebo	either physical conservative or physical liberal bashing. # No. Children are not born hating. They have to be brought up that way and carefully taught. " You					
33	32	2012	WEB	dwardmac.pit	new and active. The conservatives hate him, and fear him more than they hate; they call him bad names; they seek to destroy him. The people					
34	33	2012	WEB	inquisitr.com	because she is literally insane. Everything she has done makes no sense. I hate to see the race card pulled its almost like its your free card to be					
35	34	2012	WEB	necolebitchie	wearing. # I would prefer people to just be honest and admit that they hate her for who she is and what she has achieved. (especially since they claim					
36	35	2012	BLOG	...theamerica	people. That is just a theory to make people scared of the elite and hate them. Don't buy this crap. However the second problem is that 40%					
37	36	2012	BLOG	...ormedspec	your earlier suggestions/implications that he should have listened to the Republican Alaska Senate president who hates Palin, did it cross your mind why this person might despise the governor?					
38	37	2012	BLOG	gamasutra.co	every single one of those stupid Aztec tanks! # FRIEND: Agh, I hate Montezuma! I offered him philosophy and communism for free, and I still could					
39	38	2012	BLOG	patheos.com	horrible religion that increases the net pain in the world (like Christianity) without hating Muslims. It's an easy distinction to make. I don't like conservatism					
40	39	2012	BLOG	huffingtonpo	why he walked off the stage. He's said time and time again he hates " perfection " and plays based on the audience's response to him. Saying					
41	40	2012	BLOG	...izzypop.blo	a rollicking good pop song. And fans of a rollicking good pop song may hate thrash metal but secretly quite like Bring Your Daughter To The Slaughter by Iron Maiden					
42	41	2012	BLOG	scienceblogs	# How about a creation story which states that the Father of Everything and Everybody hates his offspring for behaving as he set them up to do, and in his					
43	42	2012	BLOG	scienceblogs	tactics. " # That would be the pot calling the kettle black. I hate to break it to you folks, but an Illinois Democrat = a Texas Republican					
44	43	2012	BLOG	zerohedge.co	" loaded with dough. " I am not of course.) and if you hate my advice... HEY PHUCKER YOU GET WHAT YOU PAY FOR! # Your strategy					
45	44	2012	BLOG	tamino.wordp	GOP, in fact, has become the anti-science party, since its Christian-right wing hates evolution and its big-business wing hates AGW theory, and any environmental finding that indicates					

HA for next week: *sneak* revisited

- At `www.english-corpora.org`, choose a corpus that can provide insight into the long-term diachrony (100+ years) of the verb *to sneak* and its past-tense variants in American English.
- Come up with a query that yields a list of the usage frequencies of the verb with all its inflected word forms.
- You should also be able to show the usage frequencies across time, e.g. more of a table with columns for decades (depending on your choice of corpus). What does the table reveal about the diachronic development of the forms sneaked versus snuck? Do you find this unexpected?
- Try different ways to visualise numerical information, e.g., just for either *sneaked* or *snuck*, generate a graph plotting the frequency of *sneaked* diachronically. Use the **display option** to change raw frequency to nomalized frequency, try the CHART display.

References I

- Brezina, Vaclav and Dana Gablasova (Feb. 2018). "The corpus method". English. In: *English Language. Description, variation, and context*. Ed. by Jonathan Culpeper et al. 2nd ed. Palgrave Macmillan. ISBN: 9781137571823.
- Stefanowitsch, Anatol (2020). *Corpus linguistics. A guide to the methodology*. Textbooks in Language Sciences 7. Berlin: Language Science Press. DOI: 10.5281/zenodo.3735822.